LangFlash: Feed-forward 3D Language Gaussian Splatting from Sparse Unposed Images

Supplementary Material

Table 6. Statistics of the processed RE10k dataset.

Metric	Value
Total frames	∼6M
Number of scenes	∼10k
Avg. masks per image	36
Avg. mask coverage per image (%)	83.5
Processing time (single NVIDIA A100)	30 days

6. RE10k Qualitative visualizations

The visual results shown in Fig. 5 provide a qualitative overview of our performance on the RE10k dataset. These examples were selected to emphasize the characteristic challenges in the dataset: numerous small and overlapping object instances, wide lighting variation, and strong viewpoint changes that stress both 2D segmentation and multi-view rendering. These qualitative results complement the numerical summary.

7. RE10k Dataset statistics

The table above (Tab. 6) summarizes the primary corpus-level statistics of the processed RE10k split used in this study. In total, we retained approximately six million frames across roughly ten thousand scenes; on average, each image contained dozens of instance masks, with mask pixels covering the majority of the image area. The reported processing time corresponds to running the full pipeline (mask extraction, per-frame cleanup, and multi-view consolidation) on a single NVIDIA A100; in practice, the pipeline is embarrassingly parallel, and the wall-clock time can be reduced by distributed execution.

8. RE10k 3D semantic segmentation

In addition to 4, we annotated five previously unseen scenes and report the per-scene mIoU as well as the average overall score in Tab. 7. The baseline methods (LSeg and LSM) struggled on several scenes, whereas our method achieved substantially higher per-scene and overall mIoU, indicating more consistent cross-view semantic aggregation. These results validate both the quality of the processed dataset and the effectiveness of our proposed approach for 3D semantic segmentation of large-scale, real-world indoor footage.

Figure 5. Additional qualitative results on RE10k. We visualize both the semantic and novel-view synthesis results.

Table 7. 3D semantic segmentation on RE10k (mIoU, %). We assign scene names (absent in the original dataset) and provide their corresponding original identifiers.

Method	*Bedroom(5aca)*	*Aisle(bc95)*	*Living room (6558)*	*Study room(89ea)*	*Pool(cd74)*	Overall
LSeg	29.60	18.64	19.12	45.84	14.48	25.53
LSM	24.39	11.67	23.33	42.95	24.77	25.42
Ours	**34.33**	**22.37**	**34.97**	**57.44**	**37.99**	**37.42**