MMOne: Representing Multiple Modalities in One Scene

Zhifeng Gu   Bing Wang[†]

The Hong Kong Polytechnic University
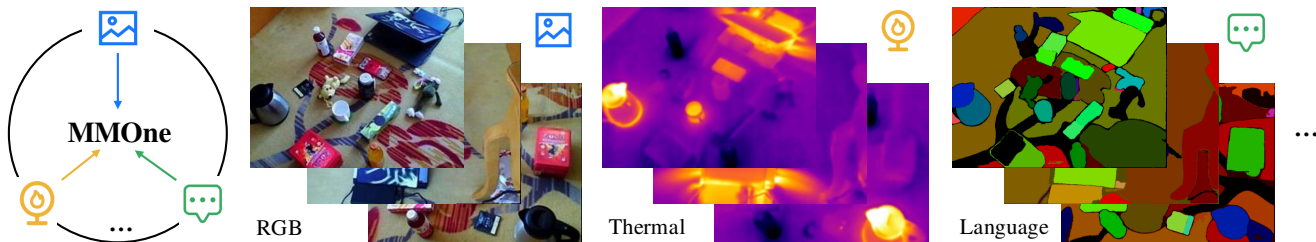
zhifeng.gu@connect.polyu.hk, bingwang@polyu.edu.hk

Figure 1. **MMOne Overview.** MMOne is a general framework designed to represent multiple modalities in one scene. By disentangling multimodal information based on the inherent differences among modalities, we achieve enhanced performance across all modalities.

Abstract

*Humans perceive the world through multimodal cues to understand and interact with the environment. Learning a scene representation for multiple modalities enhances comprehension of the physical world. However, modality conflicts, arising from inherent distinctions among different modalities, present two critical challenges: property disparity and granularity disparity. To address these challenges, we propose a general framework, **MMOne**, to represent multiple modalities in one scene, which can be readily extended to additional modalities. Specifically, a modality modeling module with a novel modality indicator is proposed to capture the unique properties of each modality. Additionally, we design a multimodal decomposition mechanism to separate multi-modal Gaussians into single-modal Gaussians based on modality differences. We address the essential distinctions among modalities by disentangling multimodal information into shared and modality-specific components, resulting in a more compact and efficient multimodal scene representation. Extensive experiments demonstrate that our method consistently enhances the representation capability for each modality and is scalable to additional modalities. The code is available at https://github.com/Neal2020GitHub/MMOne.*

1. Introduction

The 3D physical world inherently comprises multiple modalities, such as vision, thermal, and language. Humans perceive these modalities to understand and interact with the environment [46]. Scene representation is a fundamental cognitive tool for humans to comprehend the world and is crucial for spatial cognition [9, 17]. Integrating information from different modalities into a multimodal scene representation enables better comprehension of the physical world, especially in complex environments [7, 52].

Scene representation has evolved from explicit representations [13] to implicit representations [42], owing to their flexibility in geometry and appearance modeling through continuous parameterization. Neural Radiance Fields (NeRF) [27, 44] and 3D Gaussian Splatting (3DGS) [19] have demonstrated significant potential in various real-world applications, such as autonomous driving [50, 54] and robotic manipulation [41], due to their capabilities for high-fidelity scene reconstruction and real-time rendering.

Beyond RGB, modalities such as language [20, 23, 34, 40, 53] and thermal [15, 25] have been incorporated into scene representation. Current works typically employ additional feature vectors to model modality-specific properties and render these modalities similarly to RGB. Follow-up studies incorporate modality-specific designs, such as learning distinctive features [30, 48], to improve the modality representation ability. However, these approaches are limited to specific modalities and do not address the fundamental differences among different modalities. More critically, they use the same set of Gaussians to represent all modalities, which contradicts the varying levels of granularity of different modalities. These factors motivate us to consider a core problem: how to address the essential differences among modalities when representing multiple modalities simultaneously? This is inherently challenging as different modalities exhibit distinct properties and granularities.

†: Corresponding author.

Modality conflicts, which derived from the inherent differences among modalities, lead to two critical challenges when representing multiple modalities simultaneously. The **first** challenge is property disparity, which refers to the inherent distinctions in the characteristics of data from various modalities. The dimensionality and physical properties of diverse modalities vary (*e.g.*, language representation requires higher dimensionality than RGB), which requires carefully designed modality representation to model. The **second** challenge is granularity disparity, which denotes the differences in the level of granularity at which information is represented across modalities (*e.g.*, thermal exhibits coarser-grained than RGB). This issue is exacerbated when different modalities share the same geometry, resulting in redundant scene representations.

To tackle these challenges, we propose a general framework, **MMOne**, to represent multiple modalities in one scene, which can be readily extended to additional modalities. Specifically, a **modality modeling module** with modality-specific features and a novel modality indicator is proposed to capture the unique properties of each modality. The modality indicator also functions as a "switch" to selectively deactivate certain modalities during rendering. Additionally, we design a **multimodal decomposition mechanism** to separate multi-modal Gaussians into single-modal Gaussians based on modality differences, accommodating the varying levels of granularity of different modalities.

Our method addresses the essential distinctions among modalities by disentangling multimodal information into shared and modality-specific components, resulting in a more compact and efficient multimodal scene representation. Overall, our contributions are as follows:

- We propose MMOne, a general framework to represent multiple modalities in one scene, which can be extended to accommodate additional modalities.
- We design a modality modeling module to capture the modality-specific properties and a multimodal decomposition mechanism to separate multimodal information into shared and modality-specific components.
- Extensive experimental evaluations on multiple benchmarks and datasets demonstrate the effectiveness and scalability of our method, consistently enhancing the representation capability for each modality.

2. Related Work

2.1. Single-Modal Scene Representation

Single-modal scene representation models the RGB modality using Neural Radiance Fields (NeRF) [1, 12, 27, 44] and 3D Gaussian Splatting (3DGS) [4, 19], which have significantly enhanced image quality in novel view synthesis. Notably, 3DGS has been applied to various domains, such as dynamic scenes [47, 49], human avatars [32, 33],

and autonomous driving [50, 54], attributed to its explicit nature and high rendering efficiency. Numerous enhancements have been proposed to improve the rendering quality and efficiency of the RGB modality [10, 11, 24, 26]. Beyond the RGB modality, thermal image has growing research potential and promising applications across various domains [8, 14, 18, 28, 38]. Thermal3D-GS [5] represents the thermal modality with 3DGS by introducing two neural networks to model thermal infrared physical characteristics.

2.2. Dual-Modal Scene Representation

Dual-modal scene representation simultaneously represents RGB with other modalities, including language, depth, thermal, and tactility [6, 43]. For the language modality, LERF [20] and 3D-OVS [23] build NeRF-based language fields with vision language models [2, 37] to support open-vocabulary queries. GS-based methods extend 3DGS to jointly represent the RGB and language modalities, by introducing an extra language feature vector to each Gaussian [34, 36, 40, 48, 53]. Feature dimensionality reduction is often applied to adapt the high-dimensional feature space of the language modality [34, 40, 48], and a smoothed semantic indicator is introduced to disentangle the language rasterization [30]. For the depth modality, depth or normal priors are incorporated to encourage the Gaussians to align with object surfaces [3, 16, 45, 51], thus enhancing the accuracy of the scene geometry modeling. For the thermal modality, NeRF-based [15, 29, 31] and GS-based [25] methods have extended NeRF and 3DGS to represent RGB and thermal modalities simultaneously, by applying a thermal feature vector (*e.g.*, spherical harmonics [25]). However, these works typically use the same opacity and the same set of Gaussians to represent different modalities, which is suboptimal due to their distinct properties and granularities. Additionally, their modality-specific designs hinder their capability to accommodate more modalities.

2.3. Multi-Modal Scene Representation

Representing multiple (>2) modalities simultaneously is a field that remains underexplored. Recently, GLS [35] and LangSurf [22] utilize depth cues to align RGB-language Gaussians with object surfaces, facilitating precise segmentation with text queries. However, these methods are essentially "dual-modal", as they do not incorporate new properties into each Gaussian. Furthermore, they use the same Gaussians to represent all modalities, which contradicts the distinct properties and granularities among modalities, resulting in suboptimal performance and redundant scene representations. In contrast, we disentangle Gaussians into shared and modality-specific components (*i.e.*, multi-modal and single-modal Gaussians), consistently enhancing the representation capability for each modality and ensuring scalability to accommodate additional modalities.
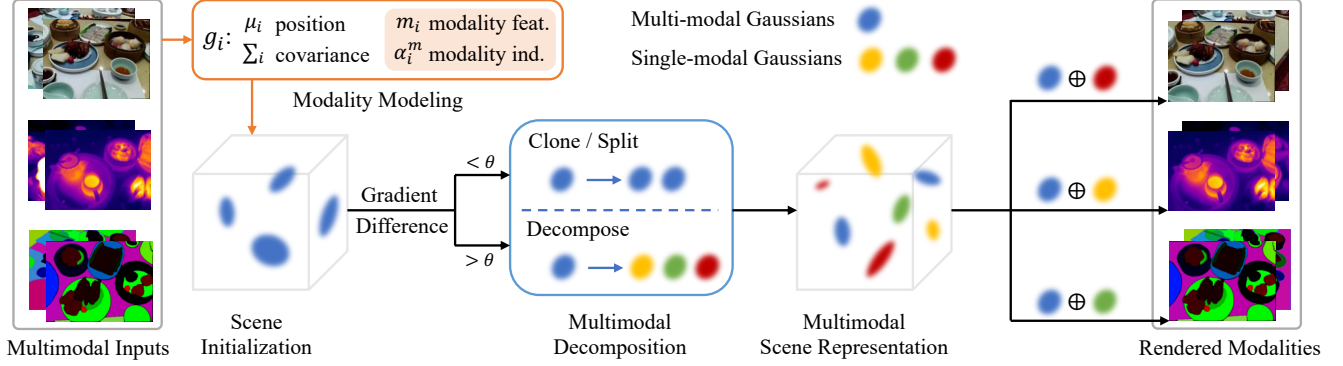
Figure 2. The General Framework of **MMOne**. Given multi-view multimodal inputs of the scene, we progressively construct a multimodal scene representation. Each modality is represented by our **modality modeling module**, which includes modality-specific features and a modality indicator. The densification process is integrated with our **multimodal decomposition mechanism**, which disentangles multimodal information based on gradient difference among modalities. Multimodal losses are combined to jointly optimize all modalities.

3. Analysis of Multimodal Representation

3.1. Preliminaries

The vanilla Gaussian Splatting (3DGS) [19] models a scene with a set of 3D Gaussians $\mathcal{G} = \{g_i | i = 1, \cdots, N\}$. Each Gaussian g_i is defined by its opacity $\alpha_i \in \mathbb{R}$, mean position $\mu_i \in \mathbb{R}^3$, covariance matrix $\Sigma_i \in \mathbb{R}^{3 \times 3}$ as $g_i(x) = e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$. Each Gaussian is also associated with color features $c_i \in \mathbb{R}^{d_c}$ (with $d_c = 3$ color dimension), modeled by spherical harmonic (SH) coefficients.

To render other modalities such as thermal [25] and language [34, 48, 53], new feature vectors are introduced to model these modalities. The rendering process for these modalities is similar to RGB rendering. 3DGS first splats the Gaussians to the 2D image plane and then applies alphablending to compute the modality value for each pixel. This process can be generally formulated as:

$$M(x) = \sum_{i=1}^{N} T_i \cdot \alpha_i \cdot g_i^{2D}(x) \cdot m_i,$$
$$\text{where} \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j \cdot g_j^{2D}(x)), \quad (1)$$

where M represents a specific modality, m_i denotes the modality feature vector (with d_m feature dimension), and $g_i^{2D}(x)$ is the 2D projection of g_i at pixel x. All attributes of the 3D Gaussians, including the modality-specific features, are learnable and optimized directly during training.

3.2. Challenges in Multimodal Representation

Modality conflicts arise when representing multiple modalities with a single scene representation. This is due to the intrinsic differences in properties and levels of granularity among different modalities. For example, as shown in Fig. 1, the thermal modality is relatively coarse, the color

modality is more detailed, and the language modality remains constant within objects or parts. Addressing these modality conflicts involves two main challenges.

Property Disparity. This refers to the inherent differences in the characteristics and attributes of data from various modalities. These disparities manifest in different forms, such as dimensionality and physical properties. For example, RGB representation requires three-dimensional features, whereas language representation necessitates a much higher dimensionality. Additionally, a piece of paper in front of a cup of hot tea would obscure the tea in the RGB and language modalities but not in the thermal modality.

Granularity Disparity. This denotes the inherent differences in the level of detail at which information is represented across various modalities. For example, the thermal modality is relatively coarse-grained, while the color modality is more fine-grained. Consequently, at the object boundaries, the thermal modality may favor fewer but larger Gaussians, whereas the color modality may require smaller but more numerous Gaussians. This challenge is exacerbated when different modalities share the same Gaussians.

4. Proposed Method

4.1. Overview

We generally formulate the problem of representing multiple modalities in one scene as learning a scene representation to simultaneously represent multiple modalities, which can be readily extended to additional modalities.

Our framework, **MMOne**, as illustrated in Fig. 2, represents multiple modalities with one scene representation. Given the multi-view multimodal inputs of the scene, we progressively construct a multimodal scene representation capable of accommodating one or more modalities. Each modality is modeled by our **modality modeling module**, which includes modality-specific features and a modality

indicator, as detailed in Sec. 4.2. The densification process is integrated with our **multimodal decomposition mechanism**, which disentangles multimodal information based on modality differences, as discussed in Sec. 4.3.

During training, each modality is rendered independently (as described in Eq. (2)), and the loss for each modality is computed separately. The overall loss is the sum of the individual losses for all modalities: $\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{M_i}$, where M_i represents a modality and m denotes the number of modalities. The form of $\mathcal{L}_{M_i}$ varies due to the unique characteristics of each modality. All attributes of Gaussians, including modality-specific features and modality indicators, are optimized simultaneously through backpropagation. During inference, we can selectively render different modalities to accommodate various application scenarios.

4.2. Representing Multiple Modalities

To address the property disparity, we propose a modality modeling module to model each modality, which can represent the distinct properties of different modalities. The module includes modality-specific features $m_i \in \mathbb{R}^{d_m}$ (with d_m feature dimension) and a modality indicator $\alpha^m \in [0, 1]$ (denoting the modality indicator of modality M). The rendering process for this modality can be formulated as:

$$M(x) = \sum_{i=1}^{N} T_i^m \cdot \alpha_i^m \cdot g_i^{2D}(x) \cdot m_i,$$

$$\text{where} \quad T_i^m = \prod_{j=1}^{i-1}(1 - \alpha_j^m \cdot g_j^{2D}(x)). \tag{2}$$

Our proposed modality indicator is designed to capture the unique properties of each modality, rather than using a shared opacity across different modalities. In addition to the weighting mechanism described in Eq. (2), the modality indicator also functions as a "**switch**" for specific modalities, enabling selective deactivation during the rendering process. When certain modalities are deactivated (*i.e.*, in an "off" state), the properties of the Gaussians (*e.g.*, location, size, and orientation) are influenced solely by the remaining "active" modalities. This approach allows us to explicitly model the varying granularities of different modalities, as each modality requires a different number of Gaussians to represent. This forms the basis of our multimodal decomposition mechanism discussed in Sec. 4.3.

The modality "switch" function is implemented by selectively skipping the rendering of certain modalities in the CUDA rasterization process, thereby freezing updates for those modalities. During training, our modality modeling module and other attributes of Gaussians are optimized simultaneously to capture modality-specific properties and distinguish information from different modalities.

As illustrated in Fig. 3, the dashed lines indicate the distinct distributions of modality indicators for different
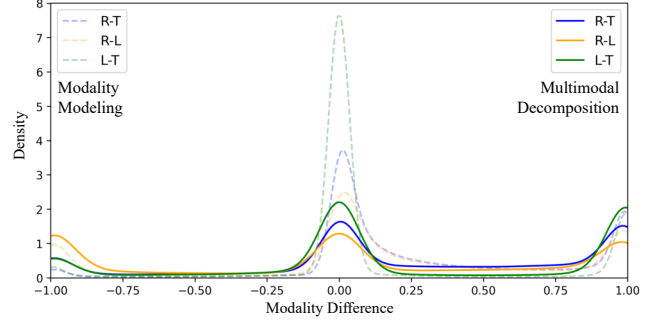

Figure 3. Distribution differences among different modality indicators using our modality modeling module ("dashed" line) and multimodal decomposition mechanism ("solid" line). "R-T" refers to $\alpha^R - \alpha^T$ in Gaussians, and so on.

modalities. These differences enable the proposed modality modeling module to effectively capture and represent the unique properties of each modality.

4.3. Multimodal Decomposition

To tackle the granularity disparity, we propose a simple yet effective multimodal decomposition mechanism to separate multi-modal Gaussians into single-modal Gaussians, accommodating the varying granularities of different modalities. This method disentangles multimodal information into shared and modality-specific components, leading to a more compact and efficient multimodal scene representation.

Multimodal Prune. 3DGS initializes Gaussians using a sparse point cloud from Structure-from-Motion (SfM) and employs an adaptive density control strategy to dynamically add and remove Gaussians. Gaussians are pruned when they exhibit low opacity or an overlarge screen space size. This approach is feasible when Gaussians are associated with one modality. However, conflicts arise when Gaussians are associated with multiple modality indicators representing different modalities. For example, if one modality indicator is low while another is high, simply pruning the Gaussian (*i.e.*, "**Hard Prune**") may adversely affect other modalities.

To address this, we propose "**Soft Prune**", which refers to pruning a specific modality rather than the entire Gaussian, as shown in Fig. 4. This is achieved by setting the corresponding modality indicator to an "off" state, thereby excluding that modality from the rendering process.

However, simply "soft pruning" modalities can lead to redundant single-modal Gaussians, which may negatively impact the mutual improvement among multiple modalities. Therefore, we increase the threshold for pruning single-modal Gaussians to reduce the number of unimportant single-modal Gaussians, thereby encouraging the learning of the shared properties among different modalities.

Multimodal Decomposition. In the densification process of 3DGS, Gaussians with gradients exceeding a certain threshold are either cloned or split based on their scales.
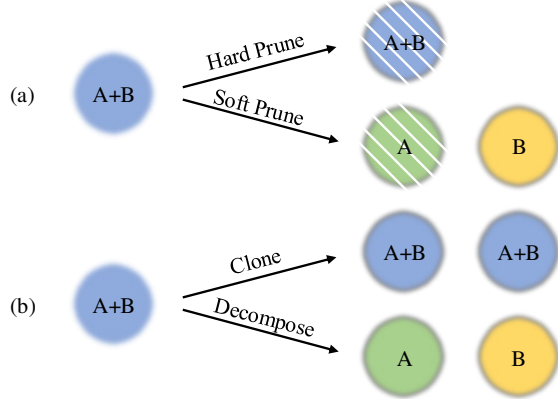
Figure 4. **Multimodal Decomposition Mechanism.** (a) Multimodal Prune: "Hard Prune" refers to directly pruning the Gaussian, while "Soft Prune" involves pruning modality "A" but retaining modality "B". (b) Multimodal Decomposition: Instead of cloning the Gaussian, multimodal decomposition disentangles a multi-modal Gaussian into multiple single-modal Gaussians.

During the learning of multimodal representation, gradients from different modalities backpropagate to the same Gaussian. Conflicts arise when these gradients cancel each other out, leading to suboptimal results for all modalities.

To address this, we propose multimodal decomposition, which disentangles the gradients from different modalities and decomposes multimodality when the gradient difference exceeds a certain threshold. Specifically, we accumulate the gradients from different modalities (*i.e.*, g_{m_i} and g_{m_j}) and the gradient difference gd_{ij} between modality m_i and modality m_j is formulated as:

$$gd_{ij} = norm(g_{m_i} - g_{m_j}). \qquad (3)$$

We decompose the identified multi-modal Gaussians into multiple single-modal Gaussians, as shown in Fig. 4. These decomposed Gaussians are then optimized separately based on their modality-specific losses. As illustrated in Fig. 3, our multimodal decomposition mechanism preserves the distinct distributions of different modality indicators and removes redundant multi-modal Gaussians.

5. Experiments

To evaluate our approach, we select three representative modalities for evaluation: RGB, thermal, and language. We conducted experiments on RGB-Thermal, RGB-Language, and RGB-Thermal-Language combinations to demonstrate the effectiveness of the proposed method.

Datasets. We conduct experimental evaluation on two real-world datasets. 1) **RGBT-Scenes dataset** [25], collected using a handheld thermal-infrared camera FLIR E6 PRO, comprises over 1000 aligned RGB and thermal images from 10 diverse scenes, including both indoor and outdoor environments with various object sizes and temperature vari-

ations. 2) **LERF dataset** [20], captured using the Polycam application of an iPhone, includes complex in-the-wild scenes. We utilize the extended version from LangSplat [34], which provides ground truth masks for text queries and additional challenging localization samples.

Metrics. We evaluate the RGB and thermal modalities using standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics assess the quality of the reconstructed images from novel views. For the language modality, we report the mean Intersection over Union (mIoU) and localization accuracy (%) for open-vocabulary semantic segmentation and object localization following LangSplat [34].

Implementation Details. Our method is implemented based on the 3DGS framework [19] using PyTorch. All experimental settings and default parameters are consistent with 3DGS. We set the threshold for multimodal decomposition to 0.0002. We modify the rasterization module of 3DGS to render thermal and language modalities similar to [25, 34]. Ground-truth semantic features are obtained following LangSplat [34] using the SAM ViT-H model [21] and the OpenCLIP ViT-B/16 model [37]. We train each comparative experiment for 30K iterations. All experiments are conducted on a single NVIDIA 4090 GPU.

5.1. Evaluation on RGB-Thermal

We compare our method with two baselines: 1) 3DGS [19], which trains the original 3DGS using RGB and thermal images separately, and 2) ThermalGaussian [25], which trains RGB and thermal jointly with a shared opacity. For a fair comparison, we adopt the same experimental settings and loss functions (*e.g.*, smoothness loss on thermal images) as used in [25]. All comparative experiments utilize the same sparse point cloud and camera poses obtained from COLMAP [39] results on RGB images.

The quantitative results of various methods are presented in Table 1. Note that we reproduce results on the "Truck" scene to align with the train/test split of other scenes and 3DGS. The table demonstrates that our method consistently outperforms the strong baseline, ThermalGaussian [25], across most scenarios. Despite approaching performance saturation, our approach shows an average PSNR improvement of 0.5dB in RGB rendering and 0.4dB in thermal rendering, highlighting the effectiveness of our method to capture modality-specific properties and granularities. Notably, our method employs one-third of the Gaussians used by ThermalGaussian, yet achieves superior performance.

The rendered RGB and thermal images are presented in Fig. 5. These qualitative results align with the quantitative findings in Tab. 1. Our modality modeling module maintains the sharpness of the RGB modality and the smoothness of the thermal modality, resulting in clearer colors and more

Table 1. Quantitative evaluation of RGB-Thermal. "R" and "T" denote RGB and Thermal. ThermalGaussian [25] is shortened as "T-GS".

M	Metric	Method	Dim	DS	Ebk	RB	Trk	RK	Bldg	II	Pt	LS	Avg.
R	PSNR ↑	3DGS	23.91	20.43	26.77	27.80	23.00	20.79	20.95	23.96	24.91	20.20	23.27
		T-GS	24.38	21.76	26.85	28.12	23.66	23.14	**24.19**	24.55	25.48	21.71	24.38
		MMOne	**24.65**	**22.05**	**27.43**	**29.03**	**23.96**	**24.12**	24.16	**25.65**	**26.01**	**21.81**	**24.89**
	SSIM ↑	3DGS	0.847	0.748	0.901	0.910	0.815	0.772	0.791	0.872	0.859	0.696	0.821
		T-GS	0.858	0.797	0.900	0.920	0.832	0.822	**0.849**	0.884	0.855	**0.739**	0.846
		MMOne	**0.862**	**0.810**	**0.918**	0.916	**0.845**	**0.842**	0.847	**0.897**	**0.876**	0.727	**0.854**
	LPIPS ↓	3DGS	**0.194**	0.299	0.171	0.201	0.238	0.217	0.228	0.188	0.183	0.280	0.220
		T-GS	**0.194**	**0.253**	0.169	**0.199**	**0.224**	0.184	**0.170**	0.186	0.195	**0.268**	**0.204**
		MMOne	0.203	0.254	**0.160**	0.235	0.226	**0.178**	0.184	**0.183**	0.178	0.291	0.209
T	PSNR ↑	3DGS	26.21	20.28	20.78	26.46	23.93	27.17	25.39	29.90	22.33	18.68	24.11
		T-GS	26.46	**22.28**	23.31	27.17	24.57	26.33	26.72	29.86	**26.16**	22.27	25.51
		MMOne	**26.90**	21.81	**23.79**	**27.39**	**25.44**	**27.65**	**27.06**	**30.27**	26.05	**22.52**	**25.89**
	SSIM ↑	3DGS	0.890	0.816	0.814	0.914	0.853	0.928	0.873	0.897	0.842	0.760	0.859
		T-GS	0.886	0.835	0.862	0.919	0.849	0.922	0.888	0.896	0.883	0.850	0.879
		MMOne	**0.894**	**0.840**	**0.874**	**0.926**	**0.870**	**0.933**	**0.902**	**0.906**	**0.895**	**0.861**	**0.890**
	LPIPS ↓	3DGS	0.126	0.240	0.314	0.213	0.160	0.126	0.223	0.088	0.265	0.383	0.214
		T-GS	0.129	0.210	0.203	**0.198**	0.155	**0.124**	0.177	0.091	0.181	0.248	0.172
		MMOne	**0.125**	**0.194**	**0.201**	0.213	**0.142**	0.127	0.198	**0.083**	0.205	0.272	0.176

defined boundaries. Furthermore, our multimodal decomposition mechanism mitigates modality conflicts, enhancing the consistency of both modalities with reduced noise.

Table 2. Quantitative evaluation of RGB-Language. "R" and "L" denote RGB and Language. LangSplat [34] is shortened as "LS*" and "LS-J" denotes the joint training baseline based on LangSplat.

M	Metric	Method	Figurines	Ramen	Teatime	Kitchen	Avg.
R	PSNR ↑	LS*	**24.31**	24.45	23.79	23.52	24.02
		LS-J	22.43	24.29	23.14	23.06	23.23
		MMOne	24.07	**24.71**	**24.03**	**24.58**	**24.35**
	SSIM ↑	LS*	**0.844**	**0.858**	**0.821**	0.893	**0.854**
		LS-J	0.818	0.849	0.802	0.880	0.837
		MMOne	0.837	0.854	0.817	**0.896**	0.851
	LPIPS ↓	LS*	**0.215**	**0.194**	**0.273**	**0.196**	**0.220**
		LS-J	0.259	0.222	0.320	0.225	0.257
		MMOne	0.242	0.222	0.307	0.203	0.244
L	mIoU ↑	LS*	45.2	46.1	53.0	46.2	47.6
		LS-J	58.6	47.8	60.0	54.6	55.3
		MMOne	**58.9**	**48.0**	**62.0**	**57.5**	**56.6**
	acc ↑	LS*	71.4	60.6	84.8	72.7	72.4
		LS-J	71.4	60.6	84.8	**77.3**	73.5
		MMOne	**80.4**	**62.0**	**86.4**	**77.3**	**76.5**

5.2. Evaluation on RGB-Language

We evaluate our method against 2 baselines: 1) LangSplat [34], which first reconstructs the scene geometry using RGB and then registers language features to the Gaussians, and 2) "LS-J", a RGB-language joint training baseline modified from LangSplat, which trains RGB and Language jointly using a shared opacity. Note that we reproduce LangSplat [34] after splitting the train/test set to evaluate both the RGB rendering and open-vocabulary queries.

As shown in Tab. 2, LangSplat [34] relies on the original 3DGS for RGB reconstruction, limiting its language understanding to this fixed RGB geometry. "LS-J" demonstrates strong performance in open-vocabulary localization and semantic segmentation, surpassing LangSplat by 7.7 in mIoU. However, due to the modality conflicts introduced by the language modality, its RGB results decline across all scenes. In contrast, our method excels in open-vocabulary queries while maintaining high RGB rendering performance. Notably, our PSNR even exceeds LangSplat, highlighting the mutual enhancements between modalities and the effectiveness of disentangling multimodality.

Both RGB and Language qualitative results are presented in Fig. 6. Our method preserves the high RGB rendering quality of 3DGS, with sharper object boundaries due to the integration of the language modality. With our modality modeling module and multimodal decomposition mechanism, the semantic masks generated from text queries are more accurately aligned with object surfaces and exhibit greater consistency within objects.

5.3. Evaluation on RGB-Thermal-Language

To evaluate our method on more than two modalities, we conduct experiments to learn the RGB-thermal-language representation. Since no existing RGB-thermal-language dataset is available, we select four scenes from the RGBT-Scenes dataset [25], which include multiple objects with aligned RGB and thermal images. We manually annotate the ground-truth semantic masks for open-vocabulary queries following LangSplat [34]. Specifically, we obtain ground-truth language features with CLIP [37] and SAM [21]. We only employ the medium level of SAM masks for the language modality learning as they best align with object-level granularity. Given the absence of established methods for representing RGB-thermal-language scenes, we compare our method with a self-implemented RGB-thermal-language joint training baseline that trains RGB,
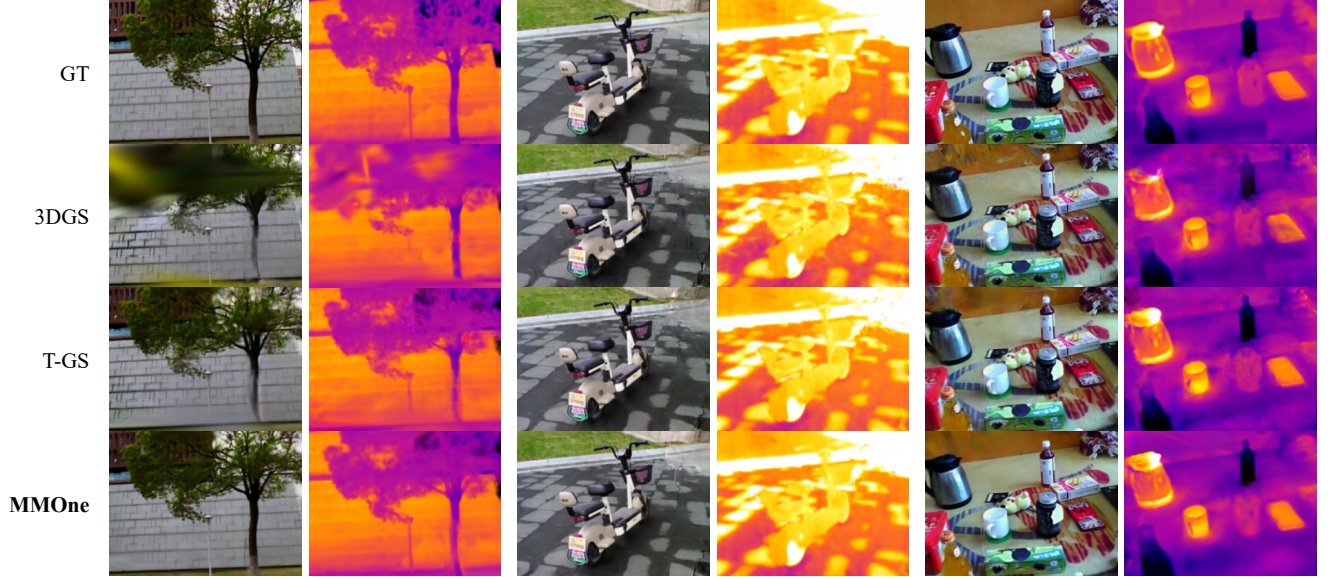
Figure 5. Qualitative results of RGB-Thermal among our MMOne, 3DGS, and ThermalGaussian (abbreviated as "T-GS").
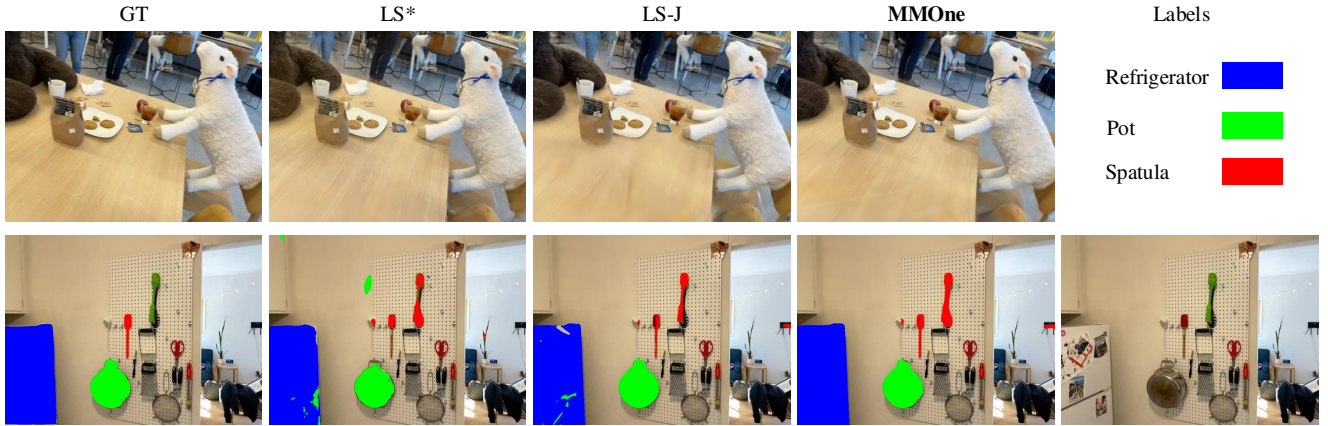

Figure 6. Qualitative results of RGB-Language among our MMOne, LangSplat ("LS*"), and joint training version of LangSplat ("LS-J"). The first row presents RGB renderings from novel views and the second row shows the open-vocabulary semantic segmentation results.

thermal, and language jointly using a shared opacity. Our experimental settings and loss functions are consistent with [25, 34] without additional modifications.

The quantitative results are presented in Tab. 3. Our method achieves superior results across all three modalities, indicating the effectiveness of our modality modeling module in modeling modality-specific properties and our multimodal decomposition mechanism that disentangle multimodality to mitigate modality conflicts. As shown in Fig. 7, the qualitative results demonstrate that our method consistently improves the rendering quality of RGB and thermal. Additionally, the masks generated from text queries are better aligned with object surfaces.

Notably, as shown in Tab. 4, the RGB and thermal rendering quality significantly degrades when incorporating language into the joint training baselines (i.e., "T-GS" and

"MM-J"), underscoring the impact of modality conflicts as the number of modalities increases. In contrast, our method not only surpasses baselines with a large margin, but also slightly enhances RGB and thermal rendering quality with the language modality. This demonstrates the effectiveness of our method to handle multiple modalities and scalability to accommodate more modalities by decomposing multimodality into shared and modality-specific components.

5.4. Ablation Studies

To evaluate the effectiveness of each component, we conduct ablation studies on representing RGB-thermal-language scenes, as modality conflicts are amplified when the number of modalities increases. As shown in Tab. 5, we begin by evaluating the impact of our modality modeling module. With the proposed modality indicator, this mod-

Table 3. Quantitative evaluation of RGB-Thermal-Language. "R", "T" and "L" denote RGB, Thermal and Language, respectively. "MM-J" represents our joint training baseline.

M	Metric	Method	Dimsum	DS	LS	Truck	Avg.
R	PSNR ↑	MM-J	23.99	20.99	20.98	23.32	22.32
		MMOne	**24.74**	**22.15**	**21.85**	**24.01**	**23.19**
	SSIM ↑	MM-J	0.854	0.782	0.721	0.827	0.796
		MMOne	**0.863**	**0.814**	**0.723**	**0.846**	**0.812**
	LPIPS ↓	MM-J	**0.199**	0.277	**0.281**	0.232	0.247
		MMOne	0.204	**0.251**	0.296	**0.228**	**0.245**
T	PSNR ↑	MM-J	26.18	21.55	21.65	24.13	23.38
		MMOne	**26.82**	**22.11**	**22.57**	**25.46**	**24.24**
	SSIM ↑	MM-J	0.886	0.828	0.840	0.842	0.849
		MMOne	**0.893**	**0.848**	**0.860**	**0.868**	**0.867**
	LPIPS ↓	MM-J	**0.130**	0.227	**0.274**	0.168	0.200
		MMOne	0.131	**0.190**	0.279	**0.147**	**0.187**
L	mIoU ↑	MM-J	56.0	26.9	44.9	52.4	45.1
		MMOne	**61.1**	**30.6**	**46.1**	**54.7**	**48.1**

Table 4. Quantitative analysis of modality conflicts. "R/T" refers to training with RGB and thermal, while "R/T/L" indicates training with RGB, thermal, and language. The bolded values highlight the changes in RGB and thermal with the introduction of language.

Methods	RGB		Thermal	
	PSNR	SSIM	PSNR	SSIM
T-GS (R/T)	22.88	0.807	23.90	0.855
MMOne (R/T)	23.12	0.811	24.17	0.866
MM-J (R/T/L)	22.32 **(-0.56)**	0.796 **(-0.011)**	23.38 **(-0.52)**	0.849 **(-0.006)**
MMOne (R/T/L)	23.19 **(+0.07)**	0.812 **(+0.001)**	24.24 **(+0.07)**	0.867 **(+0.001)**

Table 5. **Ablation Studies.** We conduct 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.

Methods	RGB			Thermal			Lang	Num
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	mIoU	×10^4
MM-J	22.32	0.796	0.247	23.38	0.849	0.200	45.1	32.9
+ MM	22.38	0.797	**0.240**	23.73	0.856	0.208	45.3	29.0
Prune (H)	22.67	0.801	0.247	23.86	0.863	**0.184**	46.9	13.4
Prune (S)	22.98	0.808	0.248	23.99	0.865	0.188	47.0	10.6
Decomp.	**23.19**	**0.812**	0.245	**24.24**	**0.867**	0.187	**48.1**	9.9

ule enhances the capability to represent modality-specific properties, resulting in consistent improvements across the RGB, thermal and language modalities.

Next, we evaluate the impact of multimodal prune, comparing "Hard Prune" and "Soft Prune". Our proposed "Soft Prune" achieves better results across all modalities, demonstrating its effectiveness in mitigating the adverse effects on other modalities during pruning. It also significantly reduces the number of Gaussians, leading to a more compact scene representation. Finally, by incorporating our multimodal decomposition mechanism, which disentangles multimodality based on the gradient differences among modalities, we achieve exceptional results across all metrics. This
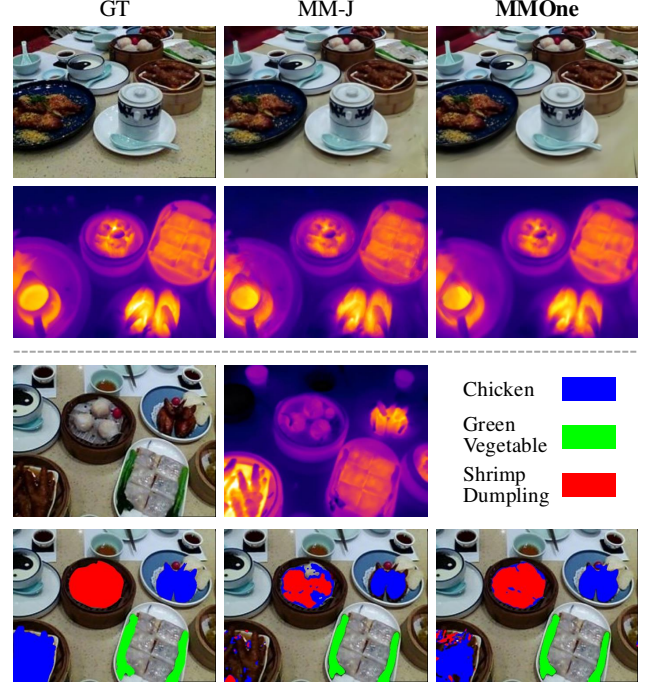

Figure 7. Qualitative results of RGB-Thermal-Language between our MMOne and the joint training baseline ("MM-J"). The first two rows present RGB and thermal renderings from novel views, the third row shows the reference images for text queries, and the fourth row displays the open-vocabulary query results.

indicates that employing different numbers of Gaussians to represent each modality can enhance overall performance, aligning with the varying levels of granularity of different modalities. Notably, our method results in a compact and efficient multimodal scene representation, using less than one-third of the number of Gaussians of the baseline while achieving superior performance.

6. Conclusion

In this work, we propose MMOne, a general framework for representing multiple modalities in one scene, which can be readily extended to additional modalities. Our method includes a modality modeling module and a multimodal decomposition mechanism, designed to capture modality-specific properties of different modalities and disentangle multimodal information into shared and modality-specific components. Extensive experiments across various modality combinations demonstrate the effectiveness and scalability of our method. Future work will focus on incorporating camera poses into the learning process and extending multimodal scene representation to dynamic scenes.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[3] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[4] Guikun Chen and Wenguan Wang. A Survey on 3D Gaussian Splatting. *arXiv preprint arXiv:2401.03890*, 2024. 2

[5] Qian Chen, Shihao Shu, and Xiangzhi Bai. Thermal3D-GS: Physics-induced 3D Gaussians for Thermal Infrared Novel-view Synthesis. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[6] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-Augmented Radiance Fields. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 2

[7] Shengshun Duan, Qiongfeng Shi, and Jun Wu. Multimodal Sensors and ML-Based Data Fusion for Advanced Robots. *Advanced Intelligent Systems*, 4(12):2200213, 2022. 1

[8] Thomas Dubail, Fidel Alejandro Guerrero Peña, Heitor Rapela Medeiros, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Privacy-Preserving Person Detection Using Low-Resolution Infrared Cameras. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[9] Russell A Epstein and Chris I Baker. Scene Perception in the Human Brain. *Annual Review of Vision Science*, 5(1): 373–397, 2019. 1

[10] Guangchi Fang and Bing Wang. Mini-Splatting: Representing Scenes with a Constrained Number of Gaussians. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[11] Guangchi Fang and Bing Wang. Mini-Splatting2: Building 360 Scenes within Minutes via Aggressive Gaussian Densification. *arXiv preprint arXiv:2411.12788*, 2024. 2

[12] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. NeRF: Neural Radiance Field in 3D Vision: A Comprehensive Review. *arXiv preprint arXiv:2210.00379*, 2022. 2

[13] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep Learning for 3D Point Clouds: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2020. 1

[14] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the Dark Spaces of Healthcare with Ambient Intelligence. *Nature*, 585(7824):193–202, 2020. 2

[15] Mariam Hassan, Florent Forest, Olga Fink, and Malcolm Mielle. ThermoNeRF: Multimodal Neural Radiance Fields for Thermal Novel View Synthesis. *arXiv preprint arXiv:2403.12154*, 2024. 1, 2

[16] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields . In *SIGGRAPH*, 2024. 2

[17] Helene Intraub. Rethinking Visual Scene Perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(1):117–127, 2012. 1

[18] B Jalil, MA Pascali, GR Leone, M Martinelli, D Moroni, O Salvetti, and A Berton. Visible and Infrared Imaging Based Inspection of Power Installation. *Pattern Recognition and Image Analysis*, 29(1):35–41, 2019. 2

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023. 1, 2, 3, 5

[20] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proceedings of the International Conference on Computer Vision*, 2023. 1, 2, 5

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the International Conference on Computer Vision*, 2023. 5, 6

[22] Hao Li, Roy Qin, Zhengyu Zou, Diqi He, Bohan Li, Bingquan Dai, Dingewn Zhang, and Junwei Han. LangSurf: Language-Embedded Surface Gaussians for 3D Scene Understanding. *arXiv preprint arXiv:2412.17635*, 2024. 2

[23] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly Supervised 3D Open-vocabulary Segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 1, 2

[24] Wenkai Liu, Tao Guan, Bin Zhu, Luoyuan Xu, Zikai Song, Dan Li, Yuesong Wang, and Wei Yang. EfficientGS: Streamlining Gaussian Splatting for Large-Scale High-Resolution Scene Representation. *IEEE MultiMedia*, 2025. 2

[25] Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermal-Gaussian: Thermal 3D Gaussian Splatting. In *Proceedings of the International Conference on Learning Representations*, 2025. 1, 2, 3, 5, 6, 7

[26] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De La Torre. Taming 3DGS: High-Quality Radiance Fields with Limited Resources. In *SIGGRAPH Asia*, 2024. 2

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[28] Roque Alfredo Osornio-Rios, Jose Alfonso Antonino-Daviu, and Rene de Jesus Romero-Troncoso. Recent Industrial Applications of Infrared Thermography: A Review.

IEEE Transactions on Industrial Informatics, 15(2):615–625, 2018. 2

[29] Mert Özer, Maximilian Weiherer, Martin Hundhausen, and Bernhard Egger. Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[30] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3D Vision-Language Gaussian Splatting. In *Proceedings of the International Conference on Learning Representations*, 2025. 1, 2

[31] Matteo Poggi, Pierluigi Zama Ramirez, Fabio Tosi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Cross-Spectral Neural Radiance Fields. In *Proceedings of the International Conference on 3D Vision*, 2022. 2

[32] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 2

[33] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 2

[34] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D Language Gaussian Splatting. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 5, 6, 7

[35] Jiaxiong Qiu, Liu Liu, Zhizhong Su, and Tianwei Lin. GLS: Geometry-aware 3D Language Gaussian Splatting. *arXiv preprint arXiv:2411.18066*, 2024. 2

[36] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. GOI: Find 3D Gaussians of Interest with an Optimizable Open-vocabulary Semantic-space Hyperplane. In *Proceedings of the International Conference on Multimedia*, 2024. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision . In *Proceedings of the International Conference on Machine Learning*, 2021. 2, 5, 6

[38] Muhamad Risqi U Saputra, Chris Xiaoxuan Lu, Pedro Porto B de Gusmao, Bing Wang, Andrew Markham, and Niki Trigoni. Graph-based Thermal-Inertial SLAM with Probabilistic Neural Networks. *IEEE Transactions on Robotics*, 38(3):1875–1893, 2021. 2

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2016. 5

[40] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[41] Olaolu Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe David Kennedy, and Mac Schwager. Splat-MOVER: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting. In *Proceedings of the Conference on Robot Learning*, 2024. 1

[42] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 1

[43] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy. Touch-GS: Visual-Tactile Supervised 3D Gaussian Splatting. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2024. 2

[44] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *SIGGRAPH*, 2023. 1, 2

[45] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025. 2

[46] Mark T Wallace. The Development of Multisensory Processes. *Cognitive Processing*, 5:69–83, 2004. 1

[47] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 2

[48] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. OpenGaussian: Towards Point-Level 3D Gaussian-based Open Vocabulary Understanding. *Advances in Neural Information Processing Systems*, 37: 19114–19138, 2024. 1, 2, 3

[49] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 2

[50] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting. In *Proceedings of the European Conference on Computer Vision*, 2024. 1, 2

[51] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian Opacity Fields: Efficient Adaptive Surface Reconstruction in Unbounded Scenes. *ACM Transactions on Graphics*, 43 (6):1–13, 2024. 2

[52] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. 1

[53] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang

Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3

[54] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. DrivingGaussian: Composite Gaussian Splatting for Surrounding Dynamic Autonomous Driving Scenes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2