POSESPECGS: POSE-ALIGNED MULTIMODAL GAUSSIAN SPLATTING WITH SPECTRUM-RESPONSE RENDERING

Anonymous authors
Paper under double-blind review

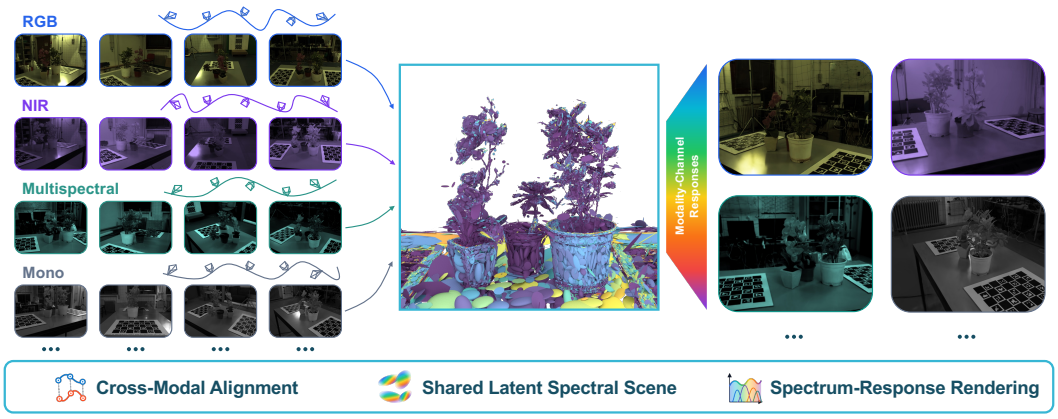

Figure 1: PoseSpecGS synthesizes multimodal novel views from RGB, near-infrared, multispectral, and monochrome observations without known rig extrinsics or synchronization. Cross-Modal Alignment registers auxiliary cameras and refines their poses; Spectrum-Response Rendering synthesizes modality-specific images from the shared Latent Spectral Scene Representation.

ABSTRACT

Multimodal novel-view synthesis aims to reconstruct a unified scene representation from images acquired by sensors with distinct image-formation characteristics. However, modality-dependent appearance and inaccurate camera poses hinder learning a compact, reusable representation. We present PoseSpecGS, a pose-aligned Gaussian splatting method with spectrum-response rendering for multimodal novel-view synthesis. Inspired by sensor spectral responses, we first introduce Spectrum-Response Rendering, which synthesizes modality-specific images from alpha-composited shared Gaussian spectral features using compact modality-channel response operators. To address pose inaccuracies, we further introduce Cross-Modal Alignment, which registers auxiliary-modality cameras to an RGB structure-from-motion backbone and jointly refines their poses and scene parameters using RGB-guided reprojection consistency, without known rig extrinsics or synchronized frame correspondences. The representation supports static and time-conditioned multimodal reconstruction. Moreover, shared response operators pretrained across scenes enable zero-shot one-to-multimodal transfer under RGB-only supervision. Experiments demonstrate state-of-the-art reconstruction quality on the evaluated benchmarks, with reduced model storage and fewer training iterations. Code will be released upon acceptance.

1 INTRODUCTION

RGB-based 3D reconstruction has made remarkable progress in recovering scene geometry and appearance (Mildenhall et al., 2020; Kerbl et al., 2023; Lu et al., 2024b), but RGB observations capture only visible-light appearance. Multimodal image collections provide complementary measurements of the same scene across spectral bands and sensing modalities, each with distinct image-formation characteristics. A unified 3D representation should therefore share scene structure across modalities while preserving modality-specific appearance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Extending RGB-centric reconstruction to practical multimodal capture presents two coupled challenges. First, existing representative approaches model modality-specific appearance through separate fields or jointly predict all modality channels from shared features using a nonlinear decoder (Lin et al., 2024; Lu et al., 2024a; Gu & Wang, 2025; Meyer et al., 2026). While joint decoding captures cross-modal dependencies implicitly, learning a compact, reusable representation that explicitly separates shared scene features from modality-channel responses remains challenging. Second, many multimodal reconstruction systems (Lu et al., 2024a; Lincetto et al., 2025) rely on synchronized acquisition or calibrated sensor rigs to establish cross-modal alignment, limiting their applicability to independently acquired observations. Even after initial alignment, residual pose inaccuracies can be absorbed by flexible appearance models, conflating modality-dependent appearance with geometric misalignment. These limitations motivate a central question: *Can we learn a compact, reusable scene representation for multimodal novel-view synthesis from overlapping heterogeneous sensor observations without using known rig extrinsics or synchronization?*

We introduce PoseSpecGS, a Gaussian splatting framework for multimodal novel-view synthesis, with Spectrum-Response Rendering for shared scene modeling and Cross-Modal Alignment for camera registration and pose refinement (Fig. 1). Inspired by the role of sensor spectral responses in image formation, we represent multimodal appearance using shared per-Gaussian spectral features. Spectrum-Response Rendering synthesizes modality-specific images by applying compact, learned modality-channel response operators to the alpha-composited feature image. This formulation allows observations across modalities to jointly constrain a shared representation while preserving modality-specific appearance. The representation further supports time-conditioned appearance modeling. Moreover, the separately parameterized modality-channel response operators provide an interface for cross-scene reuse; shared response operators pretrained across multimodal scenes enable zero-shot one-to-multimodal transfer under RGB-only reconstruction supervision.

Accurate cross-modal alignment is essential for learning this shared representation, as pose errors can substantially degrade reconstruction quality. To this end, we propose Cross-Modal Alignment, which registers auxiliary-modality cameras to an RGB structure-from-motion backbone without known rig extrinsics or synchronized frame correspondences. During Gaussian optimization, RGB-guided reprojection consistency between auxiliary-view RGB proxy images and nearby RGB observations constrains the joint refinement of auxiliary-camera poses and scene parameters. This joint optimization produces refined auxiliary-camera poses and a shared spectral scene representation, enabling high-quality multimodal novel-view synthesis through Spectrum-Response Rendering.

Our contributions are: (1) We present **PoseSpecGS**, a pose-aligned Gaussian splatting method with spectrum-response rendering for high-quality multimodal novel-view synthesis; (2) We introduce **Spectrum-Response Rendering**, which applies compact modality-channel response operators to shared spectral features for static and time-conditioned multimodal reconstruction. Moreover, shared response operators pretrained across scenes enable zero-shot one-to-multimodal transfer under RGB-only supervision; (3) We develop **Cross-Modal Alignment** to register auxiliary-modality cameras to an RGB structure-from-motion backbone and jointly refine their poses and scene parameters using RGB-guided reprojection consistency, without known rig extrinsics or synchronized frame correspondences; (4) Experiments demonstrate state-of-the-art reconstruction quality on the evaluated benchmarks, with reduced model storage and fewer training iterations.

2 Related Work

Multimodal neural rendering. NeRF and 3DGS underpin novel-view synthesis (Mildenhall et al., 2020; Kerbl et al., 2023), with cross-spectral, multispectral, and hyperspectral extensions (Poggi et al., 2022; Chen et al., 2024a). Thermal3D-GS models thermal image formation (Chen et al., 2024b), while PhysIR-Splat models emission, reflection, and atmospheric attenuation (Gao et al., 2026). ThermalGaussian and MMOne integrate RGB–thermal observations through regularization or decomposition (Lu et al., 2024a; Gu & Wang, 2025). MultimodalStudio unifies diverse sensing modalities (Lincetto et al., 2025), while SPoILeR uses multi-scene pretraining and latent regularization for RGB-supervised multimodal reconstruction (Lincetto et al., 2026). For spectral imagery, SOC-GS enforces spatial consistency, TIMS optimizes modality-specific spherical harmonics, and MS-Splatting decodes shared Gaussian features before compositing (Guo et al., 2025; Grün et al., 2025; Meyer et al., 2026). FeatSplat alpha-composites Gaussian features before screen-space decoding (Martins & Civera, 2024). Other spectral representations use latent compression or spec-

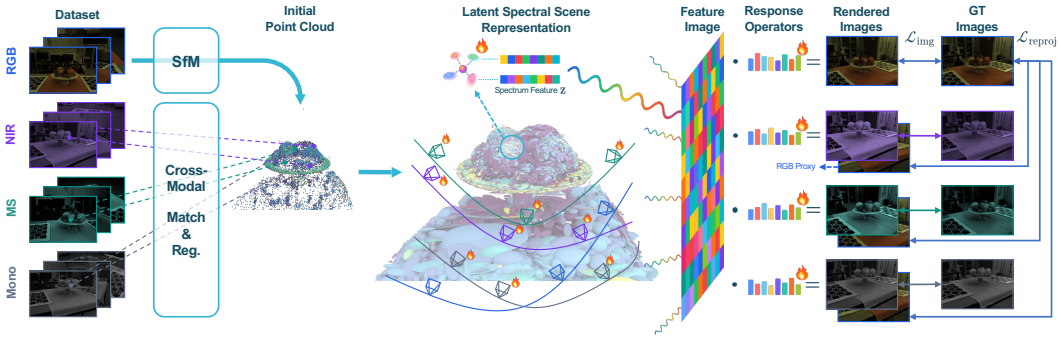

Figure 2: Overview of PoseSpecGS. Cross-Modal Alignment registers auxiliary observations to an RGB SfM backbone and refines their poses with RGB-guided reprojection, while Spectrum-Response Rendering maps the feature image of the Latent Spectral Scene Representation to modality-specific rendered images.

tral attributes (Thirgood et al., 2025; Sinha et al., 2025). NTR-Gaussian and ETGS incorporate thermodynamic evolution (Yang et al., 2025; Wang et al., 2026). PoseSpecGS combines shared spectral-feature compositing with compact linear modality-channel response operators and RGB-guided cross-modal alignment.

Cross-modal pose estimation and refinement. Cross-modal appearance differences motivate specialized matching methods (Tuzcuoğlu et al., 2024; Ren et al., 2025). Existing multimodal systems align independently reconstructed streams or jointly optimize cross-spectral poses and scene content (Poggi et al., 2022; Guo et al., 2025). Reference-based localization combines retrieval and correspondence estimation (Sarlin et al., 2019), while neural rendering methods refine uncertain cameras through image and consistency losses (Lin et al., 2021; Truong et al., 2023; Fu et al., 2024). PoseSpecGS localizes auxiliary views in an RGB SfM frame and uses RGB-guided reprojection consistency between auxiliary-view RGB proxy images and corresponding RGB observations for train-time pose refinement.

3 Method

PoseSpecGS addresses multimodal novel-view synthesis from heterogeneous sensor images:

$$\mathcal{D} = \{\mathbf{I}_m^{(n)} \mid m = 1, \dots, M,\ n = 1, \dots, N_m\}, \tag{1}$$

where $\mathbf{I}_m^{(n)} \in \mathbb{R}^{H_m^{(n)} \times W_m^{(n)} \times C_m}$ has C_m sensing channels and known intrinsics $\mathbf{K}_m^{(n)}$, with each view's camera pose $\mathbf{T}_m^{(n)} \in SE(3)$ to be estimated.

Fig. 2 outlines PoseSpecGS: Cross-Modal Alignment registers auxiliary cameras to an RGB SfM reconstruction for learning a Latent Spectral Scene Representation, and Spectrum-Response Rendering maps its composited spectral features to modality-specific rendered channels. The base formulation assumes static scenes, calibrated intrinsics, and sufficient overlap between each auxiliary view and the RGB images. We do not require synchronized frame correspondences or known cross-sensor extrinsics; observation counts and frame indices may differ across modalities.

3.1 Preliminaries: 3D Gaussian Splatting

3D Gaussian Splatting (Kerbl et al., 2023) represents a scene as a set of anisotropic Gaussian primitives $\mathcal{G} = \{G_i\}_{i=1}^{N}$. In the standard RGB formulation, each primitive is parameterized by:

$$G_i = \{\boldsymbol{\mu}_i, \mathbf{R}_i, \mathbf{s}_i, \alpha_i, \mathbf{c}_i\}, \tag{2}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the center, $\mathbf{R}_i \in SO(3)$ is the orientation, $\mathbf{s}_i \in \mathbb{R}^3_+$ is the scale, $\alpha_i \in [0, 1]$ is the opacity, and $\mathbf{c}_i$ is the color. The 3D covariance is:

$$\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top, \qquad \mathbf{S}_i = \operatorname{diag}(\mathbf{s}_i). \tag{3}$$

For a camera view, each visible Gaussian is projected to the image plane and contributes the screen-space opacity:

$$a_i(\mathbf{p}) = \alpha_i \exp\!\left(-\tfrac{1}{2}(\mathbf{p} - \boldsymbol{\mu}_i')^\top (\boldsymbol{\Sigma}_i')^{-1} (\mathbf{p} - \boldsymbol{\mu}_i')\right), \tag{4}$$

where $\boldsymbol{\mu}_i'$ and $\boldsymbol{\Sigma}_i'$ are the projected mean and covariance. In approximate front-to-back visibility order, alpha compositing gives the pixel color:

$$\mathbf{C}(\mathbf{p}) = \sum_i T_i(\mathbf{p})\, a_i(\mathbf{p})\, \mathbf{c}_i, \qquad T_i(\mathbf{p}) = \prod_{k<i} (1 - a_k(\mathbf{p})). \tag{5}$$

PoseSpecGS retains this projection, visibility ordering, and alpha-compositing rule, but replaces the RGB color $\mathbf{c}_i$ with the learned spectral feature $\mathbf{z}_i$ in Sec. 3.2.

3.2 Latent Spectral Scene Representation

PoseSpecGS represents a scene using a shared scaffold of Gaussian primitives whose geometry, opacity, and spectral features are optimized jointly from all available training modalities. Each primitive carries a D-dimensional spectral feature in a learned latent basis rather than samples of a calibrated or wavelength-continuous spectrum. We refer to the shared D-dimensional space in which Gaussian and composited spectral features are expressed as the spectral feature space.

Anchor-based primitives and spectral features. Following Scaffold-GS Lu et al. (2024b), we maintain anchors $\mathcal{A} = \{(\mathbf{a}_j, \mathbf{h}_j)\}_{j=1}^{N_a}$, where $\mathbf{a}_j \in \mathbb{R}^3$ is an anchor position and $\mathbf{h}_j \in \mathbb{R}^F$ is a learnable anchor code. Each anchor stores R local offsets and anchor scales that instantiate R local Gaussians. The anchor-decoding stage comprises separate branches for opacity, Gaussian geometry, and spectral features. For a query camera, let $\mathbf{d}_j$ be the unit direction from the camera center to anchor $\mathbf{a}_j$. The three branches produce:

$$\{\widetilde{\alpha}_{j,r}\}_{r=1}^{R} = \mathcal{F}_\alpha(\mathbf{h}_j), \quad \{(\widetilde{\mathbf{s}}_{j,r}, \mathbf{q}_{j,r})\}_{r=1}^{R} = \mathcal{F}_{\text{geo}}(\mathbf{h}_j, \mathbf{d}_j), \quad \{\mathbf{z}_{j,r}\}_{r=1}^{R} = \mathcal{F}_{\text{spec}}(\mathbf{h}_j). \tag{6}$$

Here, $\widetilde{\alpha}_{j,r}$ is a tanh opacity score, $(\widetilde{\mathbf{s}}_{j,r}, \mathbf{q}_{j,r})$ parameterizes Gaussian scale and rotation, and $\mathbf{z}_{j,r} \in \mathbb{R}^D$ is the spectral feature of local Gaussian (j, r). Only local Gaussians with positive opacity scores are retained for rasterization, and we set $\alpha_{j,r} = \widetilde{\alpha}_{j,r}$ for each retained Gaussian. These branches are shared across reflected-light modalities, so the number of scene primitives and their geometric parameters do not grow with the number of sensing channels. Thermal's distinct spectral range and image formation are addressed by the RGB–thermal extension in Appendix A.4.

Time-varying appearance. For scenes with time-varying multimodal properties, such as thermal evolution, our static formulation supports a simple and effective extension. We condition anchor codes on normalized scene time $t \in [0, 1]$:

$$\begin{aligned} \gamma(t) &= [t, \{\sin(2^k t)\}_{k=0}^{L-1}, \{\cos(2^k t)\}_{k=0}^{L-1}], \\ \mathbf{h}_j(t) &= \mathbf{h}_j + \mathbf{W}_t \gamma(t) + \mathbf{b}_t. \end{aligned} \tag{7}$$

Here L is the number of frequency pairs, giving a $(2L + 1)$-dimensional encoding. The zero-initialized affine layer produces an F-dimensional offset for spectral feature and opacity prediction, while scene geometry remains time-independent.

Spectral-feature compositing. Using the screen-space opacities and transmittances in Sec. 3.1, we composite the spectral features of retained local Gaussians $i = (j, r)$ along each camera ray:

$$\bar{\mathbf{z}}_{\mathbf{p}} = \sum_i T_i(\mathbf{p})\, a_i(\mathbf{p})\, \mathbf{z}_i, \tag{8}$$

where $\bar{\mathbf{z}}_{\mathbf{p}} \in \mathbb{R}^D$ is the composited spectral feature at pixel $\mathbf{p}$. Collecting the composited spectral features over all pixels yields the feature image $\bar{\mathbf{Z}} \in \mathbb{R}^{H \times W \times D}$. We denote the anchors, anchor codes, stored offsets and scales, and shared anchor-decoder parameters by Θ_{scene}.

3.3 Spectrum-Response Rendering

Fig. 3 illustrates the physical image-formation model that motivates our Spectrum-Response Rendering design. For an RGB, monochrome, near-infrared, or multispectral intensity channel c of modality m, the measured value can be modeled as an integral of wavelength-dependent radiance $\mathcal{L}_{\mathbf{p}}(\lambda)$ against a physical sensor response function (Kawakami et al., 2013):

$$I_{m,c}(\mathbf{p}) \propto \int \mathcal{L}_{\mathbf{p}}(\lambda)\, r_{m,c}^{\text{phys}}(\lambda)\, d\lambda, \tag{9}$$

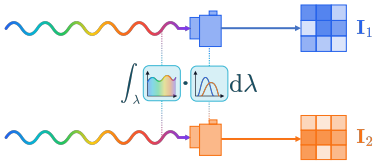

Figure 3: Image-formation model.

where $I_{m,c}(\mathbf{p})$ denotes the ground-truth channel value for channel c in modality m at pixel $\mathbf{p}$, and $r_{m,c}^{\text{phys}}(\lambda)$ is the corresponding physical sensor response function. PoseSpecGS adopts this structure as a design prior to learn a task-oriented factorization of multimodal observations, rather than to recover calibrated spectra or physical sensor response functions. For modality m with C_m channels, we collect its modality-channel response operators as $\mathbf{R}_m = [\mathbf{r}_{m,1}, \ldots, \mathbf{r}_{m,C_m}] \in \mathbb{R}^{D \times C_m}$ and its

response biases as $\mathbf{b}_m = [b_{m,1}, \ldots, b_{m,C_m}]^\top$. The predicted modality observation is then obtained by applying these response operators to the composited spectral feature:

$$\widehat{\mathbf{I}}_m(\mathbf{p}) = \boldsymbol{\psi}_m\left(\mathbf{R}_m^\top \bar{\mathbf{z}}_\mathbf{p} + \mathbf{b}_m\right), \tag{10}$$

where $\widehat{\mathbf{I}}_m(\mathbf{p}) \in \mathbb{R}^{C_m}$ is the rendered channel vector, $\mathbf{r}_{m,c}$ and $b_{m,c}$ parameterize channel c, and $\boldsymbol{\psi}_m$ applies a sigmoid componentwise. This explicit parameterization allows modality-channel response operators to be frozen or adapted independently of the scene representation.

3.4 Cross-Modal Alignment

RGB backbone localization. We first reconstruct the RGB views with COLMAP SfM Schönberger & Frahm (2016) to establish a global coordinate frame, RGB camera poses, and sparse 3D points:

$$\mathcal{B}_{\text{rgb}} = \left(\{\mathbf{T}_{\text{rgb}}^{(v)}\}_{v=1}^{N_{\text{rgb}}}, \{\mathbf{X}_\ell\}_{\ell=1}^{N_{\text{pts}}}\right), \tag{11}$$

where $\mathbf{T}_{\text{rgb}}^{(v)}$ is the world-to-camera pose of RGB image v and N_{pts} is the number of RGB-frame sparse 3D points $\mathbf{X}_\ell$. For each auxiliary image $\mathbf{I}_m^{(u)}$ with $m \neq \text{rgb}$, we retrieve overlapping RGB support views using AnyThermal (Maheshwari et al., 2026) for RGB–thermal retrieval and DINOv3 (Siméoni et al., 2025) otherwise. MINIMA (Ren et al., 2025) then establishes cross-modal 2D–2D matches. Because RGB keypoints are associated with tracks in $\mathcal{B}_{\text{rgb}}$, these matches induce 2D–3D correspondences $\mathcal{M}_u = \{(\mathbf{u}_k, \mathbf{X}_k)\}$ between auxiliary pixels and RGB-frame 3D points. We estimate an initial auxiliary world-to-camera pose with robust PnP:

$$\mathbf{T}_m^{(u,0)} = \underset{\mathbf{T}\in SE(3)}{\arg\min} \sum_{(\mathbf{u}_k,\mathbf{X}_k)\in\mathcal{M}_u} \rho\left(\left\|\mathbf{u}_k - \pi\left(\mathbf{K}_m^{(u)}, \mathbf{T}(\mathbf{X}_k)\right)\right\|_2^2\right), \tag{12}$$

where $\pi(\cdot,\cdot)$ projects a 3D point using camera intrinsics $\mathbf{K}_m^{(u)}$ and candidate pose $\mathbf{T}$, and $\rho(\cdot)$ is the robust penalty used by the PnP estimator. Offline auxiliary-only bundle adjustment jointly refines auxiliary-camera extrinsics with RGB poses and sparse points fixed, initializing the subsequent local differentiable refinement.

Bounded train-time pose refinement. Residual errors after localization can still be absorbed by the scene representation and produce blurred geometry or cross-modal artifacts. For each auxiliary view $m \neq \text{rgb}$, we therefore optimize a local six-DoF residual $\delta\boldsymbol{\xi}_m^{(n)} = [\delta\mathbf{t}_m^{(n)}, \delta\boldsymbol{\omega}_m^{(n)}] \in \mathbb{R}^6$ around its initialized world-to-camera pose $\mathbf{T}_m^{(n,0)}$:

$$\Delta\mathbf{T}_m^{(n)} = \text{Exp}_{SE(3)}\left(\delta\boldsymbol{\xi}_m^{(n)}\right), \qquad \mathbf{T}_m^{(n)} = \Delta\mathbf{T}_m^{(n)}\mathbf{T}_m^{(n,0)}, \tag{13}$$

where $\text{Exp}_{SE(3)}(\cdot)$ maps the residual to $SE(3)$ and left-multiplies the initialized world-to-camera transform. The RGB poses remain frozen, whereas auxiliary pose residuals are optimized jointly with the scene parameters by gradient descent. We clip pose gradients and clamp translation and rotation residuals using the settings in Appendix A.1.

Cross-modal RGB-guided reprojection consistency. We leverage within-modality RGB consistency as a common photometric reference for auxiliary-camera refinement. Specifically, we render an RGB proxy $\widehat{\mathbf{I}}_{\text{rgb}}^q$ and z-depth $\widehat{\mathbf{D}}_z^q$ from the shared scene at the current pose of auxiliary view $q = (m, n)$. For each nearby RGB reference $v \in \mathcal{V}_{\text{rgb}}(q)$, the depth defines a warp $\mathcal{W}_{q\to v}$ and a valid pixel set $\Omega_{q,v}$ with positive finite depth, in-frame projection, and sufficient rendered-alpha support. Let $\mathcal{V}_q^+$ contain the references with sufficient valid support; then:

$$\mathcal{L}_{\text{reproj}} = \mathbb{E}_{q\sim\mathcal{U}} \frac{1}{|\mathcal{V}_q^+|} \sum_{v\in\mathcal{V}_q^+} \frac{1}{C|\Omega_{q,v}|} \sum_{\mathbf{p}\in\Omega_{q,v}} \left\|\widehat{\mathbf{I}}_{\text{rgb}}^q(\mathbf{p}) - \mathbf{I}_{\text{rgb}}^v\left(\mathcal{W}_{q\to v}(\mathbf{p}; \widehat{\mathbf{D}}_z^q)\right)\right\|_1. \tag{14}$$

Here, $\mathcal{U}$ contains the auxiliary views sampled during training, C is the number of RGB proxy channels, and views with $\mathcal{V}_q^+ = \emptyset$ contribute zero. This provides a geometry-sensitive consistency signal in the RGB domain without directly using auxiliary-modality renderings.

We combine ℓ_1, D-SSIM, and cross-modal RGB-guided reprojection consistency:

$$\mathcal{L}_{\text{task}} = (1 - \lambda_{\text{ssim}})\mathcal{L}_1 + \lambda_{\text{ssim}}\mathcal{L}_{\text{D-SSIM}} + \lambda_{\text{reproj}}\mathcal{L}_{\text{reproj}}, \tag{15}$$

where λ_{ssim} and λ_{reproj} weight structural similarity and reprojection consistency, respectively.

Table 1: NVS results for the RGB modality on RGBT-Scenes. Best results: **first**, **second**.

Metric	Method	Dim	DS	Ebk	RB	Trk	RK	Bldg	II	Pt	LS	Avg.
PSNR ↑	3DGS	23.27	21.18	26.17	28.23	23.00	20.74	21.80	24.40	25.65	20.18	23.46
	ThermalGaussian	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**
	PhysIR-Splat	24.27	**22.25**	26.94	28.72	21.93	23.42	**24.89**	22.55	25.41	21.37	24.18
	Ours	**24.74**	**22.41**	**27.49**	**29.34**	**24.22**	**24.87**	**25.31**	**25.81**	**26.26**	**22.52**	**25.30**
SSIM ↑	3DGS	0.842	0.771	0.902	0.917	0.815	0.765	0.827	0.875	0.867	0.688	0.827
	ThermalGaussian	0.858	0.797	0.905	**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**
	PhysIR-Splat	0.856	0.809	0.907	0.907	0.821	0.831	**0.853**	0.850	0.844	0.698	0.838
	Ours	**0.861**	**0.812**	**0.919**	**0.931**	**0.853**	**0.838**	**0.878**	**0.898**	**0.872**	**0.755**	**0.862**
LPIPS ↓	3DGS	0.199	0.271	0.169	**0.197**	0.238	0.220	0.183	0.193	**0.177**	0.289	0.214
	ThermalGaussian	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
	PhysIR-Splat	**0.190**	**0.211**	**0.135**	0.225	**0.218**	**0.167**	**0.153**	0.213	0.181	0.289	**0.198**
	Ours	**0.152**	**0.207**	**0.131**	**0.182**	**0.167**	**0.150**	**0.121**	**0.148**	**0.133**	**0.207**	**0.160**

Table 2: NVS results for the thermal modality on RGBT-Scenes. Best results: **first**, **second**.

Metric	Method	Dim	DS	Ebk	RB	Trk	RK	Bldg	II	Pt	LS	Avg.
PSNR ↑	3DGS	25.99	18.71	20.61	26.55	23.93	26.45	26.83	29.69	24.09	18.48	24.13
	ThermalGaussian	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
	PhysIR-Splat	**28.56**	**22.36**	23.76	**27.74**	24.81	**27.92**	**27.96**	29.59	**27.55**	**22.70**	**26.30**
	Ours	**29.12**	**24.81**	**24.85**	**28.44**	**27.36**	**29.27**	**30.67**	**34.87**	**28.82**	**24.28**	**28.25**
SSIM ↑	3DGS	0.889	0.787	0.812	0.914	0.856	0.922	0.896	0.892	0.867	0.768	0.860
	ThermalGaussian	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
	PhysIR-Splat	**0.911**	**0.849**	**0.878**	**0.929**	0.854	**0.934**	**0.906**	0.904	**0.911**	0.843	**0.892**
	Ours	**0.926**	**0.892**	**0.900**	**0.935**	**0.901**	**0.956**	**0.938**	**0.953**	**0.929**	**0.895**	**0.923**
LPIPS ↓	3DGS	0.127	0.259	0.307	0.209	0.160	0.126	0.185	0.091	0.227	0.378	0.207
	ThermalGaussian	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
	PhysIR-Splat	**0.060**	**0.119**	**0.110**	**0.081**	**0.117**	**0.068**	**0.083**	**0.045**	**0.064**	**0.205**	**0.095**
	Ours	**0.063**	**0.082**	**0.098**	**0.091**	**0.072**	**0.047**	**0.057**	**0.029**	**0.056**	**0.147**	**0.074**

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We use three real-world datasets for static multimodal reconstruction. **RGBT-Scenes** (Lu et al., 2024a) contains over 1,000 paired RGB/thermal images from ten scenes; **MMS-DATA** (Lincetto et al., 2025) contains 32 scenes captured from 50 viewpoints in RGB, Mono, NIR, Polar, and MS; and the **MS-Splatting dataset** (Meyer et al., 2026) provides RGB, green, red, red-edge, and NIR channels. For fair comparison, we evaluate all five MMS-DATA modalities, including Polar, which lies outside our modeling scope; we do not claim physically accurate polarization modeling. We additionally evaluate the temporal extension on the **RHD dataset** (Wang et al., 2026), which provides aligned RGB and dynamic thermal observations of ten scenes.

Implementation details. We use each dataset's official split, and report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), with additional spectral metrics on the MS-Splatting dataset introduced in Appendix B.4. We construct the RGB SfM backbone with COLMAP and localize auxiliary views using cross-modal matches from MINIMA. For standard static experiments, we jointly train the scene representation and refine auxiliary-camera poses with RGB poses fixed. At test time, we refine auxiliary-camera poses with the learned scene fixed, following standard practices in pose-refining NVS evaluation (Wang et al., 2024) with RGB test-view poses remaining fixed. We set $\lambda_{\text{ssim}} = 0.2$ and $\lambda_{\text{reproj}} = 0.05$ and train for 30K iterations on one NVIDIA GeForce RTX 4090 GPU. For time conditioning, we use $L = 6$ frequency pairs. Unless otherwise specified, baselines use dataset-provided poses without pose refinement and retain their default optimization settings.

4.2 Main Multimodal Novel-View Synthesis

RGBT-Scenes. We compare PoseSpecGS with per-modality 3DGS (Kerbl et al., 2023), ThermalGaussian (Lu et al., 2024a), MMOne (Gu & Wang, 2025), and PhysIR-Splat (Gao et al., 2026). We reproduce PhysIR-Splat using its released code, as the paper does not report results for all ten scenes. Tab. 1 and Tab. 2 show that PoseSpecGS achieves the best average results across both modalities and all three metrics, with the most pronounced gains in thermal reconstruction. Fig. 4 presents qualitative comparisons, showing that PoseSpecGS reconstructs more faithful structural details and more accurate thermal intensities.

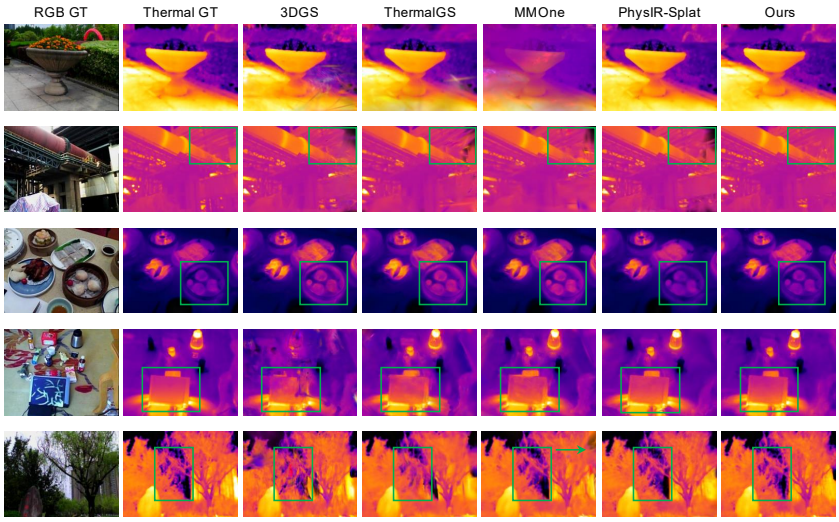

Figure 4: Qualitative RGBT-Scenes comparison.

Table 3: PSNR and SSIM results on representative scenes from MMS-DATA. Best results: **first**, **second**.

Metric	Method	Bouquet					Forest Gang 2					Fruits					Average				
		RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS
PSNR ↑	MS-Splatting	22.94	19.39	20.70	20.53	19.90	26.78	20.99	21.44	22.18	22.35	26.07	18.99	19.20	18.64	19.10	25.26	19.79	20.45	20.45	20.45
	X-NeRF	26.16	N/A	27.46	N/A	24.09	25.82	N/A	26.32	N/A	23.86	27.91	N/A	27.73	N/A	25.57	26.63	N/A	27.17	N/A	24.51
	MMS	**29.45**	**31.78**	**34.20**	**28.48**	**29.26**	**31.47**	**32.41**	**33.16**	**32.85**	**30.79**	**35.30**	**33.64**	**34.82**	**32.79**	**34.15**	**32.07**	**32.61**	**34.06**	**31.37**	**31.40**
	Ours	**36.68**	**34.55**	**36.27**	**34.13**	**34.47**	**34.89**	**33.38**	**35.20**	**33.36**	**32.85**	**37.44**	**34.51**	**34.30**	**33.19**	**34.04**	**36.34**	**34.15**	**35.26**	**33.56**	**33.79**
SSIM ↑	MS-Splatting	0.811	0.789	0.799	0.817	0.744	0.882	0.817	0.817	0.838	0.806	0.867	0.784	0.776	0.792	0.736	0.853	0.797	0.797	0.816	0.762
	X-NeRF	0.874	N/A	0.866	N/A	0.772	0.875	N/A	0.869	N/A	0.795	0.899	N/A	0.882	N/A	0.810	0.883	N/A	0.872	N/A	0.792
	MMS	**0.937**	**0.956**	**0.971**	**0.932**	**0.923**	**0.945**	**0.959**	**0.972**	**0.946**	**0.948**	**0.971**	**0.966**	**0.980**	**0.960**	**0.959**	**0.951**	**0.960**	**0.974**	**0.946**	**0.943**
	Ours	**0.977**	**0.974**	**0.981**	**0.974**	**0.966**	**0.970**	**0.972**	**0.981**	**0.970**	**0.963**	**0.980**	**0.972**	**0.966**	**0.972**	**0.965**	**0.976**	**0.973**	**0.976**	**0.972**	**0.965**

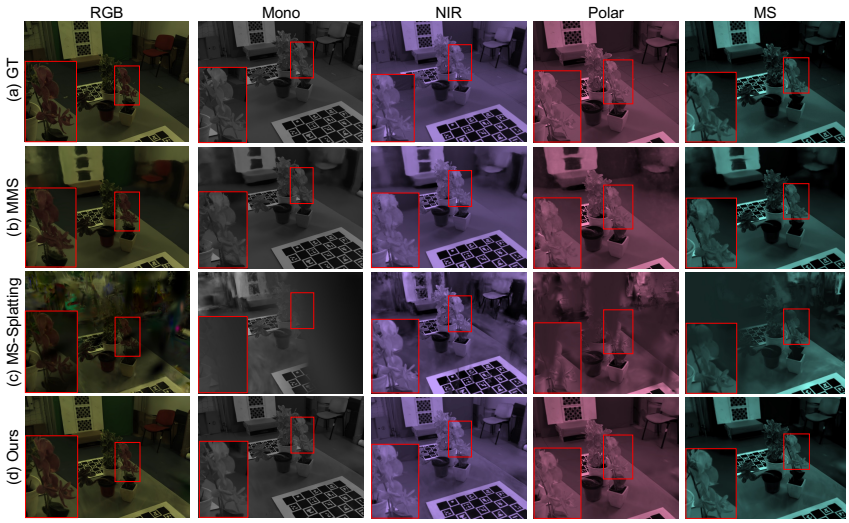

Figure 5: Qualitative MMS-DATA comparison.

MMS-DATA. We compare against X-NeRF Poggi et al. (2022), MMS (Lincetto et al., 2025), and MS-Splatting (Meyer et al., 2026). Unless otherwise specified, MMS-DATA evaluations use demosaicked images with dataset-provided foreground masks. X-NeRF is evaluated only on the modalities it supports. For visualization of MMS-DATA, Polar and MS images are averaged over their respective channels for both ground truth and predictions. Tab. 3 and Tab. 4 show that PoseSpecGS achieves the strongest average results across the five evaluated modality groups, while Fig. 5 shows fewer artifacts and sharper structures. Please see Appendix Tab. 15 and Appendix D.1 for more details.

Table 4: PSNR averaged over all 32 MMS-DATA scenes. Best results: **first**, **second**.

Method	RGB	Mono	NIR	Polar	MS
MS-Splatting	22.81	18.24	18.75	19.08	18.94
X-NeRF	25.73	N/A	26.97	N/A	23.63
MMS	**32.34**	**32.43**	**33.81**	**29.79**	**30.05**
Ours	**34.54**	**32.61**	**34.01**	**32.12**	**32.51**

Table 5: NVS results on the MS-Splatting dataset. Best results: **first**, **second**.

Method	PSNR ↑						SSIM ↑						SAM ↓	SCM ↑	SID ↓	Iterations ↓
	All	RGB	G	R	RE	NIR	All	RGB	G	R	RE	NIR				
3DGS, per modality	23.62	20.86	22.92	25.23	23.33	25.75	0.721	0.623	0.667	0.769	0.730	0.815	0.146	0.812	0.044	30K × 5
ThermalGaussian(MS)	24.18	20.17	24.04	26.09	24.14	26.50	0.718	0.564	0.695	0.788	0.732	0.814	0.127	0.843	0.035	**120K**
TIMS	24.99	**21.29**	24.62	26.61	25.07	27.42	0.755	**0.647**	0.719	0.810	0.766	0.834	0.116	0.860	0.031	**120K**
FeatSplat(MS)	25.20	20.75	25.12	27.49	25.13	27.57	**0.766**	0.641	**0.733**	**0.828**	**0.781**	0.848	0.114	0.879	0.026	**120K**
MS-Splatting	**25.65**	21.17	**25.42**	**27.79**	**25.79**	**28.13**	0.763	0.633	0.729	0.827	0.780	**0.849**	**0.109**	**0.888**	**0.024**	**120K**
Ours	**26.52**	**21.85**	**26.53**	**28.59**	**26.70**	**28.94**	**0.798**	**0.659**	**0.783**	**0.855**	**0.816**	**0.877**	**0.094**	**0.902**	**0.017**	**30K**

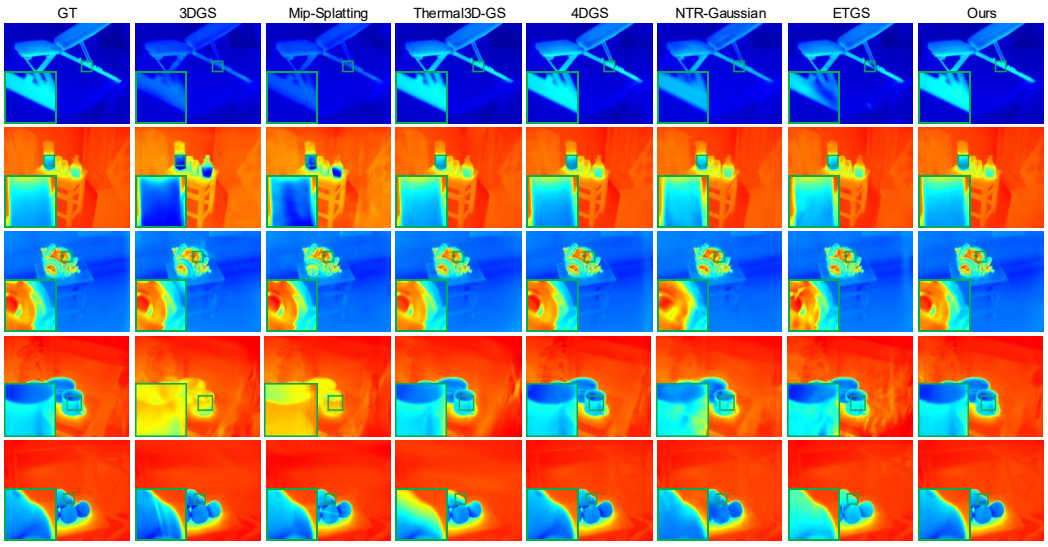

Figure 6: Qualitative thermal comparison on the RHD dataset (pseudocolor visualization). **Please refer to the supplementary video for dynamic results.**

For the comparison in Tab. 6, all three methods use the same RGB-backbone localization on the representative MMS-DATA scenes in Tab. 3. We equip per-modality 3DGS and MS-Splatting with our train-time pose-refinement scheme, and apply the test-time evaluation protocol in Sec. 4.1 to all methods. Tab. 6 shows that PoseSpecGS retains the best overall quality among these pose-refining methods. Please see Appendix A.3, Appendix C.2, and Appendix C.7 for more details.

Table 6: Average PSNR and SSIM on representative MMS-DATA scenes with pose refinement for all methods.

Metric	Method	RGB	Mono	NIR	Polar	MS
PSNR ↑	3DGS, per modality	31.87	29.17	29.75	29.04	28.44
	MS-Splatting	35.44	31.79	34.31	32.76	32.90
	Ours	**36.34**	**34.15**	**35.26**	**33.56**	**33.79**
SSIM ↑	3DGS, per modality	0.949	0.948	0.949	0.946	0.916
	MS-Splatting	0.973	0.967	**0.976**	0.969	0.959
	Ours	**0.976**	**0.973**	**0.976**	**0.972**	**0.965**

MS-Splatting dataset. Tab. 5 compares PoseSpecGS with per-modality 3DGS, TIMS (Grün et al., 2025), MS-Splatting (Meyer et al., 2026), ThermalGaussian(MS), and FeatSplat(MS) (Martins & Civera, 2024), with (MS) denoting multispectral adaptations implemented by MS-Splatting. PoseSpecGS achieves the best PSNR/SSIM and spectral metrics with 30K iterations, compared with 120K for the joint baselines. Please see Appendix Tab. 16 and Appendix D.2 for more details.

4.3 Extensibility to Time-Varying RGB–Thermal Scenes

We evaluate the time-conditioned extension on the RHD dataset for joint RGB and time-varying thermal reconstruction. We use the camera poses from the dataset and disable pose optimization. Tab. 7 compares PoseSpecGS with 3DGS, Mip-Splatting (Yu et al., 2024), Thermal3D-GS (Chen et al., 2024b), 4DGS (Wu et al., 2024), NTR-Gaussian (Yang et al., 2025), and ETGS (Wang et al., 2026) for thermal reconstruction. PoseSpecGS achieves the best average thermal results across all three metrics, while Fig. 6 shows fewer floaters and more accurate dynamic thermal intensities, demonstrating the representation's extensibility to time-varying RGB–thermal scenes. Please see Appendix Tab. 17 and Appendix D.3 for more details.

Table 7: Average dynamic NVS results for the thermal modality on the RHD dataset. Best results: **first**, **second**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS	32.16	0.978	0.078
Mip-Splatting	31.51	0.976	0.085
Thermal3D-GS	34.68	0.983	0.072
4DGS	33.94	0.972	0.076
NTR-Gaussian	34.96	0.981	0.089
ETGS	**40.68**	**0.989**	**0.050**
Ours	**43.01**	**0.992**	**0.038**

4.4 Cross-Scene One-to-Multimodal Transfer under RGB-Only Supervision

Table 8: Representation ablations averaged over representative MMS-DATA scenes. Parentheses show changes from Ours (red: lower; green: equal or higher).

Configuration	PSNR ↑					SSIM ↑				
	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS
Intensity compositing	34.73 (-1.61)	31.78 (-2.37)	34.29 (-0.97)	32.70 (-0.86)	32.49 (-1.30)	0.973 (-0.003)	0.966 (-0.007)	0.975 (-0.001)	0.969 (-0.003)	0.959 (-0.006)
Modality-specific feature decoding	33.85 (-2.49)	31.67 (-2.48)	33.67 (-1.59)	32.19 (-1.37)	31.85 (-1.94)	0.970 (-0.006)	0.964 (-0.009)	0.974 (-0.002)	0.967 (-0.005)	0.954 (-0.011)
Joint response MLP	35.44 (-0.90)	33.29 (-0.86)	35.19 (-0.07)	32.91 (-0.65)	32.76 (-1.03)	0.974 (-0.002)	0.971 (-0.002)	0.979 (+0.003)	0.970 (-0.002)	0.962 (-0.003)
Ours	**36.34**	**34.15**	**35.26**	**33.56**	**33.79**	**0.976**	**0.973**	**0.976**	**0.972**	**0.965**

Table 9: Alignment ablations averaged over representative MMS-DATA scenes. Each row below Ours additionally removes its named component; parentheses show changes from Ours (red: lower; green: equal or higher).

Configuration	PSNR ↑					SSIM ↑				
	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS
Ours	**36.34**	**34.15**	**35.26**	**33.56**	**33.79**	**0.976**	**0.973**	**0.976**	**0.972**	**0.965**
– RGB-guided reprojection consistency	35.34 (-1.00)	34.09 (-0.06)	34.59 (-0.67)	33.39 (-0.17)	33.13 (-0.66)	0.975 (-0.001)	0.971 (-0.002)	0.972 (-0.004)	0.972 (+0.000)	0.964 (-0.001)
– Pose refinement	34.23 (-2.11)	32.00 (-2.15)	29.00 (-6.26)	27.59 (-5.97)	25.56 (-8.23)	0.968 (-0.008)	0.957 (-0.016)	0.926 (-0.050)	0.939 (-0.033)	0.884 (-0.081)
– RGB backbone localization	21.18 (-15.16)	21.99 (-12.16)	25.62 (-9.64)	23.27 (-10.29)	22.16 (-11.63)	0.755 (-0.221)	0.801 (-0.172)	0.875 (-0.101)	0.840 (-0.132)	0.740 (-0.225)

To evaluate our method's ability to encode transferable cross-modal information, we follow SPoILeR's supervision, regularization, and evaluation protocol on MMS-DATA (Lincetto et al., 2026): pretraining on RGB, Mono, NIR, and nine-channel MS from 27 scenes, followed by synthesis of unobserved modalities in 5 held-out scenes under RGB-only supervision. The only modification is a shared spatial basis field with scene-specific coefficients; all decoder dimensions and modality-channel response operators remain unchanged. As shown in Tab. 10, PoseSpecGS outperforms SPoILeR in RGB/Mono/NIR, while trailing by only 0.28 dB in MS PSNR. The small MS gap may stem from the greater sensitivity of our compact response operators to variations in the spectral feature distribution when generalizing highly diverse MS channels from RGB-only supervision; the MS response operator has only 0.29k trainable parameters, whereas SPoILeR's three-layer MLP has 5.97k. These results demonstrate competitive cross-scene transfer with substantially fewer pretraining iterations. Additional details and pairwise reuse experiments are provided in Appendix B.1 and Appendix C.4.

Table 10: Average cross-scene NVS results on five held-out MMS-DATA scenes.

Metric	Method	RGB	Mono	NIR	MS
PSNR ↑	SPoILeR	30.07	25.78	26.55	**25.45**
	Ours	**31.66**	**26.55**	**27.25**	25.17
SSIM ↑	SPoILeR	N/A	0.880	0.870	N/A
	Ours	N/A	**0.889**	**0.905**	N/A

4.5 Ablation Studies

Effectiveness of representation design. We evaluate three rendering variants on the representative MMS-DATA scenes in Tab. 3. Intensity compositing uses view-dependent per-modality color decoders and alpha-composites their sigmoid-mapped Gaussian intensities. Modality-specific feature decoding assigns each modality a separate spectral-feature decoder while sharing anchor codes and geometry. Joint response MLP maps 32-dimensional composited spectral features jointly to all output channels. Tab. 8 shows that the full model achieves the best overall reconstruction quality among these variants. Please see Appendix A.2 for implementation details of the variants.

Effectiveness of Cross-Modal Alignment. On the same scenes, we first remove cross-modal RGB-guided reprojection consistency, then additionally disable train- and test-time pose refinement, and finally remove RGB backbone localization and use the dataset-provided poses directly. Tab. 9 shows that overall reconstruction quality declines as these components are cumulatively removed, demonstrating the contribution of Cross-Modal Alignment to NVS quality.

5 Conclusion

We presented PoseSpecGS, a pose-aligned Gaussian splatting method with spectrum-response rendering for multimodal novel-view synthesis. Spectrum-Response Rendering synthesizes modality-specific images from shared Gaussian spectral features using compact modality-channel response operators. Cross-Modal Alignment registers auxiliary cameras and jointly optimizes poses and scene parameters with RGB-guided reprojection consistency, without known rig extrinsics or synchronization. The representation supports static and time-conditioned multimodal reconstruction. Moreover, response operators pretrained across scenes enable zero-shot one-to-multimodal transfer under RGB-only supervision. Experiments demonstrate state-of-the-art reconstruction quality on the evaluated benchmarks with reduced model storage and fewer training iterations. As a limitation, time-conditioned reconstruction assumes fixed geometry. Future work includes extending PoseSpecGS to deforming scenes and scaling multimodal pretraining to improve zero-shot cross-scene one-to-multimodal generalization.

AI USE STATEMENT

AI tools were used to improve the clarity and readability of the manuscript and assist with literature search. The authors take full responsibility for the final work.

ETHICS STATEMENT

Our study focuses on multimodal novel-view synthesis. All experimental evaluations are conducted using publicly available datasets, curated to avoid sensitive or private content. We assert that this research has been carried out in accordance with the code of ethics.

REPRODUCIBILITY STATEMENT

In order to support reproducibility and verification, we present implementation and evaluation details in our paper, and will release the associated source code upon acceptance.

REFERENCES

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4160–4169, June 2023.

Gerry Chen, Sunil Kumar Narayanan, Thomas Gautier Ottou, Benjamin Missaoui, Harsh Muriki, Cédric Pradalier, and Yongsheng Chen. Hyperspectral neural radiance fields. *arXiv preprint arXiv:2403.14839*, 2024a.

Qian Chen, shihao Shu, and Xiangzhi Bai. Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis. In *European Conference on Computer Vision*, 2024b.

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20796–20805, June 2024.

Jingyuan Gao, Yumeng Hu, Fei Gao, and Mingjin Zhang. Physir-splat: Physically consistent thermal infrared radiative transfer in 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11818–11828, 2026.

Josef Grün, Lukas Meyer, Maximilian Weiherer, Bernhard Egger, Marc Stamminger, and Linus Franke. Towards Integrating Multi-Spectral Imaging with Gaussian Splatting. In Bernhard Egger and Tobias Günther (eds.), *Vision, Modeling, and Visualization*. The Eurographics Association, 2025. ISBN 978-3-03868-294-3. doi: 10.2312/vmv.20251237.

Zhifeng Gu and Bing Wang. Mmone: Representing multiple modalities in one scene. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1088–1098, October 2025.

Haipeng Guo, Huanyu Liu, Jiazheng Wen, and Junbao Li. Cross-spectral gaussian splatting with spatial occupancy consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3229–3237, 2025. doi: 10.1609/aaai.v39i3.32333.

Rei Kawakami, Hong-Xun Zhao, Robby T. Tan, and Katsushi Ikeuchi. Camera spectral sensitivity and white balance estimation from sky images. *International Journal of Computer Vision*, 105:187 – 204, 2013. URL https://api.semanticscholar.org/CorpusID:255101319.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5741–5751, October 2021.

Yvette Y Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. ThermalNeRF: Thermal radiance fields. In *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2024.

Federico Lincetto, Gianluca Agresti, Mattia Rossi, and Pietro Zanuttigh. Multimodalstudio: A heterogeneous sensor dataset and framework for neural rendering across multiple imaging modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Federico Lincetto, Gianluca Agresti, Mattia Rossi, Piergiorgio Sartor, and Pietro Zanuttigh. Learning spectral and polarimetric clues for one-to-multimodal novel view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermalgaussian: Thermal 3d gaussian splatting. *arXiv preprint arXiv:2409.07200*, 2024a.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024b.

Parv Maheshwari, Jay Karhade, Yogesh Chawla, Isaiah Adu, Florian Heisen, Andrew Porco, Andrew Jong, Yifei Liu, Santosh Pitla, Sebastian Scherer, and Wenshan Wang. Anythermal: Towards learning universal representations for thermal perception, 2026. URL https://arxiv.org/abs/2602.06203.

T. Berriel Martins and Javier Civera. Feature splatting for better novel view synthesis with low overlap. *British Machine Vision Conference (BMVC)*, 2024.

Lukas Meyer, Josef Grün, Maximilian Weiherer, Bernhard Egger, Marc Stamminger, and Linus Franke. Multi-spectral gaussian splatting with neural color representation. *Computer Graphics Forum*, 45(2), 2026. doi: 10.1111/cgf.70337. URL https://meyerls.github.io/ms_splatting.

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. In *ECCV*, 2020.

Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T. Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM Trans. Graph.*, 2023.

Matteo Poggi, Pierluigi Zama Ramirez, Fabio Tosi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Cross-spectral neural radiance fields. In *Proceedings of the International Conference on 3D Vision*, 2022. 3DV.

Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Saptarshi Neil Sinha, Holger Graf, and Michael Weinmann. Spectralgaussians: Semantic, spectral 3d gaussian splatting for multi-spectral scene representation, visualization and analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:789–803, 2025. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2025.06.008. URL http://dx.doi.org/10.1016/j.isprsjprs.2025.06.008.

Christopher Thirgood, Oscar Mendez, Erin Ling, Jon Storey, and Simon Hadfield. Hypergs: Hyperspectral 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5970–5979, June 2025.

Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4190–4200, June 2023.

Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4275–4286, 2024.

Weiyao Wang, Pierre Gleize, Hao Tang, Xingyu Chen, Kevin J Liang, and Matt Feiszli. Icon: Incremental confidence for joint pose and radiance field optimization. *CVPR*, 2024.

Zhongwen Wang, Han Ling, Weihao Zhang, Yinghui Sun, and Quansen Sun. Etgs: Explicit thermodynamics gaussian splatting for dynamic thermal reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, June 2024.

Kun Yang, Yuxiang Liu, Zeyu Cui, Yu Liu, Maojun Zhang, Shen Yan, and Qing Wang. Ntr-gaussian: Nighttime dynamic thermal reconstruction with 4d gaussian splatting based on thermodynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 691–700, June 2025.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

APPENDIX

The appendix provides implementation details in Appendix A, method and evaluation details in Appendix B, additional quantitative analyses in Appendix C, and qualitative results in Appendix D.

A IMPLEMENTATION DETAILS

A.1 CROSS-MODAL ALIGNMENT

We first associate the RGB-side keypoints in the MINIMA correspondences with triangulated RGB keypoints using mutual nearest-neighbor matching within a radius of 6 pixels in the RGB images. We then estimate the camera poses using PnP with RANSAC, with the reprojection error threshold set to 3 pixels.

Appendix Tab. 11 summarizes the specific settings used for train-time pose optimization in the static experiments. During optimization, only the extrinsics of the auxiliary cameras are updated, while the RGB camera poses and the intrinsics remain fixed and define the reference coordinate frame.

Table 11: Train-time pose refinement settings for Cross-Modal Alignment.

Optimization Aspect	**Configuration**
Optimized parameters	Auxiliary-camera extrinsics
Pose residual	Local 6-DoF, $SE(3)$
RGB camera pose	Fixed (not optimized)
Translation learning rate	1×10^{-3}
Rotation learning rate	5×10^{-4}
Gradient-norm clipping threshold	0.1
Translation residual bound	± 0.25 scene-coordinate units per component
Rotation residual bound	± 0.25 rad per component
Depth used for RGB reprojection	Rendered z-depth

A.2 SCENE REPRESENTATION AND SPECTRAL RENDERING

Anchors and decoder branches. Appendix Tab. 12 provides the detailed network architecture configurations and the number of channels used for each modality. The anchors are initialized by voxelizing the sparse point cloud produced by the RGB SfM backbone in Eq. 11. The three branches in Eq. 6 are implemented as lightweight MLPs, each consisting of a single hidden layer with ReLU activation.

Table 12: Configuration of the scene representation. $\mathrm{L}(a \to b)$ denotes a linear layer.

Component	**Specification**
Anchor code dimension F	32
Local Gaussians per anchor R	10
Spectral feature dimension D	32
Anchor initialization	Voxelized RGB SfM points, with voxel size 10^{-3}
Opacity branch	$\mathrm{L}(F \to F)$, ReLU, $\mathrm{L}(F \to R)$, $\tanh$
Geometry branch	$\mathrm{L}(F+3 \to F)$, ReLU, $\mathrm{L}(F \to 7R)$
Spectral feature branch	$\mathrm{L}(F \to F)$, ReLU, $\mathrm{L}(F \to RD)$
MMS-DATA channels	RGB 3, Monochrome 1, NIR 1, Polarization 4, Multi-spectral 9; 18 in total
MS-Splatting channels	RGB 3, G 1, R 1, RE 1, NIR 1; 7 in total
Thermal imaging channels	3(RGBT-Scenes), 1(RHD)

Representation ablation configurations. Intensity compositing replaces the shared spectral-feature branch and post-compositing response operators with a view-dependent Scaffold-style color MLP for each modality. Its 35-dimensional input concatenates the 32-dimensional anchor code and three-dimensional viewing direction; a 32-unit ReLU hidden layer is followed by a sigmoid output whose

size is 10 times that modality's channel count. Its output provides the channel intensities for the anchor's 10 local Gaussians, which are then alpha-composited.

Modality-specific feature decoding uses a separate spectral-feature decoder for each modality, mapping the shared anchor codes to modality-dependent per-Gaussian spectral features while retaining shared Gaussian geometry. These features are alpha-composited before the original modality-channel response operators and sigmoid are applied.

Joint response MLP retains feature compositing but replaces the compact linear channel-response operators in Eq. 10 with a two-layer response MLP. Applied pixelwise to the composited 32-dimensional feature image, it has dimensions $32 \to 32 \to 18$, with a ReLU hidden activation and a sigmoid output; the channels of the queried modality are then selected.

Spectral feature space. The spectral feature space is designed to provide a shared spectral basis for all channels of the reflected-light modalities, while modeling the modality-channel response operators as separately parameterized components. Keeping these factors explicit provides an interface for fitting a new channel using only its $D + 1$ response parameters (a D-dimensional operator and a scalar bias), while the existing scene representation and anchor decoder remain fixed. We exploit these separately parameterized mappings for reuse in Appendix C.4 and adaptation in Appendix C.5. The model compactness reported in Appendix Tab. 22 primarily stems from the unified Gaussian spectral representation across modalities, with an anchor-based structure and shared base geometry and opacity parameters.

Channel response operators. The learned response operators $\mathbf{r}_{m,c} \in \mathbb{R}^D$ and response biases $b_{m,c}$ are independently parameterized for each modality and channel, shared across all pixels and views of a scene, and shared across scenes in the reuse experiments of Appendix B.2.

Rendering feature images. The feature image $\bar{\mathbf{Z}}$ is generated by the Gaussian rasterizer. The rasterizer can synthesize at most 32 channels in a single pass. When the dimensionality of the feature space exceeds 32, we rasterize the features in blocks of 32 channels. All blocks share the same projected Gaussian parameters and visibility ordering, and their outputs are subsequently concatenated. The response operators are then applied in screen space.

Rendering depth. The renderer additionally accumulates the camera-space z coordinate using the same compositing weights as those used for feature rendering, and divides the accumulated value by the accumulated alpha. For pixels receiving no Gaussian contribution, the rendered depth is set to zero. The rendered depth is used solely as an internal geometric quantity. It is neither compared against any depth observations nor subjected to any form of direct depth supervision.

Thermal modeling. Thermal differs from the reflected-light modalities in both its emissive image formation and spectral range, motivating a separate branch with implementation details in Appendix A.4 and modeling scope in Appendix B.3.

A.3 Optimization Settings

Data processing. Each dataset's official split defines reconstruction training and evaluation. For standard reconstruction, each iteration samples one view uniformly without replacement from the multimodal training-view pool. Once every view in the pool has been visited, the pool is randomly reshuffled and a new sampling cycle begins. Images from each modality are normalized according to their native bit depth by dividing the pixel values by the maximum representable value of the corresponding integer data type. For RHD, grayscale thermal images are replicated across three channels for training; no pseudocolor mapping is applied to the training images.

Training objective. In addition to the loss terms defined in Eq. 15, we retain the Gaussian volume regularization term used in Scaffold-GS Lu et al. (2024b). This term is defined as the mean product of the scales along the three axes of the local Gaussians retained for the current rendered view:

$$\mathcal{L}_{\text{vol}} = \frac{1}{|\mathcal{G}_{\text{vis}}|} \sum_{i \in \mathcal{G}_{\text{vis}}} \prod_{k=1}^{3} s_{i,k}, \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{vol}} \mathcal{L}_{\text{vol}}, \tag{16}$$

where $\mathcal{G}_{\text{vis}}$ denotes the set of local Gaussians retained for the current rendered view, and $s_{i,1}, s_{i,2}, s_{i,3}$ denote the scales of the i-th Gaussian along the three axes. We set $\lambda_{\text{vol}} = 0.01$ and keep it fixed across all experiments.

Training schedule. Appendix Tab. 13 and Appendix Tab. 14 report the training configurations and learning-rate settings for the different parameter groups, respectively. For parameters whose learning rates are decayed, we linearly interpolate the learning rate in log space over 30K training iterations. Both the ℓ_1 loss and the D-SSIM loss are averaged over the pixel and channel dimensions. The scene-space scaling factor in Appendix Tab. 14 is defined as $s_{\text{scene}} = 1.1 \max_v \|\mathbf{o}_v - \bar{\mathbf{o}}\|_2$, where $\mathbf{o}_v$ denotes the camera center of each training camera across all modalities under its initial pose, and $\bar{\mathbf{o}}$ denotes the mean of these camera centers.

Table 13: Standard static training and reprojection settings.

Configuration	**Value**
Number of training iterations	30K
Loss weights $(\lambda_{\text{ssim}}, \lambda_{\text{reproj}}, \lambda_{\text{vol}})$	(0.2, 0.05, 0.01)
Adam parameters (β_1, β_2)	(0.9, 0.999)
Adam ϵ	10^{-15}
Maximum number of RGB reference views per query view	3
Alpha threshold for valid reprojection pixels	0.05
Minimum fraction of valid pixels per reference view	1%

Train-time pose refinement. For standard static reconstruction, we jointly optimize the scene representation and auxiliary-camera poses for 30K iterations using only training views, while RGB poses remain fixed. No learned scene or rendering parameters are optimized using test-view images in this protocol. RHD instead uses the dataset-provided poses with pose optimization disabled.

During train-time pose refinement, the auxiliary camera poses are updated using a separate Adam optimizer, with the specific settings given in Appendix Tab. 11. When enabled, pose refinement and RGB-guided reprojection consistency are active from the first iteration. On MMS-DATA, the RGB reference views for each auxiliary view are retrieved using DINOv3 during Cross-Modal Alignment and selected from the RGB training views. For each query view, at most three reference views are retained. The alpha threshold and valid-pixel-ratio threshold used to determine whether a reference view is included in $\mathcal{V}_q^+$ in Eq. 14 are listed in Appendix Tab. 13. Both the RGB proxy images and the camera-space depth used for warping are differentiably rendered from the shared scene representation. As a result, gradients from $\mathcal{L}_{\text{reproj}}$ propagate to both the scene parameters and the auxiliary-camera pose residuals, while the RGB reference observations are always treated as constants. The comparison with pose-refined baselines is described in Sec. 4.2. Specifically, the pose-refined 3DGS baseline in Tab. 6 is trained independently for each modality and therefore does not use cross-modal reprojection consistency during pose refinement.

Test-time pose refinement. Following prior work on test-time camera alignment for NVS evaluation Lin et al. (2021); Park et al. (2023); Truong et al. (2023); Bian et al. (2023); Fu et al. (2024); Wang et al. (2024), we refine test-camera poses photometrically to reduce misalignment effects on reconstruction metrics. For each auxiliary-modality test view, we optimize its local six-DoF extrinsic residual (three translation and three rotation components) from the localization initialization using the $SE(3)$ parameterization in Eq. 13. All scene and rendering parameters, camera intrinsics, and other views' poses remain frozen; RGB poses are always fixed. We use Adam for 300 iterations with translation and rotation learning rates of 1×10^{-3} and 5×10^{-4}, respectively, as in training (Appendix Tab. 11). Following CamP Park et al. (2023), the objective is the mean squared error between the rendered image and the corresponding-modality test image, without SSIM or RGB-guided reprojection terms. We then render with the refined pose to compute PSNR, SSIM, and LPIPS. The RGB-supervised cross-scene reconstruction experiments in Appendix B.1 instead use fixed poses; RHD does not use test-time pose refinement.

A.4 Static RGB–Thermal Reconstruction

In the thermal branch, each local Gaussian (j, r) carries a thermal feature $\mathbf{z}_{j,r}^{\text{th}} \in \mathbb{R}^{D_{\text{th}}}$. This feature is decoded from the anchor code $\mathbf{h}_j$, with $D_{\text{th}} = 32$. We implement a Beer–Lambert-style distance-dependent attenuation model, under which the thermal feature is attenuated along the viewing path

as $\mathbf{z}_{j,r}^{\mathrm{th}} \odot \exp(-\boldsymbol{\beta}_j d_j)$, where $\odot$ denotes element-wise multiplication and d_j denotes the distance from the camera center to the anchor $\mathbf{a}_j$. The learned per-channel attenuation coefficients $\boldsymbol{\beta}_j = \mathrm{softplus}(\mathcal{F}_\beta(\mathbf{h}_j)) \in \mathbb{R}^{D_{\mathrm{th}}}$ are predicted by a two-layer network with width 32.

The thermal branch additionally predicts a thermal-specific opacity component, together with residual adjustments to the Gaussian scale and rotation. A two-layer network of width 32 produces three scale-logit residuals $\Delta\widetilde{\mathbf{s}}_{j,r}$ and four quaternion-component residuals $\Delta\mathbf{q}_{j,r}$ per local Gaussian, each bounded elementwise by $0.1\tanh(\cdot)$. For the shared covariance-decoder outputs $(\widetilde{\mathbf{s}}_{j,r}, \mathbf{q}_{j,r})$ in Eq. 6 and anchor log-scales $\boldsymbol{\ell}_j$, the thermal geometry is:

$$\begin{aligned} \mathbf{s}_{j,r}^{\mathrm{th}} &= \exp(\boldsymbol{\ell}_j) \odot \mathrm{sigmoid}(\widetilde{\mathbf{s}}_{j,r} + \Delta\widetilde{\mathbf{s}}_{j,r}), \\ \mathbf{q}_{j,r}^{\mathrm{th}} &= \frac{\mathbf{q}_{j,r} + \Delta\mathbf{q}_{j,r}}{\|\mathbf{q}_{j,r} + \Delta\mathbf{q}_{j,r}\|_2}. \end{aligned} \tag{17}$$

Thus, the bound applies to the raw residual components, not to the final scale or rotation angle. During thermal rendering, only local Gaussians with opacity greater than 2×10^{-3} are retained. Gradients from the thermal branch are allowed to propagate into the shared scene representation. For the RGBT-Scenes dataset specifically, we supervise the thermal branch with three-channel pseudocolor thermal images and compute PSNR, SSIM, and LPIPS on the same pseudocolor images, the same as ThermalGaussian Lu et al. (2024a). The learning rate of the attenuation network follows the setting used for the spectral feature decoder in Appendix Tab. 14, and the covariance-residual network uses the learning-rate setting of the geometry decoder. All remaining settings follow Appendix A.3.

Table 14: Learning-rate schedules for standard static MMS-DATA training. An arrow denotes log-linear decay from the initial to the final learning rate; a single value denotes a constant learning rate. The spatial multiplier s_{scene} is defined in Appendix A.3.

Parameter Group	**Learning-rate schedule**
Local offsets	$10^{-2} s_{\mathrm{scene}} \rightarrow 10^{-4} s_{\mathrm{scene}}$
Anchor codes	7.5×10^{-3}
Anchor log-scales	7×10^{-3}
Opacity decoder	$2 \times 10^{-3} \rightarrow 2 \times 10^{-5}$
Geometry decoder	4×10^{-3}
Spectral-feature decoder, response operators, and biases	$8 \times 10^{-3} \rightarrow 5 \times 10^{-5}$

B Method and Evaluation Details

B.1 RGB-Supervised Cross-Scene Reconstruction

Tab. 10 follows SPoILeR's MMS-DATA split, with 27 pretraining scenes and five held-out scenes: Teddy Bear, Toys, Bird House, Bouquet, and Fruits Lincetto et al. (2026).

Data and camera poses. We approximately align the point clouds and camera poses of all 32 scenes in a common coordinate system, following SPoILeR Lincetto et al. (2026), which combines camera calibration with scene-wise normalization of scale, position, and orientation during preprocessing. SIFT/FLANN matches between RGB images with the same view index are lifted to 3D correspondences through COLMAP observation tracks, with unreliable matches and points with large reprojection errors removed. We then estimate a global similarity transform $\mathbf{X}' = s\mathbf{R}\mathbf{X} + \mathbf{t}$ using RANSAC and refit it on the inliers. The same transform is applied consistently to each scene's point cloud and all camera poses, with intrinsics unchanged.

Shared representation. The only change to the main pipeline's representation in this cross-scene experiment is the anchor-feature parameterization described below; the color decoder retains the same dimensions as in the main pipeline. Each anchor has an 8-dimensional geometry code and 12 scene-specific radiance coefficients. A shared spatial basis field combines the anchor coordinates with these coefficients to produce radiance features. The field comprises three sets of triplanes with tensor shapes $3 \times 8 \times r \times r$ for $r \in \{96, 192, 384\}$, and an eight-channel volumetric grid of shape $1 \times 8 \times 64 \times 64 \times 64$. The basis-field decoder is an $83 \rightarrow 256 \rightarrow 256 \rightarrow 192$ MLP with ReLU hidden activations. A separate 12-dimensional global base-coefficient vector is initialized to all ones. The spatial basis field and shared decoders are learned across all the pretraining scenes.

Pretraining and RGB-only reconstruction. We pretrain on RGB, Mono, NIR, and nine-channel MS observations using all views of the 27 pretraining scenes, following SPoILeR Lincetto et al. (2026). The reported results use the checkpoint at 10K iterations per scene from a schedule configured for 30K iterations per scene. Following SPoILeR, pretraining uses LSG, modality-to-luma, and inverse-latent Lincetto et al. (2026) regularization with weights 0.1, 1.0, and 1.0; MS channels are treated separately in the modality-to-luma regularizer. For each held-out scene, we freeze the shared spatial basis field and shared decoders and optimize only scene-specific parameters for 30K iterations using RGB supervision. During this stage, LSG is disabled, and the modality-to-luma and inverse-latent weights are reduced to 0.02 and 0.01.

Evaluation. Same as in SPoILeR, PSNR is evaluated on mosaicked images, whereas SSIM is evaluated only on NIR and Mono modalities, because SSIM is not meaningful on mosaicked multi-channel images. Both metrics use the official foreground masks and are averaged equally across the five held-out scenes.

Baseline results. SPoILeR values are taken from its paper; we did not reproduce its training because of the memory demands of its multi-scene strategy. SSIM values are available only for Mono and NIR in this comparison, so the remaining entries are marked N/A. The approximately 37K pretraining iterations per scene normalize SPoILeR's 1M total iterations by 27 scenes.

B.2 Spectral Feature Space Reuse and Adaptation

PoseSpecGS and the 3DGS references follow the camera initialization, train-time pose refinement, and test-time evaluation protocol of Tab. 6, with implementation details in Sec. 4.1.

Cross-scene reuse. Appendix Tab. 20 evaluates Forest Gang 1 $\rightarrow$ Forest Gang 2 and Bouquet $\rightarrow$ Orchid, with source and target scenes captured using the same sensor configuration. We first train PoseSpecGS on source-scene RGB, Mono, NIR, and MS observations, then transfer its anchor decoder, modality-channel response operators, and response biases to the target scene. These components remain frozen while the target scene's anchor positions, anchor codes, local offsets, and scales are optimized from RGB, Mono, and MS observations. The target scene thus learns its own spectral features in the source-trained spectral feature space through the frozen anchor decoder. Target-scene NIR images provide no reconstruction supervision during cross-scene reuse; NIR renderings use the transferred response operator and bias. Per-modality 3DGS is trained directly on each target modality and serves as a target-supervised reference.

Few-shot channel adaptation. For Appendix Tab. 21, we use Bouquet, Forest Gang 2, and Fruits and train the Latent Spectral Scene Representation using RGB, Mono, and MS observations. We then freeze the representation and existing modality-channel response operators and fit a new NIR response operator and bias using 12.5% of the NIR training views, selected to maximize frame-index separation from the test views. All configurations, including NIR-only 3DGS, use the same NIR evaluation set. Adaptation runs for 1,000 iterations and takes approximately 4 minutes; the all-modal model is a fully supervised reference.

B.3 Thermal and Polarization Scope

Thermal modeling. The thermal branch models appearance separately while sharing anchor geometry and propagating gradients to it, as detailed in Appendix A.4. The model remains an empirical model of observed intensity: attenuation along the viewing direction follows the functional form of Beer–Lambert extinction, while the extinction coefficient is treated as a free and uncalibrated parameter. PoseSpecGS does not recover radiometric temperature, emissivity, or atmospheric parameters, nor does it impose any heat-transfer model.

Polarization. MMS-DATA provides four polarization-analyzer intensity channels, which we include for fair comparison. The scalar response integral in Eq. 9 is not used as a physical model for these measurements, and PoseSpecGS does not explicitly model camera-roll dependence or recover Stokes/Mueller quantities.

B.4 Multispectral Similarity Metrics

Following MS-Splatting Meyer et al. (2026), the spectral metrics in Tab. 5 compare a vector of B ground-truth multispectral channels, $\mathbf{r} \in \mathbb{R}^B$, with the corresponding rendered-channel vector

$\mathbf{t} \in \mathbb{R}^B$ at each evaluated pixel. With a small numerical stabilizer $\epsilon > 0$, spectral angle mapper (SAM) and spectral correlation mapper (SCM) are:

$$\mathrm{SAM}(\mathbf{r}, \mathbf{t}) = \arccos\left(\mathrm{clip}\left(\frac{\mathbf{r}^\top \mathbf{t}}{(\|\mathbf{r}\|_2 + \epsilon)(\|\mathbf{t}\|_2 + \epsilon)}, -1, 1\right)\right), \tag{18}$$

$$\mathrm{SCM}(\mathbf{r}, \mathbf{t}) = \frac{(\mathbf{r} - \bar{r}\mathbf{1})^\top (\mathbf{t} - \bar{t}\mathbf{1})}{(\|\mathbf{r} - \bar{r}\mathbf{1}\|_2 + \epsilon)(\|\mathbf{t} - \bar{t}\mathbf{1}\|_2 + \epsilon)}, \tag{19}$$

where $\bar{r}$ and $\bar{t}$ are the channel-wise means of the ground-truth and rendered vectors, respectively. For spectral information divergence (SID), we form band-wise distributions $p_b = (r_b + \epsilon)/\sum_j (r_j + \epsilon)$ and $q_b = (t_b + \epsilon)/\sum_j (t_j + \epsilon)$ and compute:

$$\mathrm{SID}(\mathbf{r}, \mathbf{t}) = \sum_{b=1}^{B} p_b \log \frac{p_b}{q_b} + \sum_{b=1}^{B} q_b \log \frac{q_b}{p_b}. \tag{20}$$

We use the dataset-level aggregation of the MS-Splatting evaluation protocol for the reported results.

C Additional Quantitative Results

C.1 Reconstruction Quality across Scenes

MMS-DATA. Appendix Tab. 15 reports PoseSpecGS PSNR and SSIM for all 32 MMS-DATA scenes, complementing the representative-scene results in Tab. 3 and the average PSNR in Tab. 4.

Table 15: Per-scene PoseSpecGS test PSNR and SSIM on all 32 MMS-DATA scenes.

Scene	PSNR ↑					SSIM ↑				
	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS
African Art	36.88	35.02	36.56	35.46	34.52	0.982	0.976	0.976	0.984	0.974
Aloe	34.99	31.95	34.52	32.04	33.87	0.969	0.959	0.963	0.963	0.962
Bird House	35.62	32.91	32.11	34.21	34.43	0.981	0.971	0.955	0.983	0.972
Book	34.38	30.18	33.35	30.83	30.95	0.967	0.948	0.952	0.944	0.937
Bouquet	36.68	34.55	36.27	34.13	34.47	0.977	0.974	0.981	0.974	0.966
Chess	38.23	34.29	36.14	33.04	34.75	0.978	0.966	0.967	0.962	0.959
Clock	34.50	32.62	33.86	31.32	32.30	0.973	0.966	0.966	0.963	0.962
Easter Egg	33.14	30.84	32.77	31.07	32.28	0.973	0.968	0.970	0.966	0.965
Fan	35.59	35.95	37.41	33.26	34.58	0.966	0.973	0.976	0.964	0.966
Forest Gang 1	34.89	32.40	34.90	33.45	31.79	0.971	0.965	0.967	0.964	0.961
Forest Gang 2	34.89	33.38	35.20	33.36	32.85	0.970	0.972	0.981	0.970	0.963
Fruits	37.44	34.51	34.30	33.19	34.04	0.980	0.972	0.966	0.972	0.965
Gamepads	38.31	34.52	35.50	33.57	35.00	0.981	0.969	0.967	0.971	0.966
Glass Clock	31.20	30.20	31.65	28.48	29.75	0.959	0.959	0.958	0.948	0.948
Globe	38.95	35.89	37.27	35.68	36.40	0.982	0.974	0.971	0.976	0.972
Laptop	34.93	33.95	37.10	32.01	32.95	0.969	0.972	0.976	0.972	0.961
Laurel Wreath	35.08	31.57	34.29	31.31	31.61	0.973	0.958	0.959	0.957	0.951
Lego Ship	35.32	33.12	34.77	31.69	32.39	0.976	0.966	0.965	0.967	0.956
Makeup	36.03	30.92	34.49	32.12	33.39	0.972	0.939	0.960	0.963	0.958
Orchid	34.81	36.64	37.31	36.26	35.27	0.975	0.983	0.980	0.986	0.977
Pillow	29.84	27.98	29.49	26.70	27.25	0.952	0.947	0.952	0.915	0.927
Plant	35.07	34.33	35.98	34.47	34.25	0.977	0.972	0.974	0.978	0.970
Steel Pot	30.95	30.35	29.63	27.95	29.11	0.956	0.953	0.921	0.944	0.941
Teddy Bear	35.95	32.89	34.87	34.97	34.06	0.977	0.972	0.971	0.976	0.966
Tin Box 1	25.75	25.70	26.53	24.03	24.90	0.909	0.909	0.912	0.907	0.865
Tin Box 2	28.54	29.19	31.28	27.88	26.44	0.961	0.961	0.962	0.961	0.942
Toys	33.48	30.82	32.96	30.72	31.00	0.974	0.957	0.959	0.967	0.956
Trophies	31.58	31.36	32.90	28.36	29.97	0.962	0.960	0.962	0.945	0.949
Truck	35.48	34.54	34.70	34.23	33.30	0.980	0.970	0.966	0.980	0.963
Vases	36.96	34.32	33.93	34.24	35.31	0.974	0.970	0.939	0.967	0.965
Watering Can 1	36.13	33.70	34.79	34.37	34.03	0.981	0.972	0.970	0.982	0.967
Watering Can 2	33.63	32.80	31.34	33.46	33.25	0.975	0.972	0.970	0.979	0.963
Average	34.54	32.61	34.01	32.12	32.51	0.970	0.964	0.963	0.964	0.957

MS-Splatting dataset. Appendix Tab. 16 complements the aggregate comparison in Tab. 5 with complete per-scene PoseSpecGS results on the MS-Splatting dataset.

Table 16: Per-scene PoseSpecGS results on all seven MS-Splatting scenes.

Scene	PSNR ↑						SSIM ↑						SAM ↓	SCM ↑	SID ↓
	All	RGB	G	R	RE	NIR	All	RGB	G	R	RE	NIR			
Bud Swelling	25.75	20.34	25.08	28.63	25.35	29.31	0.788	0.659	0.746	0.860	0.799	0.877	0.108	0.870	0.019
Fruit Trees	23.92	18.42	24.36	26.98	23.14	26.70	0.717	0.549	0.690	0.796	0.727	0.826	0.129	0.849	0.030
Garden	28.68	23.94	28.68	27.54	32.20	31.02	0.813	0.703	0.797	0.809	0.876	0.882	0.066	0.963	0.008
Golf	30.65	24.02	31.09	32.35	31.11	34.70	0.867	0.682	0.891	0.931	0.885	0.944	0.056	0.955	0.005
Lake	24.09	22.74	24.48	27.62	22.36	23.26	0.756	0.678	0.752	0.843	0.718	0.790	0.114	0.869	0.023
Single Tree	25.83	20.26	25.59	27.03	26.49	29.79	0.760	0.525	0.728	0.826	0.813	0.907	0.100	0.875	0.019
Solar	26.75	23.24	26.46	29.97	26.27	27.83	0.884	0.816	0.878	0.921	0.891	0.914	0.081	0.936	0.012
Average	26.52	21.85	26.53	28.59	26.70	28.94	0.798	0.659	0.783	0.855	0.816	0.877	0.094	0.902	0.017

RHD dataset. Appendix Tab. 17 expands the average thermal results in Tab. 7 with per-scene thermal comparisons and RGB reconstruction results from the same jointly trained model. Baseline results are taken from ETGS Wang et al. (2026).

Table 17: Per-scene RHD results; scene names are abbreviated. Thermal best results: **first**, **second**.

Metric	Method	CC	WB	CD	CB	CE	HT	WP	HW	WC	WW	Avg.
PSNR ↑	3DGS	36.27	32.37	29.78	29.74	34.66	39.47	30.48	34.47	24.43	29.94	32.16
	Mip-Splatting	36.48	32.44	28.78	29.13	35.16	37.45	32.74	30.06	26.18	26.67	31.51
	Thermal3D-GS	39.47	33.61	32.89	32.11	38.77	39.83	34.12	34.93	30.96	30.06	34.68
	4DGS	33.58	30.52	34.50	35.61	33.99	39.75	34.64	30.36	34.15	32.34	33.94
	NTR-Gaussian	41.82	34.63	35.18	35.56	38.09	38.99	32.44	33.65	28.18	31.01	34.96
	ETGS	**44.73**	**40.23**	**37.63**	**38.25**	**41.67**	**42.27**	**41.78**	**36.70**	**39.41**	**44.16**	**40.68**
	Ours Thermal	**46.05**	**45.16**	**39.44**	**41.75**	**43.09**	**43.87**	**45.31**	**36.69**	**42.65**	**46.08**	**43.01**
	Ours RGB	31.91	32.98	26.35	28.96	33.14	30.35	32.23	31.87	31.60	31.42	31.08
SSIM ↑	3DGS	0.978	0.988	0.968	0.960	0.980	**0.990**	0.987	0.988	0.968	0.974	0.978
	Mip-Splatting	0.974	0.988	0.961	0.958	0.980	0.986	0.988	0.979	0.972	0.969	0.976
	Thermal3D-GS	0.983	0.989	0.981	0.970	0.983	**0.990**	0.990	0.989	0.978	0.976	0.983
	4DGS	0.963	0.979	0.976	0.967	0.962	0.986	0.984	0.960	0.975	0.970	0.972
	NTR-Gaussian	0.986	0.987	0.980	0.974	0.974	0.987	0.984	0.982	0.974	0.978	0.981
	ETGS	**0.987**	**0.992**	**0.989**	**0.983**	**0.985**	**0.991**	**0.994**	**0.990**	**0.989**	**0.994**	**0.989**
	Ours Thermal	**0.992**	**0.994**	**0.990**	**0.988**	**0.987**	**0.991**	**0.995**	**0.991**	**0.992**	**0.995**	**0.992**
	Ours RGB	0.954	0.977	0.945	0.948	0.968	0.953	0.957	0.961	0.960	0.959	0.958
LPIPS ↓	3DGS	0.066	0.087	0.072	0.075	0.031	0.057	0.091	0.031	0.131	0.137	0.078
	Mip-Splatting	0.068	0.080	0.093	0.078	0.031	0.075	0.073	0.043	0.126	0.178	0.085
	Thermal3D-GS	0.060	0.095	0.050	0.063	0.029	0.062	0.092	0.030	0.110	0.128	0.072
	4DGS	0.101	0.075	0.047	0.076	0.065	0.056	0.064	0.073	0.105	0.099	0.076
	NTR-Gaussian	0.072	0.092	0.052	0.092	0.098	0.092	0.104	0.055	0.133	0.102	0.089
	ETGS	**0.054**	**0.072**	**0.030**	**0.050**	**0.028**	**0.054**	**0.060**	**0.028**	**0.085**	**0.042**	**0.050**
	Ours Thermal	**0.034**	**0.048**	**0.028**	**0.041**	**0.024**	**0.046**	**0.043**	**0.021**	**0.064**	**0.035**	**0.038**
	Ours RGB	0.064	0.060	0.134	0.128	0.075	0.074	0.073	0.066	0.074	0.065	0.081

C.2 Reconstruction Quality with Pose Refinement

Appendix Tab. 18 provides per-scene results for the comparison in Tab. 6 in Sec. 4.2. All methods use the same localization initialization, train-time pose-refinement scheme, and test-time evaluation protocol described in Sec. 4.2 and Sec. 4.1.

Table 18: PSNR and SSIM on representative MMS-DATA scenes with train-time and test-time pose refinement for all methods.

Metric	Method	Bouquet					Forest Gang 2					Fruits					Average				
		RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS	RGB	Mono	NIR	Polar	MS
PSNR ↑	3DGS, per modality	31.60	30.60	32.48	29.56	29.26	30.43	26.84	28.08	28.32	28.18	33.58	30.06	28.69	29.23	27.89	31.87	29.17	29.75	29.04	28.44
	MS-Splatting	35.75	33.19	35.79	33.07	33.38	34.25	29.49	32.05	32.88	31.83	36.33	32.69	**35.08**	32.33	33.49	35.44	31.79	34.31	32.76	32.90
	Ours	**36.68**	**34.55**	**36.27**	**34.13**	**34.47**	**34.89**	**33.38**	**35.20**	**33.36**	**32.85**	**37.44**	**34.51**	34.30	**33.19**	**34.04**	**36.34**	**34.15**	**35.26**	**33.56**	**33.79**
SSIM ↑	3DGS, per modality	0.952	0.950	0.965	0.949	0.919	0.935	0.943	0.932	0.940	0.910	0.960	0.952	0.948	0.948	0.920	0.949	0.948	0.949	0.946	0.916
	MS-Splatting	0.973	0.968	0.979	0.969	0.959	0.968	0.963	0.973	0.968	0.957	0.979	0.970	**0.977**	0.969	0.960	0.973	0.967	**0.976**	0.969	0.959
	Ours	**0.977**	**0.974**	**0.981**	**0.974**	**0.966**	**0.970**	**0.972**	**0.981**	**0.970**	**0.963**	**0.980**	**0.972**	0.966	**0.972**	**0.965**	**0.976**	**0.973**	**0.976**	**0.972**	**0.965**

C.3 RGBT-Scenes Reconstruction with Matched Pose Refinement

Comparing a method that refines test-view poses with baselines that do not can confound reconstruction quality with the benefit of test-time camera alignment. To address this evaluation asymmetry, we extend the controlled comparison in Tab. 6 to all ten RGBT-Scenes scenes. We equip Thermal-Gaussian Lu et al. (2024a), MMOne Gu & Wang (2025), and PhysIR-Splat Gao et al. (2026) with our full camera pipeline: RGB-backbone localization, the same train-time pose-refinement scheme

including RGB-guided reprojection consistency, and the same pose-refining NVS evaluation protocol described in Sec. 4.1. At test time, the learned scene is frozen and only auxiliary test-camera poses are refined photometrically before computing the metrics; RGB poses remain fixed. The baselines are trained for 30K iterations and evaluated on the official test split. PoseSpecGS results are reused unchanged from Tab. 1 and Tab. 2.

Appendix Tab. 19 reports equal-weight averages over the ten scenes for both RGB and thermal. PoseSpecGS retains the best average PSNR, SSIM, and LPIPS in both modalities. Its average PSNR exceeds the strongest baseline by 0.52 dB for RGB and 1.77 dB for thermal. Together with Tab. 6, these results show that the reconstruction advantage persists beyond MMS-DATA when camera initialization, train-time pose refinement, and test-time alignment follow the same protocol. The gains therefore cannot be attributed solely to giving PoseSpecGS access to test-time pose refinement, supporting the effectiveness of its multimodal scene representation and rendering design under matched camera processing.

Table 19: Average RGB and thermal NVS results on RGBT-Scenes with matched pose refinement. Best results: **first**, **second**.

Method	PSNR ↑		SSIM ↑		LPIPS ↓	
	RGB	Thermal	RGB	Thermal	RGB	Thermal
ThermalGaussian	24.38	25.60	0.849	0.886	**0.169**	0.105
MMOne	**24.78**	25.83	**0.855**	0.891	0.179	0.108
PhysIR-Splat	24.44	**26.48**	0.840	**0.892**	0.195	**0.093**
Ours	**25.30**	**28.25**	**0.862**	**0.923**	**0.160**	**0.074**

C.4 Cross-Scene Reuse of the Spectral Feature Space

Beyond reconstruction quality, we evaluate cross-scene reuse of the same full PoseSpecGS model. Using the same sensor configuration, we fit a new scene to RGB, Mono, and MS observations while keeping the source-trained anchor decoder, modality-channel response operators, and response biases fixed. Appendix Tab. 20 shows that PoseSpecGS outperforms the NIR-supervised 3DGS reference on both scene pairs without target-scene NIR reconstruction supervision. Successful reuse of the frozen mappings suggests that the spectral feature space captures cross-modal structure whose utility extends beyond fitting the source scene. Appendix B.2 details the transfer protocol, and Appendix C.5 evaluates efficient adaptation to held-out channels.

Table 20: Cross-scene transfer on MMS-DATA. Gray marks NIR, excluded from PoseSpecGS target-scene reconstruction supervision.

Metric	Configuration	Forest Gang 1 → Forest Gang 2				Bouquet → Orchid			
		RGB	Mono	NIR	MS	RGB	Mono	NIR	MS
PSNR ↑	3DGS, per modality	30.43	26.84	28.08	28.18	33.64	31.64	33.01	31.40
	Ours, transfer	**33.30**	**30.85**	**31.11**	**31.31**	**36.92**	**35.25**	**34.97**	**34.10**
SSIM ↑	3DGS, per modality	0.935	0.943	0.932	0.910	0.963	0.963	0.973	0.944
	Ours, transfer	**0.968**	**0.966**	**0.958**	**0.960**	**0.984**	**0.981**	**0.980**	**0.976**

C.5 Efficient Channel Adaptation

We learn a Latent Spectral Scene Representation from RGB, Mono, and MS, then freeze it and the existing modality-channel response operators. Following the protocol in Appendix B.2, we use 12.5% of NIR training views to fit a new NIR response operator and bias in 1,000 iterations (approximately 4 minutes). Appendix Tab. 21 shows that this lightweight adaptation outperforms NIR-only 3DGS on average, suggesting that the spectral feature space captures cross-modal structure that extends beyond the training modality combination.

Table 21: Few-shot NIR adaptation on representative MMS-DATA scenes. The all-modal model is a fully supervised reference; bold marks better results against NIR-only 3DGS.

Configuration	PSNR ↑				SSIM ↑			
	Bouquet	Forest Gang 2	Fruits	Average	Bouquet	Forest Gang 2	Fruits	Average
3DGS, NIR only	32.48	28.08	28.69	29.75	0.965	0.932	0.948	0.949
Ours, 12.5% NIR training views	**32.58**	**32.06**	**32.57**	**32.40**	**0.970**	**0.972**	**0.961**	**0.968**
Ours, all modalities	36.27	35.20	34.30	35.26	0.981	0.981	0.966	0.976

C.6 Storage Efficiency

Appendix Tab. 22 reports model storage in MB for the representative MMS-DATA scenes in Tab. 3. PoseSpecGS uses less model storage than both per-modality 3DGS and MS-Splatting on all evaluated scenes, demonstrating the compactness of the shared representation.

Table 22: Model storage (MB) on representative MMS-DATA scenes.

Method	Bouquet	Forest Gang 2	Fruits
3DGS, per modality	269.02	265.91	195.71
MS-Splatting	426.36	975.42	590.06
Ours	**90.68**	**71.36**	**53.05**

C.7 Alignment Robustness to Incomplete Observations

MMS-DATA poses. MMS-DATA derives its camera poses from per-frame poses of a reference RGB camera and fixed extrinsics of the multi-camera rig Lincetto et al. (2025). Tab. 3 and Tab. 6 report reconstruction results on the same representative scenes under different pose protocols for MS-Splatting; the former evaluates with the supplied MMS-DATA poses, while the latter evaluates with our RGB-backbone localization and train- and test-time pose refinement. The substantial gap between MS-Splatting's results across these two tables indicates that the pose protocol has a strong effect on GS reconstruction quality. Appendix Fig. 7 depicts this change within MS-Splatting on two RGB and two NIR test views: the supplied-pose renderings show blurred plant boundaries and floating artifacts around the plants and tabletop, whereas the pose-refined renderings recover sharper structures. Because the reconstruction method and scene are held fixed in this visual comparison, the observed discrepancy cannot be attributed to a difference between GS representations and is therefore most likely due to inconsistent RGB camera poses and the rigs. The supplied calibration is adequate for the NeRF-based MMS model, but these results suggest residual pose errors that are consequential for GS; they do not establish absolute pose accuracy. We therefore do not treat the supplied RGB poses or rigs as ground truth for the pose evaluation below.

Rig consistency under incomplete observations. Because the supplied MMS-DATA poses are insufficiently reliable as camera-pose ground truth, we evaluate pose initialization through recovered rig consistency. For a fixed sensor rig, the relative transform between RGB and each auxiliary camera should remain constant across capture positions.

Both methods register the modalities as independent image collections, without a rig model or fixed inter-camera constraints. We retain a fixed half-subset of observations from each modality to reduce cues from densely sampled, capture-indexed observations. Both methods use the same fixed RGB SfM backbone, including its camera poses and sparse point cloud, and estimate auxiliary-camera poses in the resulting world coordinate frame and reconstruction scale. The SfM baseline reconstructs each auxiliary modality independently and aligns its sparse point cloud to the RGB backbone using a $\mathrm{Sim}(3)$ transform initialized from camera centers with matching capture indices. Our initialization instead uses DINOv3-based RGB reference retrieval, cross-modal matching, PnP, and auxiliary-only bundle adjustment. All poses are evaluated before Gaussian training or subsequent pose refinement.

For same-rig RGB–auxiliary pairs registered by both methods, let T_i^{rgb} and T_i^m denote the world-to-camera transforms of the RGB camera and auxiliary modality m, respectively, at capture index i. Their relative transform is:

$$T_i^{m\leftarrow\mathrm{rgb}} = T_i^m (T_i^{\mathrm{rgb}})^{-1}.$$

For each scene–modality group, we obtain the mean relative rotation by projecting the arithmetic mean of the rotation matrices onto $\mathrm{SO}(3)$, and compute the mean relative translation by arithmetic averaging. Rotation variance is the mean squared geodesic distance from the mean rotation, measured in deg^2. Translation variance is the mean squared Euclidean deviation from the mean translation, normalized by the mean trajectory radius of the shared RGB backbone and reported in $\%^2$.

Appendix Tab. 23 reports these statistics on identical frame subsets, with 3–7 common pairs per group. Overall, our initialization recovers a markedly more stable inter-camera rig than the SfM

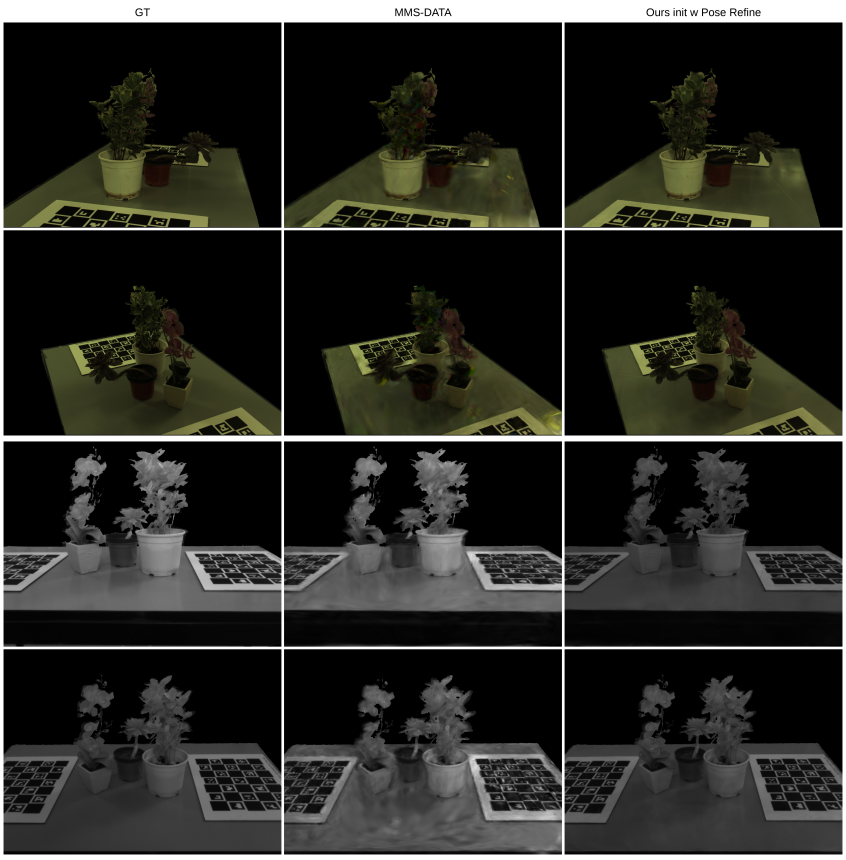

Figure 7: MS-Splatting test-view comparison on MMS-DATA. The top two rows show RGB views and the bottom two show NIR views. From left to right: ground truth, MS-Splatting with the dataset-provided poses, and the same method with RGB-backbone initialization and train- and test-time pose refinement. The supplied-pose renderings exhibit blur and floaters.

baseline. Without any rig prior, it yields lower rotation variance in all 11 comparable groups and lower translation variance in 9 of 11, while producing valid initializations for all 12 groups.

Agreement with the final reconstruction. We further compare both initializations with the rig recovered by full-observation PoseSpecGS after train-time and test-time pose refinement. Because the refined reconstruction retains the same fixed RGB poses, all three pose sets share the same RGB-anchored coordinate frame and reconstruction scale; no additional alignment or scale fitting is applied for evaluation. The refined poses serve only as a reconstruction-derived reference, not as independent ground truth.

Using the same-rig RGB–auxiliary pairs shared by SfM, Ours, and Refined gives 3–7 pairs in each of the 11 comparable scene–modality groups. Appendix Tab. 23 reports both rig variance and the discrepancy between each initialized mean rig transform and the refined mean transform. Averaged over these groups, the rotation variances for SfM, Ours, and Refined are 2.046, 0.024, and $0.009\,\mathrm{deg}^2$, respectively, while the corresponding translation variances are 15.478, 1.205, and $0.653\,\%^2$. The mean rotation discrepancy from Refined is $2.785°$ for SfM and $0.566°$ for Ours; the corresponding translation discrepancies are 4.689% and 0.797% of the RGB trajectory radius.

Bouquet MS is reported separately using the six pairs shared by Ours and Refined and is excluded from these averages because SfM fails to produce an initialization. The low rig variance and small discrepancy from the refined reconstruction show that, despite using only half of the observations and no rig constraints, our initialization recovers inter-camera geometry close to that obtained after full train- and test-time refinement. This comparison does not establish absolute pose accuracy or equivalent rendering quality, since the refined reconstruction is itself not independent ground truth.

Table 23: Initialized versus refined rig, including the SfM baseline. Rotation and translation variances are in deg^2 and $\%^2$. ΔR (°) and Δt (%) compare each initialized mean rig transform with the refined mean. All three pose sets share the same RGB coordinate frame and scale. Each comparable row uses the same n pairs for all three pose sets; Refined is an internal comparator.

Scene	Modality	n	Rot. variance ↓			Trans. variance ↓			ΔR ↓		Δt ↓	
			SfM	Ours	Refined	SfM	Ours	Refined	SfM	Ours	SfM	Ours
Bouquet	Mono	7	0.930	0.021	0.004	1.463	0.705	0.360	1.945	1.082	2.602	1.218
	NIR	4	0.626	0.006	0.005	0.964	0.864	0.215	2.770	0.288	5.440	0.198
	Polar	4	0.192	0.018	0.006	0.994	0.376	0.397	2.413	0.123	4.360	0.308
	MS	6*	–	0.012*	0.009*	–	0.708*	0.361*	–	0.929*	–	0.317*
Forest Gang 2	Mono	7	0.287	0.010	0.008	1.324	1.477	1.193	2.360	1.283	3.485	2.232
	NIR	4	6.533	0.056	0.004	13.582	2.167	1.146	2.184	0.370	4.748	0.341
	Polar	6	0.201	0.032	0.005	1.038	1.784	0.908	2.215	0.115	3.786	0.617
	MS	6	0.755	0.007	0.009	1.614	0.880	0.660	4.228	0.892	5.837	0.407
Orchid	Mono	7	0.146	0.017	0.008	1.525	0.773	0.359	1.787	0.993	1.972	1.205
	NIR	5	1.604	0.050	0.025	5.717	2.801	1.567	2.997	0.264	6.281	1.003
	Polar	3	0.247	0.009	0.006	1.506	0.744	0.073	2.673	0.116	5.265	0.416
	MS	5	10.986	0.041	0.018	140.528	0.687	0.303	5.059	0.698	7.804	0.819
Common-group average (11 groups)			2.046	0.024	0.009	15.478	1.205	0.653	2.785	0.566	4.689	0.797
Success (all 12 scene–modality groups): SfM 11/12; Ours **12/12**.												

*SfM failed on Bouquet MS. This supplementary row uses six pairs shared by Ours and Refined and is excluded from all comparative averages. Refined denotes the final reconstruction, not independent ground truth.

D Additional Qualitative Results

Error maps throughout this section directly show pixelwise differences between rendered and ground-truth images without per-image contrast stretching; darker regions indicate smaller errors.

D.1 MMS-DATA Qualitative Results

Appendix Fig. 8–11 complement the MMS-DATA evaluation in Appendix Tab. 15 with qualitative examples across the five evaluated modalities.

D.2 MS-Splatting Dataset Qualitative Results

Appendix Fig. 12 complements the MS-Splatting dataset evaluation in Appendix Tab. 16 with qualitative examples of joint RGB and multispectral reconstruction.

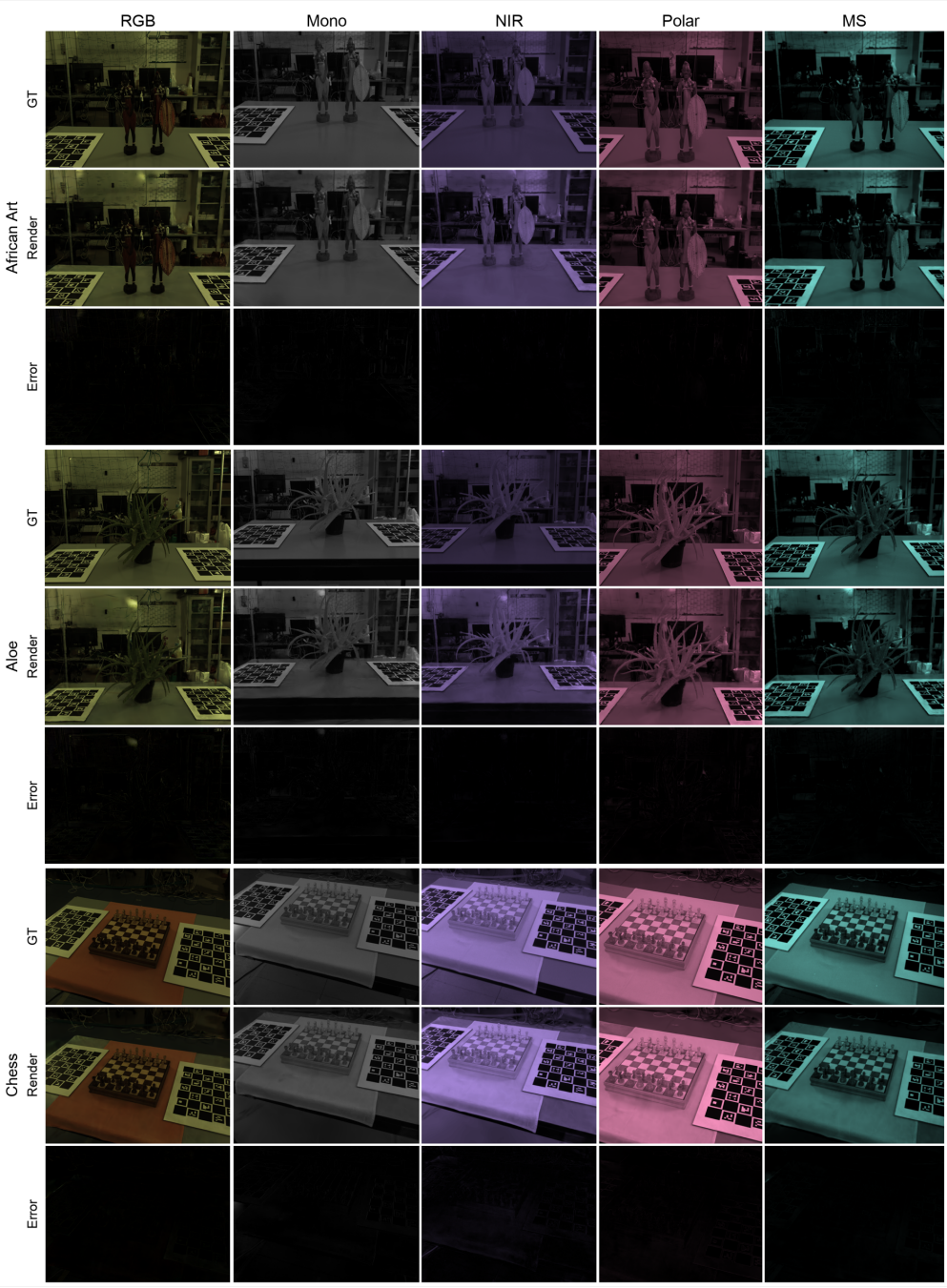

Figure 8: Qualitative MMS-DATA results (part 1): ground truth, PoseSpecGS renderings, and error maps for RGB, Mono, NIR, Polar, and MS.

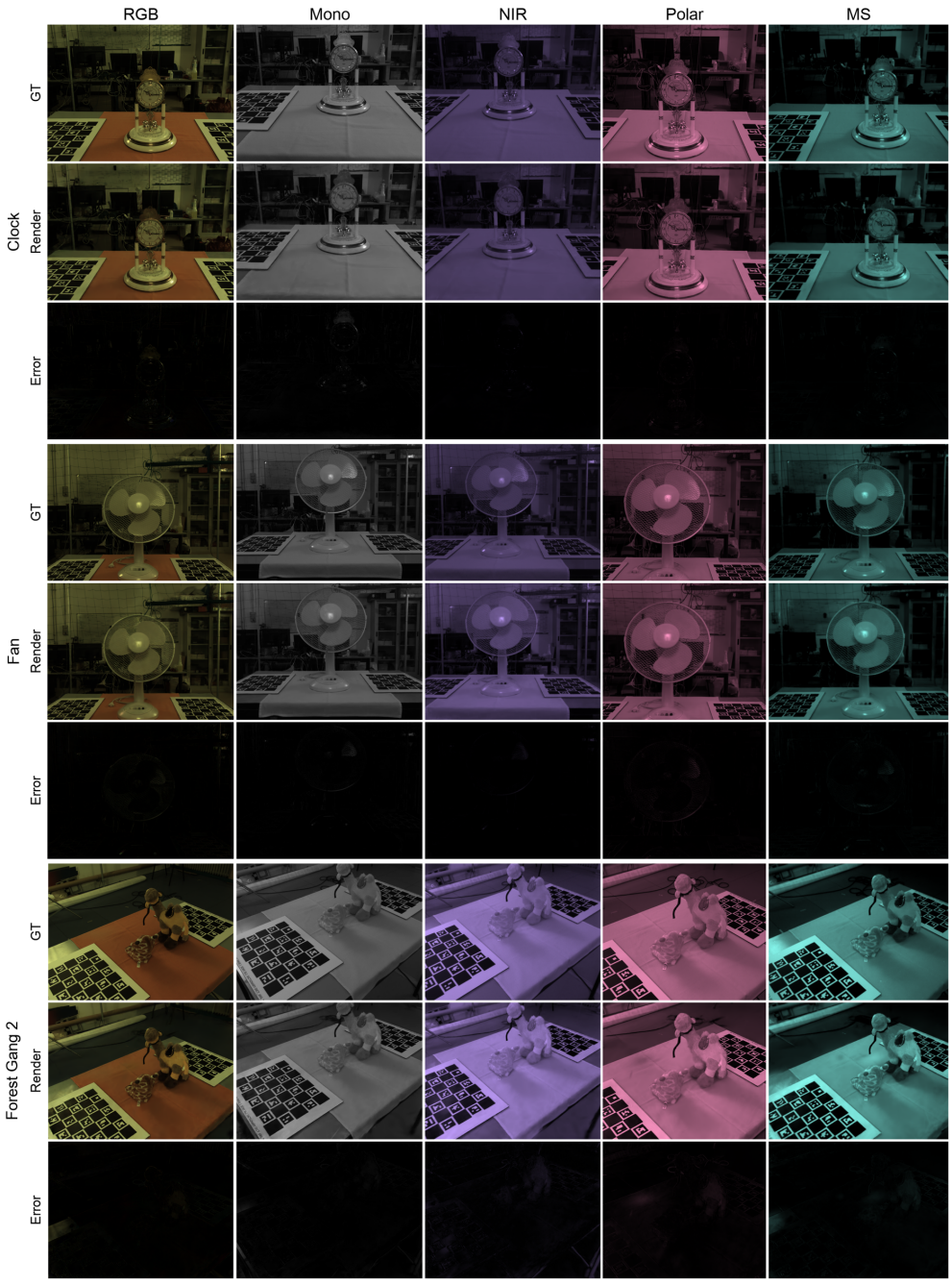

Figure 9: Qualitative MMS-DATA results (part 2): ground truth, PoseSpecGS renderings, and error maps for RGB, Mono, NIR, Polar, and MS.

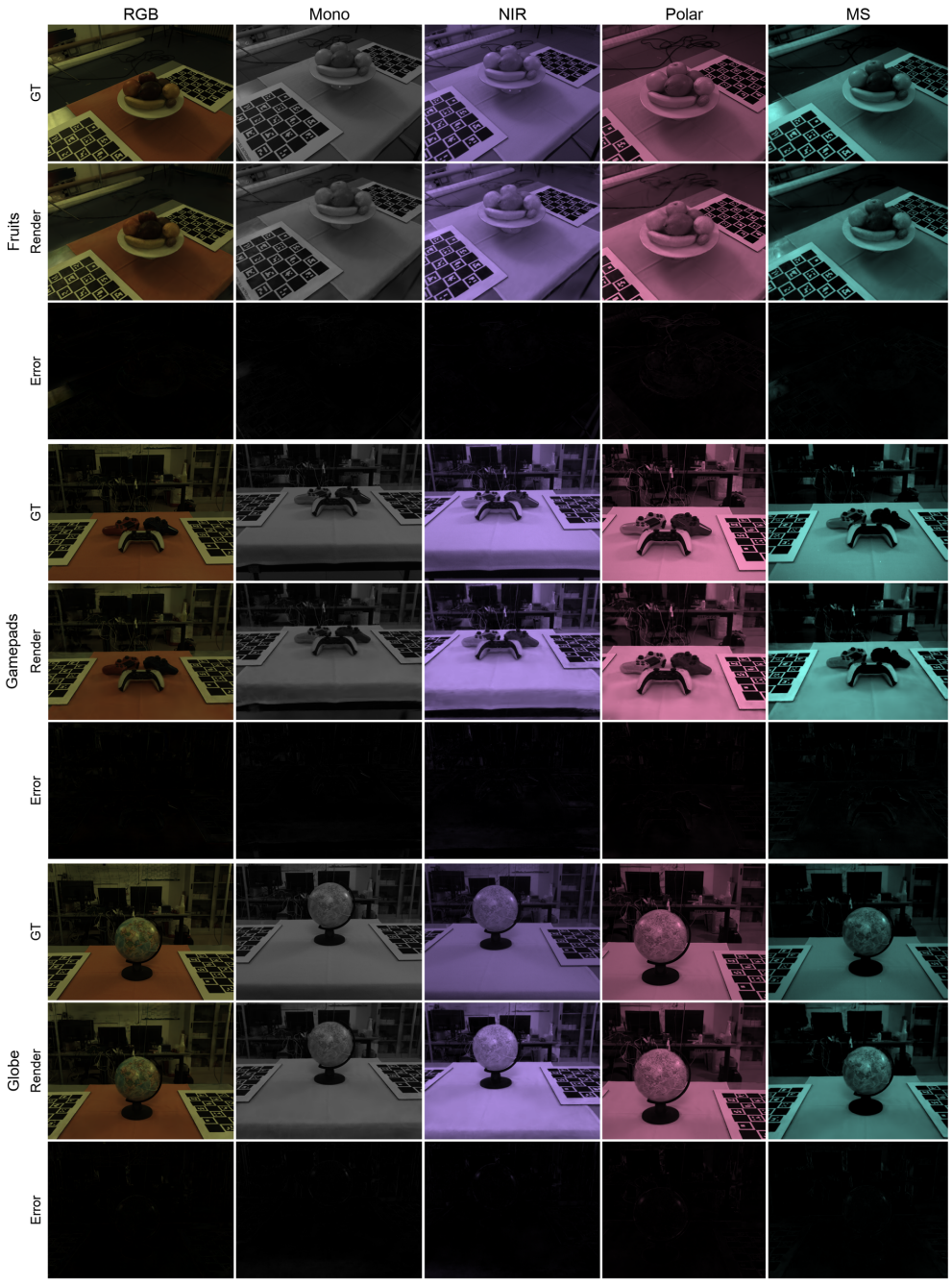

Figure 10: Qualitative MMS-DATA results (part 3): ground truth, PoseSpecGS renderings, and error maps for RGB, Mono, NIR, Polar, and MS.

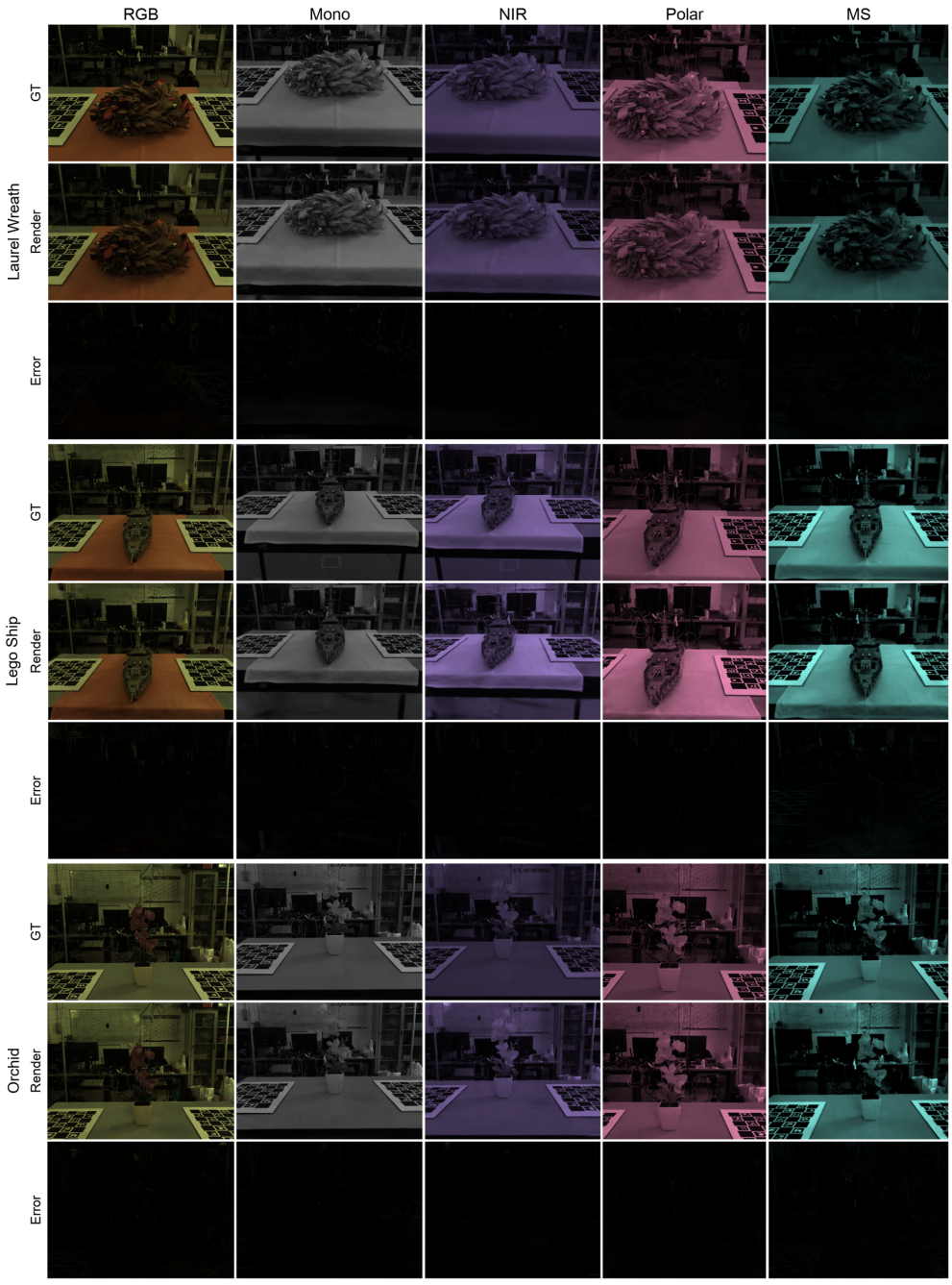

Figure 11: Qualitative MMS-DATA results (part 4): ground truth, PoseSpecGS renderings, and error maps for RGB, Mono, NIR, Polar, and MS.

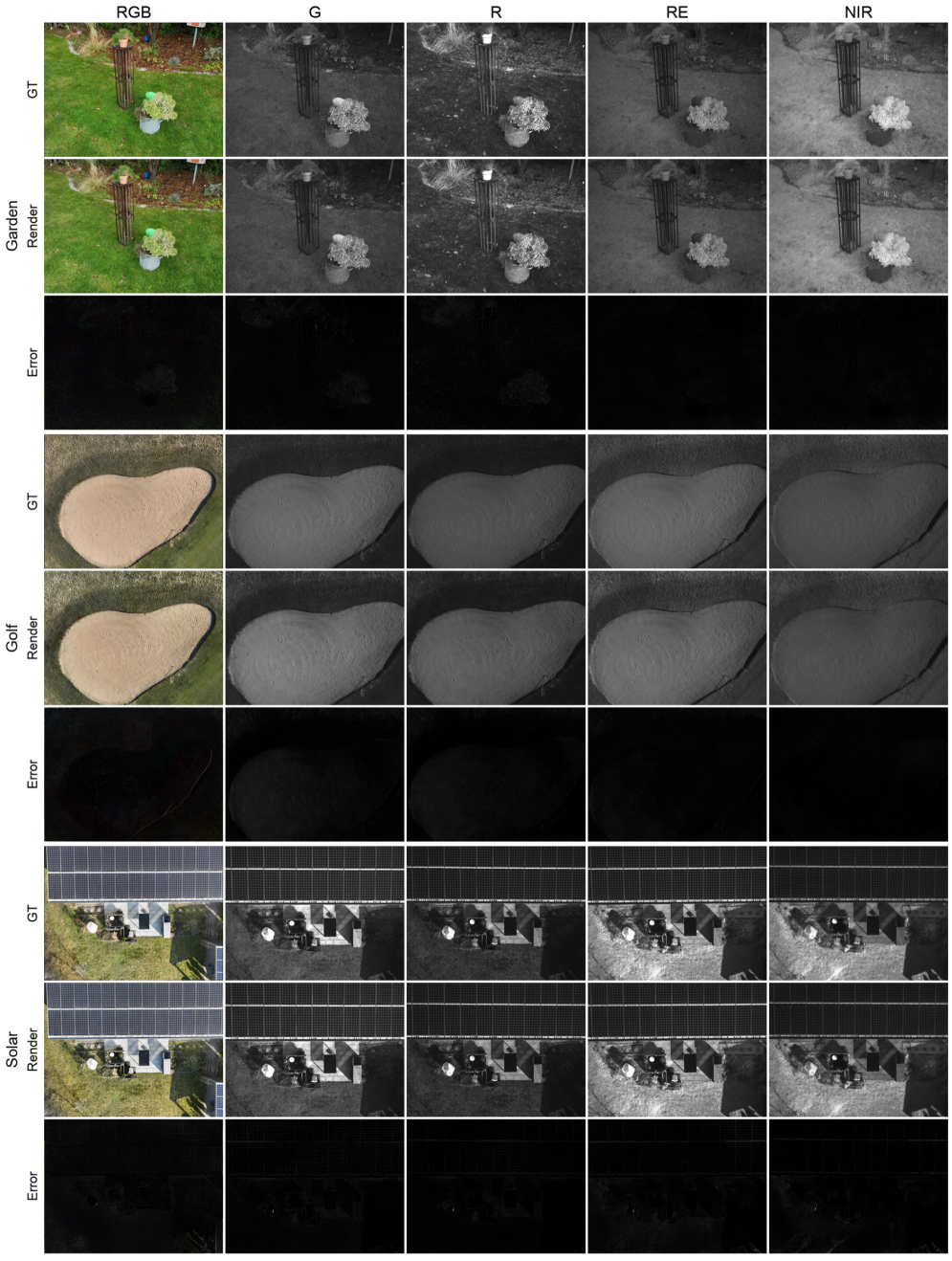

Figure 12: Qualitative MS-Splatting dataset results: ground truth, PoseSpecGS renderings, and error maps for RGB, G, R, RE, and NIR.

D.3 RHD Dataset Qualitative Results

Appendix Fig. 13 and Appendix Fig. 14 complement the RHD evaluation in Appendix Tab. 17 with qualitative examples of joint RGB–thermal reconstruction. The grayscale thermal ground-truth images and renderings are shown in pseudocolor for better visualization. The six scenes are Warming Bottles, Cooling Bench, Cooling E-bike, Heat Transfer, Warming Peaches, and Warming Cups.

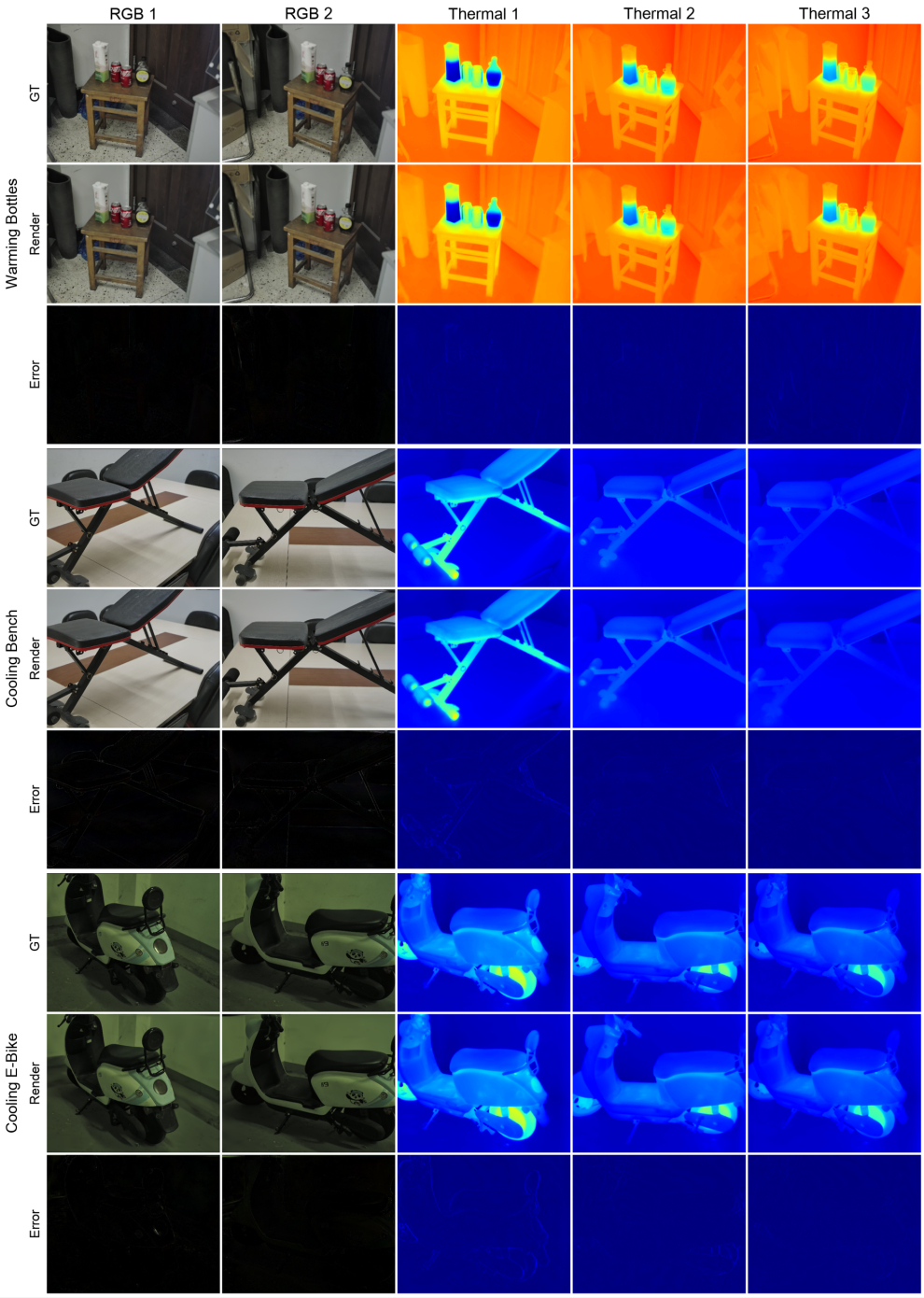

Figure 13: Qualitative RHD results (part 1): ground truth, PoseSpecGS renderings, and error maps for RGB and thermal (pseudocolor visualization).

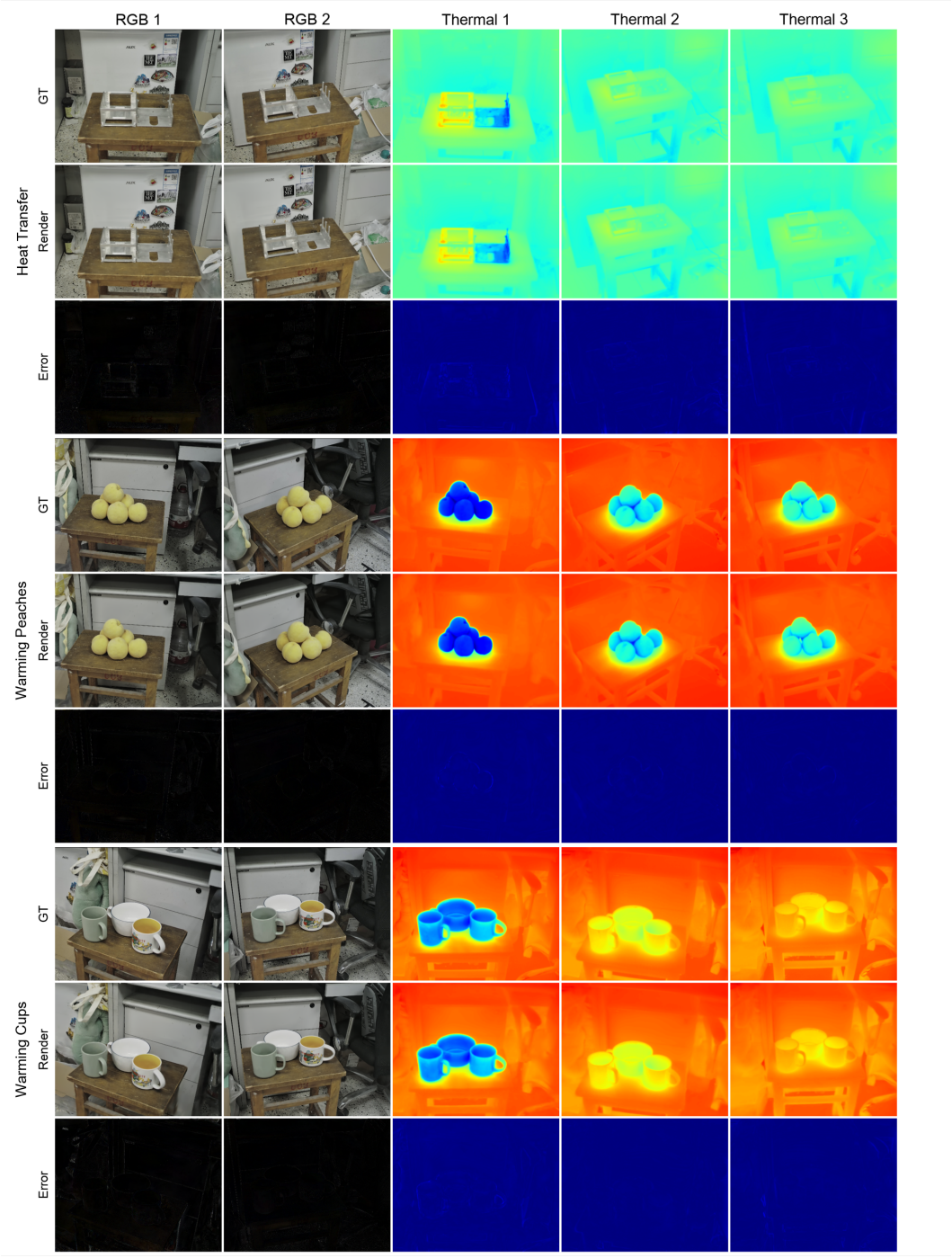

Figure 14: Qualitative RHD results (part 2): ground truth, PoseSpecGS renderings, and error maps for RGB and thermal (pseudocolor visualization).

D.4 Cross-Modal Alignment Visualizations

Trajectory overview. We visualize the cross-modal alignment results across all scenes in MMS-DATA, as demonstrated in Appendix Fig. 15. Our method never observes the rig extrinsics and registers each auxiliary-modality view on its own, yet the rig relationship appears roughly recovered in the visualization.

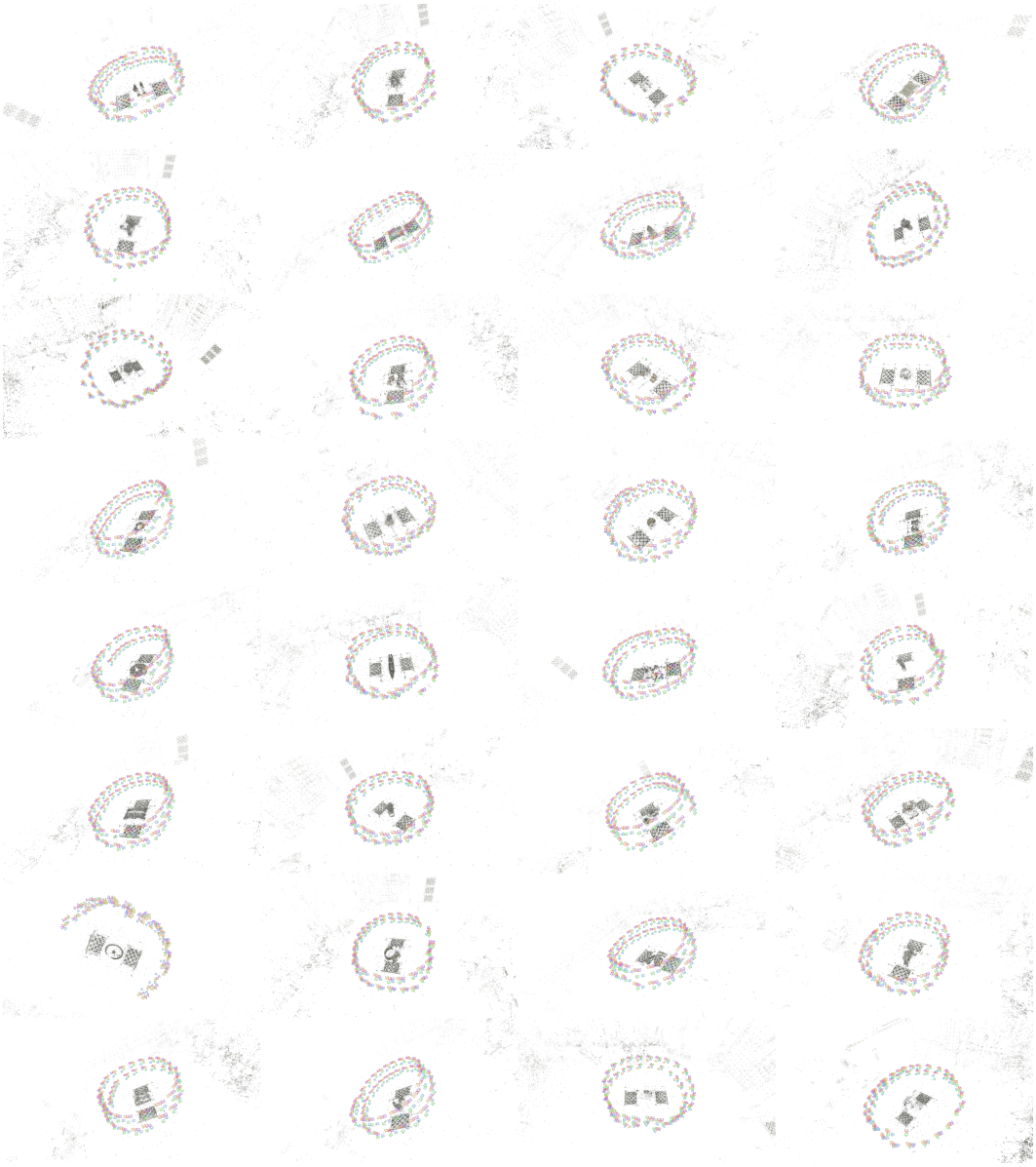

Figure 15: Registered camera poses and RGB SfM point clouds for all 32 MMS-DATA scenes after Cross-Modal Alignment. Camera modality: RGB, Mono, NIR, Polar, MS.