A. More Qualitative Results

We show more qualitative results of Long-LRM in Fig. 6 and project page (http://arthurhero.github.io/projects/llrm/), including a video comparison with optimization-based 3D GS methods.

B. More Hyperparameter Settings

Gaussian parameters. We use SH degree 3 for predicted Gaussian colors. We apply a bias -6.9 and a maximum cap -1.2 to the Gaussian scales before sending them to the exponential function. We apply a bias -2.0 to the opacity values before sending them to the sigmoid function. We align the per-pixel Gaussians to the camera rays originated from the pixels.

Camera pose normalization. In the dataloader, we calculate the average pose of the input cameras by averaging the forward-, downward- and rightward camera directions, and use the cross-product method to obtain an orthonormal rotation matrix along with the average camera positions in the world coordinate. We use the inverse of this average camera pose to normalize all cameras. Finally, we rescale the camera positions to a [-1,1] bounding box.

C. Experiment Details for Ablation Studies on Model Architecture

In Table 4, we present the model architecture ablation studies with different length of input sizes. We train all variants on DL3DV-10K and evaluate on DL3DV-140. The number of training steps are empirically decided based on the model convergence. We study the model behavior under four different settings: 4 input views at 256×256, 32 input views at 256×256, and 32 input views at 512×512, and our extreme setting: 32 input views at 960×540.

For these ablation studies, we use a shorter frame range during evaluation for fair comparisons among each experiments. In details, we choose the first 96 frames from the original video frame sequence, then uniformly sample 8 test views. The 4 to 32 training views are then uniformly sampled from the rest views, i.e., not overlapping to the testing views. We keep the same set of training and testing views for different experimental setups. The input images are resized and center-cropped to squares except for the last row.

D. Additional Experiment Results

D.1. More results on RealEstate10K.

We report the performance of Long-LRM with and without depth and opacity losses on RealEstate10K in Tab. 10. We observe that the losses are less essential for this two-view setup, as the opacity loss aims to save the number of Gaussians and the depth loss helps stabilize training for our long input setup. We also turn off the token merging module for this setup as saving resources is no longer necessary. We show qualitative comparisons in Fig. 7, where Long-LRM demonstrates better rendering of details than baselines.

Method	Init	PSNR↑	Chamfer↓	F-Score↑
2D GS	COLMAP	19.25	0.193	0.272
Long-LRM(2D)	/	**23.27**	**0.135**	**0.414**

Table 8. Novel view synthesis (PSNR) and mesh reconstruction quality (Chamfer and F-Score) comparison with optimization-based 2D GS on ScanNetv2 test split.

Input Views	Depth Supervision	Zero-shot	Method	Abs Diff↓	Abs Rel↓	Sq Rel↓	$\delta<1.25$↑
80∼800 (every 10th)	✓	✗	COLMAP [41]	0.264	0.137	0.138	83.4
			Atlas [38]	0.123	0.065	0.045	93.6
			VoRTX [43]	0.092	0.058	0.036	93.8
32	✗	✓	Long-LRM	0.119	0.073	0.051	94.0

Table 9. We evaluate Long-LRM(2D)'s ability for zero-shot full-scene geometry reconstruction on the ScanNetv2 test split [9]. With only 32 input images, we render median-depth maps from the reconstructed 2D Gaussians and evaluate against the ground-truth depth at **all** frames in each scene sequence. To put into context, we also list performance of past MVS approaches under the same evaluation settings that use much denser input views and are trained on ScanNet training split with ground-truth depth.

Method	Loss Type	PSNR↑	% Gaussians w/ opacity>0.001
GS-LRM	rendering-only	28.10	99.9
Long-LRM	rendering-only	**28.54**	99.9
Long-LRM	+opacity+depth	28.44	44.7

Table 10. Ablation studies on training objectives with 2-view 256-resolution setup on RealEstate10K.

Layers	PSNR↑	Iter Time (sec)	Memory (GB)
{1T7M}×3	21.62	2.9	35
{7M1T}×3	21.58	2.6	35
3T21M	diverged	3.8	35
21M3T	diverged	3.8	35

Table 11. Ablation study on the layer combination of transformer and Mamba2.

D.2. Zero-shot results on ScanNetv2.

We present several zero-shot reconstruction results on the ScanNetv2 test split with our model trained solely on DL3DV-10K. We add aspect ratio augmentation in stage 3 of training, where the input image are randomly cropped between 1:1 to 1.77:1. During inference, we sample 32 input images from each scene of ScanNet and resize the images so that the height is 540.

We report in Table 8 the novel view synthesis and the mesh reconstruction quality comparison between Long-LRM(2D GS ver.) and the optimization-based 2D GS. For novel view synthesis, we evaluate on every 80-th frame, given the denser image sequence of ScanNet. For mesh reconstruction, we render median-depth maps from the reconstructed 2D Gaussians at every 10th camera and use TSDF fusion (voxel 4cm) to construct the mesh. Note that the 32-view setup is relatively sparse for ScanNet scenes, leading to poor performance of optimization-based 2D GS, even with the extra COLMAP initialization, stressing the value of the prior knowledge learned by our model.

Table 9 reports quantitative results in terms of depth map

Figure 5. Zero-shot reconstruction results on ScanNetv2 from Long-LRM (2D GS ver.). We render both RGB images and depth maps from the predicted 2D Gaussians. The depth maps exhibit fine object details, demonstrating Long-LRM's capability for instant geometry reconstruction on complicated novel scenes.

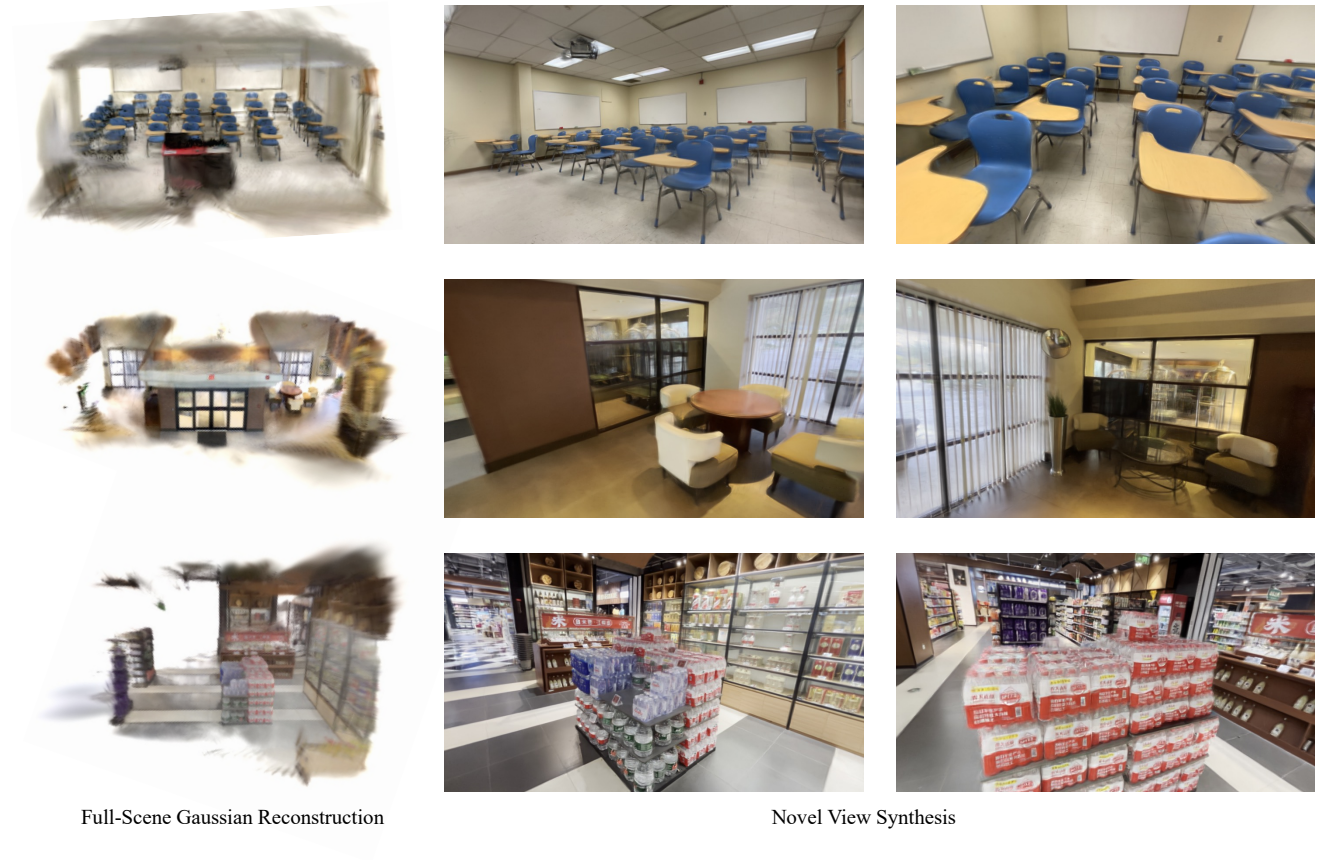

Figure 6. More qualitative results from Long-LRM's wide-coverage scene reconstruction. The left column illustrates the overlook of the reconstructed Gaussians, and the right columns show high-quality synthesized novel views from different perspectives. These examples demonstrate Long-LRM's ability to handle diverse and complex scenes, accurately reconstructing fine-level details, and generating photorealistic views from multiple angles, effectively capturing both geometric and appearance variations.

quality metrics. We render depth maps from the predicted 2D Gaussians at all keyframes in the video sequence at the original resolution of 1272×948. To put the results in context, we list the performance of a few classic multi-view stereo approaches that are trained on the ScanNet train split with ground-truth depth supervision under the same evaluation settings. Fig. 5 shows the qualitative results of the color and depth renderings from the predicted 2D Gaussians.

D.3. Ablation studies on layer combination.

Long-LRM uses a hybrid architecture of Mamba2 and transformer blocks. In Table 11 we study the impact of different

Figure 7. Qualitative comparison on RealEstate10K.

block combination configurations on model performance. We use the 4-input settings at image resolution 256×256 without token merging. We found that if the transformer blocks are even distributed among the Mamba2 blocks, then the model training is more stable than transformer blocks being concentrated at the beginning or the end of the model. The PSNR curves during training do not differ much for the configurations in the table, including the ones that diverge at the end.

E. Limitations

We now briefly discuss the limitations. While we have successfully scaled the model to support 32 high-resolution views and achieved wide-coverage large-scale GS reconstruction, we observe only marginal performance improvements when further increasing the number of input views. Specifically, increasing the input to 64 views only leads to less than 1 dB PSNR improvement. Notably, 64 high-res images correspond to extremely long sequences, exceeding 500K in context length, which presents a significant challenge for current sequence processing models. Addressing this limitation will require future work to better manage ultra-long sequences. Additionally, since the entire DL3DV training set contains images with a fixed wide field of view (FOV), we found that our model struggles to generalize on test sets with significant FOV variations (e.g., the MipNeRF360 dataset with a much smaller FOV). We suspect this limitation is due to the use of Mamba2 blocks, as differing FOVs can alter the meaning of tokens at different positions. Developing models that can generalize effectively across varying FOVs may require more diverse datasets with a range of various FOVs, at a scale similar to DL3DV.