ImViD: Immersive Volumetric Videos for Enhanced VR Engagement

Supplementary Material

A. Overview

Within the supplementary material, we provide:

- A more detailed introduction and analysis of existing datasets for Dynamic Novel View Synthesis (NVS) tasks in Appendix B.
- More benchmark results and discussion in Appendix C.
- Additional experiments details and STG++ implement details in Appendix D.
- Some clarifications and more descriptions of technical details regarding capture rig in Appendix E.
- Real-time immersive volumetric video demos and other dynamic scene reconstruction results are in our video. Please refer to the attached .mp4 file.

B. Comprehensive Summary of Datasets for Dynamic Novel View Synthesis Tasks

The earliest studies on dynamic reconstruction have naturally focused on human digital avatars. Datasets such as Human3.6M [18], Panoptic Sports [19], ZJ-Mocap [44], and Tensor-4D [48] primarily focus on depicting simple human actions but do not include backgrounds, which is crucial to the immersive application experiences. We will introduce more complex datasets that include environments from monocular based and multi-view based.

Monocular acquisition systems are popular due to their low cost and ease of construction. Datasets such as HyperNeRF [43], Dynamic Scene Dataset [60], and D2NeRF [56] use a mobile phone as devices, capturing dynamic scenes by waving the phone. However, these datasets suffer from resolutions below 1080p, limited capture space (similar to fixed-point shooting), and durations under one minute. Although NeRF On-the-go [45] allows for larger capture ranges by walking while shooting, high-quality reconstructions are confined to the vicinity of the capture path, and the small field of view (FOV) limits prolonged observations of specific scene positions.

Multi-camera data collection has gained significant attention due to its ability to provide a larger FOV and richer details. For instance, the Immersive Light Field dataset [6] employs 46 cameras to capture 15 indoor and outdoor scenes, while Technicolor [46] uses a 4×4 camera rig for 12 indoor sequences. The UCSD Dynamic Scene Dataset [34] consists of 96 outdoor videos focused on single-person activities captured by 10 cameras. The Plenoptic Dataset [28] uses 21 cameras for 6 indoor scenes. Similarly, datasets like [27, 33, 54] utilize 13, 18, and 24 cameras, respectively, to capture dynamic scenes. However, all these setups remain static during capture, limiting them to frontal views and hindering 360° reconstruction. Additionally, the video sequences are typically short, with a maximum duration of 2 minutes (often less than 30 seconds) and a maximum resolution of 3840×2160, which is insufficient for immersive VR experiences.

Moreover, the previously mentioned datasets, whether monocular or multi-view, lack sound recordings, despite the importance of multimodality for immersion. The Replay dataset [49] addresses this by focusing on long sequences with professional actors in familiar settings. It employs a ring of 8 static DSLR cameras paired with binaural microphones and 3 head-mounted GoPro cameras, providing 46 videos at 4K. However, aside from the head-mounted cameras, which can rotate slightly with head movements, all other cameras remain static. Furthermore, the DSLR arrangement does not align with human viewing habits in VR, making them unsuitable as benchmarks for novel view synthesis tasks. The latest work [9] presents a dataset of 28 scenes captured with a 360° camera, each including multiple audio and video sequences. However, this dataset is constrained by a fixed-point shooting strategy, resulting in sparse viewpoints that hinder the reconstruction of high-quality dynamic scenes. Further comparisons between our work and these datasets can be found in Tab. 8.

C. More Benchmark Results and Analysis

For fair evaluation, all of our experiments use the default parameters recommended by these works.

C.1. Quantitative & Qualitative Results

Quantitative Results. Table 6 is an extension of Table 5. Table 7 displays the train view performance of baselines and STG++ on each 60-frame segment across different scenes.

Qualitative Results. Figure 8 shows the results of the test view, while Figure 9 presents more results of the train view. Here, we also include a comparison with the latest work Ex4DGS [*], whose code was released shortly before our submission, limiting our ability to investigate thoroughly. Figure 10 presents the Ex4DGS's results from the same viewpoints as Figure 5 showed in the main text. It performs slightly worse in the challenging motion area due to incomplete dynamic and static partitioning.

C.2. Analysis

4DGS proposes a two-stage training approach. In the first stage, the algorithm initializes a static scene using the 0th

[*]Lee J et al. Fully Explicit Dynamic Gaussian Splatting. Advances in Neural Information Processing Systems, 2024, 37: 5384-5409.

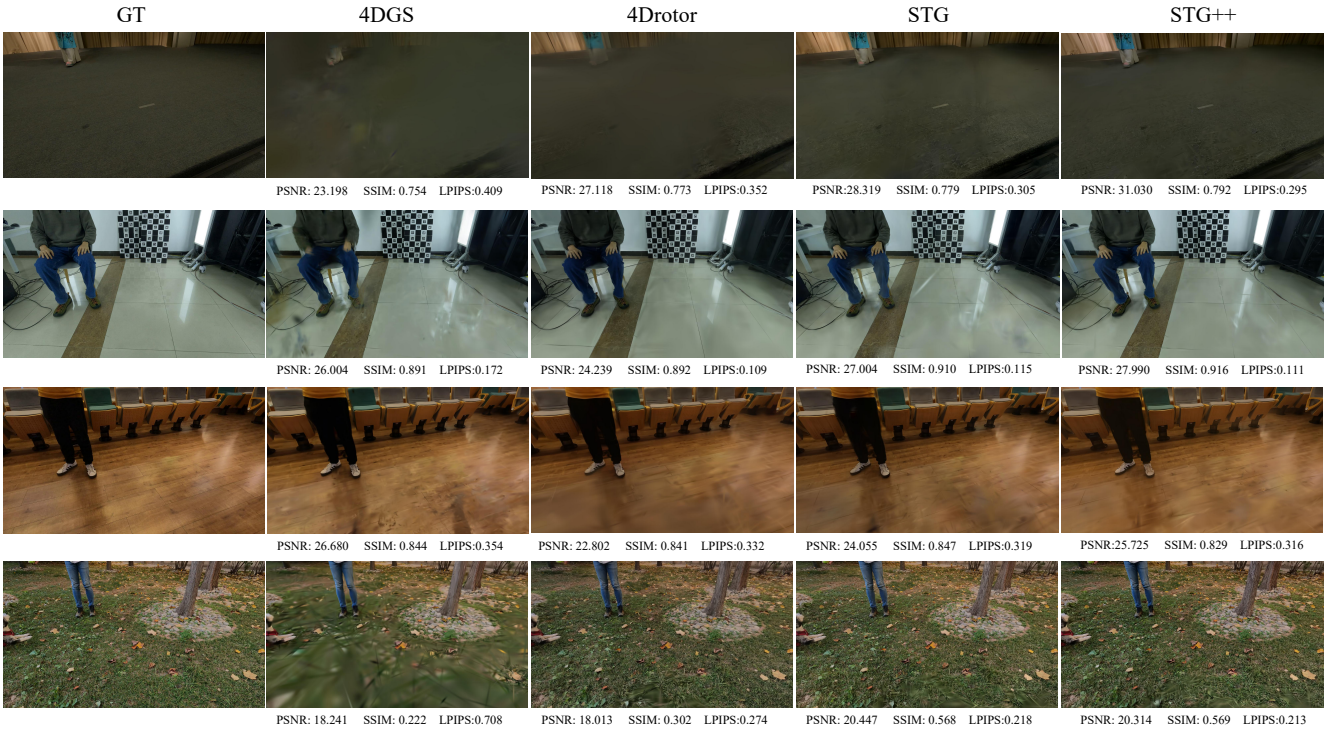


Figure 8. Test view results of three baselines and STG++ on Scene1, Scene2, Scene5, Scene6.

frame and limits the number of final points. It maintains the number and color attributes of the Gaussians while only predicting changes in their positions, rotations, and scaling. This results in minimal model storage, leading to better performance in the static parts of the scene compared to other baselines, with reduced flickering. However, its performance declines significantly in scenes requiring more points for detailed representation, and it cannot address the floaters caused by inconsistent colors in adjacent views. The fitting of larger and faster motions and suddenly-appear/disappear objects is particularly poor.

4Drotor uses dense point clouds as input, which increases memory requirements, especially in large scenes, leading to longer training times and a higher risk of memory overflow. However, by introducing rotors to extend 3D Gaussians to 4D, the authors can directly adapt the density control strategy of the original 3DGS to the t-dimensional space. Consequently, it may perform better in areas with significant motion, such as Figure 9 *Scene2 Laboratory* around human hands.

D. STG++ (Color Mapping) Details

Although STG is not the smallest in terms of storage among all baselines and cannot directly train a model with 300 frames (requiring splitting into multiple 60-frame segments), it achieves better results under the train-views compared to other baselines. Therefore, we delve deeper into its study, hoping it can serve as the foundational architecture for our initial implementation of immersive volumetric video. However, when viewing in SIBR_Viewer, we notice two significant drawbacks:

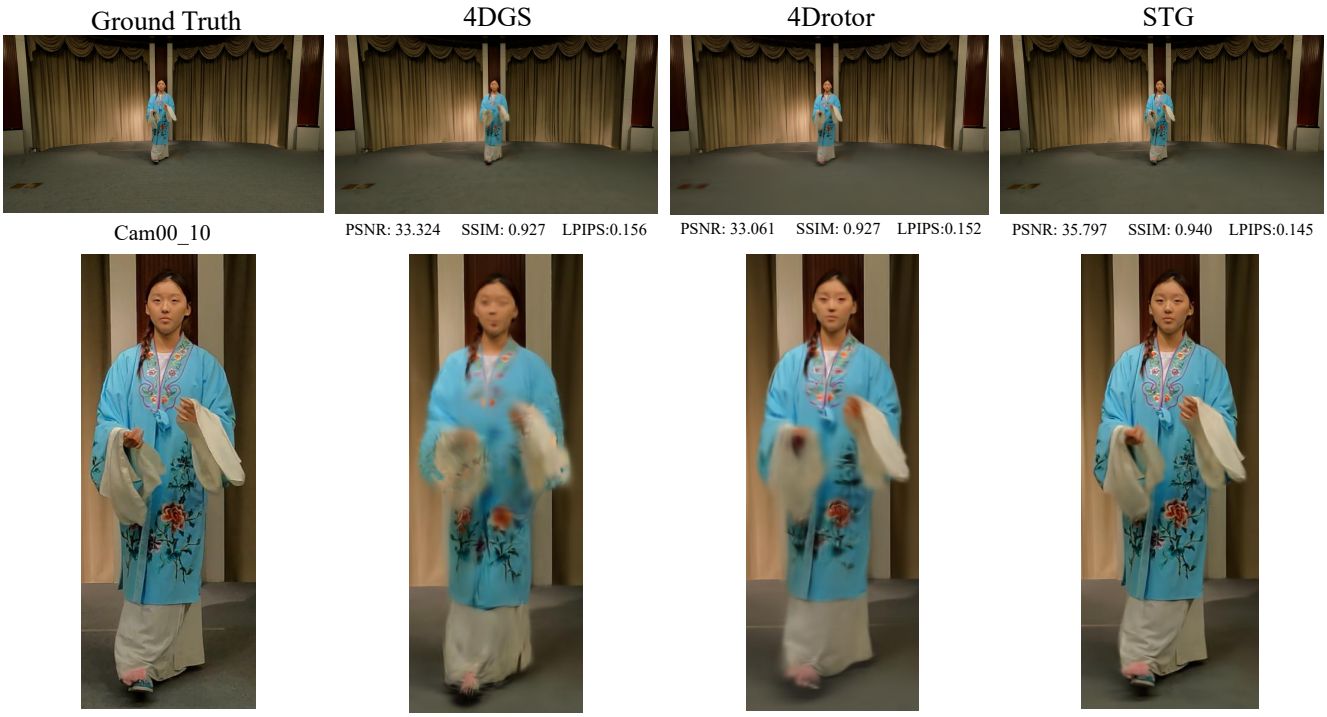
1) In each 60-frame segment, when the viewpoint changes, there is a noticeable flickering of scene points and the presence of floaters, especially when the ground truth of the train-views shows significant color differences due to lighting occlusions and other objective reasons.

2) Besides the color inconsistency during viewpoint changes within each segment, when we modify SIBR_Viewer to continuously load multiple segments, the transitions between segments become even more abrupt. This is a drawback of segmented training, as the appearance of the Gaussians cannot remain consistent between segments.

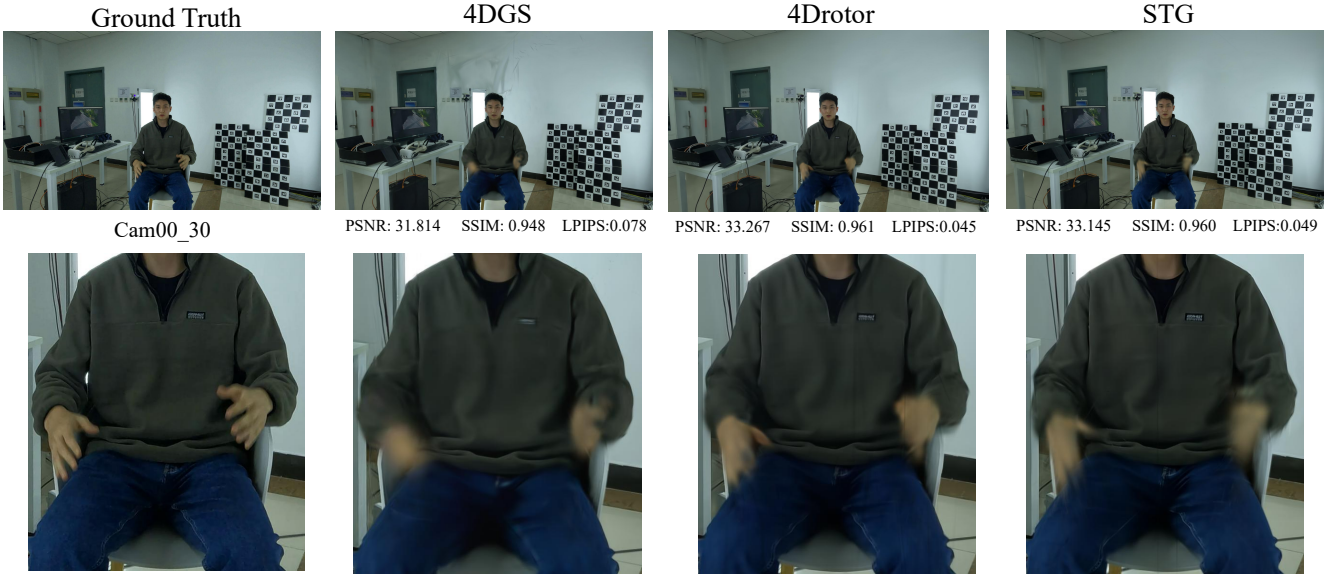
Thus, we propose a learnable viewpoint-dependent affine color transformation function $\phi_i(W, T)$ and maintain its values across different segments. Here, i is the index of the camera, W is a 3×3 transformation matrix, and T is a (1,3) offset vector. Just like the affine transformation, the colors in rendered images C_i' are related to the colors in real scenes (SIBR_Viewer) C_i as follows:

$$C_i' = \mathbf{W} \cdot C_i + \mathbf{T} \tag{13}$$

The loss is calculated between the rendered images C_i' and



(a) The results of train views for three baselines on *Scene1 Opera_girl*.



(b) The results of train views for three baselines on *Scene2 Laboratory*.

Figure 9. More benchmark results visualization (Part 1).

the ground truth as:

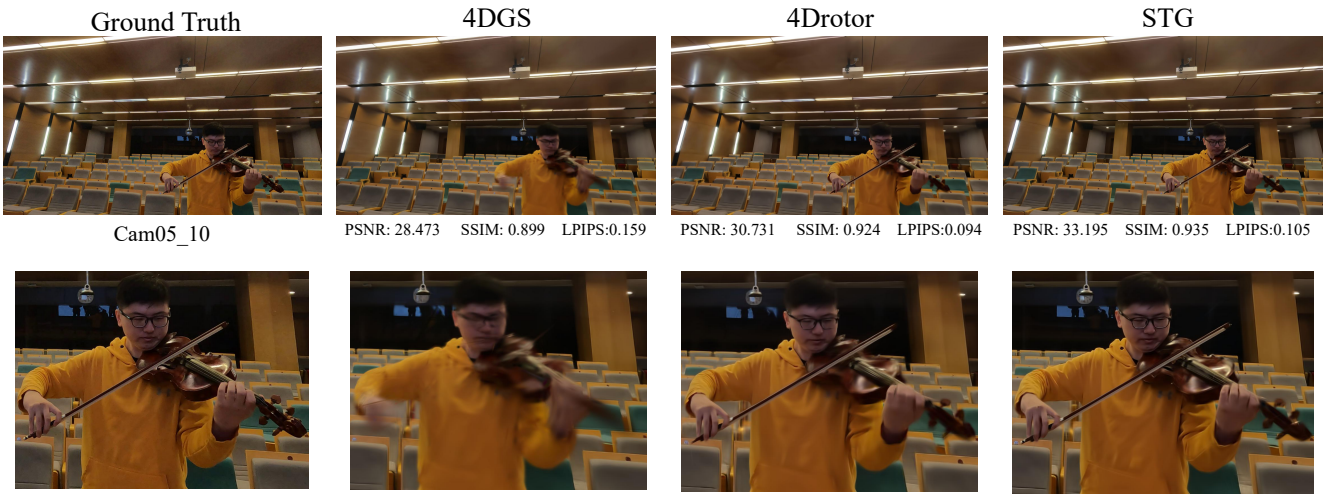
$$Loss = (1 - \lambda_1)L_1(gt, C_i') + \lambda_1 D_{SSIM}(gt, C_i') \quad (14)$$

You can get a more intuitive sense of the improvement from the video in supplementary materials.

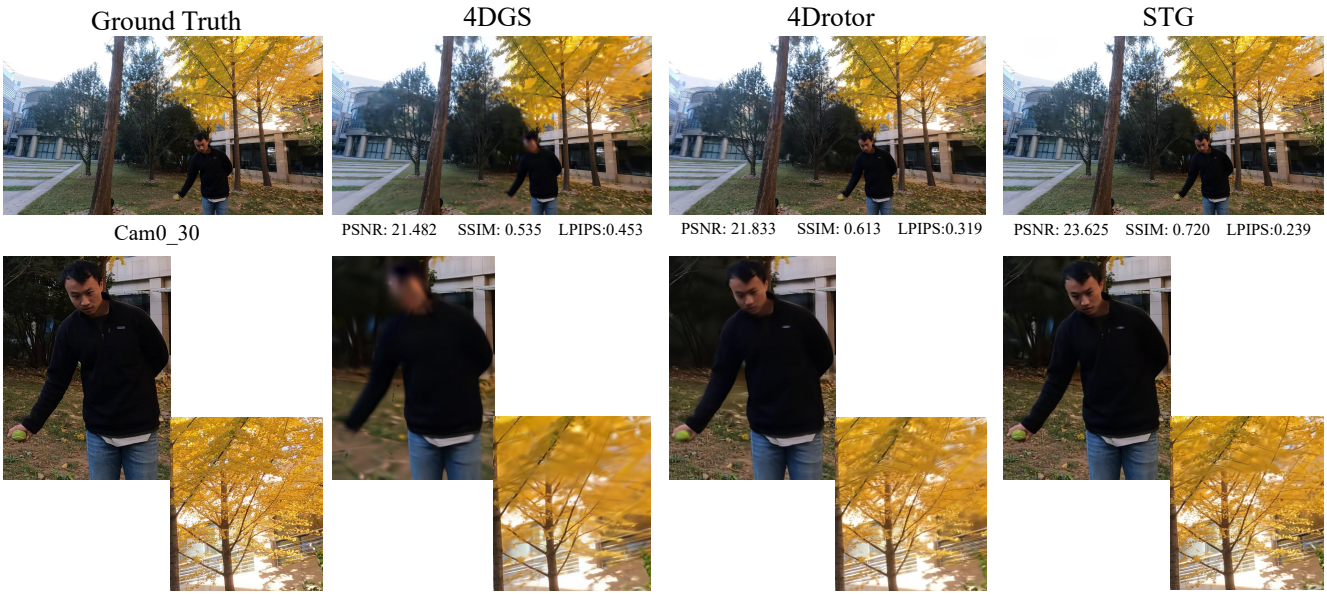
E. Clarifications and Technical Details.

E.1. About "Volumetric" Term

In fact, both academia and industry have yet to provide a clear definition of volumetric video capture methods. Our



(c) The results of train views for three baselines on *Scene5 Rendition*.



(d) The results of train views for three baselines on *Scene6 Puppy*.

Figure 9. More benchmark results visualization (Part 2).

original intention was to define videos that use 3D reconstruction technologies and provide a 6-DOF experience as volumetric video, which is the future of media. Most existing volumetric videos are constructed from an outside-looking-in manner, often lacking natural backgrounds and lighting, which reduces immersion. Inspired by Google's work [6], we aim to develop videos offering a multi-modal, inside-looking-out 6-DoF experience, and thus call it "immersive volumetric video".

E.2. STG++ Limitations

We introduce viewpoint-based color transformation to address global color inconsistencies caused by varying lighting conditions between cameras in real-world scenes (see supp.video 03:41). However, local flickering remains a complex issue due to variations in materials and environmental lighting changes. It is still a significant challenge for the current community, requiring more adaptive and fine-grained processing. We will consider this in future work.

Figure 10. Ex4DGS's performance on the same views as Figure 5 showed in the main text.

Method	Direction Score	Distance Score
w/o SFR	1.17	1.67
SFR+distance	1.69	3.02
SFR+direction	3.81	3.03
SFR+direction+distance	**3.91**	**3.07**

Figure 11. Ablation Studies Illustrate the Effectiveness of Our Proposed Sound Field Reconstruction (SFR) Baseline. A total of 58 participants participated in the user study, rating their sense of direction and distance based on results from various algorithms, using a scale from 1 to 5 (1 being low and 5 being high).

E.3. More Validation of Sound Field Reconstruction

It is worth noting that, AV-NeRF [31] is the closest existing work to our goal of sound field reconstruction (SFR), which means it also targets sound synthesis in a novel location. However, it focuses solely on sound synthesis and uses professional binaural audio acquisition equipment (which is entirely different from our capture system) to collect sound from various locations in space. As a result, its SFR method cannot directly leverage data collected by our camera rig. We will try to modify and adapt its approach for comparison with our baseline in the future.

But to further assess the effectiveness of each module in our proposed SFR method, we have conducted an ablation study based on user feedback, as shown in Figure 11. The average scores for both metrics indicate that incorporating direction and distance modeling in sound field reconstruction significantly enhances participants' immersive video experiences, further demonstrating our SFR algorithm as a practical baseline.

E.4. Capture Rig Setups and Calibration

Time-Synchronized. We used GoPro's official QR control app on mobile phone, enabling each camera to scan a dynamically updating QR code for time synchronization.

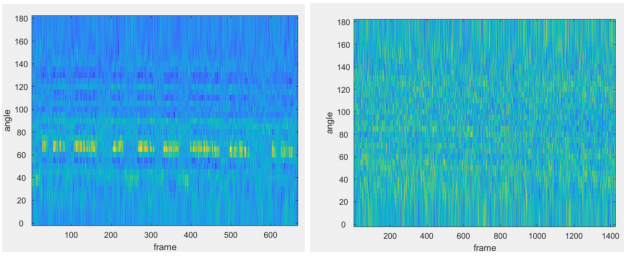


Figure 12. Sound Localization w & w/o Noise Reduction. The yellow areas in the image represent the highest sound intensity in each frame, clearly indicating the sound source's angle change relative to a specific camera.

Noise Reduction and Cart Speed. Our cart only generates noise from axle rotation during sharp turns. The description in main text L.250-253 mitigates this noise and greatly reduces the impact of environmental sounds (e.g., wind) on sound quality. As shown in Figure 12, the denoised sound achieves clearer localization (left), while the noisy sound results in more divergent localization (right). This indicates that the sound quality we collected can not only construct multi-modal volumetric videos but also contribute to sound field, inspiring future work on sound localization and reconstruction from multiple sound sources. Additionally, while the cart experiences slight shaking on uneven terrain, it moves as a rigid body, so its speed is not limited. This prior knowledge can even accelerate our calibration process. The slow speed in this work is primarily due to safety considerations, and we plan to collect faster-moving data in the future.

Calibration. Although we have completed the calibration of the data intended for release, including both fixed-point and mobile shooting, before the submission deadline, it is important to note that, for moving shots, we have tested various open-source algorithms, but none offer an efficient solution for moving multi-view data. Using the original COLMAP takes days to calibrate poses for each frame in long videos. A feasible approach to speed up may refer to [†]. We also look forward to working with colleagues in this community to explore more efficient and accurate calibration solutions using this dataset.

E.5. Continuously Updated Dataset

Currently, other segments in Scene1 include high-speed motions, as shown in Figure 13. And we will continue to update the dataset, increase its richness to make more contributions to the development of the community.

[†]Bernhard Kerbl et al. A Hierarchical 3D Gaussian Representation for Real-Time Rendering of Very Large Datasets. TOG, 2024.1

Table 6. Performance of three baseline methods and STG++ on the **ImViD** Dataset. All methods selected **cam10** as the test view.

Method	Scene1 Opera_girl			Scene2 Laboratory			Scene5 Violin			Scene6 Puppy		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
4DGS	23.227	0.753	0.410	25.798	0.889	0.176	26.586	0.842	0.356	18.121	0.222	0.711
4DRotor	27.263	0.775	0.328	28.007	0.918	0.098	24.083	0.850	0.296	17.916	0.298	0.331
STG	28.482	0.786	0.287	26.306	0.910	0.114	23.144	0.846	0.317	20.497	0.594	0.211
STG++	31.240	0.799	0.277	27.581	0.916	0.107	25.747	0.834	0.310	20.533	0.598	0.202

Table 7. Comparison of average metrics for three baselines and STG++ across four scenes. Due to its smaller model size, 4DGS [55] can train 300 frames at once, so there are no segmented results.

		Frames1-60			Frames60-120			Frames120-180			Frames180-240			Frames240-300			Avarage		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Scene1 Opera	4DGS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**35.005**	**0.930**	**0.156**
	4DRotor	33.502	0.912	0.142	33.735	0.913	0.137	33.749	0.913	0.137	31.460	0.893	0.155	33.718	0.913	0.139	**33.233**	**0.909**	**0.142**
	STG	34.915	0.920	0.127	35.343	0.922	0.125	35.478	0.924	0.124	35.496	0.922	0.125	34.913	0.921	0.126	**35.229**	**0.922**	**0.125**
	STG++	35.195	0.922	0.125	35.603	0.923	0.123	35.822	0.924	0.122	35.738	0.924	0.123	35.738	0.925	0.123	**35.619**	**0.924**	**0.123**
Scene2 Laboratory	4DGS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**32.701**	**0.949**	**0.078**
	4DRotor	36.207	0.967	0.049	36.519	0.967	0.046	34.593	0.96	0.067	36.484	0.968	0.046	36.679	0.967	0.049	**36.096**	**0.966**	**0.051**
	STG	33.405	0.949	0.077	33.257	0.949	0.078	33.641	0.951	0.077	33.140	0.948	0.081	33.298	0.950	0.078	**33.348**	**0.949**	**0.078**
	STG++	33.450	0.950	0.079	33.666	0.951	0.076	33.616	0.951	0.077	33.452	0.950	0.079	33.748	0.952	0.073	**33.586**	**0.951**	**0.076**
Scene5 Rendition	4DGS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**33.645**	**0.918**	**0.183**
	4DRotor	33.398	0.935	0.135	33.361	0.935	0.133	33.255	0.934	0.136	33.398	0.935	0.132	32.638	0.932	0.136	**33.210**	**0.934**	**0.134**
	STG	34.508	0.930	0.158	34.029	0.929	0.161	33.900	0.928	0.165	34.178	0.929	0.163	34.222	0.929	0.161	**34.167**	**0.929**	**0.162**
	STG++	34.426	0.928	0.160	34.277	0.929	0.163	34.169	0.928	0.165	34.407	0.929	0.161	34.203	0.927	0.159	**34.296**	**0.928**	**0.162**
Scene6 Puppy	4DGS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**21.117**	**0.450**	**0.561**
	4DRotor	21.902	0.643	0.301	21.988	0.646	0.297	21.719	0.629	0.319	21.884	0.644	0.297	21.844	0.645	0.300	**21.867**	**0.641**	**0.303**
	STG	23.307	0.714	0.247	23.381	0.717	0.243	23.387	0.718	0.241	23.331	0.717	0.245	23.413	0.718	0.241	**23.364**	**0.716**	**0.243**
	STG++	23.316	0.714	0.246	23.423	0.719	0.240	23.392	0.719	0.241	23.314	0.716	0.246	23.438	0.719	0.240	**23.377**	**0.717**	**0.242**

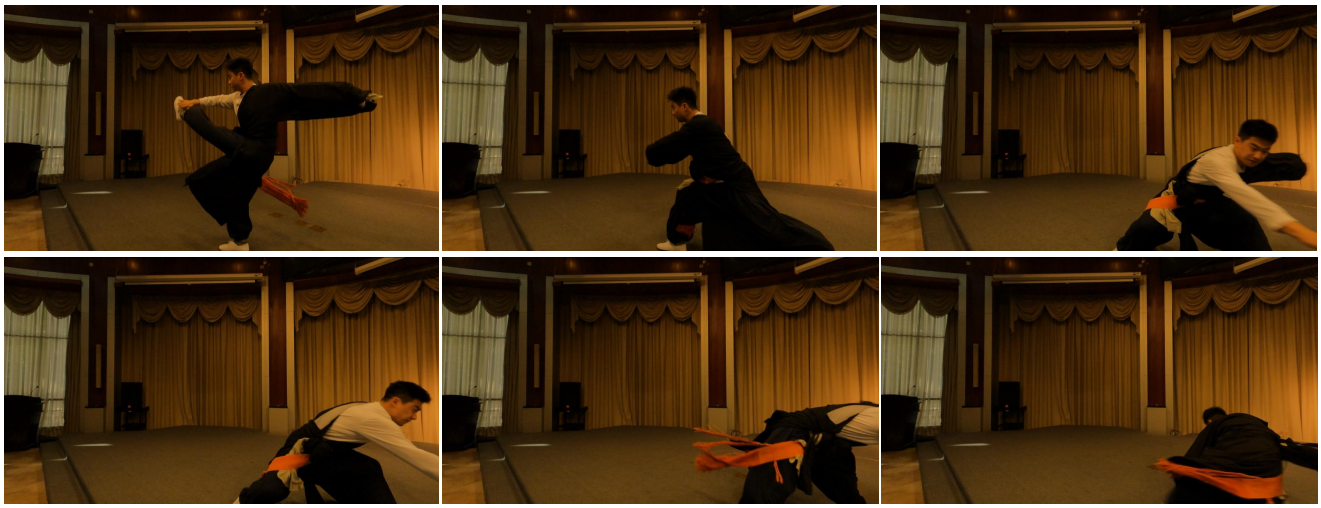


Figure 13. High-Speed Motions Data in Our Dataset ImViD. Scene 1: opera boy spinning kick.

Table 8. Existing real-world datasets for dynamic novel view synthesis.

Datasets	No.Scene	Outdoor/Indoor	Cameras	Mobility	Resolution	Angles	Duration	FPS	Multimodality	Content
PanopticSports [19]	65	Indoor	480 cameras	Static	640×480	360°	5mins	25	✗	Human-centric actions
Technicolor [46]	12	Indoor	16 cameras	Static	2048×1088	Frontal	2s	30	✗	Has a number of close-ups sequences, captured medium angle scenes and other animated scenes where the movement does not come from a human.
Immersive-Lightfield [6]	15	both	46 cameras	Static	2560×1920	Frontal	10-30s	30	✗	Simple and slow motion of human,animals,objects
HyperNeRF [43]	17	Indoor	1 hand-held phone	Fixed-point Waving	1920×1080	Frontal	30-60s	30	✗	Waving a mobile phone in front of a moving scene, object-centric
Dynamic Scene Datasets (NVIDIA) [60]	8	Outdoor	1 Mobile phone /12 cameras	Fixed-point Waving /Static	1920×1080	Frontal	5s	60	✗	Simple body motions (facial, jump, etc)
UCSD Dynamic [34]	96	Outdoor	10 cameras	Static	1920×1080	Frontal	1-2mins	**120**	✗	Various visual effects and human interactions
ZJU-Mocap [44]	10	Indoor	21 cameras	Static	1024×1024	360°	20s	50	✗	Simple body motions (punch, kick, etc.)
Plenoptic Dataset (DyNeRF/Neural 3D) [28]	6	Indoor	21 cameras	Static	2704×2028	Frontal	10-30s	30	✗	Contains high specularity, translucency and transparency objects, motions with changing topology, selfcast moving shadows, volumetric effects, various lighting conditions and multiple people moving around in open living room space
D2NeRF [56]	10	Indoor	dual-hold phone	Fixed-point Waving	1920×1080	Frontal	5s	30	✗	Contains more challenging scenarios with rapid motion and non-trivial dynamic shadows
iPhone Datasets [15]	14	both	1 hand-held phone /2 cameras	Fixed-point Waving /Static	640×480	Frontal	8-15s	30/60	✗	Featuring non-repetitive motion, from various categories such as generic objects, humans, and pets
Meetroom Datasets [27]	4	Indoor	13 cameras	Static	1280×720	Frontal	10s	30	✗	One or three persons have discussion, working, trimming in a meeting room
ENeRF-Outdoor [33]	4	Outdoor	18 cameras	Static	1920×1080	Frontal	20-40s	30	✗	Complex human motions
Replay [49]	46	Indoor	12 cameras	Static	3840×2160	360°	5mins	30	✓(Audio)	Dancing, chatting, playing video games, unwrapping presents, playing ping pong
Campus Datasets [54]	6	Outdoor	24 cameras	Static	3840×2160	Frontal	5-10s	30	✗	Includes more realistic observations such as pedestrians, moving cars, and grasses with people playing
MoDGS [36]	6	both	1 cameras	Static	–	Frontal	–	–	✗	Contains diverse subjects like skating, a dog eating food, YOGA, etc.
DiVa-360 [37]	53	Indoor	53 cameras	Static	1280×720	Frontal	51s	120	✓(Audio)	For Object-centric tasks. Contains dynamic objects and intricate hand-object interactions.
NeRF On-the-go [45]	12	both	1 hand-held phone	Moveable	4032×3024	360°	5-10s	30	✗	Including 10 outdoor and 2 indoor scenes, features a wide range of dynamic objects including pedestrians, cyclists, strollers, toys, cars, robots, and trams), along with diverse occlusion ratios ranging from 5% to 30%
360+X [9]	28	both	1 360°cameras and 1 Spectacles cameras	Static	5760×2880	360°	10s (2152 sequence)	30	✓(Audio)	Capture in 17 cities across 5 countries.Panoptic perspective to scene understanding with audio
ImViD(Ours)	7	**both**	**46 cameras**	**Moveable**	**5312×2988**	**Frontal and 360°**	**1-5mins**	60	**✓(Audio)**	Seven common indoor and outdoor scenes in daily life, including opera, face-to-face communication, teaching, discussion, music performance, interaction with pets, and playing. Each scene has high-quality synchronized multi-view video and audio with a duration of more than 1 minute, and contains rich elements such as various small objects, glass, and changes in light and shadow