

Supplementary Material for SV-GS: Sparse View 4D Reconstruction with Skeleton-Driven Gaussian Splatting

Jun-Jee Chao
University of Minnesota
chao0107@umn.edu

Volkan Isler
The University of Texas at Austin
isler@cs.utexas.edu

In this supplementary material, we first quantify the difficulty of our downsampled datasets in Section 1. Then we provide additional implementation details in Section 2 and additional results in Section 3.

1. Dataset Analysis

To quantify the difficulty of our problem setup, we measure the Angular Effective Multi-View Factor (Angular EMF) [1] for both the original datasets and the downsampled versions used in our experiments. The Angular EMF measures the average angular change of the camera around a centered target per unit time, which is used as an approximation of how much multi-view information is available in a monocular video sequence. A higher value indicates more viewpoint variation within a short time period (an easier reconstruction problem), while a lower value corresponds to a more challenging setting. Note that this metric only considers the average camera motion but not the target motion, nor does it consider the number of observations and their temporal sparsity.

We compute the Angular EMF for the sequences in the original datasets and for our downsampled versions. Table 1 shows that the Angular EMF of our downsampled datasets is an order of magnitude lower than that of the original datasets, demonstrating the increased difficulty of our setting. Even without considering scene motion or the reduced number of observations, our setup is significantly more challenging than those used in most existing works as indicated by the Angular EMF metric.

Table 1. We present the Angular EMF metric [1] on both the original dataset and the downsampled dataset used in our experiments.

	D-NeRF [9]	DG-Meshh [3]	ZJU-MoCap [7]
Original	2132.3	2700.9	1395.2
Ours (downsampled)	217.8	265.4	271.7

2. Implementation Details

2.1. Initial Static 3D Reconstruction from multi-view images.

In the main experiments, we use the same set of multi-view images at the first time step for all methods to ensure a fair comparison. Since the original datasets do not provide such multi-view images, we render them from the training camera viewpoints using the pretrained checkpoints released by [14]. These multi-view images, together with the sparse temporal observations, are used to supervise the baselines throughout their entire optimization process. In contrast, our method only uses the multi-view images during initialization but not during the deformation optimization. The deformation field in our approach is learned solely from sparse observations.

In our initialization stage, we uniformly sample 3D points within the scene bounds as the initial Gaussian centers. We then follow the original 3DGS [2] optimization process, including Gaussian densification and pruning, for 80,000 steps using the multi-view images by minimizing $\mathcal{L}_{\text{perceptual}}$.

2.2. Replacing the Multi-View Initialization with a Pre-trained Generative Model.

Score Distillation Sampling (SDS). To relax the need for multi-view initialization at the first time step, we leverage a learned generative prior to provide an initial 3D reconstruction from a single image. We adopt the Score Distillation Sampling (SDS) technique introduced by DreamFusion [8], which optimizes a NeRF [5] representation under the guidance of a pretrained 2D diffusion model. The key idea is to align the rendered images from NeRF with the generative prior without direct multi-view supervision. Since 2D diffusion model does not guarantee to generate consistent 3D object across different viewpoints, the output cannot be directly applied to supervise a NeRF representation like the traditional multi-view setup. Instead of direct image supervision, SDS guides the NeRF optimization through gradients computed in the denoising space. Specifically, an im-

age rendered from NeRF is perturbed with Gaussian noise, mimicking the forward diffusion process. Then the noised image is passed into the pretrained diffusion model to predict the added noise. Finally, the difference between the predicted and actual noise is then backpropagated to update the NeRF parameters.

Image to 3DGS. Following [11], we adopt Zero-1-to-3 [4] as our 2D diffusion prior. Zero-1-to-3 [4] trains a 2D Diffusion model conditioned on a single input image and a relative camera pose to generate novel views with explicit camera control. To initialize the static 3DGS at the first time step from a single observation I^r , we optimize the Gaussian parameters with $\mathcal{L}_{perceptual}$ only at the corresponding viewpoint, and employ the \mathcal{L}_{SDS} [8] to guide the optimization for all other viewpoints without observations. Specifically, the SDS formulation is:

$$\nabla_{\mathcal{G}} \mathcal{L}_{SDS} = \mathbb{E}_{t,p,\epsilon} [w(t)(\epsilon_{\phi}(I^p; t, I^r, \Delta p) - \epsilon) \frac{\partial I^p}{\partial \mathcal{G}}] \quad (1)$$

where $t \sim \mathcal{U}[0.02, 0.98]$ is the diffusion time step sampled uniformly at random between 0.02 and 0.98, $w(t)$ is the weighting function from DDIM [10]. $\epsilon_{\phi}(\cdot)$ is the predicted noise from the pretrained Zero-1-to-3 [4], conditioned on the sampled time step t , reference image I^r and the relative camera pose Δp which transforms the reference viewpoint r to the rendering viewpoint p . We follow the configuration used in [11] and run the optimization for 1,000 steps to initialize the canonical static 3DGS.

2.3. Deformation Field Optimization.

Our learnable deformation field consists of the joint local pose predictor MLP_{Θ} , the bone influence radii r_j , the skinning correction field MLP_{Φ} , and the detail deformation field MLP_{Ψ} . All MLPs use 8 linear layers with a hidden dimension of 256. To stabilize optimization, the final layer weights of MLP_{Θ} and MLP_{Ψ} are initialized with a zero-mean Gaussian distribution ($\sigma = 1e-5$), and their biases are set to zero. This enforces the networks to predict small initial displacements from the canonical Gaussians, which correspond to the first time step. When optimizing these parameters, we keep the Gaussian parameters in the canonical frame fixed. In the multi-view initialization setup, we set $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 1$, and run the optimization for 40,000 steps with the ADAM optimizer [6] to minimize the total loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{perceptual} + \lambda_2 \mathcal{L}_{motion} + \lambda_3 \mathcal{L}_{detail} \quad (2)$$

In the experiment where the pre-trained diffusion model replaces the multi-view initialization, we add \mathcal{L}_{SDS} to Equation (2) with $\lambda_{SDS} = 1, \lambda_1 = 1e5, \lambda_2 = 2e4, \lambda_3 = 1$ for optimizing the deformation field. The reference image I^r in \mathcal{L}_{SDS} is updated at each iteration to the corresponding

observation of the sampled time step. The relative camera pose Δp is uniformly sampled from $[-180^\circ, 180^\circ]$ in azimuth and $[-30^\circ, 30^\circ]$ in elevation, keeping the same camera distance. We optimize the deformation field for 60,000 steps where the Gaussian parameters are fixed during the first 40,000 steps to stabilize motion estimation. Since the canonical Gaussians are initialized from only a single image at the first time step during the initialization stage, we allow the canonical Gaussians to update during the last 20,000 steps such that other observations can refine the canonical Gaussians after the motion estimation has converged.

3. Additional Results

Results with multi-view initialization at the first time step. In Fig. 1, we show the input training views at 6 sparse time steps, along with the corresponding renderings from all methods. To visualize the motion, we also render from a fixed front-view camera at those 6 time steps. While existing methods reproduce the training views well, their reconstruction and estimated motion suffer from artifacts when viewed from novel viewpoints, likely due to sparse observations and heavy occlusions. Similarly, we present more front-view qualitative results in Fig. 2 and Fig. 3.

Results without multi-view initialization. We present quantitative comparison on the *Jumpingjacks* scene from the D-NeRF dataset [9] in Table 2. All methods are evaluated with and without the multi-view initialization at the initial time step. By leveraging a pre-trained generative model, our method outperforms the baselines that have access to multi-view information, despite relying only on sparse observations with a single image per sparse time step. However, the diffusion-based initialization does not always produce good initial reconstruction, particularly when the reference image suffers from heavy self-occlusion or when the initial target pose involves significant part overlap (e.g., crossed limbs). In future work, we plan to explore skeleton-

Table 2. Quantitative results on the *Jumpingjacks* scene from D-NeRF dataset [9] downsampled at 0.1 intervals (11 frames). Results are reported for all methods with and without multi-view initialization at the first time step.

w/ multi-view initialization at the first time step			
Method	SSIM \uparrow	PSNR \uparrow	LPIPS ($\times 100$) \downarrow
4DGS [13]	0.881	21.19	8.83
SK-GS [12]	0.908	21.94	7.95
RigGS [14]	0.812	20.40	10.64
Ours	0.950	25.30	4.73
w/o multi-view initialization			
Method	SSIM \uparrow	PSNR \uparrow	LPIPS ($\times 100$) \downarrow
4DGS [13]	0.839	18.84	10.35
SK-GS [12]	0.861	19.25	12.11
RigGS [14]	0.819	19.67	14.04
Ours	0.935	22.62	6.58

conditioned generative models or category-specific priors to improve robustness.

References

- [1] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. [1](#)
- [2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [1](#)
- [3] Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. In *The Thirteenth International Conference on Learning Representations*, 2025. [1](#)
- [4] 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. [2](#)
- [5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. [1](#)
- [6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015. [2](#)
- [7] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021. [1](#)
- [8] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. [1](#), [2](#)
- [9] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. [1](#), [2](#)
- [10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [2](#)
- [11] Jiayang 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. [2](#)
- [12] Diwen Wan, Yuxiang Wang, Ruijie Lu, and Gang Zeng. Template-free articulated gaussian splatting for real-time reposable dynamic view synthesis. *Advances in Neural Information Processing Systems*, 37:62000–62023, 2024. [2](#)
- [13] 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)*, pages 20310–20320, 2024. [2](#)
- [14] Yuxin Yao, Zhi Deng, and Junhui Hou. Riggs: Rigging of 3d gaussians for modeling articulated objects in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5592–5601, 2025. [1](#), [2](#)

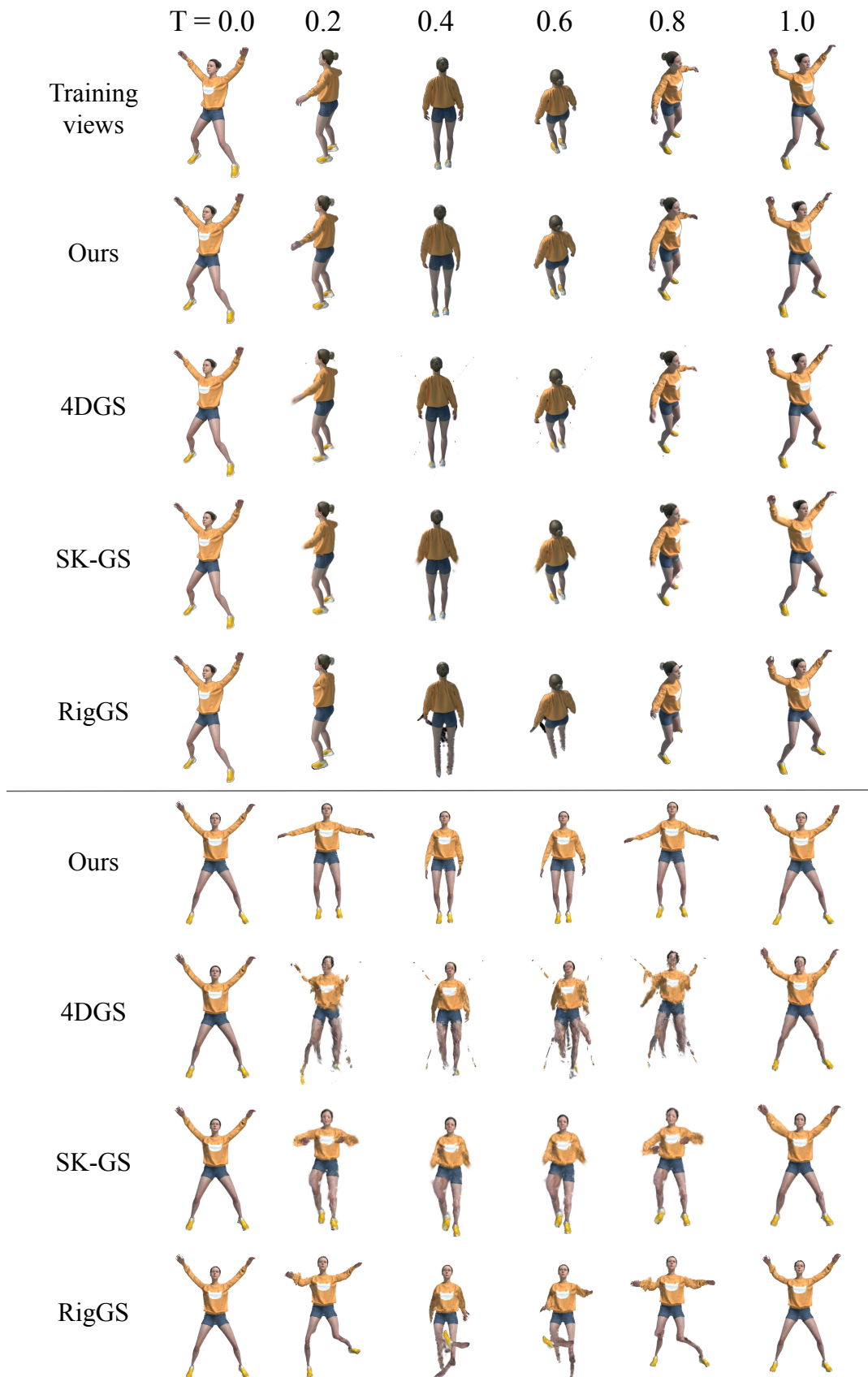


Figure 1. We present additional results on the Jumpingjack scene, showing the input training views and the rendered front views at 6 discrete time steps.

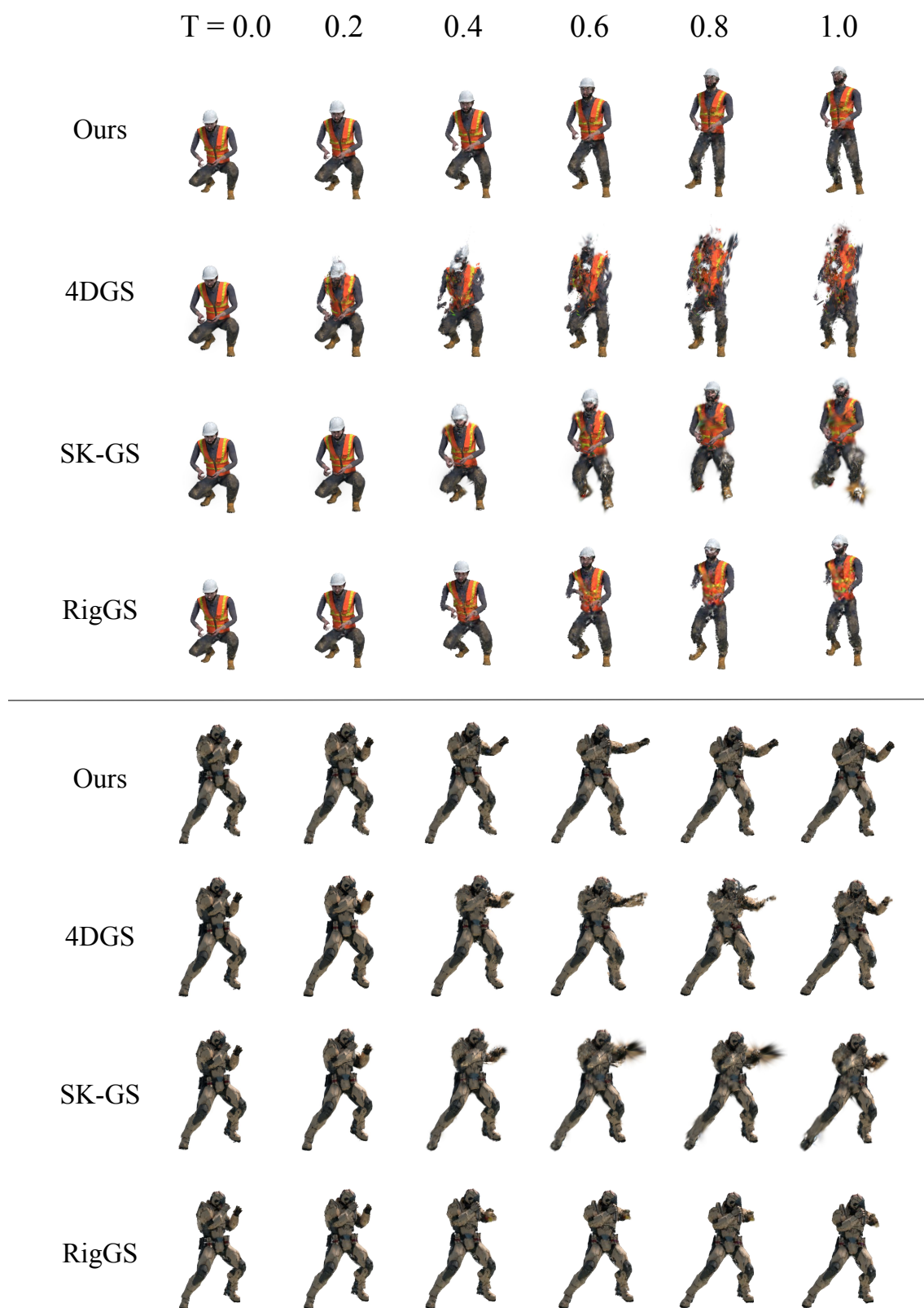


Figure 2. We present the front view rendering at 6 discrete time steps on the D-NeRF dataset.

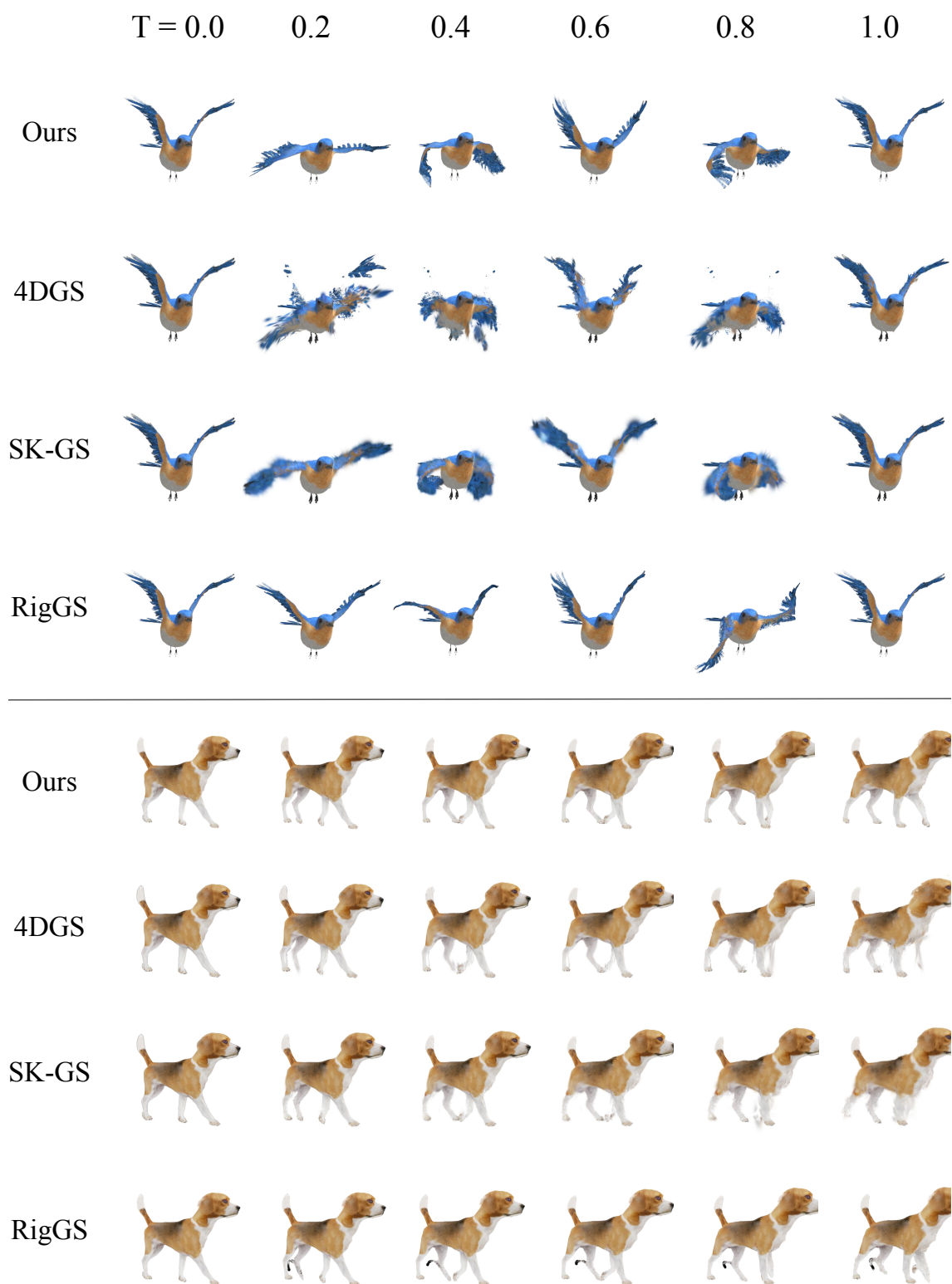


Figure 3. We present the front view rendering at 6 discrete time steps on the DG-Mesh dataset.