LiveStre4m: Feed-Forward Live Streaming of Novel Views from Unposed Multi-View Video

Supplementary Material

A. Detailed Per-Scene Results

In this section, additional qualitative and quantitative results are presented across all scenes rendered, using both Neural3DVideo [14] and MeetRoom [13] datasets. Leveraging 2 unposed input video streams to generate the output video from the target viewpoint at 1024×768 image resolution. Figure 5 shows qualitative results for frame 150 of the generated video alongside the ground truth images. Per-scene quantitative results are summarized in Tab. 7, where scores are averaged over all frames of each generated novel-view video.

Table 7. Quantitative results on the scenes described in Sec. 4.1 using LiveStre4m, evaluated at two image resolutions in a single NVIDIA H100 GPU.

Output Res. (p)	**Scene**	**PSNR ↑**	**Runtime (s) ↓**
1024×768	cut	20.43	0.074
	sear	21.79	0.074
	discussion	18.75	0.074
	trimming	17.94	0.074
	vrheadset	19.24	0.074
512×384	cut	21.16	0.062
	sear	23.72	0.062
	discussion	20.92	0.062
	trimming	18.80	0.062
	vrheadset	18.24	0.061

B. VRU-Basketball Dataset

To validate the robustness of LiveStr4m in highly dynamic scenarios, it is evaluated on the VRU-Basketball dataset [32, 33]. This dataset comprises multi-view recordings of professional basketball games, captured by 34 static cameras, providing a challenging benchmark due to rapid player motion and complex scene dynamics.

Consistent with the experimental setup described earlier in this paper, the central camera is selected as the target viewpoint, while the two nearest cameras serve as input views. These unposed inputs are fed into LiveStre4m to generate the full video sequence from the target viewpoint. Quantitative results, including PSNR and runtime, are reported in Tab. 8, and qualitative comparisons between synthesized frames and ground-truth images are shown in Fig. 6.

Table 8. Quantitative results on the VRU-Basketball dataset [32, 33] obtained with two feed-forward methods, the proposed LiveStre4m (ours) and FLARE [41]. Results obtained in a single H100 GPU.

Model	**VRU-Basketball**		
	Runtime (s) ↓	PSNR ↑	Resolution ↑
FLARE [41]	0.248	17.42	512×384
LiveStre4m (Ours)	0.076	17.88	1024×768
LiveStre4m (Ours)	0.061	18.68	512×384

C. Pose Estimation Metrics

Although LiveStr4m was not explicitly developed for camera pose prediction, this component plays an important role in downstream novel view synthesis. As described in in Sec. 3, the estimated camera poses guide 3D scene reconstruction, making accurate predictions essential. Table 9 reports quantitative results for camera pose estimation, comparing LiveStre4m with FLARE [41]. Employing standard camera pose accuracy metrics evaluated across the five scenes described in Sec. 4.1, LiveStr4m achieves performance comparable to the baseline, indicating that reliable pose estimation can be obtained even though it is not explicitly optimized for this task..

Table 9. Quantitative comparison of camera pose prediction accuracy. Metrics reported are RRA@5°, RTA@5°, and the combined AUC@30° (average of rotation and translation AUC).

Model	**Resolution**	**RRA@5° ↑**	**RTA@5° ↑**	**AUC@30° ↑**
FLARE [41]	512×384	100.00	60.00	91.66
LiveStre4m	1024×768	100.00	60.00	92.09
LiveStre4m	512×384	100.00	60.00	83.00

D. Real World Deployment

This paper shows that LiveStre4m is capable of generating high resolution novel-view videos at 13fps using a minimal buffer of only 3 input frames. However, several limitations remain for real-world deployment in live streaming scenarios, such as sports broadcasts or concerts. Namely, powerful hardware is required and the frame rate of the novel-viewpoint video is still lower than the desirable 24fps. Finally, as shown in Tab. 1, the visual quality of the generated video is lower than slower state-of-the-art 3D reconstruction methods.

Figure 5. Qualitative comparison of reference and predicted viewpoints at frame 150 across all five scenes, rendered at 1024×768 resolution.

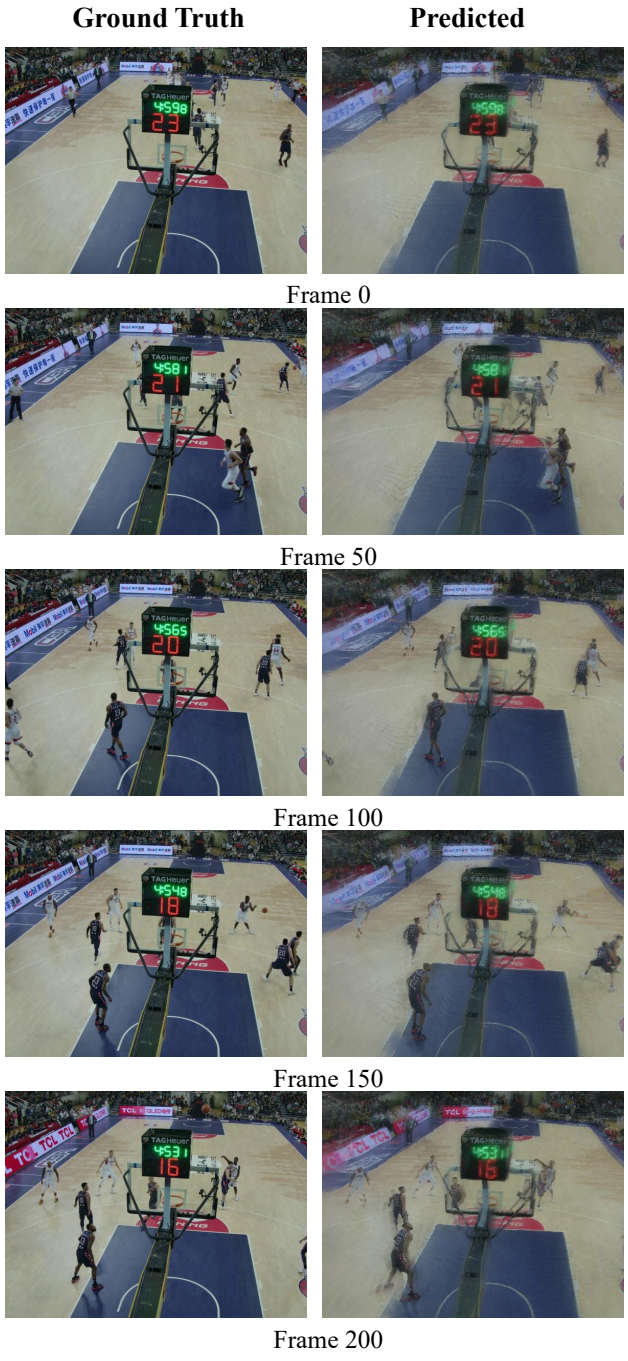

Figure 6. Qualitative results comparing different time steps of the generated video of the GZ scene of the VRU-Basketball dataset [32, 33] with the expected outputs rendered at 1024×768 resolution.