DynMF: Neural Motion Factorization for Real-time Dynamic View Synthesis with 3D Gaussian Splatting

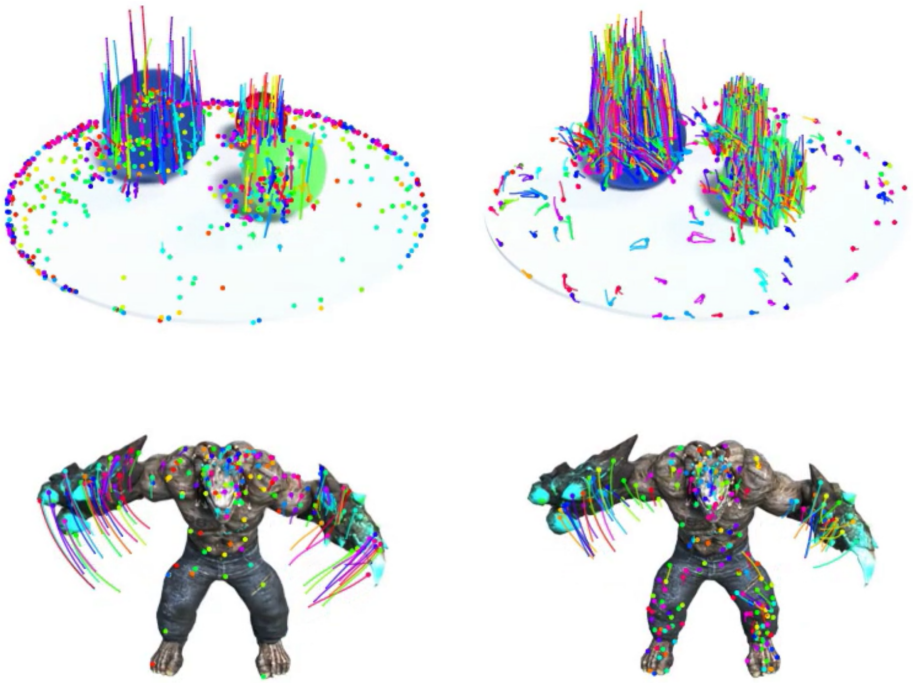
In this part, we present a few more important ablation studies that are essential for the understanding of our method's functionality. First and foremost, this section is accompanied by a video presentation that briefly explains the methodology of our framework and demonstrates a plethora of dynamic rendering results and ablations. This summarizes and demonstrates our results in rendering, tracking, and decomposition in the best possible way. Next, we offer a series of ablations studies. We show a few visual comparisons between using or not the $\mathcal{L}_1$ loss regularization on the motion coefficients, we present a big ablation on the usage of each one of the three aforementioned losses, and we finally visualize the 10 basis trajectories a scene is comprised of, in the case where they are linearly combined to produce motions, independently combine to produce motions (sparsity) or explicitly defined by the Fourier series. At the end of the supplementary material, we demonstrate a few more rendering results and present analytic quantitative results for every scene on each dataset for the three main metrics, PSNR, SSIM, and LPIPS.

1 Ablations

$\mathcal{L}_1$ **Loss Regularization:** As described in the main paper, Gaussians tend to choose a trajectory and move, even if they are part of the background. This is feasible without hurting the photometric accuracy. We regularize the motion coefficients to penalize unnecessary movements. Figure 1 and 2 show some qualitative comparisons between enforcing or not an $\mathcal{L}_1$ regularization. We observe that the $\mathcal{L}_1$ loss is enough to enforce points on the background to remain static during the process.

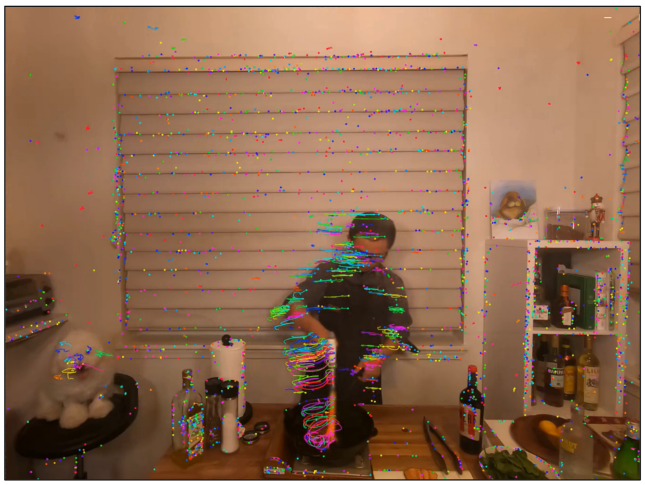
Rigid Loss: We strongly believe our method encourages locality and rigidity properties between points in the dynamic scene, given a small but adequate number of basis trajectories. We conduct experiments with and without the rigid loss and observe that the latter offers no additional expressiveness in the rendering quality of the dynamic scenes, while at the same time, it severely increases the training time especially when the number of Guassians is substantial. For the quantitative results refer to Table 1.

Fourier basis: Instead of having a learned basis, we decided to experiment with an explicit basis as well; namely the Fourier series. We demonstrated that Fourier can somehow model the complex motions of a dynamic scene, which is expected from a global trajectory approximator. Nonetheless, it falls behind in terms of

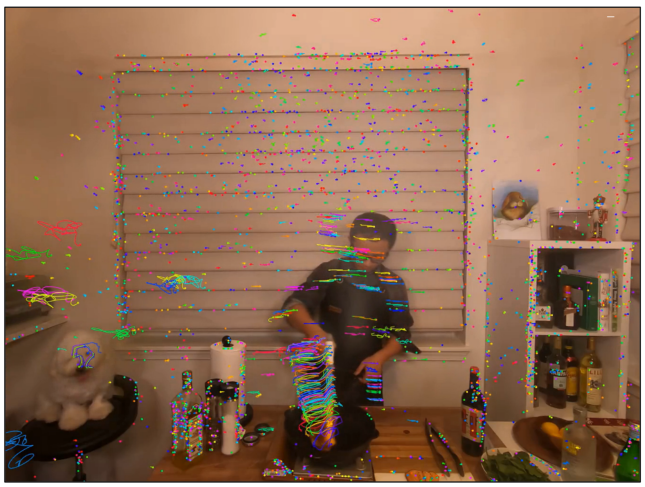


(a) with $\mathcal{L}_1$ Loss (b) without $\mathcal{L}_1$ Loss

Fig. 1: Qualitative comparison between using or not the $\mathcal{L}_1$ loss for regularizing the motion coefficients of each Gaussian. Results for the D-NeRF dataset.



(a) with $\mathcal{L}_1$ Loss



(b) without $\mathcal{L}_1$ Loss

Fig. 2: Qualitative comparison between using or not the $\mathcal{L}_1$ loss for regularizing the motion coefficients of each Gaussian. Results for the N3DV dataset.

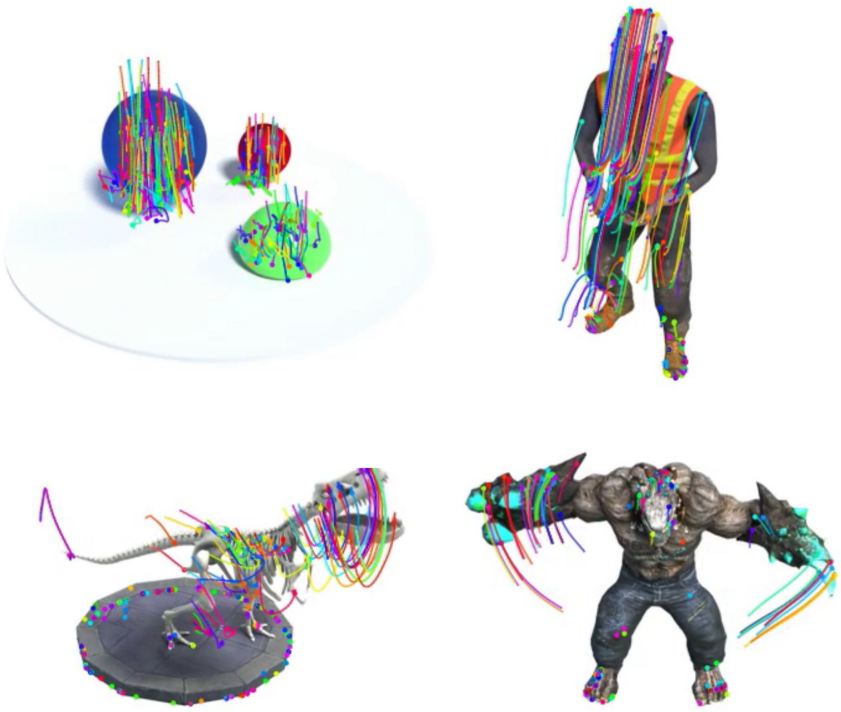
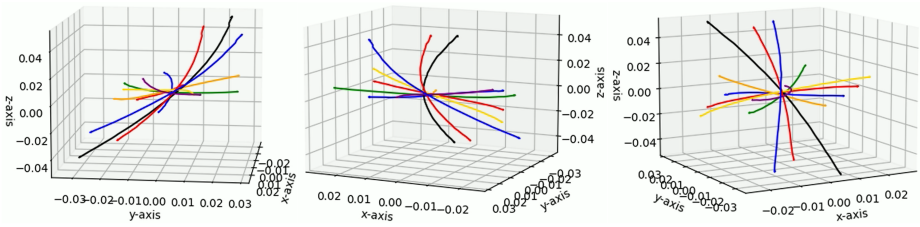
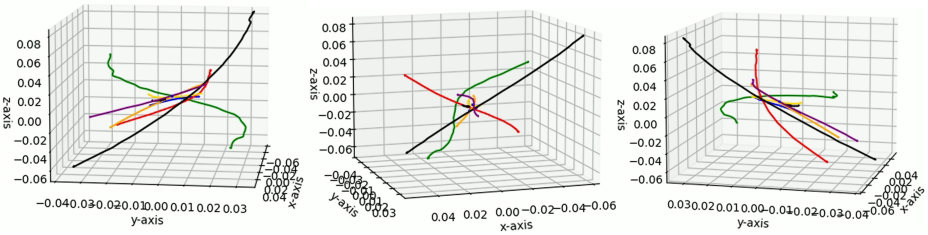


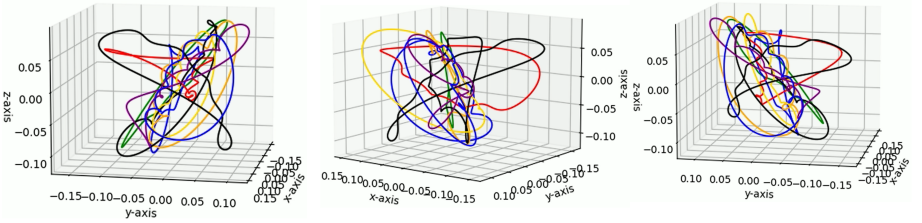
Fig. 3: More visualizations on per-point trajectory tracking for the synthetic D-NeRF dataset.



(a) Linear combination of basis trajectories per Gaussian



(b) Sparse selection of basis trajectories per Gaussian



(c) Explicit Fourier basis trajectories per Gaussian

Fig. 4: Visualization of the basis trajectories' trace, along time, in the 3D world space. The results refer to the 'Mutant' scene on the D-NeRF dataset. (a) The 10 basis trajectories without applying a sparsity loss and allowing each Gaussian to linearly combine all 10 of them to model its motion. (b) The 10 basis trajectories with the application of the sparsity loss, to enforce each Gaussian to choose only one trajectory for its motion. (c) The 10 basis trajectories explicitly modeled by the Fourier series instead of a smalled learned MLP.

expressivity and rendering quality compared to a learned basis through a small MLP. Figure 4 depicts the learned or explicit trajectories that can model the dynamics of a scene. Specifically, each Gaussian can choose one or more of these 10 trajectories to model its unique motion in the dynamic scene. When we allow a linear combination of the 10 trajectories, then the basis functions are uniformly spread in the 3D world as depicted in Figure 4 (a). This is because each Gaussian can model its unique motion by linearly combining these motions, which lets it uniformly move in the space of the dynamic scene. If we restrict each Gaussian to choose only one trajectory, then these become more odd and specific to the motion needs of the corresponding scene, as in Figure 4 (b). Finally, in Figure 4 (c) we clearly see the periodic trace of the Fourier basis functions that are explicitly defined instead of learned by a small efficient MLP. We hypothesize that the complexity of the formed motion produced by the Fourier signals does not allow a sufficient convergence to an expressive enough solution. To model the roughly 3D linear trajectories produced by the MLP a much higher number of frequencies would be needed, making the optimization problem even more challenging.

	PSNR ↑								
DynMF	Hell Warrior	Mutant	Hook	Balls	Lego	T-Rex	Stand Up	Jumping Jacks	Mean
B=10 w $\mathcal{L}_{decomp}$	31.60	40.01	29.89	41.05	24.49	32.93	38.52	33.61	34.01
B=64 w $\mathcal{L}_{decomp}$	34.70	40.16	31.83	42.14	25.58	33.96	38.97	35.53	35.36
B=10 w $\mathcal{L}_1$	36.60	41.00	31.30	41.01	25.27	35.10	41.16	35.75	35.90
B=64 w $\mathcal{L}_1$	**37.51**	**41.68**	33.91	**41.95**	**25.51**	**35.82**	41.00	37.74	**36.89**
B=10 w $\mathcal{L}_{rigid}$	36.10	40.11	32.01	39.85	25.20	34.89	40.50	34.74	35.43
B=64 w $\mathcal{L}_{rigid}$	37.10	40.64	33.15	41.41	25.31	34.99	40.79	37.41	36.35
B=10 w no Losses	36.51	40.79	31.36	40.03	25.29	35.08	41.10	35.74	35.74
B=64 w no Losses	37.31	41.32	**33.94**	41.95	25.37	35.19	**41.16**	**38.04**	36.79

Table 1: Ablation study on the $\mathcal{L}_1$ regularization loss, on the $\mathcal{L}_{decomp}$ sparsity loss and on the $\mathcal{L}_rigid$ rigid loss. B refers to the number of basis trajectories for the scene.

References

1. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. CVPR (2023)
2. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes (2024)
3. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers (2022)
4. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: CVPR (2023)

Table 2: Per-scene quantitative comparisons on D-NeRF monocular synthetic dynamic scenes.

Method	Hell Warrior			Mutant			Hook			Bouncing Balls		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
D-NeRF [9]	25.02	0.95	-	31.29	0.97	-	29.25	0.96	-	32.80	0.98	-
K-Planes [4]	25.70	0.952	-	33.79	0.983	-	28.50	0.954	-	41.22	0.992	-
TiNeuVox [3]	28.17	0.97	0.07	33.61	0.98	0.03	31.45	0.97	0.05	40.73	0.99	0.04
V4D [5]	27.03	0.96	-	36.27	0.99	-	31.04	0.97	-	**42.67**	0.99	-
[6]	34.15	0.948	-	37.45	0.982	-	33.19	0.967	-	33.29	0.983	-
[12]	36.85	-	-	39.26	-	-	33.33	-	-	36.30	-	-
4DGS [11]	28.71	0.973	0.04	37.59	0.988	0.02	32.73	0.976	0.03	40.62	0.994	0.02
DynMF (ours)	**37.51**	**0.975**	**0.02**	**41.68**	**0.996**	**0.01**	**33.91**	**0.979**	**0.02**	41.95	**0.994**	**0.01**

Method	Lego			T-Rex			Stand Up			Jumping Jacks		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
D-NeRF [9]	21.64	0.83	-	31.75	0.97	-	32.79	0.98	-	32.80	0.98	-
K-Planes [4]	25.48	0.948	-	31.79	0.981	-	33.72	0.983	-	32.64	0.977	-
TiNeuVox [3]	25.03	0.92	0.07	32.70	0.98	0.03	35.43	0.99	0.02	34.23	0.98	0.03
V4D [5]	**25.62**	**0.95**	-	34.53	0.99	-	37.20	0.99	-	35.36	0.99	-
[6]	22.21	0.837	-	26.22	0.964	-	39.10	0.987	-	30.95	0.970	-
[2]	25.24	-	-	31.24	-	-	38.89	-	-	33.37	-	-
4DGS [11]	25.03	0.938	0.04	34.23	0.985	0.02	38.11	0.990	0.01	35.42	0.986	0.01
DynMF (ours)	25.51	0.941	**0.03**	**35.82**	**0.991**	**0.01**	**41.00**	**0.993**	**0.01**	**37.74**	**0.992**	**0.01**

Method	Split-Cookie			Lemon			Chicken			3D Printer		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HyperNeRF [8]	31.5	-	0.220	31.4	-	0.230	**28.8**	-	**0.145**	20.2	-	**0.109**
TiNeuVox [3]	28.4	0.801	0.281	28.0	0.811	0.271	28.3	0.778	0.371	22.8	0.611	0.542
DynMF (ours)	**32.5**	**0.928**	**0.141**	**32.3**	**0.920**	**0.144**	27.7	**0.829**	0.254	**23.1**	**0.698**	0.301

Table 3: Per-scene quantitative comparisons on HyperNeRF monocular real dynamic scenes.

Table 4: Per-scene quantitative comparisons on N3DV multi-view real dynamic scenes.

Method	Coffee Martini			Cook Spinach			Cut Beef		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DyNeRF [7]	-	-	-	-	-	-	-	-	-
MixVoxels-L [10]	29.36	0.946	-	31.61	0.965	-	31.30	0.965	-
K-Planes [4]	**29.99**	**0.943**	-	32.60	**0.966**	-	31.82	**0.966**	-
Hexplane [1]	-	-	-	32.04	-	**0.08**	32.55	-	**0.08**
4DGS [12]	28.33	-	-	**32.93**	-	-	**33.85**	-	-
DynMF (ours)	29.26	0.931	0.156	32.56	0.960	0.117	32.50	0.963	0.117

Method	Flame Salmon			Flame Steak			Sear Steak		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DyNeRF [7]	29.58	0.961	-	-	-	-	-	-	-
MixVoxels-L [10]	29.92	0.945	-	31.21	0.970	-	31.43	0.971	-
K-Planes [4]	**30.44**	**0.945**	-	32.38	0.970	-	32.58	**0.974**	-
Hexplane [1]	29.47	-	**0.08**	32.08	-	**0.07**	32.39	-	**0.07**
4DGS [12]	29.38	-	-	**34.03**	-	-	**33.51**	-	-
DynMF (ours)	30.11	0.937	0.149	32.78	**0.970**	0.103	32.69	0.972	0.108

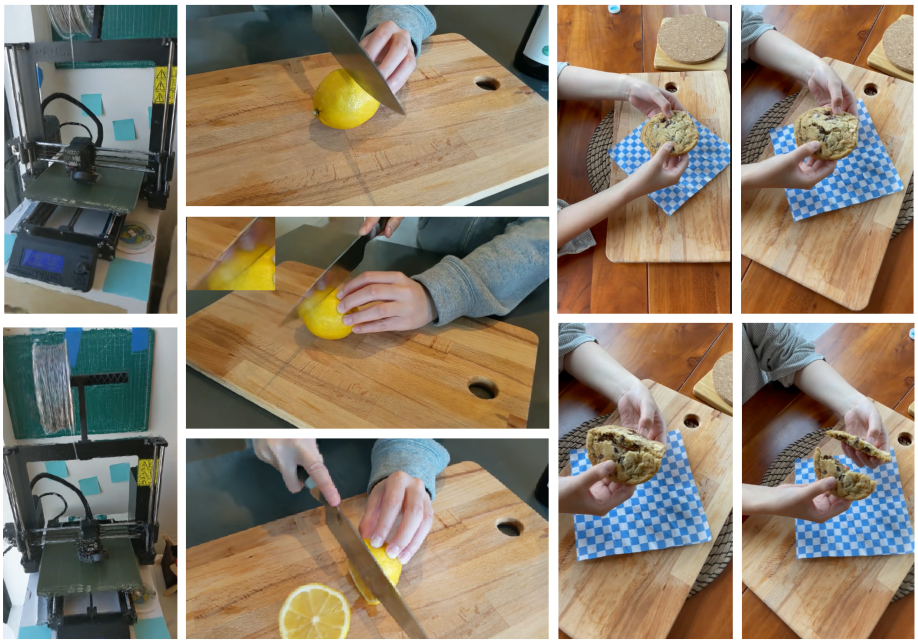


Fig. 5: More qualitative rendering results on the HyperNeRF dataset from novel test views.

Fig. 6: More qualitative rendering results on the N3DV dataset from novel test view.

5. Gan, W., Xu, H., Huang, Y., Chen, S., Yokoya, N.: V4d: Voxel for 4d novel view synthesis (2022)
6. Katsumata, K., Vo, D.M., Nakayama, H.: An efficient 3d gaussian representation for monocular/multi-view dynamic scenes (2023)
7. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., Lv, Z.: Neural 3d video synthesis from multi-view video. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5511–5521. IEEE Computer Society, Los Alamitos, CA, USA (jun 2022)
8. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021)
9. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
10. Wang, F., Tan, S., Li, X., Tian, Z., Liu, H.: Mixed neural voxels for fast multi-view video synthesis. arXiv preprint arXiv:2212.00190 (2022)
11. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Xinggang, W.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
12. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv 2310.10642 (2023)