

Object-Aware 4D Human Motion Generation

Shurui Gui*^{1,2†} Deep Patel*¹ Xiner Li² Martin Renqiang Min¹

¹NEC Laboratories America ²Texas A&M University

{shurui.gui, lxe}@tamu.edu {dpatel, renqiang}@nec-labs.com

*Equal contribution.

Abstract

Recent advances in video diffusion models have enabled the generation of high-quality videos. However, these videos still suffer from unrealistic deformations, semantic violations, and physical inconsistencies that are largely rooted in the absence of 3D physical priors. To address these challenges, we propose an object-aware 4D human motion generation framework grounded in 3D Gaussian representations and motion diffusion priors. With pre-generated 3D humans and objects, our method, Motion Score Distilled Interaction (MSDI), employs the spatial and prompt semantic information in large language models (LLMs) and motion priors through the proposed Motion Diffusion Score Distillation Sampling (MSDS). The combination of MSDS and LLMs enables our spatial-aware motion optimization, which distills score gradients from pre-trained motion diffusion models, to refine human motion while respecting object and semantic constraints. Unlike prior methods requiring joint training on limited interaction datasets, our zero-shot approach avoids retraining and generalizes to out-of-distribution object aware human motions. Experiments demonstrate that our framework produces natural and physically plausible human motions that respect 3D spatial context, offering a scalable solution for realistic 4D generation.

1. Introduction

Recent advancements in video generation have led to impressive results in generating realistic and semantically rich visual content. Video diffusion models [5, 8, 9, 32, 33, 49, 60] have achieved high visual quality on diverse tasks. Despite the progress, state-of-the-art models, including large-scale systems like Sora [30], still face persistent challenges such as unrealistic deformation, object penetration, and semantic violations. These issues often stem from the lack of explicit physical and spatial constraints, which are difficult to capture in purely 2D representations [1, 51].

To address these limitations, there has been a growing interest in incorporating 3D priors into generative modeling. The success of methods like DreamFusion [38] has demonstrated that distilling 2D priors from pre-trained diffusion models can guide 3D content generation, which has motivated many 3D and 4D generation works [11, 13, 22, 26, 28, 35, 58, 61]. However, 4D generation methods that rely solely on video diffusion models inherit the spatial ambiguity and semantic misalignment problems. For instance, prompts involving spatial relations (e.g., “a dog under the bed”) often produce incorrect visual arrangements. To mitigate this, compositional 4D generation approaches [3, 4, 52] have been proposed to combine multiple priors and synthesize novel distributions. Yet, these methods still face a fundamental bottleneck: human motions generated from pre-trained video models often suffer from distortions and fail to respect the physical constraints of interactions with static objects.

In this work, we tackle the challenging problem of zero-shot object-aware 4D human motion generation. Specifically, we aim to generate realistic 3D human motion interacting with a 3D static object over time, without requiring additional training on paired human-object data. Unlike prior methods [7, 12, 48, 53] that rely on training dedicated models with limited joint human-object datasets, our framework leverages a compositional approach with strong generalization capability. Our method, Motion Score Distilled Interaction (MSDI), builds on recent advances in 3D Gaussian representations, i.e., motion diffusion models, and spatial reasoning with large language models (LLMs).

Specifically, we first generate high-fidelity human and object 3D Gaussians using HumanGaussian [27] and DreamGaussian [43], respectively. To control the temporal motion, we propose to guide human trajectories using LLM-generated spatial instructions, which provide coarse but plausible global motion plans. Then, instead of directly sampling from pre-trained motion diffusion models, which are often unreliably out of distribution, MSDS distills guidance from the motion diffusion model to form an optimization process that adjusts human poses and trajectories to

[†]Work done during internship at NEC Laboratories America.

align with both learned motion priors and interaction constraints. Furthermore, we formulate a constrained optimization framework that combines MSDS loss with smoothness, trajectory alignment, and collision-avoidance terms. This allows us to generate motion sequences that are realistic, smooth, and physically plausible with respect to the static object. Our zero-shot formulation ensures that the system can benefit from future improvements in motion diffusion models without the need for retraining, offering a scalable path toward generalizable and realistic object-aware 4D human motion generation. Experiments on multiple zero-shot prompts show, that our generated 4D scenes produce realistic motions with high physical constraint obedience ability while previous 4D generation methods can only generate unnatural distortions without plausible motion.

2. Related Work

Video generation. Video generation models have been widely used to generate realistic videos. Although the video diffusion models [5, 8, 9, 32, 33, 49, 60] have shown promising results in various areas, unrealistic deformation, twisted, penetration, and semantics violations still exist even in large video generation model Sora [30]. These issues are often considered as the lack of physics information learned [2, 47, 54]. Despite many studies in addressing these issues by using trajectory tracking [14], occlusion masks [19], and semantic masks [34], we argue that it is not feasible to solve this problem in 2D space without introducing extra information equivalent to physical information in 3D space, and the natural and intrinsic way to tackle these challenges should lie in the use of 3D space.

3D generation. While 3D generation has been explored in recent years [13, 26, 28, 41, 43, 46, 61], one of the most popular and convincing directions is to extract prior knowledge from 2D diffusion models. Specifically, DreamFusion [38], the first method introducing Score Distillation Sampling (SDS), generates 3D content by leveraging information from 2D image diffusion models. This work inspires a significant amount of following works [10, 50, 56] on improving 3D content quality, optimization efficiency, and human avatar generations [16, 21, 27].

4D Generation. Aligning with the philosophy of extracting information from pre-trained image diffusion models, many 4D generation works adopt pre-trained video diffusion models [25, 55, 59], to tackle challenges in image to 4D [39] and video to 4D [11, 17, 22, 23, 31, 35, 57, 58] tasks. However, the generation ability of these studies cannot go beyond the original distribution of the pre-trained diffusion models, and shares the same limitations as the original 2D pre-trained diffusion models. For example, most of the pre-trained video diffusion models face difficulties in understanding spatial information, *e.g.*, generating with a prompt "a human walks towards the table" can produce unrealistic results, such as

deformed bodies, poor framing that shows only the legs, or the human being omitted from the scene entirely. In order to solve this challenge, one convincing direction is to apply compositional 4D generation, which incorporates multiple prior distributions and combine them to generate samples with novel distributions. Recently, 4DFy [4], Comp4D [52], and TC4D [3] have shown promising results on 4D compositional generation. However, although generating contents in 3D space helps with spatial information/trajectory planning, all the motion information from the pre-trained video diffusion model inherits its original distortions, especially on human-related motions. Motivated by this problem, we consider distilling information from dedicated motion models to guide the motion optimization process.

Different from interaction generations between humans and objects/scenes [7, 48, 53], our method is zero-shot, which does not need to train specific dedicated models; thus, it widens the application range. Human-object interaction generations like InterDiff [53] and CG-HOI [12] require the joint distribution of humans and objects for training, while the sizes of these datasets are still limited, which cannot extend to any out-of-distribution scenarios. Since our work focuses on the interaction between humans and static objects, our setting is more similar to the HUMAN-ISE task [7, 48]. Compared with them, our setting eliminates the object-locating phase and focuses on the human trajectory and human motion generation with the static object. While these works require training extra models for human trajectory and human motion generation just for the additional object, our method does not require any additional motion diffusion model training and can achieve realistic interactions between humans and static objects with motion diffusion model score distillation sampling (MSDS). This zero-shot behavior enables this framework to improve as the motion diffusion model iterates in the future without extra distribution and retraining requirements.

3. Preliminaries

3.1. 3D Gaussian Splatting

3D Gaussian Splatting [20] (3DGS) is a dominating representation in the 3D field, due to its explicit 3D space representation and high efficient optimization. The individual units of 3DGS are 3D Gaussian ellipsoids, where each 3D Gaussian is parameterized by position μ , anisotropic covariance Σ as its shape, and opacity α and spherical harmonic coefficients sh as its optical characteristics, where sh is a view-dependent property. The shape of the 3D Gaussian Σ can be considered as the composition of a scaling and a rotation as follows:

$$\Sigma = RSS^T R^T, \quad (1)$$

where the scaling matrix S can be denoted as a 3D vector s , and the rotation matrix R as a quaternion $q \in \mathbf{SO}(3)$.

Therefore, the formal definition of a Gaussian centered at point μ is:

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

where \mathbf{x} is a random variable in 3D space.

To render 3D Gaussians into a 2D image, 3DGS considers the additional opacity α and spherical harmonic coefficients by utilizing a tile-based rasterizer and point-based α -blend rendering. For each pixel u , its color $C(u)$ is rendered under the following calculation:

$$C(u) = \sum_{i \in N} T_i c_i \alpha_i \mathcal{SH}(sh_i, v), \quad (3)$$

$$T_i = G(\mathbf{x}, \mu_i) \prod_{j=1}^{i-1} (1 - \alpha_j G(\mathbf{x}, \mu_j)),$$

where T_i denotes the transmittance for the i -th Gaussian, \mathcal{SH} denotes the spherical harmonic function, and v represents the viewing direction. The 3D Gaussian optimization process includes adjusting all 3D Gaussian properties $\{\mu, q, s, \sigma, c\}$ and the high-level 3D Gaussian density modifications using densifying and pruning processes.

3.2. SMPL-X

SMPL models [37] represent a human by transforming the human mesh of a standard pose, the canonical model, into the observation space, using pose parameter θ , shape parameter β , and expression parameter ϕ :

$$M(\beta, \theta, \phi) = \text{LBS}(T(\beta, \theta, \phi), J(\beta), \theta, \mathcal{W}), \quad (4)$$

$$T(\beta, \theta, \phi) = \mathbf{T} + B_s(\beta) + B_e(\phi) + B_p(\theta),$$

where M is the function mapping parameters to a transformed human mesh model; T represents the transformed human key points/vertices adjusted by different human shapes, expressions, and poses through corresponding functions B_s , B_e , and B_p , respectively. Given the transformed vertices, the skins of the human mesh need to be adjusted according to the transformations of several nearby joints, which is done by the linear blend skinning function $\text{LBS}(\cdot)$ where \mathcal{W} stands for blend weights that determine the effects from different joints. Specifically, the LBS process is defined as follows:

$$\mathbf{v}_o = \mathcal{G} \cdot \mathbf{v}_c, \quad \mathcal{G} = \sum_{k=1}^K w_k \mathcal{G}_k(\theta, j_k), \quad (5)$$

where the vertices \mathbf{v}_o in the observation space is deformed from the canonical pose vertices \mathbf{v}_c by the deformation \mathcal{G} . The deformation is determined by the affine deformation $\mathcal{G}_k(\theta, j_k)$ that merges the warping effects from K neighboring joints, simulating the smooth position changes of vertices.

4. Motion Score Distilled Interaction

Our object-aware human motion generation (OAHM) framework addresses the challenging zero-shot generation problem by leveraging pre-trained motion priors within an explicit optimization paradigm. In this section, we first introduce the overall OAHM generation pipeline, followed by a description of our spatially-aware coarse motion generation strategy. Finally, we detail our Motion Score Distilled Interaction (MSDI) method, where we propose motion score distillation sampling (MSDS) and incorporate spatial and physical constraints to optimize human motion trajectories, enabling the synthesis of physically plausible and natural motion sensitive to object.

4.1. OAHM Generation Framework

With expressive 3D representations as the foundation, we employ HumanGaussian [27] to generate high-fidelity 3D human Gaussians G_h from textual prompts, and utilize DreamGaussian [43] to synthesize 3D objects from an initial shape-e geometry [18], as illustrated in Figure 1. Given a motion sequence X , we establish a correspondence between the motion trajectory and the human Gaussian points G_h , enabling dynamic Gaussian-based human motion.

Concretely, we initialize the SMPL-X model in a rest pose, consistent with the canonical configuration of the human Gaussian points. Each Gaussian point is mapped to the nearest barycentric coordinate on the corresponding SMPL-X mesh face. By preserving these fixed barycentric correspondences, any transformation applied to the SMPL-X mesh is faithfully propagated to the associated Gaussian points, ensuring coherent deformation of the 3D human representation.¹ Note that this mapping is differentiable, enabling gradients back propagation.

With a controllable Gaussian human and an independently generated Gaussian object co-located in the same coordinate system, we render interactive sequences via Gaussian splatting. While high-quality 3D human and object representations can be readily obtained, the availability of joint 4D human-object distributions remains limited [6], making it infeasible to train generative models directly on such data. To this end, we propose a zero-shot OAHM generation framework. Constructing this framework and achieving realistic object-aware human motion remains highly challenging due to the need for temporally consistent, physically plausible, and semantically appropriate interactions. Therefore, this framework presents two major challenges: (1) extracting meaningful and realistic human motion distributions, and (2) enforcing human-object interaction constraints on generated motions.

¹Recalculating or interpolating linear blend skinning (LBS) weights for the Gaussian points is a viable alternative; however, we focus on the barycentric mapping approach for clarity.

4.2. Spatial-Aware Coarse Motion Generation

Motion diffusion models. To address the first challenge, we avoid relying on video diffusion models, as prior work [4] has shown that distilling human motion from such models often leads to unrealistic results. Instead, we leverage dedicated human motion diffusion models (MDMs) [40, 45], which are currently state-of-the-art for generating plausible human motions.

Our motion representation consists of an N -length motion sequence $X = \{x^i\}_{i=1}^N$, where each $x^i \in \mathbb{R}^{3+6+J \times 6}$ encodes the pose parameters $\theta^i \in \mathbb{R}^{J \times 6}$, global translation $r^i \in \mathbb{R}^3$, and 6D global orientation $\gamma^i \in \mathbb{R}^6$. Other parameters, such as body shape, are omitted for simplicity. During the MDM process, the N -frame motion sequence X is subject to T steps of Gaussian noise:

$$q(X_t | X_{t-1}) = \mathcal{N}(\sqrt{\alpha_t} X_{t-1}, (1 - \alpha_t)I), \quad (6)$$

where $t \in \{1, \dots, T\}$ denotes the diffusion step and $X_T \sim \mathcal{N}(0, I)$. The MDM is trained to predict the clean motion \hat{X}_0 given a noisy motion X_t and a text condition c encoded by a CLIP-based text encoder.

LLM-based trajectory generation. Despite the advantages of MDMs, directly sampling from these models does not guarantee meaningful object-aware motions, as they lack explicit spatial awareness necessary for modeling relationships between humans and objects. Attempts to use guidance from 2D image and video diffusion models also failed to yield reliable spatial supervision signals.

To overcome this, we harness the spatial reasoning capabilities of LLMs. Given the initial coordinates of the human and object, along with a textual motion instruction, the LLM generates a coarse global trajectory for the human. This LLM-derived trajectory, denoted as r_{LLM}^i for $i = 1, \dots, N$, can be further refined using trajectory interpolation and collision detection, enabling the system to produce physically plausible paths, such as automatic detours around obstacles. For example, when instructed to “walk four meters toward a table two meters away,” the LLM can synthesize a motion that navigates around the object.

The LLM-generated trajectory is used to initialize the global translation in the MDM framework [40]. With estimated time/frames and extracted pure motion prompt as two additional inputs, the MDM can yield a coarse motion sequence that incorporates spatial awareness. While the resulting motions may lack fine-grained realism, they provide a strong starting point for subsequent optimization. Detailed examples of LLM prompts are provided in the Appendix.

4.3. Constrained Motion Optimization

The core challenge in generating realistic object-aware human motion lies not only in producing plausible human motion, but also in enforcing physical constraints such as

collision avoidance and trajectory fidelity. These challenges are not easily addressed by existing generative models. As mentioned above, two major issues must be overcome: (1) generating meaningful, in-distribution human motion sequences, and (2) ensuring these motions respect spatial and physical constraints posed by objects in the environment.

While our LLM-guided approach and MDM address the extraction of plausible motion trajectories, these solutions alone cannot guarantee realistic interactions. Specifically, directly applying the LLM-generated coarse trajectories often results in infeasible or unnatural motions, as these trajectories may violate object penetration constraints or fall outside the motion distribution captured by the pre-trained MDM. Moreover, existing diffusion models are not inherently designed to encode or enforce collision and spatial constraints.

To overcome these limitations, we introduce a constrained motion optimization framework, namely, motion score distilled interaction (MSDI), that jointly refines human motion by leveraging the prior knowledge encoded in motion diffusion models, while explicitly enforcing trajectory, smoothness, and collision-avoidance constraints.

Motion Diffusion Score Distillation Sampling (MSDS). Instead of generating human motions directly from the diffusion model, we extract the score (gradient) information from the MDM to guide the optimization of both trajectories and poses under physical constraints. Specifically, we propose *Motion Diffusion Score Distillation Sampling* (MSDS), which optimizes human motion X by maximizing the log-likelihood under the MDM prior. The gradient of the MSDS objective is given by:

$$\nabla_X \mathcal{L}_{\text{MSDS}}(\phi) \triangleq \mathbb{E}_{t, \epsilon} [w(t) (X - \text{MDM}_\phi(X_t, t, c))], \quad (7)$$

where MDM_ϕ denotes the pre-trained MDM and $w(t)$ is a weighting function over diffusion steps. This process aligns the optimized motion with the learned distribution of human poses and trajectories.

Constrained Optimization Objectives. To ensure the resulting motions are physically plausible and interact naturally with the object, we further introduce explicit constraints:

- **Trajectory Alignment.** We regularize the optimized trajectory to remain close to the LLM-generated coarse trajectory. The trajectory loss is defined as:

$$\mathcal{L}_{\text{traj}} = \lambda_{\text{middle}} \cdot \sum_{i=2}^{N-1} \|r^i - r_{LLM}^i\|_2^2 + \lambda_{\text{end}} \cdot \sum_{i \in \{1, N\}} \|r^i - r_{LLM}^i\|_2^2, \quad (8)$$

where λ_{middle} and λ_{end} control the fidelity at middle and endpoint frames, respectively.

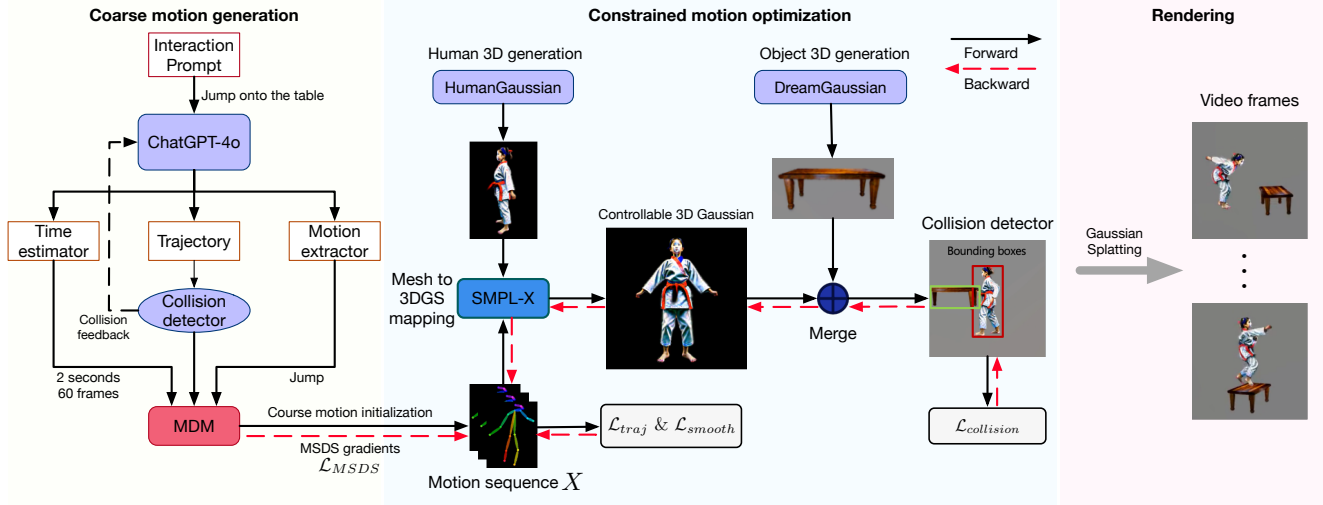


Figure 1. **Method overview.** The framework includes 4 components: human and object 3D generation, coarse trajectory generation, constrained motion optimization, and rendering.

- **Motion Smoothness.** To prevent unnatural or abrupt changes in motion, we introduce a jerk (third derivative) regularization:

$$\mathcal{L}_{smooth} = \sum_{i=1}^N \left\| \frac{d^3 r^i}{dt^3} \right\|_2^2, \quad (9)$$

where, in practice, the derivative is approximated using finite differences over adjacent frames.

- **Collision Avoidance.** To prevent human-object penetration, we employ a two-stage collision detection and penalty scheme. First, we compute the intersection C of the 3D bounding boxes for the human and object. If C is non-empty, we evaluate pairwise collisions between object points $o_i \in C$ and their nearest human points $h_j \in C$. The collision loss is then given by:

$$\mathcal{L}_{collision} = \max(\mathbf{n}_j \cdot (h_j - o_j), -\epsilon_c), \quad (10)$$

where \mathbf{n}_j is the normal vector at h_j and ϵ_c is a collision margin hyperparameter.

MSDI Objective. The final loss function for human motion optimization combines the above terms:

$$\mathcal{L} = \lambda_{MSDS} \cdot \mathcal{L}_{MSDS} + \lambda_{traj} \cdot \mathcal{L}_{traj} + \lambda_{smooth} \cdot \mathcal{L}_{smooth} + \lambda_{collision} \cdot \mathcal{L}_{collision}, \quad (11)$$

where λ_{MSDS} , λ_{traj} , λ_{smooth} , and $\lambda_{collision}$ are hyperparameters balancing the different objectives.

This MSDI constrained optimization aggregates all gradients to the motion sequence X and updates it. Through this optimization, we ensure that the generated motions X are not only realistic according to the learned motion diffusion prior but also spatially and physically consistent with the surrounding environment and objects.

5. Experiments

In this section, we evaluate the effectiveness of our proposed MSDI framework for object aware 4D human motion generation. Our experiments include both qualitative and quantitative analyses, benchmarking against the state-of-the-art 4Dfy method. We report results across a suite of objective metrics designed to assess motion realism, diversity, and physical plausibility. Ablation studies further demonstrate the importance of main components within our pipeline.

5.1. Metrics

To quantitatively assess the quality of generated object-aware human motion, we adapt several metrics that collectively measure pose realism, motion diversity, and temporal dynamics. We evaluate using established metrics like Optical Flow Score [29] and we introduce three metrics designed to assess motion dynamics: Pose Plausibility, Pose Variation and Trajectory Length. We propose this suite of metrics because there is no single universally accepted metric to quantify the perceptual quality of human motion. It is important to consider these metrics in combination, as any individual metric can be trivially satisfied by a degenerate solution (e.g., a high trajectory score with a static, implausible pose). Since the metrics operate on different scales and cannot be combined arithmetically, their value lies in the holistic, comparative assessment of different methods. A model can only be judged to produce high-quality motion if it demonstrates strong and balanced performance across this entire suite.

For each generated video, we extract per-frame 3D human meshes using HMR2.0 [15], which estimates SMPL parameters [37] for every detected human instance using a

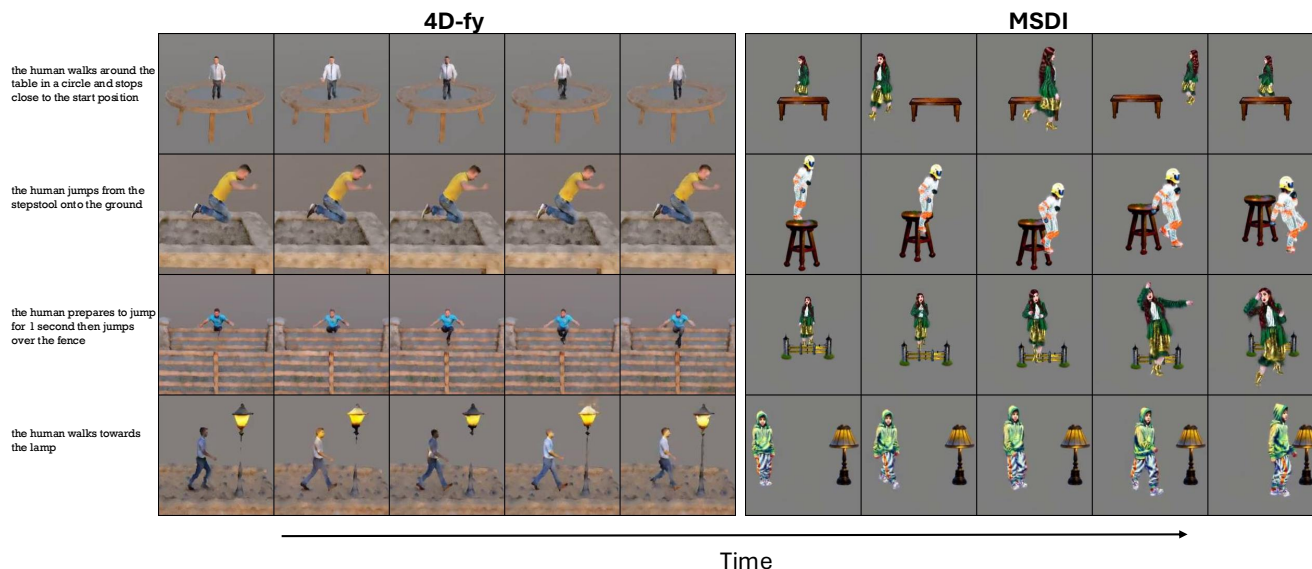


Figure 2. **Qualitative Results.** Generated videos from 4Dfy and MSDI across various text prompts. Each row corresponds to a different prompt. Within each row, columns display frames sampled at incremental timesteps from the generated video, illustrating temporal progression and motion characteristics. The frames are center cropped for better visibility.

ViTDet detector [24]. From these per-frame SMPL models and multi-view RGB frames, we compute the following metrics:²

Pose Plausibility. We evaluate the realism of each human pose using VPoser [36], a variational autoencoder trained on large-scale pose data. For each frame t , we convert the predicted SMPL body pose parameters θ_t^{pose} into VPoser-compatible axis-angle representation $\phi_t \in \mathbb{R}^{N_V \times 3}$, then encode these to obtain a posterior $q(z_t | \phi_t)$ over the latent pose space. The plausibility for each frame is quantified by the KL divergence to a standard normal prior $p(z)$.

Pose Variation. To quantify diversity and motion magnitude, we measure the temporal standard deviation of the pose vector ϕ_t (flattened dimension $K = N_V \times 3$) across all frames. High variation reflects diverse and dynamic motions.

Trajectory Length. To assess the extent of global character movement within the 3D space, we calculate the trajectory length of the root joint. For each frame t , HMR2.0 provides the 3D keypoint coordinates. The total trajectory length is the sum of Euclidean distances between the root joint positions in consecutive frames. A longer trajectory length suggests more substantial displacement of the character over time.

Optical Flow Score. To quantify the amount of motion and temporal dynamics, we compute an Optical Flow Score [29]. For each of the $N_{cv} = 4$ views, we estimate dense optical flow between consecutive frames using RAFT [44].

²Further details on metrics and user study are provided in the supplementary material.

The score for each view is the average magnitude of these flow vectors across all pixels and frames. The final Optical Flow Score is the average of these per-view scores. A higher score signifies a more pronounced motion.

User Study. To complement our quantitative metrics, we also conducted a user study to qualitatively assess the performance of our method against 4D-fy. The study was designed to measure human perception of motion quality, physical plausibility, and overall realism. We followed the human evaluation setup established by 4D-fy and MAV3D [42].

5.2. Results

We conduct a comprehensive set of experiments to evaluate our method. The results demonstrate that MSDI consistently outperforms baseline 4D-fy method overall across all metrics.

Quantitative and Qualitative Analysis. As shown in our quantitative analysis (Figure 3), MSDI shows a clear advantage. Specifically, it achieves substantially higher scores in both Pose Variation and Optical Flow, indicating diverse and larger motion. Furthermore, MSDI produces better Pose Plausibility and longer Trajectory Lengths for most prompts, while remaining comparable on others.

This numerical advantage translates directly to visually perceptible improvements, as shown in our qualitative comparisons in Figure 2. In contrast, 4Dfy often produces videos where frames appear largely similar, with only minor arm or leg movements, and the human subject frequently remains

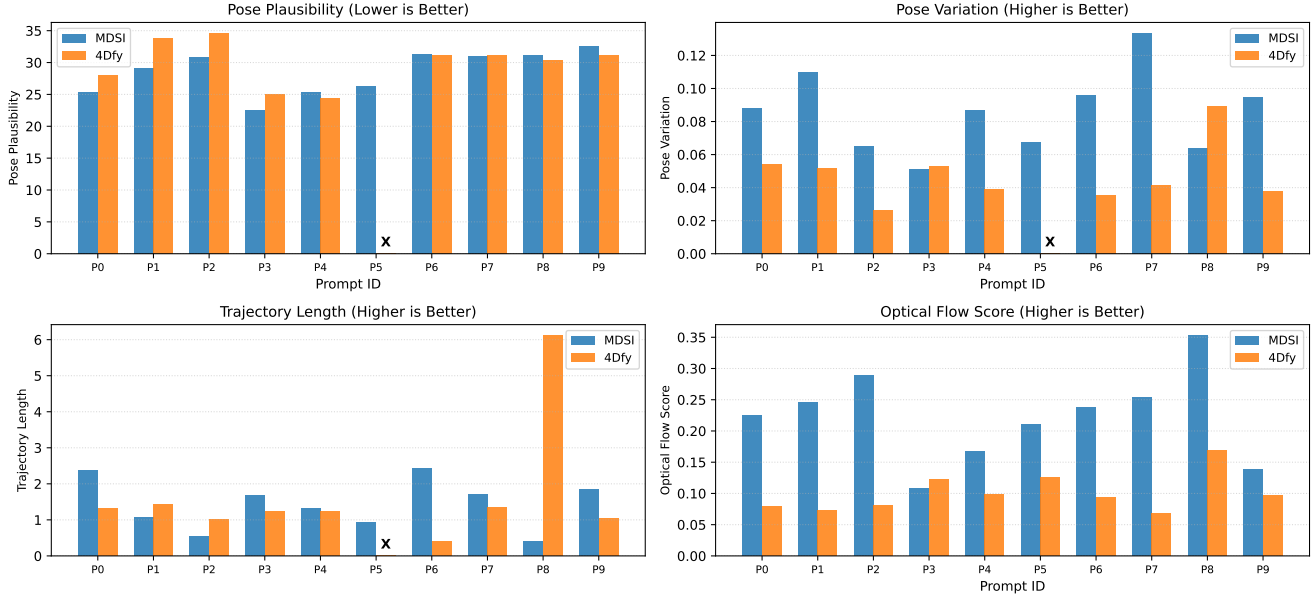


Figure 3. **Quantitative Results.** Quantitative comparison of MSDI and 4Dfy. The bar chart displays scores for 4 key metrics across 10 text prompts. An 'X' marker indicates that the metric failed to detect any humans in all four generated views for that particular prompt.

static in the same position. It’s limited human motion could be attributed to the inherent constraints of the underlying video diffusion model, such as VideoCrafters [8] used for Score Distillation Sampling. MSDI using Motion SDS generates coherent and physically-grounded interactions with human subject showing larger change in their positions over time.

User Study. To validate that our quantitative and qualitative findings align with human judgment, we performed a formal user study comparing MSDI against 4D-fy. The results, summarized in Table 1, show a decisive preference for MSDI across all categories. Crucially, the preference in **Motion Quality (MQ)** is an overwhelming **87%**, directly validating our core technical contribution. This result also confirms that the higher Pose Plausibility and Variation captured by our metrics correspond to motions that humans perceive as significantly more natural and realistic. The high preference for **Appearance (AQ)** (**79%**), **3D Structure (SQ)** (**71%**) and **Text Alignment (TA)** (**75%**) further suggests that physically plausible motion enhances overall visual fidelity and text alignment. The **80%** overall preference underscores that generating believable 4D videos hinges not just on appearance, but critically on the quality of the motion itself.

5.3. Ablation study

We conduct qualitative ablation studies to demonstrate the importance of key components in our proposed method. We focus on the prompt “the human jumps onto the table”

Preferred Method	AQ	SQ	MQ	TA	Overall
MSDI (%)	79%	71%	87%	75%	80%
4D-fy (%)	21%	29%	13%	25%	20%

Table 1. User study results comparing MSDI with 4D-fy. We report the percentage of times users preferred our method. MSDI significantly outperforms 4D-fy across all metrics, with all results being statistically significant ($p < 0.001$).

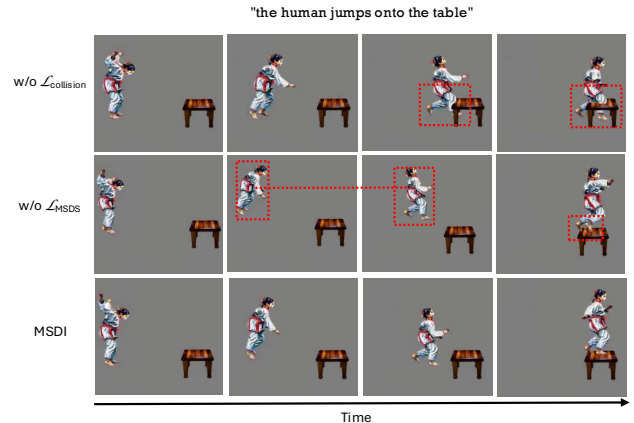


Figure 4. **Ablation study on key components of MSDI.** We visualize the impact of removing our main loss terms for the prompt “the human jumps onto the table”.

to highlight specific failure modes when certain losses are excluded. For all variants, the high-level LLM planned

trajectory remains consistent. Visualizations are provided in Figure 4.

Effect of Collision Loss ($\mathcal{L}_{\text{collision}}$). Without collision loss $\mathcal{L}_{\text{collision}}$, the optimization fails to enforce physical non-penetration constraints. The generated human visibly penetrates or pierces into the table surface during the landing phase of the jump.

Effect of Motion Diffusion Score Distillation Sampling ($\mathcal{L}_{\text{MSDS}}$). Excluding the MSDS $\mathcal{L}_{\text{MSDS}}$, significantly degrades the quality of the human motion and object interaction, particularly contact points. Without $\mathcal{L}_{\text{MSDS}}$, the human appears to float during the jump and makes unnatural contact, with improperly planted feet.

6. Conclusion

In this work, we introduced Motion Score Distilled Interaction (MSDI), a novel zero-shot framework for object-aware human motion generation. Our approach uniquely combines the strengths of 3D Gaussian representations for high-fidelity visuals, motion diffusion models for realistic human movement priors, and large language models for spatial reasoning and initial trajectory planning. A key component of our framework is Motion Diffusion Score Distillation Sampling (MSDS), which allows us to refine human motion by leveraging gradients from pre-trained motion diffusion models. This, coupled with our constrained optimization strategy that considers trajectory alignment, motion smoothness, and collision avoidance, enables the generation of interactions that are not only natural but also physically plausible and respectful of object presence.

Unlike previous methods that often require extensive training on specific datasets, MSDI operates in a zero-shot manner. This means it can generalize to novel interactions without retraining, making it a scalable and adaptable solution. Our experiments have shown that MSDI can produce realistic human motions interacting with static 3D objects, overcoming common issues like unnatural distortions and physical violations seen in outputs from methods relying solely on video diffusion models. We believe MSDI offers a promising direction for creating more dynamic and believable 4D content by effectively integrating 3D physical and semantic priors into the generation process.

References

- [1] Sherwin Bahmani, Jeong Joon Park, Despoina Paschali-dou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. *arXiv preprint arXiv:2206.14797*, 2022. 1
- [2] Sherwin Bahmani, Jeong Joon Park, Despoina Paschali-dou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. *arXiv preprint arXiv:2206.14797*, 2022. 2
- [3] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024. 1, 2
- [4] 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. 1, 2, 4
- [5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghai Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2
- [6] 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. 3
- [7] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1855–1866, 2024. 1, 2
- [8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 2, 7
- [9] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentrion: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024. 1, 2
- [10] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 2
- [11] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dream-scene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 1, 2
- [12] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 1, 2
- [13] Matt Deitke; Dustin Schwenk; Jordi Salvador; Luca Weihs; Oscar Michel; Eli VanderBilt; Ludwig Schmidt; Kiana Ehsani; Aniruddha Kembhavi; Ali Farhadi. Objaverse: A universe of annotated 3d objects. *IEEE*, 1314. 1, 2
- [14] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompt-

- ing: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 2
- [15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 5
- [16] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20418–20431, 2024. 2
- [17] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 $\{\deg\}$ dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 2
- [18] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 3
- [19] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Occlusion-aware video object inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14468–14478, 2021. 2
- [20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2
- [21] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 505–515, 2024. 2
- [22] Yao-Chih Lee, Yi-Ting Chen, Andrew Wang, Ting-Hsuan Liao, Brandon Y Feng, and Jia-Bin Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024. 1, 2
- [23] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2
- [24] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 6
- [25] Hanwen Liang, Yuyang Yin, Dejie Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 2
- [26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2
- [27] Xian Liu, Xiaohang Zhan, Jiayang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6646–6657, 2024. 1, 2, 3
- [28] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 2
- [29] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejiong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 5, 6
- [30] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1, 2
- [31] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2
- [32] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 1, 2
- [33] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 1, 2
- [34] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2019. 2
- [35] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024. 1, 2
- [36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 6
- [37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3, 5
- [38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2
- [39] Jiawei Ren, Liang Pan, Jiayang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2
- [40] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 4

- [41] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2
- [42] 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. 6
- [43] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 2, 3
- [44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6
- [45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 4
- [46] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 2
- [47] Kaihong Wang, Kumar Akash, and Teruhisa Misu. Learning temporally and semantically consistent unpaired video-to-video translation through pseudo-supervision from synthetic optical flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2477–2486, 2022. 2
- [48] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 1, 2
- [49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohe Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1, 2
- [50] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 2
- [51] Tianyi Xie, Zeshun Zong, Yuxin Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 1
- [52] Dejjia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024. 1, 2
- [53] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 1, 2
- [54] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Towards physically plausible video generation via vlm planning. *arXiv preprint arXiv:2503.23368*, 2025. 2
- [55] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion²: Dynamic 3d content generation via score composition of orthogonal diffusion models. *arXiv preprint arXiv:2404.02148*, 2024. 2
- [56] Taoran Yi, Jiemin Fang, Zanwei Zhou, Junjie Wang, Guan-jun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Xinggang Wang, and Qi Tian. Gaussiandreamerpro: Text to manipulable 3d gaussians with highly enhanced quality. *arXiv preprint arXiv:2406.18462*, 2024. 2
- [57] 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. 2
- [58] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024. 1, 2
- [59] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2
- [60] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1, 2
- [61] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 1, 2