MMOne: Representing Multiple Modalities in One Scene

Supplementary Material

1. Additional Implementation Details

In this section, we introduce the implementation details of the RGB-Thermal, RGB-Language, and RGB-Thermal-Language experiments. We also introduce the hyperparameter settings in our proposed module.

RGB-Thermal. We adhere to the experimental settings of ThermalGaussian for a fair comparison. Specifically, we use spherical harmonic coefficients to model the thermal modality and adjust our thermal rasterization to render thermal images akin to RGB images. The loss function mirrors that of, incorporating a smoothness loss for the thermal modality with λ_{smooth} set to 0.6. The weights for both RGB and thermal losses are set to 0.5.

RGB-Language. Following LangSplat, we modify our language rasterization to render three-dimensional language features. The language loss is defined as the L1 loss between the ground-truth and rendered feature maps. The weights for both RGB and language losses are set to 0.5. For open-vocabulary localization and semantic segmentation, we adopt the same procedure as LangSplat.

RGB-Thermal-Language. We employ the same thermal and language rasterization process as used in the two-modality evaluations. The same thermal and language losses are applied. The weights for RGB and thermal losses are set to 0.5, and the weight for language loss is set to 0.2.

Hyperparameters. In our proposed "Soft Prune", we set the pruning threshold for single-modal Gaussians to 0.5, effectively removing unimportant Gaussians and resulting in a more compact scene representation. For our multi-modal decomposition mechanism, we employ the L2 norm to calculate gradient differences among modalities. This decomposition is integrated into the densification process. If the gradient difference between two modalities exceeds 0.0002, we decompose the multi-modal Gaussian into multiple single-modal Gaussians.

2. Additional Ablation Studies

To further investigate the sensitivity of the threshold setting of multimodal decomposition, we conduct an additional ablation study. As shown in Tab. 1, our chosen threshold consistently achieves superior performance across all metrics. Moreover, we observe only a slight performance drop when the threshold is adjusted, highlighting the robustness of our proposed method.

We also present the complete ablation results for each scene in Tab. 2. Our full method (Decomp.) consistently outperforms other approaches across most scenes, demonstrating the effectiveness of our multimodal decomposition mechanism. Moreover, the proposed "Soft Prune" method consistently surpasses "Hard Prune" in most scenes, highlighting its advantage in mitigating conflicts associated with pruning entire Gaussians. While our modality modeling module serves as the foundation for the multimodal decom-

Table 1. Ablation for the threshold of multimodal decomposition.

Threshold	RGB			Thermal			Lang	Num
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	$\times 10^4$
0.0001	23.11	0.809	0.246	23.94	0.864	0.190	47.1	**9.5**
0.0002	**23.19**	**0.812**	**0.245**	**24.24**	**0.867**	**0.187**	**48.1**	9.9
0.0003	23.13	0.810	0.246	24.04	0.865	**0.187**	47.3	10.1
0.0004	23.03	0.808	0.246	24.15	0.866	**0.187**	47.0	10.3

Table 2. Full ablation studies on RGB-Thermal-Language by gradually adding components to our joint training baseline "MM-J". "MM" refers to our modality modeling module. "H" and "S" denote "Hard" and "Soft", respectively.

M	Metric	Method	Dimsum	DS	LS	Truck	Avg.
R	PSNR ↑	MM-J	23.99	20.99	20.98	23.32	22.32
		+ MM	24.16	21.41	20.92	23.02	22.38
		Prune (H)	24.14	21.45	21.80	23.27	22.67
		Prune (S)	24.69	21.99	21.78	23.45	22.98
		Decomp.	**24.74**	**22.15**	**21.85**	**24.01**	**23.19**
	SSIM ↑	MM-J	0.854	0.782	0.721	0.827	0.796
		+ MM	0.856	0.795	0.716	0.820	0.797
		Prune (H)	0.855	0.801	0.718	0.831	0.801
		Prune (S)	**0.864**	0.812	0.720	0.835	0.808
		Decomp.	0.863	**0.814**	**0.723**	**0.846**	**0.812**
	LPIPS ↓	MM-J	0.199	0.277	0.281	0.232	0.247
		+ MM	**0.196**	0.255	**0.274**	0.236	**0.240**
		Prune (H)	0.207	0.259	0.288	0.235	0.247
		Prune (S)	0.204	0.253	0.298	0.235	0.248
		Decomp.	0.204	**0.251**	0.296	**0.228**	0.245
T	PSNR ↑	MM-J	26.18	21.55	21.65	24.13	23.38
		+ MM	26.35	**22.25**	21.42	24.89	23.73
		Prune (H)	26.35	21.73	22.55	24.80	23.86
		Prune (S)	26.62	21.44	22.43	25.45	23.99
		Decomp.	**26.82**	22.11	**22.57**	**25.46**	**24.24**
	SSIM ↑	MM-J	0.886	0.828	0.840	0.842	0.849
		+ MM	0.885	0.847	0.837	0.855	0.856
		Prune (H)	0.891	0.838	**0.862**	0.862	0.863
		Prune (S)	0.892	0.837	0.861	**0.868**	0.865
		Decomp.	**0.893**	**0.848**	0.860	**0.868**	**0.867**
	LPIPS ↓	MM-J	0.130	0.227	0.274	0.168	0.200
		+ MM	0.149	0.193	0.333	0.158	0.208
		Prune (H)	**0.123**	0.197	**0.264**	0.152	**0.184**
		Prune (S)	0.129	0.200	0.278	**0.145**	0.188
		Decomp.	0.131	**0.190**	0.279	0.147	0.187
L	mIoU ↑	MM-J	56.0	26.9	44.9	52.4	45.1
		+ MM	55.9	29.0	44.3	51.9	45.3
		Prune (H)	59.4	30.3	44.6	53.6	46.9
		Prune (S)	59.4	28.7	**46.3**	53.6	47.0
		Decomp.	**61.1**	**30.6**	46.1	**54.7**	**48.1**

Table 3. Number of Gaussians ($\times 10^4$) for each scene in RGB-Thermal. ThermalGaussian is shortened as "T-GS".

Method	Dim	DS	Ebk	RB	Trk	RK	Bldg	II	Pt	LS	Avg.
T-GS	32.2	27.0	19.8	9.3	27.6	43.3	66.5	45.9	20.0	35.7	32.7
MMOne	**7.5**	**5.5**	**10.4**	**4.4**	**9.8**	**13.6**	**23.2**	**14.4**	**8.4**	**12.2**	**12.2**

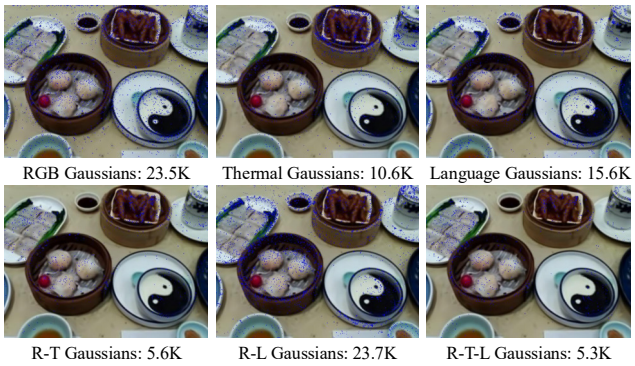

Figure 1. **Gaussian Distributions.** The total number of Gaussians is 89.5K. "T-L Gaussians" are omitted for visual clarity.

position mechanism, its performance remains suboptimal without the disentangling of modalities, due to the varying levels of granularity among them.

Table 4. Number of Gaussians ($\times 10^4$) in RGB-Language.

Method	Figurines	Ramen	Teatime	Kitchen	Avg.
LS*	92.2	58.6	182.0	168.0	125.2
LS-J	55.1	31.1	119.0	105.0	77.6
MMOne	**29.5**	**16.4**	**28.9**	**42.9**	**29.4**

3. Additional Qualitative Results

To analyze the distributions of multimodal and single-modal Gaussians, we use the "Dimsum" scene as an example. As shown in Fig. 1, different modalities require varying number of Gaussians.

We also present the qualitative results of modality conflicts in Fig. 3. Both "T-GS" and "MM-J" refer to joint training of multiple modalities with a shared opacity. The rendering results of "MM-J" show significant blurring, which severely degrades the quality of both RGB and thermal renderings. This suggests that, without our proposed modality decomposition, modality conflicts become more pronounced as the number of modalities increases. This observation aligns with our intuition, as different modalities possess distinct properties. In contrast, our methods, trained on two or three modalities, consistently deliver superior results. Notably, the introduction of the language modality does not degrade the performance of RGB and thermal modalities, due to our modality modeling module and multimodal decomposition mechanism, which ensure scalability to additional modalities.

4. Additional Quantitative Results

We present the number of Gaussians in the RGB-Thermal and RGB-Language experiments to further highlight the effectiveness of our method in achieving a compact representation. As shown in Tab. 3, our method uses approximately one-third of the Gaussians utilized by ThermalGaussian. Similarly, Tab. 4 demonstrates that our method employs only 25% of the Gaussians used by LangSplat and 40% of those used by the joint training baseline modified from LangSplat. These results underscore that our multi-modal decomposition mechanism effectively eliminates redundant Gaussians, leading to a more compact and efficient scene representation.

For RGB-Language experiments, we additionally include Feature-3DGS as another joint training baseline with a different language rasterizer. As shown in Tab. 5, due to modality conflicts, the performance of "F-GS" and "LS-J" drops 0.7%mIoU and 0.8dB for language and RGB, respectively. In contrast, 9.0%mIoU and 0.3dB improvements are achieved by our MMOne.

To further demonstrate the benefits of incorporating thermal information, we conduct additional RGB-Language experiments on the "Dimsum" scene. As shown in Tab. 6, the inclusion of thermal data leads to improvements in both RGB rendering quality and open-vocabulary segmentation accuracy, highlighting its effectiveness in enhancing multi-modal scene understanding.

Table 5. Additional quantitative comparisons on RGB-Language. Feature-3DGS is shortened as "F-GS".

Method	RGB			Lang	
	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	acc ↑
LS*	24.02	**0.854**	**0.220**	47.6	72.4
F-GS	<u>24.16</u>	<u>0.851</u>	<u>0.232</u>	46.9	71.7
LS-J	23.23	0.837	0.257	<u>55.3</u>	<u>73.5</u>
MMOne	**24.35**	<u>0.851</u>	0.244	**56.6**	**76.5**

Table 6. Quantitative comparisons between RGB-Language and RGB-Thermal-Language.

Method	RGB			Lang
	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑
MMOne(R/L)	24.46	0.861	**0.204**	57.1
MMOne(R/T/L)	**24.74**	**0.863**	**0.204**	**61.1**

Table 7. Quantitative results of incorporating monocular depth.

Method	RGB			Thermal			Lang
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑
MMOne(R/T/L)	24.74	**0.863**	**0.204**	26.82	**0.893**	0.131	**61.1**
MMOne(R/T/L/D)	**24.75**	**0.863**	0.206	**26.84**	**0.893**	**0.130**	**61.1**

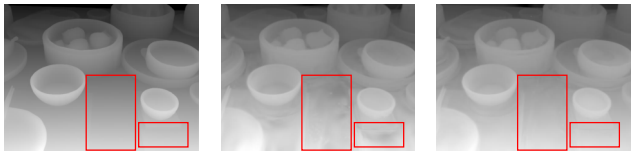

Figure 2. Qualitative results of incorporating monocular depth.

5. Additional Experiments on Scalability

We further validate scalability by incorporating monocular depth in the "Dimsum" scene. The results in Fig. 2 and Tab. 7 show that the rendered depth quality is enhanced, particularly on flat surfaces, without compromising the performance of RGB, thermal, and language.

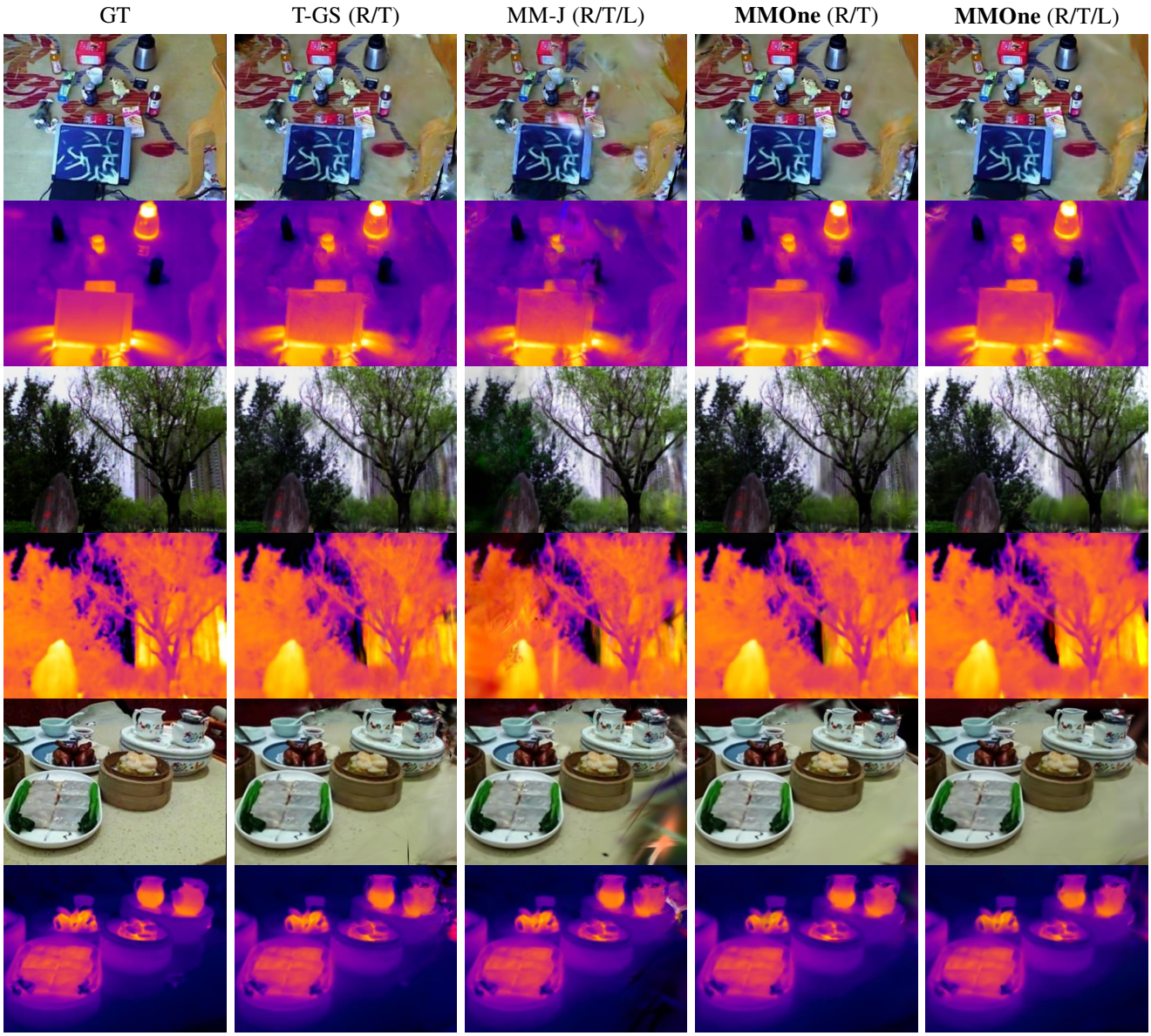

Figure 3. Qualitative results of the modality conflicts caused by the introduction of language. "T-GS" refers to ThermalGaussian and "MM-J" denotes our RGB-Thermal-Language joint training baseline.