

Learning Explicit Continuous Motion Representation for Dynamic Gaussian Splatting from Monocular Videos

Supplementary Material

1. Additional Details

Initiation of motion bases. We use the 2D tracklets τ obtained from the pre-trained model [2] to initialize the motion bases.

$$\tau = [(p_t, v_t)]_{t=1}^{N_T}, \quad (1)$$

where p_t and v_t represent the position and visibility of the 2D tracklet, respectively. We lift the 2D tracklet position into 3D using depth reprojection. For training view v , given the 2D tracklet p_v , we obtain the corresponding 3D position q_v as follows:

$$q_v = P_v \pi_K^{-1}(p_v, D_v[p_v]), \quad (2)$$

where π_K^{-1} denotes the back-projection using the camera intrinsics K , P_v denotes the camera extrinsics of training view v , and D_v represents the depth map of the same view. Since the depth values at the locations corresponding to invisible 2D tracklets do not reflect the true depth after reprojection, these invisible tracklets degrade the initialization of the motion bases. To mitigate this issue, we replace the 3D reprojection values of invisible tracklets with those of the nearest visible 3D position q_v using linear interpolation.

Then, we initialize the orientation of all 3D points with the identity matrix I . Finally, we uniformly sample 3D points over time as control points for the SE(3) B-spline motion bases.

Details of multi-view diffusion model. We use Zero123-xl-diffusers (Stable-Diffusion v1.5) as the multi-view diffusion model, where the conditioning inputs consist of a reference image and the corresponding relative camera pose. Following DreamScene4D [1], the reference image contains only the foreground region of the input. We achieve view sampling by applying a small random perturbation to the center of the training camera.

Zero123-xl-diffusers can only provide an object-centric multi-view SDS prior, a naive application of object-centric multi-view SDS prior will inevitably suffer from a domain gap issue. However, we mitigate this by the following operations: (i) similar to DreamScene4D [1], we apply SDS loss only over foreground objects, thereby reducing the scene-level task to a foreground object modeling problem. (ii) we perform view sampling near the training views to ensure a reliable diffusion prior. (iii) we use rendered and reference images from the same time frame to increase geometric consistency of foreground objects.

Motion smoothness loss. Similar to MoSca [4], we maintain motion smoothness and propagate the visible informa-



Figure 1. Visual comparison of novel view synthesis.



Figure 2. Effect of pruning strategy.

Table 1. Ablation on the iPhone and NVIDIA datasets. “w/ Random” denotes randomly pruning a control point with error below ϵ_{prune} , while “w/ All” prunes all control points with errors below ϵ_{prune} .

Method	iPhone			NVIDIA		
	mPSNR \uparrow	mSSIM \uparrow	mLPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
w/ Random	19.84	0.721	0.283	27.59	0.853	0.062
w/ All	19.37	0.716	0.293	27.26	0.848	0.065
Ours	20.17	0.729	0.274	27.81	0.871	0.049

tion to the unknowns by optimizing a physics-inspired as-rigid-as-possible (ARAP) loss. Given two timestamps separated by a time interval Δ , we define the ARAP loss \mathcal{L}_{arap} as:

$$\mathcal{L}_{arap} = \sum_{m=1}^{N_m} \sum_{n \in K(m)} \left\| t_t^{(m)} - t_t^{(n)} \right\| - \left\| t_{t+\Delta}^{(m)} - t_{t+\Delta}^{(n)} \right\| + \left\| Q_t^{-1(n)} t_t^{(m)} - Q_{t+\Delta}^{-1(n)} t_{t+\Delta}^{(m)} \right\|, \quad (3)$$

where $n \in K(m)$ denotes that motion base n is one of the k -nearest neighbors of motion base m . The first term encourages the preservation of local distances within the neighborhood, while the second term preserves local coordinates by involving the local frame Q in the optimization. Simi-

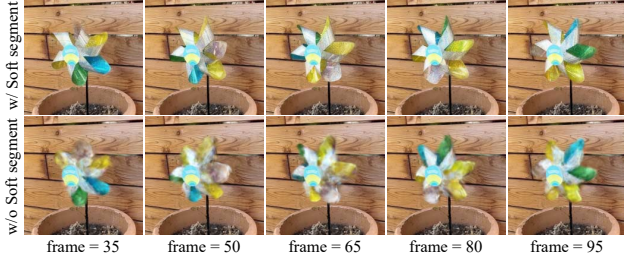


Figure 3. Effect of soft segmentation reconstruction.

lar to SoM [7], we maintain dynamic Gaussian deformation smoothness by optimizing the track loss. Specifically, we additionally render the 2D tracks $\hat{p}_{t \rightarrow t'}$ for a pair of randomly sampled timestamps t (query) and t' (target). We supervise these rendered correspondences with the lifted long-range 2D track:

$$\mathcal{L}_{track} = \|p_{t \rightarrow t'} - \hat{p}_{t \rightarrow t'}\|_2^2. \quad (4)$$

2. Additional Comparison Results

Comparison with Per-Point Deformation Methods. As shown in Figure 1, both our motion representation and per-point deformation field (PPDF) based methods (4DGaussians and MarbleGS) struggle to handle large non-rigid motion, but our method produces a slightly better result. The reason is that, our method allows us to fuse dynamic Gaussians across all frames to construct the target frame’s dynamic foreground, while PPDF is highly prone to overfitting to the training view despite its strong representation flexibility of motion.

More qualitative results. Figures 4, 5, 6, and 7 provide more visual comparison of novel view synthesis results on the iPhone [3] and NVIDIA [8] dataset. As shown, our method clearly outperforms previous methods.

3. Additional Ablation Results

Adaptive control mechanism. We conduct an ablation study to evaluate the effectiveness of our control point pruning strategy. Our method selects the control point with the smallest pruning error among those whose errors are below the threshold ϵ_{prune} . We compare it with two alternatives: (i) randomly pruning one control point whose pruning error is below ϵ_{prune} (w/ Random), and (ii) pruning all control points whose pruning errors are below ϵ_{prune} (w/ All). As shown in Table 1, our strategy achieves the best performance.

Soft segment reconstruction. We conduct an ablation study to evaluate the effectiveness of our soft segment reconstruction. As shown in Figure 3, soft segment reconstruction helps to reconstruct dynamic paper-windmill (cropped view) with better long-term consistency.

References

- [1] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dream-scene4d: Dynamic multi-object scene generation from monocular videos. *arXiv*, 2024. 1
- [2] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, Joao Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *ACCV*, 2024. 1
- [3] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *NeurIPS*, 2022. 2, 3, 4
- [4] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *CVPR*, 2025. 1, 3, 4, 5, 6
- [5] Yiming Liang, Tianhan Xu, and Yuta Kikuchi. Himor: Monocular deformable gaussian reconstruction with hierarchical motion representation. *CVPR*, 2025. 3, 4, 5, 6
- [6] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. *CVPR*, 2025. 3, 4, 5, 6
- [7] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *ICCV*, 2024. 2, 3, 4, 5, 6
- [8] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020. 2, 5, 6

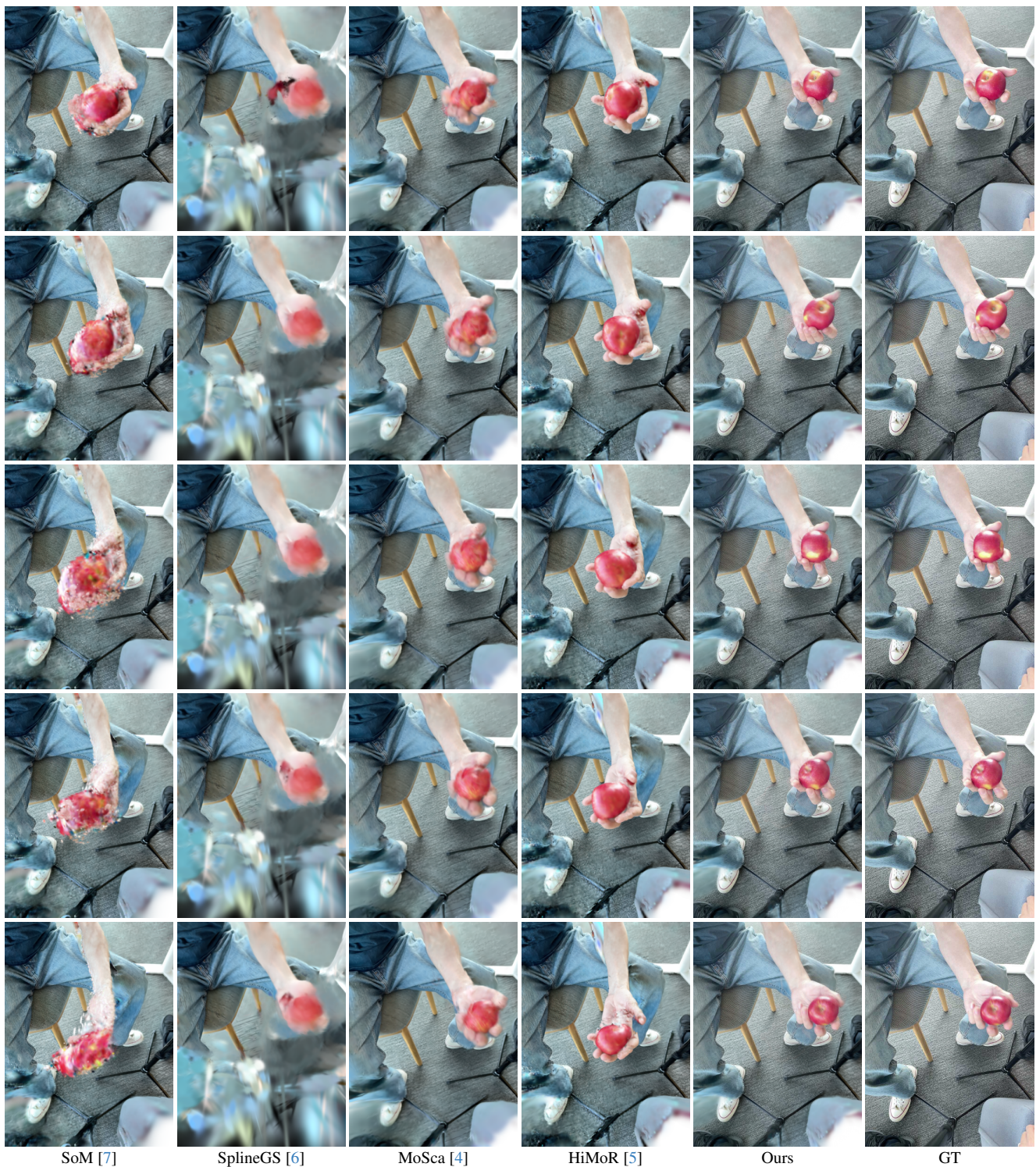


Figure 4. **Visual comparison of novel view synthesis on the scene “Apple” of the iPhone dataset [3].** The time interval between adjacent images is ten frames.

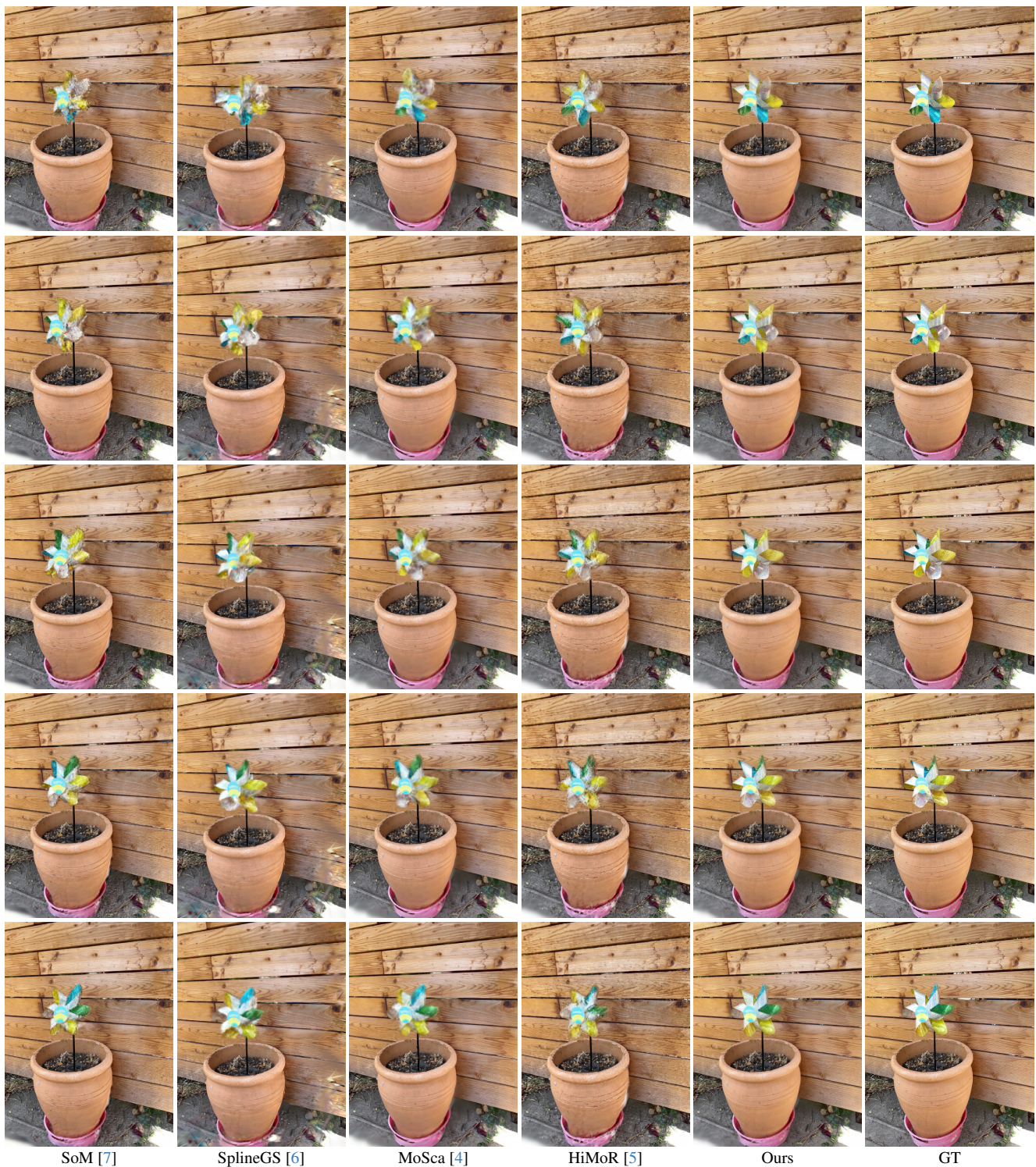


Figure 5. **Visual comparison of novel view synthesis on the scene “Paper-windmill” of the iPhone dataset [3].** The time interval between adjacent images is ten frames.

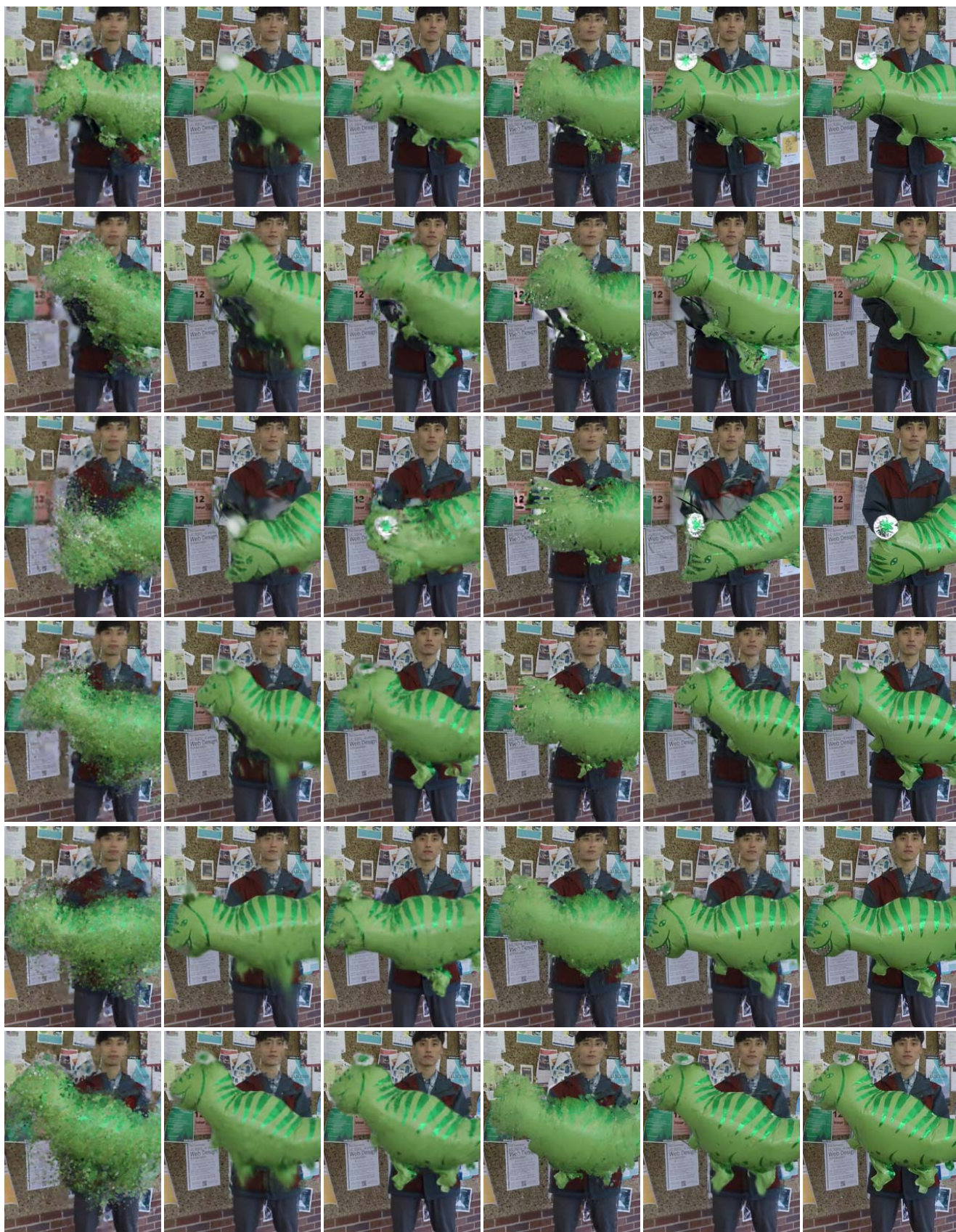


Figure 6. **Visual comparison of novel view synthesis on the scene “Balloon1” of the NVIDIA dataset [8].** The time interval between adjacent images is two frames.



SoM [7]

SplineGS [6]

MoSca [4]

HiMoR [5]

Ours

GT

Figure 7. **Visual comparison of novel view synthesis on the scene “Umbrella” of the NVIDIA dataset [8].** The time interval between adjacent images is two frames.