YOGA SPLAT: AN ACTION CONDITIONED APPROACH TO 4D GAUSSIAN GENERALIZATION

Gaurav S. Gaonkar, Mindy Kim, Preetish Juneja, Bumjin Joo, Calvin Luo, Chen Sun
Department of Computer Science
Brown University
Providence, RI 02912, USA
{guarav_suhas_gaonkar, mindy_kim, preetish_juneja, bumjin_joo, calvin_luo, chen_sun4}@brown.edu

ABSTRACT

While video diffusion models can generate physically plausible motion through various means, their reliance on 2D representations limit their ability to capture complex 3D geometry and dynamics. 4D Gaussian Splatting (4DGS) addresses this by modeling dynamics scenes, but current models remain passive and lack support for controllable action-conditioned synthesis. In this work, we propose an action conditioned 4DGS framework with the ability to generate diverse, semantically-directed videos from a generalized model. Our results demonstrate that conditioning enables generalization to multiple actions within a single model, and that the learned deformation captures underlying motion structure rather than simply mimicking seen data. As a result, this work opens the door to controllable, physically grounded video generation from compact 3D representations.

1 INTRODUCTION

Recent advances in video generation have been driven by the success of video diffusion models, which have managed to achieve advanced high-fidelity video synthesis by learning to model spatiotemporal consistency. Models such as PhysDiff as introduced by Yuan et al. (2023) have demonstrated that injecting physical priors can enable a diffusion model to capture plausible physical motion using 2D images. However, since diffusion models are trained on information obtained from the 2D space (i.e. pixels from the images), they are fundamentally limited by their lack of access to explicit 3D scene geometry despite their abilities to model short-term dynamics well while maintaining coherence across frames.

In order to overcome potential barriers to physically accurate motion presented by operating in 2D space, the field has increasingly tried to incorporate methods that work with 3D scenes and representations that encode rich spatial information. One such technique that we found extremely promising is Gaussian splatting as popularized by Kerbl et al. (2023), which has recently emerged as a powerful and efficient method for rendering. Gaussian splatting represents scenes using collections of 3D Gaussians with learnable properties such as position, scale, and color that are optimized using stochastic gradient descent with a specified reconstruction loss.

Building on this framework, Yang et al. (2024) extend the 3D Gaussian representation by incorporating time as an explicit dimension, allowing each Gaussian to evolve smoothly over time. Recent works on 4D Gaussian Splatting (4DGS), including those discussed below, have shown impressive results in reconstructing dynamic scenes from both monocular and multi-view videos. However, little attention has been given to generalizing a single 4DGS model across different scenes. In the aforementioned work by Yang et al. (2024), each unique video or motion sequence requires training a new set of Gaussians from scratch. This limitation arises due to the tightly coupled optimization of the two core components of the 4DGS framework: the Gaussian model, which learns a static point cloud at time $t = 0$, and the Deformation Network, which models how the Gaussians deform over

time. Because these components are jointly trained and interdependent, the resulting model cannot generalize to new scenes without retraining.

Moreover, current 4DGS models are inherently passive, they can reconstruct or interpolate motion from observations but lack the capacity for controllable generation based on user-provided cues, such as actions. In contrast, the video generation literature has made significant strides in this direction, with several works demonstrating successful action-conditioned video generation using diffusion models. Motivated by this, we propose integrating action conditioning into the 4DGS framework. This enhancement would not only enable controllable video generation but also ensure that outputs remain grounded in a physically coherent 3D representation.

In this work, we take the first step towards bridging this gap by proposing an action conditioned approach to 4D Gaussian generalization by combining the spatiotemporal expressiveness of 4D Gaussians with the controllability found in diffusion-based generative models. The rest of the paper is organized as follows. Section 2 briefly introduces past works in literature that aid our methodology. Section 3 describes our additions and modifications to past works to achieve the desired goal. Section 4 discusses our results and Section 5 concludes our work as well as addresses potential future directions for additional research.

2 Related Works

3D Gaussians as a Scene Representation and Real-Time Rendering

A *3-D Gaussian* represents a scene element by a mean $\boldsymbol{\mu} \in \mathbb{R}^3$, a covariance decomposed into an orientation $R \in SO(3)$ and axis–aligned scale $\boldsymbol{\sigma}$, an RGB colour and an opacity weight. When rasterising, every ellipsoid is projected to an ellipse in the image plane; the resulting densities are blended in back-to-front order, yielding closed-form, differentiable colour and alpha. Compared with volumetric NeRFs, Gaussians require *no ray marching* and run at faster speeds while retaining photorealistic quality. Because each Gaussian is a learnable point, the cloud can be sparsified, densified or regularised (e.g. spatial sparsity, local rigidity, material priors) in a straightforward, gradient-based manner.

3D Gaussian splatting has significantly advanced the process of rendering when compared with its predecessor NeRF in terms of both quality, speed, and efficiency. As mentioned earlier in the paper, Kerbl et al. (2023) introduced the method for modern Gaussian splatting that represents scenes using 3D Gaussians. Lin et al. (2025) introduce a physics-inspired method that they call OmniPhysGS with the goal of producing more general and realistic physical dynamics with a focus on different types of materials.

4D Gaussian Splatting for Video Synthesis

To capture *dynamic* scenes, Yang *et al.* Yang et al. (2024) extend static 3-D Gaussian splatting into the **4-D domain**: each Gaussian is assigned a frame-wise rigid motion (rotation & translation), enabling time-varying view synthesis. Building on this idea, Liu *et al.* Wu et al. (2024) introduce a two-stage pipeline that *first* coarsely fits a static initial set of 3-D Gaussians and *then* jointly trains that cloud with a dedicated *Deformation Network*. The network predicts per-Gaussian changes in position, scale and orientation that transform the canonical composition at $t{=}0$ into any target timestep t. It follows an encoder–decoder design: a spatial–temporal encoder produces a single latent code, and a multi-headed decoder regresses each deformation modality independently. We visualize their architecture in Figure 1.

Subsequent work has pushed 4-D Gaussians toward editability and physical realism. SC-GS Huang et al. (2024) adds sparse user-defined control points that let artists manipulate dynamic content, while OmniPhysGS-4D Lin et al. (2025) couples Gaussians with constitutive material models to generate physically-plausible motion in heterogeneous substances.

Taken together, these advances showcase the evolution of Gaussian splatting from a fast static renderer into a versatile, learnable 4-D representation capable of real-time, physically-aware video synthesis. Our work follows this trajectory: we retain the coarse-to-fine strategy of Wu et al. (2024) but

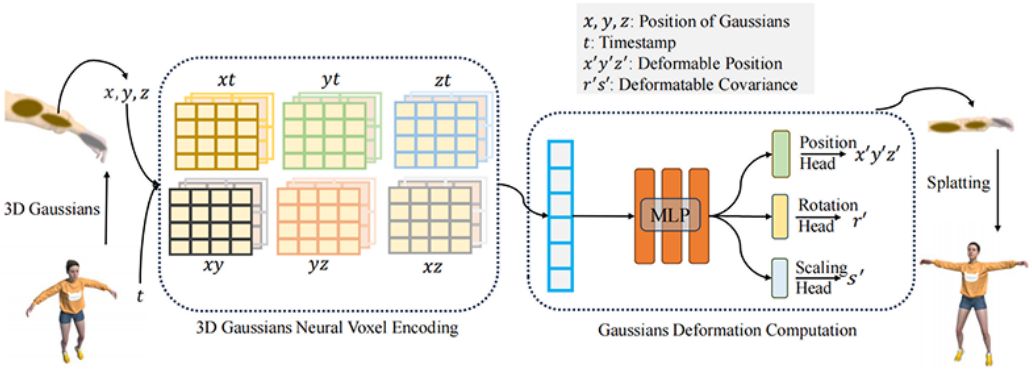


Figure 1: Overview of the deformation–aware 4-D Gaussian splatting pipeline of Wu et al. (2024).

further **jointly optimize a single deformation model across multiple video sequences**, enabling shared understanding of motion patterns that generalises beyond any individual clip.

Action-Conditioned Video Generation

As aforementioned, the generalization of a single 4D-Gaussian model to multiple scenes conditioned on **action** or **text-prompt** remains a relatively challenging and unexplored area.

Within the 2D diffusion model literature, we recall that works such Yuan et al. (2023) introduce PhysDiff a physics constrained based diffusion model, exploring the generation of physically plausible human motions through text-conditioning. Additionally, further advancements such as the Action-Conditioned Video Generation (ACVG) framework by Sarkar & Ghose (2024) in the realm of robotics further explores the potential for action-conditioning to improve training on various tasks within and outside robotics through generated video data. These works highlight two complementary insights: (i) conditioning on high-level actions or text greatly expands the controllability of generative models, and (ii) injecting even coarse physical priors helps maintain realism.

Our method to train generalizable 4D-Gaussians inherits both lessons: we condition a shared 4-D Gaussian deformation model on discrete action labels while regularising trajectories with local-rigidity constraints, enabling a single network to render diverse, physically-plausible motions across multiple scenes. We demonstrate its generalisation capability on a human-motion **Yoga** data set MOYO (Tripathi et al., 2023) and therefore refer to our model as **Yoga Splat**.

3 Methodology

Data

SMPL-X

In an effort to accurately model and analyze human actions, Pavlakos et al. (2019) introduce SMPL-X as a realistic and complex 3D human body model which is able to integrate body, hand, and face expressions in physically consistent manners. SMPL-X achieves high quality models by adding specialized body and joint prior objectives, as well as physical realism penalties.

MOYO

The **MOYO** data set Tripathi et al. (2023) offers a uniquely detailed motion-capture archive of yoga practice. It records *multi-view* video of a wide variety of poses performed by several distinct practitioners and, for every frame, supplies a fully optimised SMPL-X mesh together with per-frame physical metrics. This rich combination of geometry and appearance is precisely why we choose MOYO for our study: the high-fidelity meshes of extreme, self-occluding yoga postures form a demanding benchmark that pushes a dynamic 3-D representation to its limits.

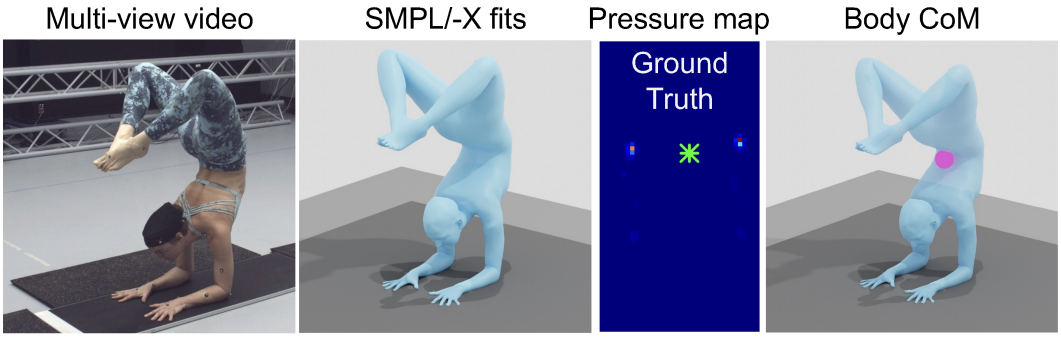


Figure 2: Richness of features of a single sample from the MOYO dataset by Tripathi et al. (2023)

Throughout our experiments we therefore use MOYO's SMPL-X meshes as *input geometry*. The data follow the D-NeRF directory structure, each sequence is discretized into time-stamped frames, and every frame is accompanied by camera intrinsics and extrinsics (represented by a single transformation matrix) and the corresponding RGB image. Having the camera angle, translation and rotation recorded for every u as well as aligning the mesh at the center of the image allowed us to render consistent synthetic views and to supervise the deformation network with precise ground-truth poses.

Evaluating Complexity of Action Conditioning

Sequential Fine Tuning

To assess the generalization ability of an action-conditioned 4-D Gaussian model, we conducted a transfer-learning experiment. We first trained the full 4DGS architecture—comprising a static 3-D Gaussian cloud $\mathcal{G}$ and a deformation network $\mathcal{D}$—on a yoga pose with action embedding $[0, 1]$. Following the schedule from the original paper, we trained $\mathcal{G}$ alone for 3,000 coarse iterations and then jointly optimised $\mathcal{G}$ and $\mathcal{D}$ for 20,000 fine iterations. To test generalization, we transferred the model to a second pose with similar initial yoga form and passed action embedding $[1, 0]$. We duplicated the learned point cloud to obtain $\mathcal{G}'$, reused the deformation network $\mathcal{D}$ from the previous training, and fine-tuned both components while skipping the coarse stage. However, after fine-tuning, the shared deformation model failed to accurately represent the first pose leading to early catastrophic forgetting with just 2 actions, suggesting that sequential fine-tuning of a shared deformation network is insufficient as shown in Figure 3. This result motivates our approach of jointly training on multiple sequences.

Joint Training without Action Condition

To investigate further, we trained a single 4DGS model on both scenes simultaneously, this time without any explicit action conditioning. In this setup, we initialised both scenes with the same Gaussian point cloud at $t = 0$, coarse-tuned the model using only scene 1, and then jointly fine-tuned the entire model on both scenes. However, in the absence of action conditioning, the model failed to accurately reproduce either of the motions as seen in Figure 4. This suggests that the deformation network struggled to disentangle the dynamics of the two sequences without explicit supervision, reinforcing the need for action-aware conditioning when learning generalizable motion representations.

We run this experiment to particularly demonstrate the necessity of jointly training both $\mathcal{G}$ and $\mathcal{D}$ upon a singular dataset.

Building an Action-Conditioned 4D Gaussian Splatting Model

As shown in Figure 5, we build upon the 4DGS pipeline outlined in Figure 1 by establishing a framework to incorporate action conditions in the latent representation within the Deformation Network. Our final iteration of the model adds the action embedding to the output of the Spatial-Temporal

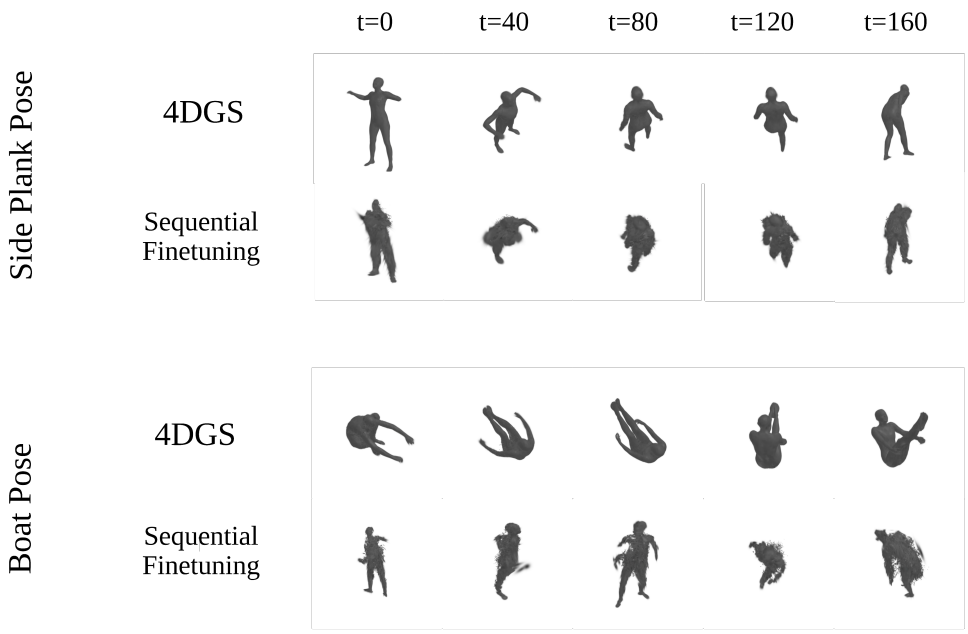


Figure 3: Rendering of Sequential Fine-tuning

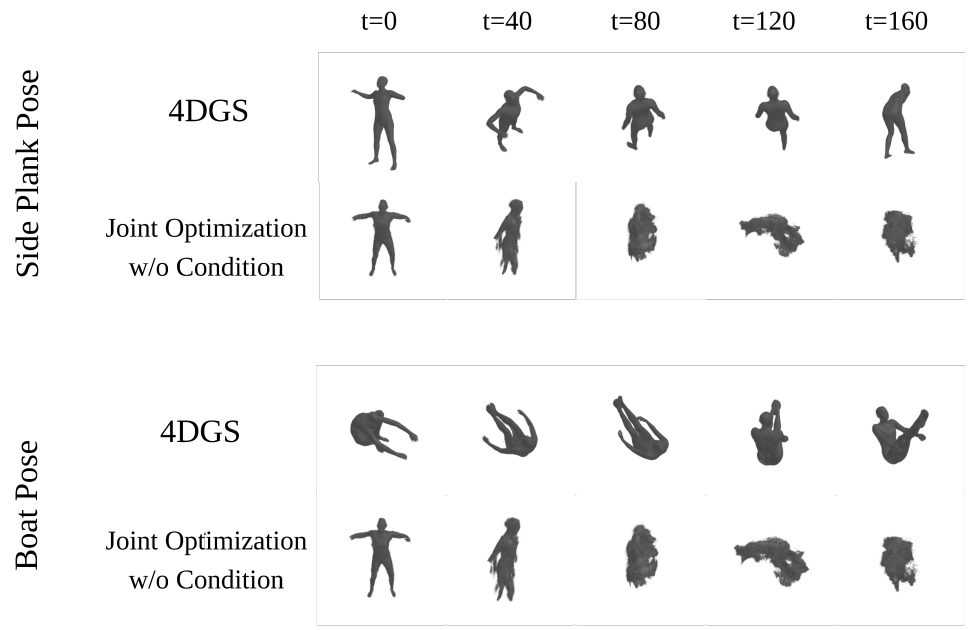


Figure 4: Rendering of Joint Optimization without Action

Encoder, which is made up of the 3D Gaussians' Neural Voxel Encoder and an MLP that downsamples the outputted encoding. This downsampled embedding is sent to each deformation head to calculate the transformation of position, scale, and orientation of the set of initial Gaussians. Dur-

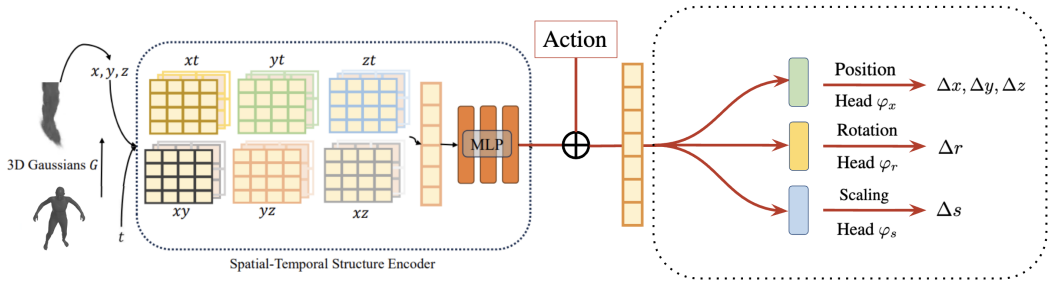


Figure 5: Overview of the proposed architecture that builds upon Figure 1

ing experimentation with this model, there were a couple obstacles we encountered in our attempts to condition the model with a given action, revealing potential limitations of the original model architecture. Firstly, the 4DGS model was extremely sensitive to changes in hyperparameters, so increasing the output latent size of the Spatial-Temporal encoding by a factor of 2 greatly decreased performance of the model and produced nonsensical outputs even without any changes made to the architecture itself. Thus, adding the action embedding directly to the output of the Neural Voxel encoding before sending it into the MLP feature encoder proved almost impossible, which we hypothesized was due to the sensitivity of the deformation heads to the input representation. Secondly, the initial plan was to embed the actions using a pretrained LLM since the hope was for the deformation network to be able break down the yoga poses into basic movements and potentially generalize to new movements. However, the embeddings ended up quite large and resulted in the model being unable to converge, resulting in the simple one-hot encoding representation of the different actions. This issue could also be traced back to the complexity of the MOYO dataset, so using a simpler dataset with less intricate movements may produce better results in the continuous action space 6.

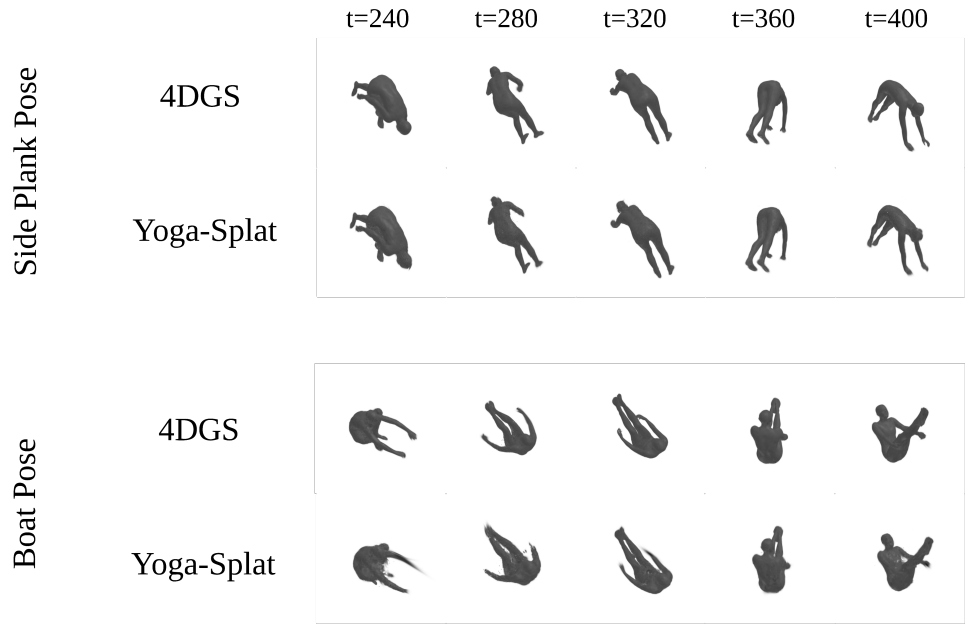
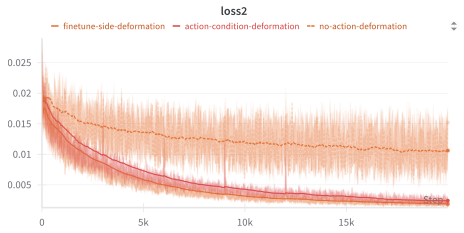


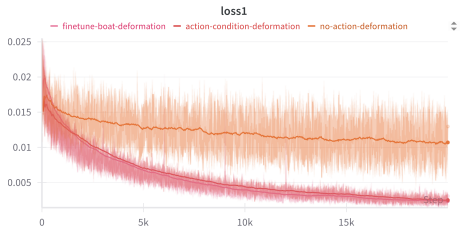
Figure 6: Comparison of Yoga Splat generated videos with 4DGS model instances trained on a single pose

4 RESULTS

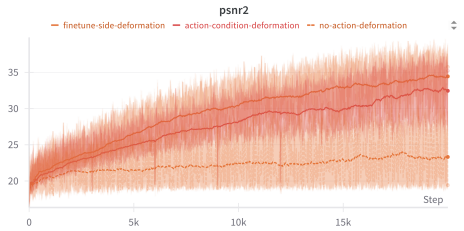
As discussed earlier, we conducted three experiments to evaluate the effectiveness of different training strategies: (1) sequential fine-tuning, (2) joint training without action conditioning, and (3) joint training with action conditioning. We visualise the training dynamics using loss and PSNR metrics in Figure **??**. In the sequential fine-tuning setup, we observed that although the PSNR increased and the loss remained low during fine-tuning on the second scene, the model exhibited catastrophic forgetting producing inaccurate and degraded renderings for the first scene. In the joint training without action conditioning, the model failed to generalize across both scenes, as evidenced by consistently high loss values and low PSNR, leading to poor rendering quality. In contrast, our proposed method—joint training with action conditioning—achieved the best results. It maintained low loss and high PSNR across both scenes, enabling the model to render accurate and high-quality outputs for both motion sequences. We were able to match the PSNR score of approximately 34 mentioned in the paper for good-quality rendering. This demonstrates the importance of explicit action conditioning for learning generalizable and faithful 4D Gaussian representations across multiple dynamic scenes.



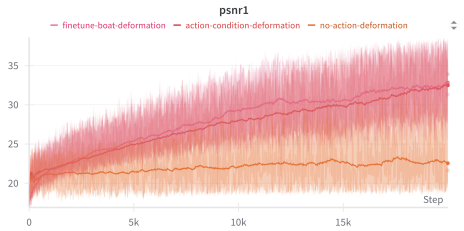
(a) Loss - Side Plank



(b) Loss - Boat Pose



(c) PSNR - Side Plank



(d) PSNR - Boat Pose

Figure 7: Quantitative comparison of rendering performance across two yoga poses: Side Plank and Boat Pose.

We also found that when the model is trained jointly with action conditioning, it begins to capture the semantics of each action. As a result, when presented with an out-of-distribution action embedding during inference, the model becomes confused and produces collapsed or structurally inconsistent deformations, as it attempts to simultaneously render aspects of both scenes as seen in Figure 8.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel action-conditioned 4D Gaussian framework which enhances the 4D Gaussian Splatting Model by Wu et al. (2024) to generate videos of mutliple distinct actions with a singular, unified model. To achieve this, we add an action condition into the deformation prediction heads to allow the model to learn action-dependent deformations on a consistent set of 3D Gaussians. Experiments without action conditions demonstrate the necessity of injecting this prior into model predictions for convergence onto different action poses. Furthermore, we validate the Action Conditioned 4D Gaussian model's conditioning on the action prior by demonstrating unexpected outputs on unseen actions.

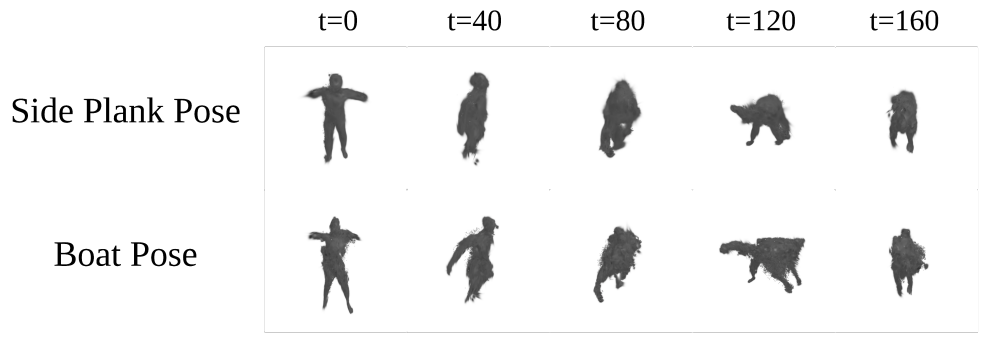


Figure 8: Renderings of Yoga Splat generated on OOD action conditions

In future works, further evaluation of adherence to physical plausibility would provide interesting insight into the model's capability of learning the laws of physics. Moreover, a natural extension of this work would be to consider continuous, general action spaces. For instance, utilizing LLM embeddings on a chain of thought instruction for a given action might provide a generalizable framework to visualize a broad variety of distinct actions.

6 Division of Labor

Our division of labor is summarized below

- Preetish was responsible for exploring prior work on video diffusion models, and worked on adapting the 3D Gaussian representation of data for use in our model.
- Bumjin focused on identifying suitable datasets and preprocessing them into mesh format. He also contributed significantly to the architectural engineering of the model.
- Gaurav played a central role in designing the overall approach, implementing core parts of the architecture, and conducting experimental evaluations.
- Mindy supported the engineering process by transforming mesh data into a format usable by the model and also contributed to running and monitoring experiments.

References

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes, 2024. URL https://arxiv.org/abs/2312.14937.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation, 2025. URL https://arxiv.org/abs/2501.18982.

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)*, pp. 10975–10985, 2019.

Meenakshi Sarkar and Debasish Ghose. Action-conditioned video data improves predictability, 2024. URL https://arxiv.org/abs/2404.05439.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4713–4725, 2023. URL https://ipman.is.tue.mpg.de.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, June 2024.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting, 2024. URL https://arxiv.org/abs/2310.10642.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model, 2023. URL https://arxiv.org/abs/2212.02500.