SYNCGS: SELF-ALIGNING DYNAMIC RGB AND THERMAL GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic reconstruction from RGB and Thermal streams benefits from their complementary appearance and thermal cues. In practice, however, independently recorded streams often exhibit residual temporal offsets and viewpoint discrepancies, which complicate reliable cross-modal reconstruction. We present SyncGS, a self-aligning framework for dynamic Gaussian reconstruction from RGB and Thermal streams. SyncGS uses cross-modal motion cues to build a motion cost volume with local spatial tolerance and continuously queries this volume to estimate a sequence-level temporal offset. During optimization, Thermal image supervision uses the current temporal estimate to learn an effective spatial correction shared across the sequence under a shared virtual-camera model. To limit mutual compensation between calibration and scene variables, we introduce Observable-Owned Optimization, which assigns motion, image-alignment, and reconstruction objectives to the temporal, spatial, and scene variables, respectively, while retaining a shared forward state. Experiments on DynamicRGBT-Scenes show improved reconstruction quality in both RGB and Thermal modalities, recovery of imposed temporal offsets, and resilience to controlled pose-initialization perturbations.

1 INTRODUCTION

Thermal imaging captures heat-related patterns that are not directly available in visible-light images, while RGB observations provide detailed appearance and structural cues. Their complementarity supports multimodal scene reconstruction for applications such as infrastructure inspection, industrial monitoring, and dynamic scene analysis (Hassan et al., 2024). Gaussian splatting provides an explicit scene representation with efficient differentiable rendering (Kerbl et al., 2023), and its dynamic extensions model time-varying geometry and appearance from video (Wu et al., 2024; Bae et al., 2024). In RGB and Thermal reconstruction, ThermalGaussian learns a multimodal Gaussian representation from calibrated observations (Lu et al., 2025), ThermalGaussian++ further considers spatially non-aligned inputs and thermal-resolution enhancement (Lu et al., 2026b), and Dynamic Thermal Gaussians extends multimodal Gaussian reconstruction to moving scenes with changing thermal patterns (Lu et al., 2026a).

In practical acquisition, RGB and Thermal streams are often recorded independently, even when the cameras are rigidly mounted on a common moving platform. Coarse manual synchronization can therefore leave a residual temporal offset, while the relative camera transformation remains fixed within each sequence. The resulting offset may vary across capture sessions, so frames with the same nominal index need not correspond to the same dynamic state. DynamicRGBT-Scenes provides a representative instance of this setting, using two viewpoints with a paired RGB and Thermal camera at each viewpoint.

Temporal and spatial misalignment are tightly coupled in dynamic reconstruction. A temporal offset changes the dynamic state compared across modalities, whereas a viewpoint discrepancy changes the projected spatial correspondence. The flexibility of dynamic Gaussian representations allows these errors to be partially absorbed by deformation, geometry, or modality-specific appearance. Consequently, reducing reconstruction error alone does not guarantee correct temporal or spatial correspondence. Existing methods for unsynchronized reconstruction estimate temporal correspondence through learnable offsets, frame matching, motion cues, or cross-view constraints (Kim et al.,

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.

2024; Xu et al., 2026; Choi et al., 2025; Liu et al., 2025a), while multimodal Gaussian methods primarily focus on cross-modal representation and rendering. A method that estimates both sequence-level temporal correspondence and effective spatial alignment within dynamic RGB and Thermal reconstruction is therefore still needed.

The central challenge is therefore not only to estimate temporal and spatial alignment, but also to determine which observations should update which variables. Because the dynamic scene representation is sufficiently flexible to absorb calibration errors, unrestricted joint optimization can reduce reconstruction loss without recovering the correct alignment. This motivates an optimization strategy that preserves the interaction among temporal, spatial, and scene variables during forward rendering while separating their update paths during backward optimization.

We present SyncGS, a self-aligning framework for dynamic Gaussian reconstruction from asynchronous and spatially misaligned RGB and Thermal streams. SyncGS extracts cross-modal motion cues and constructs a motion cost volume with local spatial tolerance. Continuous queries into this fixed volume estimate a sequence-level temporal offset. During each forward pass, the current temporal estimate determines the dynamic scene state used for Thermal rendering, while SyncGS learns one effective spatial correction for each viewpoint's paired cameras and shares this correction across all frames of the corresponding sequence under a shared virtual-camera model. Observable-Owned Optimization assigns motion-based synchronization, image-alignment, and reconstruction objectives to the temporal, spatial, and scene variables, respectively. The current estimates of all variable groups are used in shared forward rendering, while gradients remain restricted to the variable group assigned to each objective. The scene representation is initialized from an RGB-only dynamic Gaussian reconstruction and retains shared and modality-specific Gaussian components.

Our contributions are summarized as follows:

- We formulate dynamic reconstruction from asynchronous and spatially misaligned RGB and Thermal streams as a self-aligning rendering problem with a sequence-level temporal offset and frame-shared effective spatial corrections.
- We develop a continuous temporal calibration method based on a cross-modal motion cost volume with local spatial tolerance, together with effective spatial alignment under a shared virtual-camera model.
- We introduce Observable-Owned Optimization, which assigns synchronization, image-alignment, and reconstruction objectives to the temporal, spatial, and scene variables, respectively, while preserving a shared forward state.
- Experiments on DynamicRGBT-Scenes demonstrate improved RGB and Thermal reconstruction quality, accurate recovery of controlled temporal offsets, and robustness to controlled pose-initialization perturbations.

2 Related Work

2.1 Multimodal RGB and Thermal Reconstruction

ThermoNeRF jointly renders RGB and Thermal views with a multimodal neural radiance field and uses paired observations to learn scene density and modality-specific outputs (Hassan et al., 2024). ThermalGaussian introduces a multimodal 3D Gaussian representation for RGB and Thermal rendering, using calibrated cameras, registered image pairs, and multimodal regularization (Lu et al., 2025). ThermalGaussian++ extends this framework to non-aligned multimodal image pairs through multimodal pose optimization and further adds a joint super-resolution module to improve low-resolution thermal reconstruction (Lu et al., 2026b). BiNAR reconstructs non-aligned RGB and infrared observations with a bi-modal Gaussian framework (Wang et al., 2026a). For unpaired RGB and thermal inputs, Cordonnier et al. (2026) independently estimate camera poses for each modality, then align the resulting pose sets through cross-modal feature matching and Procrustes registration.

Beyond cross-modal appearance, thermal radiation can also vary over time. ThermalGS learns time-conditioned thermal attributes from RGB-guided geometry (Liu et al., 2025b). ETGS equips Gaussian primitives with thermodynamic parameters governed by a heat-transfer ordinary differential equation and models temperature evolution from temporally aligned observations (Wang et al.,

2026b). The Dynamic Thermal Gaussians framework jointly represents scene motion, RGB appearance, and Thermal observations through shared and modality-specific Gaussian components (Lu et al., 2026a). SyncGS reconstructs dynamic scenes from independently captured RGB and thermal streams by estimating their sequence-level temporal offset and spatial alignment corrections.

2.2 Dynamic Gaussian Scene Reconstruction

Gaussian splatting represents a scene with explicit anisotropic primitives and supports efficient differentiable rendering (Kerbl et al., 2023). Prior dynamic Gaussians methods model in different ways: Dynamic 3D Gaussians maintains persistent primitives over time, Deformable 3D Gaussians learns a deformation field, and 4D Gaussian Splatting represents dynamic content in a spatiotemporal feature space (Luiten et al., 2024; Yang et al., 2024; Wu et al., 2024). E-D3DGS further assigns per-Gaussian embeddings to capture heterogeneous motion (Bae et al., 2024). Together, these methods expand the capacity to represent dynamic scenes, but motion modeling alone does not ensure temporal and cross-view consistency. SyncGS addresses this issue by explicitly associating alignment signals to the temporal and spatial variables that they directly constrain.

2.3 Temporal and Spatial Alignment

Synchronization from unsynchronized videos has been studied with learnable temporal offsets, frame-level correspondence, motion-based initialization, and cross-view geometric constraints (Kuo et al., 2013; Furgale et al., 2013; Zhi et al., 2022; Wu et al., 2019; Haresh et al., 2021). In neural rendering, BARF jointly refines camera poses and a radiance field, establishing camera optimization as part of scene reconstruction (Lin et al., 2021a). Sync-NeRF jointly optimizes temporal offsets and a dynamic radiance field (Kim et al., 2024), while Dynamic Gaussian Scene Reconstruction from Unsynchronized Videos combines frame-level matching with continuous temporal refinement (Xu et al., 2026). Human motion can provide both temporal and camera calibration cues (Choi et al., 2025), and Visual Sync derives synchronization from cross-view object tracks and geometric constraints (Liu et al., 2025a).

Dynamic reconstruction additionally requires correspondence along the scene timeline. An incorrect temporal offset changes the dynamic state being compared, while sensor displacement changes where the same motion is observed in the two modalities. SyncGS constructs a locally tolerant motion cost volume across RGB and Thermal observations, learns a session-level temporal offset, and estimates frame-shared effective spatial corrections at the calibrated scene time.

2.4 Optimization of Coupled Variables

Camera pose estimation in differentiable reconstruction can be performed against a fixed scene representation or jointly with scene optimization. iNeRF estimates camera pose by inverting a pretrained NeRF (Lin et al., 2021b), whereas BARF, SPARF, and NoPe-NeRF jointly optimize camera poses and neural scene representations using photometric or geometric constraints (Lin et al., 2021a; Truong et al., 2023; Bian et al., 2023). Gaussian-based methods further extend this paradigm to explicit scene representations by jointly refining camera parameters and Gaussian primitives (Fu et al., 2024; Yao et al., 2025; Gao et al., 2026). In these joint formulations, camera and scene parameters are coupled through differentiable rendering, with reconstruction residuals providing supervision for both parameter groups.

More generally, when multiple objectives act on shared or coupled parameters, multi-task optimization methods address competing objectives by balancing or modifying their gradients. MGDA seeks a common multi-objective update direction (Sener & Koltun, 2018), GradNorm balances gradient magnitudes across tasks (Chen et al., 2018), and PCGrad and CAGrad modify conflicting task gradients (Yu et al., 2020; Liu et al., 2021). In dynamic multimodal reconstruction, Observable-Owned Optimization instead assigns each objective to the variable group directly constrained by its observable. Each objective updates only its associated variables, while all variable groups contribute their current values to the same forward rendering.

3 METHOD

As illustrated in Figure 1, we present SyncGS, a unified framework for self-aligning dynamic reconstruction from independently captured RGB and Thermal streams. The two streams may use different temporal references and viewpoints, so the same nominal frame index need not correspond to the same scene state or viewing geometry. SyncGS integrates temporal calibration, cross-modal spatial alignment, and multimodal Gaussian reconstruction within a common optimization framework. It estimates a sequence-level temporal offset and a sequence-constant relative transformation, while assigning each objective to the variables it directly constrains.

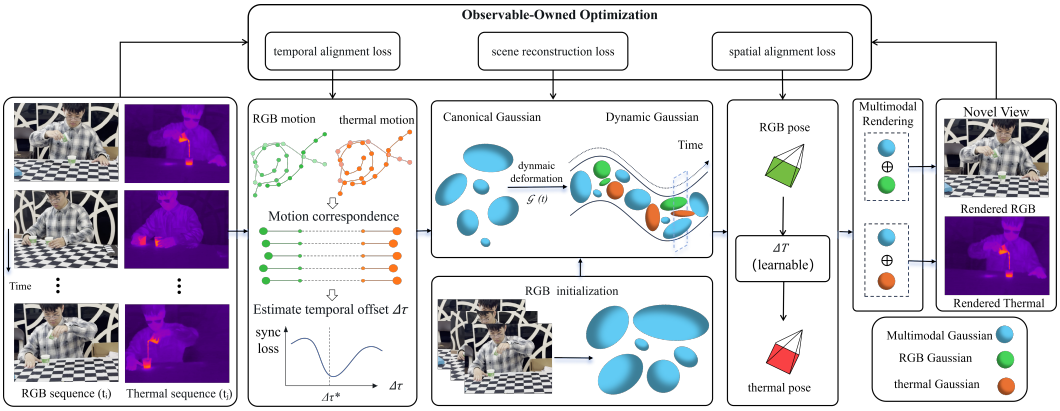

Figure 1: Overview. SyncGS estimates a sequence-level temporal offset from cross-modal motion, learns an effective spatial correction for Thermal rendering, and coordinates these variables with a shared/private dynamic Gaussian representation through objective-specific gradient ownership.

3.1 PRELIMINARIES

SyncGS adopts the multimodal dynamic Gaussian representation introduced by Dynamic Thermal Gaussians (Lu et al., 2026a). Each Gaussian uses shared geometric parameters across modalities while retaining modality-specific attributes. For the i-th Gaussian, the covariance is factorized as $\mathbf{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$, and its multimodal state is represented by

$$G_i = \left(\boldsymbol{\mu}_i, \mathbf{R}_i, \mathbf{S}_i, \{\mathbf{f}_i^m, \sigma_i^m, \mathbf{e}_i^m\}_{m\in\{\mathrm{RGB},\mathrm{Thermal}\}}\right), \tag{1}$$

where $\boldsymbol{\mu}_i$, $\mathbf{R}_i$, and $\mathbf{S}_i$ denote the shared position, rotation, and scale, while $\mathbf{f}_i^m$, σ_i^m, and $\mathbf{e}_i^m$ denote modality-specific appearance, opacity, and deformation-related attributes.

Let $\mathcal{G}_0$ denote the canonical Gaussian representation. A time-conditioned deformation model produces the dynamic scene state

$$\mathcal{G}_\theta(t) = \mathcal{D}_\theta(\mathcal{G}_0, t), \tag{2}$$

where θ denotes the learnable scene parameters. For modality $m \in \{\mathrm{RGB}, \mathrm{Thermal}\}$, the image rendered at scene time t and camera pose P_m is written as

$$\widehat{I}_m(t; P_m) = \mathcal{R}_m(\mathcal{G}_\theta(t), P_m), \tag{3}$$

where $\mathcal{R}_m$ denotes differentiable Gaussian rendering with the corresponding modality-specific attributes and projection settings. This representation separates the dynamic scene state from modality-specific rendering, allowing temporal and spatial correspondence to be optimized within the same scene model.

3.2 MOTION-BASED CONTINUOUS TEMPORAL CALIBRATION

We estimate temporal correspondence from cross-modal motion rather than raw appearance. Although RGB and Thermal images have different intensity distributions, moving structures produce

related temporal changes in both streams. For modality m, we construct a motion map from consecutive observations as

$$M_m(t) = \mathrm{Blur}\left(\left|\mathrm{Norm}(I_m(t+1)) - \mathrm{Norm}(I_m(t))\right|\right), \tag{4}$$

where Norm denotes contrast normalization and Blur suppresses high-frequency noise.

Because the two camera views are not spatially registered, the same moving structure may appear at displaced image locations. We therefore compare motion maps within a local translation neighborhood:

$$C(t,u) = \max_{\delta \in \mathcal{N}} \mathrm{NCC}\left(M_{\mathrm{Thermal}}(t), \mathcal{W}_\delta\left(M_{\mathrm{RGB}}(u)\right)\right), \tag{5}$$

where $\mathcal{W}_\delta$ shifts the RGB motion map by a local translation δ, and NCC denotes normalized cross-correlation. The resulting correlation volume measures the temporal agreement between a Thermal transition at t and an RGB transition at u.

We define the temporal offset such that a positive $\Delta\tau$ associates a Thermal transition at time t with an RGB transition at $t + \Delta\tau$. By continuously querying the correlation volume along the RGB time axis, the offset can take sub-frame values. We estimate the sequence-level offset by minimizing the negative average correlation:

$$\mathcal{L}_{\mathrm{sync}}(\Delta\tau) = -\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} C(t, t+\Delta\tau), \tag{6}$$

where $\mathcal{T}$ is the set of valid Thermal transition times. The same $\Delta\tau$ is shared by the camera pairs within one capture session and is estimated separately for different sessions.

The temporal query is optimized with a coarse-to-fine schedule. The motion maps and correlation volume are computed from the observed images and kept fixed during optimization. Only the query offset is updated by $\mathcal{L}_{\mathrm{sync}}$, while the current estimate of $\Delta\tau$ is used in the subsequent spatial and scene rendering steps. Thus, temporal calibration is optimized jointly with spatial alignment and scene reconstruction through alternating objective-specific updates.

3.3 Cross-Modal Spatial Alignment

Using the current temporal estimate, we learn an effective spatial correction for each RGB and Thermal camera pair. Let $P_{\mathrm{RGB}}(t)$ denote the world-to-camera transformation of the RGB camera at time t. Under the column-vector convention, we define $\Delta T \in SE(3)$ as a transformation from RGB-camera coordinates to Thermal-camera coordinates. The effective Thermal pose is

$$P_{\mathrm{Thermal}}(t) = \Delta T\, P_{\mathrm{RGB}}(t + \Delta\tau), \qquad \Delta T \in SE(3). \tag{7}$$

The correction ΔT is shared across all frames of the same camera pair and initialized as the identity transformation, which imposes a sequence-constant spatial correction.

Both modalities are rendered under a shared virtual-camera model with fixed projection settings. The current temporal estimate determines the dynamic scene state and the corresponding point on the RGB camera trajectory, while ΔT adjusts the effective Thermal viewpoint. The Thermal rendering is therefore written as

$$\widehat{I}_{\mathrm{Thermal}}(t) = \mathcal{R}_{\mathrm{Thermal}}\left(\mathcal{G}_\theta(t + \Delta\tau), P_{\mathrm{Thermal}}(t)\right). \tag{8}$$

The Thermal image residual computed from this rendering provides the spatial-alignment objective. During a spatial update, the current temporal estimate and scene parameters remain in the forward pass, while the gradient is assigned to ΔT. The updated spatial correction is then used together with the latest temporal estimate in subsequent forward passes. Temporal and spatial variables are consequently optimized together with the dynamic Gaussian scene rather than in separate calibration stages.

3.4 Observable-Owned Optimization

The temporal offset, effective Thermal transformation, and dynamic Gaussian parameters jointly affect the projected observations. To prevent one variable from absorbing the error of another, SyncGS formulates the training objective as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{scene}} + \lambda_{\mathrm{sync}}\mathcal{L}_{\mathrm{sync}} + \lambda_{\mathrm{pose}}\mathcal{L}_{\mathrm{pose}}. \tag{9}$$

Table 1: Per-scene PSNR comparison on DynamicRGBT-Scenes. T-GS denotes ThermalGaussian, and DTG denotes Dynamic Thermal Gaussians with multimodal reconstruction regularization. Complete reconstruction metrics are reported in the appendix.

Modality	Metric	Method	Heating Table	HotBar	HeatGun	Bacon	IcePacks	HairDryer	HairDryer Dark	Ironing Cloth	Candles	HotWater	Foam	Covers	Avg.
Thermal	PSNR ↑	D-3DGS	18.13	28.31	23.80	22.57	29.09	24.34	23.17	28.33	34.02	29.54	24.90	21.57	25.65
		4DGaussians	24.83	31.28	30.00	26.60	32.33	27.39	29.49	30.82	33.17	30.98	36.64	29.33	30.24
		E-D3DGS	25.81	30.93	30.56	29.46	31.01	24.50	30.81	32.27	32.20	31.56	38.11	31.43	30.72
		T-GS	17.99	19.88	22.12	19.32	25.06	21.33	21.04	23.37	27.50	23.53	22.84	18.81	21.90
		MMOne	18.05	20.33	22.35	21.01	25.29	22.43	20.76	23.55	27.83	23.86	23.97	19.06	22.37
		DTG	26.45	33.69	33.58	31.54	33.91	27.55	30.90	32.28	34.94	35.40	38.57	33.38	32.68
		Ours	**28.18**	**34.64**	**33.94**	**32.00**	**35.97**	**33.25**	**33.17**	**36.26**	**37.43**	**36.72**	**41.35**	**36.06**	**34.92**
RGB	PSNR ↑	D-3DGS	18.11	22.73	22.36	18.32	15.08	14.35	27.10	16.80	22.95	16.45	15.52	18.17	18.99
		4DGaussians	19.87	27.19	26.23	20.30	18.03	15.50	30.09	21.99	24.72	19.07	19.02	23.74	22.15
		E-D3DGS	21.55	27.68	28.00	23.25	18.60	15.66	31.73	23.75	26.84	22.39	22.45	25.38	23.94
		T-GS	15.96	16.61	18.81	11.89	11.43	13.20	21.64	13.49	14.26	11.70	12.58	15.57	14.76
		MMOne	16.14	16.96	18.42	12.45	11.63	12.69	21.34	13.35	14.43	11.79	12.65	15.29	14.76
		DTG	21.03	27.63	28.02	23.87	19.40	16.93	31.92	24.09	27.63	22.44	22.95	26.57	24.37
		Ours	**22.09**	**28.36**	**29.42**	**24.87**	**19.89**	**18.26**	**33.06**	**25.88**	**28.96**	**23.46**	**25.04**	**26.59**	**25.49**

Here, $\mathcal{L}_{\text{scene}}$ contains the RGB and Thermal reconstruction losses together with the regularization terms of the shared/private Gaussian representation.

We assign each objective to the variables directly constrained by its corresponding observation:

$$\mathcal{U}_{\text{sync}} = \{\Delta\tau\}, \qquad \mathcal{U}_{\text{pose}} = \{\Delta T\}, \qquad \mathcal{U}_{\text{scene}} = \{\theta\}. \tag{10}$$

Thus, $\mathcal{L}_{\text{sync}}$ updates $\Delta\tau$, $\mathcal{L}_{\text{pose}}$ updates ΔT, and $\mathcal{L}_{\text{scene}}$ updates θ. Variables outside the assigned set are held fixed during differentiation, while all current estimates remain in the forward rendering. The optimization therefore preserves the coupling among temporal, spatial, and scene variables in the forward pass while separating their parameter updates.

We initialize the optimization with an RGB-only dynamic Gaussian representation. Synchronization, pose, and scene updates alternate throughout joint training, and each update uses the latest shared state while modifying only its assigned variables. The updated estimates are then used in subsequent forward passes. Detailed update budgets are provided in Section 4.1.

4 Experiments

4.1 Implementation Details

We evaluate SyncGS on DynamicRGBT-Scenes (Lu et al., 2026a), which comprises 12 scenes with diverse motion patterns, thermal variations, and illumination conditions. The dataset was captured using two camera rigs rigidly mounted on the same moving platform. Within each rig, the RGB and thermal streams were recorded independently, with a sequence-specific temporal offset. We follow Dynamic Thermal Gaussians in the scene naming and ordering.

We evaluate reconstruction quality for both modalities using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. We initialize the optimization with an RGB-only dynamic Gaussian representation trained for 30,000 updates, followed by alternating calibration and scene updates in Stage 2. The temporal offset is initialized at 0.

4.2 Dynamic RGB and Thermal Reconstruction

Table 1 compares SyncGS with D-3DGS (Yang et al., 2024), 4DGaussians (Wu et al., 2024), E-D3DGS (Bae et al., 2024), ThermalGaussian (Lu et al., 2025), MMOne (Gu & Wang, 2025), and Dynamic Thermal Gaussians on DynamicRGBT-Scenes (Lu et al., 2026a).

Across the twelve scenes, Ours yields higher PSNR than the Dynamic Thermal Gaussians baseline in both modalities. The gains are particularly clear in scenes with substantial thermal changes or

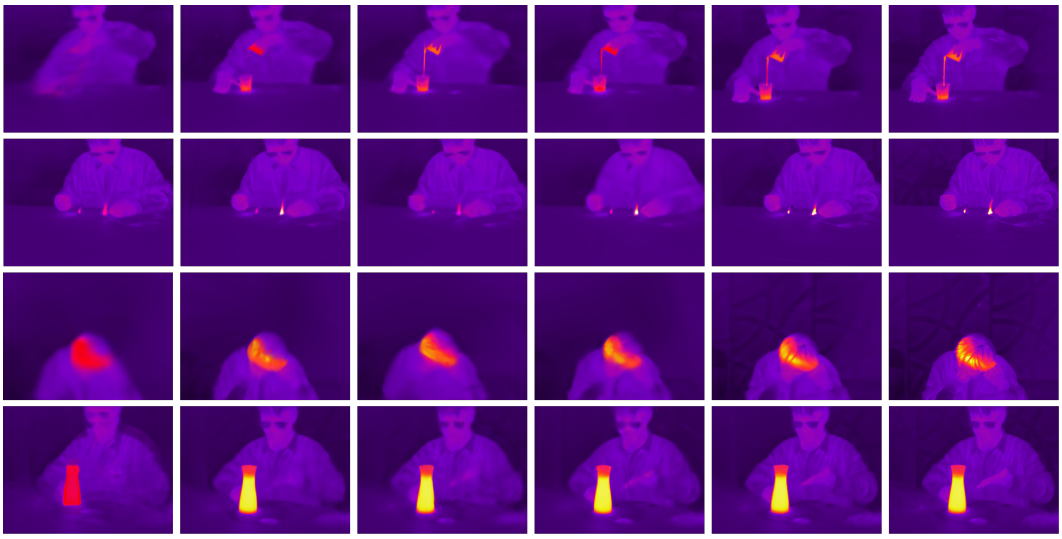

Figure 2: Qualitative comparison of thermal novel-view synthesis on DynamicRGBT-Scenes.

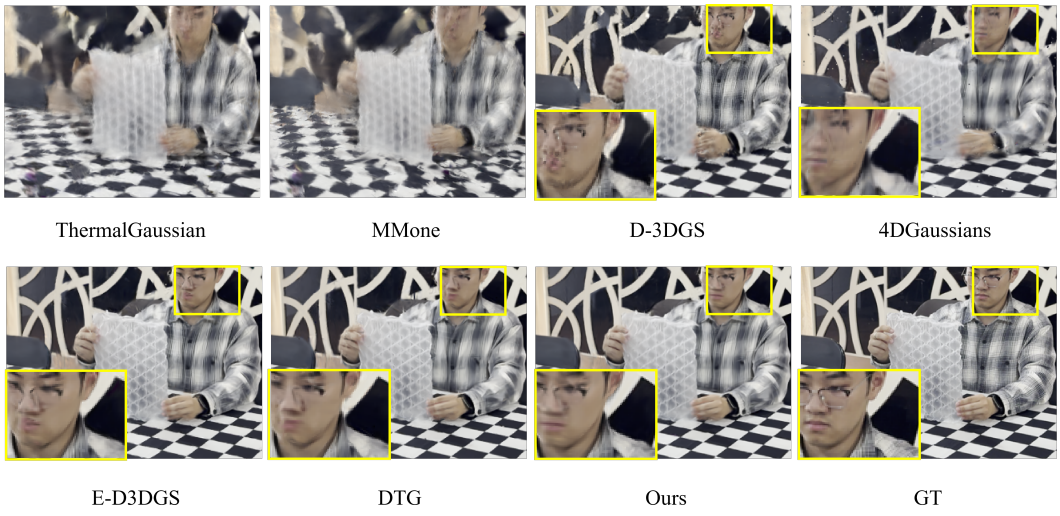

Figure 3: Qualitative comparison of RGB novel-view synthesis on DynamicRGBT-Scenes.

pronounced motion. This per-scene consistency supports the benefit of jointly calibrating temporal correspondence and cross-modal geometry. Figures 2 and 3 compare Thermal and RGB novel-view synthesis at matched viewpoints and times. The qualitative results complement Table 1: Ours preserves moving boundaries, localized thermal responses, and fine RGB structures more faithfully across the selected scenes.

4.3 Temporal Calibration and Gradient Ownership

We first evaluate the recoverable range of temporal calibration on the Covers scene by gradually increasing the injected temporal offset. Covers contains pronounced object motion and thermal variation, providing clear cross-modal motion cues. The learnable temporal offset is initialized at zero, and all settings use the same initialization, training views, and optimization budget. For each injected offset, the recovery error is reported as the absolute difference between the estimated and injected offsets.

Table 2: Temporal recovery under increasing injected offsets on the Covers scene. Thermal reconstruction metrics are reported for each setting, and the recovery error is measured in frames.

Temporal Recovery			Thermal Reconstruction		
Injected Offset	Estimated Offset	Recovery Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓
0	-0.002	0.002	35.87	0.957	0.080
10	9.988	0.012	35.77	0.956	0.080
20	19.996	0.004	35.77	0.958	0.077
30	30.005	0.005	35.82	0.959	0.077
40	18.108	21.892	32.93	0.943	0.102

Table 3: Controlled temporal perturbation on the Covers scene. All variants use a 20-frame shift, and the recovery error is measured in frames.

		RGB			Thermal		
Variant	Recovery Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fixed Clock	20.00	26.53	0.889	0.122	34.51	0.952	0.089
Joint All-Gradient	27.79	26.45	0.889	0.129	34.16	0.951	0.096
Ours	**0.01**	**26.62**	**0.892**	**0.119**	**35.77**	**0.958**	**0.077**

For injected offsets up to 30 frames, SyncGS recovers the temporal shift with an error no larger than 0.012 frames, while the Thermal reconstruction metrics remain stable. At 40 frames, the estimated offset deviates substantially from the injected value, with the recovery error increasing to 21.892 frames. The accompanying degradation in PSNR, SSIM, and LPIPS places the observed recovery boundary between 30 and 40 frames under the tested settings.

We next isolate the effect of gradient ownership using a fixed 20-frame temporal perturbation. Fixed Clock disables temporal calibration and holds the offset at its zero initialization. Joint All-Gradient enables temporal calibration but allows the reconstruction objective to propagate through all differentiable variables, including the temporal parameter. Ours routes each objective to its corresponding parameter group while preserving the shared state used during forward rendering.

As shown in Table 3, Ours recovers the imposed temporal shift with an error of only 0.01 frames, whereas Fixed Clock cannot compensate for the perturbation and Joint All-Gradient converges to an incompatible temporal solution with an error of 27.79 frames. Ours also achieves the strongest reconstruction performance, particularly in the Thermal modality. These results suggest that routing motion-based supervision to the temporal variable limits interference from reconstruction gradients, while retaining the shared forward state required for scene optimization.

4.4 Spatial Alignment Evaluation

We assess spatial-alignment robustness on HairDryer, where moving structures provide clear cross-modal boundary cues. We perturb the initial rotation and translation of the shared Thermal camera transformation while keeping the temporal offset fixed at zero. Translation perturbations are normalized by the scene extent, and all settings use the same perturbation directions and training budget.

As shown in Table 4, Thermal reconstruction remains stable up to the $30^\circ/0.6$ initialization, where PSNR changes by only 0.56 dB and SSIM remains at or above 0.945. At $50^\circ/1.0$, PSNR decreases to 28.04 dB and LPIPS increases to 0.232, while stronger perturbations from $60^\circ/1.0$ to $100^\circ/5.0$ become numerically unstable after Gaussian densification. The tested results therefore place the observed stability boundary between $50^\circ/1.0$ and $60^\circ/1.0$.

4.5 Ablation Studies

We evaluate the contributions of temporal calibration and effective spatial alignment using four matched configurations. Dynamic Thermal Gaussians serves as the baseline, while Temporal-only and Spatial-only activate the two alignment components separately. Ours combines both components under the same reconstruction and evaluation protocol.

Table 4: Thermal reconstruction under increasing spatial initialization perturbations on HairDryer.

Initialization		Thermal Reconstruction		
Rotation	Translation	PSNR ↑	SSIM ↑	LPIPS ↓
5°	0.1	34.55	0.952	0.126
10°	0.2	34.47	0.950	0.142
20°	0.4	34.34	0.949	0.138
30°	0.6	33.99	0.945	0.153
50°	1.0	28.04	0.907	0.232

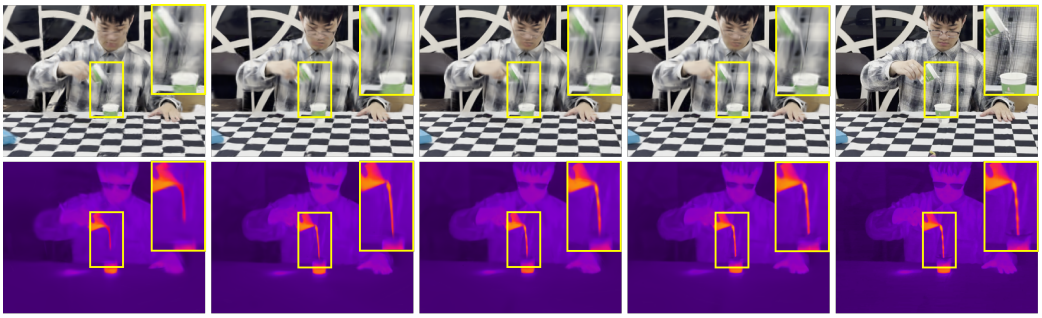

Figure 4: Qualitative ablation results on RGB and Thermal renderings.

Table 5: Ablation of temporal calibration and spatial alignment, averaged over 12 scenes.

	RGB			Thermal		
Variant	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DTG	24.37	0.809	0.232	32.68	0.949	0.109
Temporal-only	24.96	0.829	0.200	33.10	0.953	0.096
Spatial-only	25.40	0.834	0.197	34.34	0.959	0.083
Ours	**25.49**	**0.835**	**0.196**	**34.92**	**0.963**	**0.080**

Table 5 shows that both temporal calibration and spatial alignment improve reconstruction over the baseline, with spatial alignment providing the larger single-component gain, particularly in Thermal PSNR. Combining the two yields the best results: SyncGS achieves 25.49 dB RGB PSNR and 34.92 dB Thermal PSNR, together with the highest SSIM and lowest LPIPS in both modalities. Adding temporal calibration to Spatial-only further improves Thermal PSNR by 0.58 dB, indicating complementary benefits. The qualitative results in Fig. 4 are consistent, showing more coherent dynamic structures and more consistent RGB and Thermal renderings. We analyze objective-specific gradient ownership separately in Sec. 4.3.

5 Conclusion

We presented SyncGS, a self-aligning framework for dynamic reconstruction from temporally asynchronous and spatially misaligned RGB and Thermal streams. SyncGS jointly optimizes sequence-level temporal correspondence and frame-shared effective spatial correction within a shared/private dynamic Gaussian representation. Its Observable-Owned Optimization routes motion, alignment, and reconstruction objectives to their corresponding variables while maintaining a shared forward state. Experiments on DynamicRGBT-Scenes show improved reconstruction in both modalities, accurate temporal-offset recovery within the tested range, and robustness to the tested pose initializations. Overall, SyncGS provides a unified framework for self-aligning dynamic RGB and Thermal reconstruction.

AI USE STATEMENT

Large Language Models (LLMs) were used solely to assist with the linguistic refinement of this manuscript, including grammar checking, sentence rephrasing, readability improvement, and polishing of the presentation. The LLMs were not involved in the research ideation, problem formulation, methodology design, experimental design, data collection, data analysis, or interpretation of the results. All scientific contributions, experiments, analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this manuscript, including any text that was assisted or refined by LLMs.

REFERENCES

Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3D gaussian splatting. In *Computer Vision – ECCV 2024*, 2024. URL https://arxiv.org/abs/2404.03613.

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*, pp. 4160–4169, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Bian_NoPe-NeRF_Optimising_Neural_Radiance_Field_With_No_Pose_Prior_CVPR_2023_paper.html.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Changwoon Choi, Jeongjun Kim, Geonho Cha, Minkwan Kim, Dongyoon Wee, and Young Min Kim. Humans as a calibration pattern: Dynamic 3D scene reconstruction from unsynchronized and uncalibrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6598–6608, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Choi_Humans_as_a_Calibration_Pattern_Dynamic_3D_Scene_Reconstruction_from_ICCV_2025_paper.html.

Jean Cordonnier, Chenghao Xu, Olga Fink, and Malcolm Mielle. Unpaired RGB-Thermal gaussian-splatting using visual geometric transformers, 2026. URL https://arxiv.org/abs/2606.05491.

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*, pp. 20796–20805, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Fu_COLMAP-Free_3D_Gaussian_Splatting_CVPR_2024_paper.html.

Paul Furgale, Jörn Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1280–1286, 2013. doi: 10.1109/IROS.2013.6696514.

Yu Gao, Lutong Su, Ruixiang Huang, Tianji Jiang, Jiadong Tang, Yufeng Yue, and Yi Yang. Energy-GS: Image energy-guided pose alignment gaussian splatting with redesigned pose gradient flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7319, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gao_Energy-GS_Image_Energy-guided_Pose_Alignment_Gaussian_Splatting_with_redesigned_pose_CVPR_2026_paper.html.

Zhifeng Gu and Bing Wang. MMOne: Representing multiple modalities in one scene. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1088–1098, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Gu_MMOne_Representing_Multiple_Modalities_in_One_Scene_ICCV_2025_paper.html.

Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N. Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5548–5558, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Haresh_Learning_by_Aligning_Videos_in_Time_CVPR_2021_paper.html.

Mariam Hassan, Florent Forest, Olga Fink, and Malcolm Mielle. ThermoNeRF: Joint RGB and thermal novel view synthesis for building facades using multimodal neural radiance fields, 2024. URL https://arxiv.org/abs/2403.12154.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433.

Seoha Kim, Jeongmin Bae, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Sync-NeRF: Generalizing dynamic NeRFs to unsynchronized videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:2777–2785, 2024. doi: 10.1609/aaai.v38i3.28057.

Thomas Kuo, Santhoshkumar Sunderrajan, and B. S. Manjunath. Camera alignment using trajectory intersections in unsynchronized videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1121–1128, 2013. doi: 10.1109/ICCV.2013.143.

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*, pp. 5741–5751, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Lin_BARF_Bundle-Adjusting_Neural_Radiance_Fields_ICCV_2021_paper.html.

Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1323–1330, 2021b. doi: 10.1109/IROS51168.2021.9636708. URL https://doi.org/10.1109/IROS51168.2021.9636708.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Shaowei Liu, David Yifan Yao, Saurabh Gupta, and Shenlong Wang. Visual sync: Multi-camera synchronization via cross-view object motion, 2025a. URL https://arxiv.org/abs/2512.02017.

Yuxiang Liu, Xi Chen, Shen Yan, Zeyu Cui, Huaxin Xiao, Yu Liu, and Maojun Zhang. ThermalGS: Dynamic 3D thermal reconstruction with gaussian splatting. *Remote Sensing*, 17(2):335, 2025b. doi: 10.3390/rs17020335. URL https://www.mdpi.com/2072-4292/17/2/335.

Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. ThermalGaussian: Thermal 3D gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.07200.

Rongfeng Lu, Lifeng Lin, Xiaobao Wei, Quan Chen, Ming Lu, Yitian Xue, Yaoqi Sun, Yuhan Gao, Anke Xue, and Chenggang Yan. Dynamic thermal gaussians: Multimodal 4D gaussian splatting, 2026a. URL https://arxiv.org/abs/2609.24531.

Rongfeng Lu, Chi Zhu, Quan Chen, Le Zhang, Ming Lu, Tingyu Wang, Haofan Ren, Yitian Xue, Yunfei Guo, and Chenggang Yan. ThermalGaussian++: Improving alignment and resolution for ThermalGaussian. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9), 2026b. doi: 10.1109/TPAMI.2026.3689388.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision*, pp. 800–809, 2024. doi: 10.1109/3DV62453.2024.00044.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

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*, pp. 4190–4200, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Truong_SPARF_Neural_Radiance_Fields_From_Sparse_and_Noisy_Poses_CVPR_2023_paper.html.

Zhongwen Wang, Han Ling, Weihao Zhang, Yinghui Sun, and Quansen Sun. BiNAR: A bi-modal framework for non-aligned RGB-IR 3D reconstruction via gaussian splatting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4407–4416, 2026a. URL https://openaccess.thecvf.com/content/WACV2026/html/Wang_BiNAR_A_Bi-Modal_Framework_for_Non-Aligned_RGB-IR_3D_Reconstruction_via_WACV_2026_paper.html.

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*, 2026b. URL https://openreview.net/forum?id=P2Nw2LMkjH.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004. URL https://ece.uwaterloo.ca/~z70wang/publications/ssim.html.

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*, pp. 20310–20320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_4D_Gaussian_Splatting_for_Real-Time_Dynamic_Scene_Rendering_CVPR_2024_paper.html.

Xinyi Wu, Zhenyao Wu, Yujun Zhang, Lili Ju, and Song Wang. Multi-video temporal synchronization by matching pose features of shared moving subjects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2729–2738, 2019. doi: 10.1109/ICCVW.2019.00334.

Zhixin Xu, Hengyu Zhou, Yuan Liu, Wenhan Xue, Hao Pan, Wenping Wang, and Bin Wang. Dynamic gaussian scene reconstruction from unsynchronized videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11469–11477, 2026. doi: 10.1609/aaai.v40i14.38129.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Deformable_3D_Gaussians_for_High-Fidelity_Monocular_Dynamic_Scene_Reconstruction_CVPR_2024_paper.html.

Yuxin Yao, Yan Zhang, Zhening Huang, and Joan Lasenby. SmallGS: Gaussian splatting-based camera pose estimation for small-baseline videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4560–4569, 2025. doi: 10.1109/CVPRW67362.2025.00437. URL https://openaccess.thecvf.com/content/CVPR2025W/PBVS/html/Yao_SmallGS_Gaussian_Splatting-based_Camera_Pose_Estimation_for_Small-Baseline_Videos_CVPRW_2025_paper.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Xiangyang Zhi, Jiawei Hou, Yiren Lu, Laurent Kneip, and Sören Schwertfeger. Multical: Spatiotemporal calibration for multiple IMUs, cameras and LiDARs. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2446–2453, 2022. doi: 10.1109/IROS47612.2022.9982031.

A APPENDIX

This appendix reports the complete per-scene reconstruction metrics on DynamicRGBT-Scenes. DTG denotes Dynamic Thermal Gaussians with multimodal reconstruction regularization, T-GS denotes ThermalGaussian, and Avg. denotes the average across all 12 scenes.

Table 6: Complete per-scene reconstruction results in the Thermal modality on DynamicRGBT-Scenes.

Metric	Method	Heating Table	Hot Bar	Heat Gun	Bacon	Ice Packs	Hair Dryer	Hair Dryer Dark	Ironing Cloth	Candles	Hot Water	Foam	Covers	Avg.
PSNR↑	D-3DGS	18.13	28.31	23.80	22.57	29.09	24.34	23.17	28.33	34.02	29.54	24.90	21.57	25.65
	4DGaussians	24.83	31.28	30.00	26.60	32.33	27.39	29.49	30.82	33.17	30.98	36.64	29.33	30.24
	E-D3DGS	25.81	30.93	30.56	29.46	31.01	24.50	30.81	32.27	32.20	31.56	38.11	31.43	30.72
	T-GS	17.99	19.88	22.12	19.32	25.06	21.33	21.04	23.37	27.50	23.53	22.84	18.81	21.90
	MMOne	18.05	20.33	22.35	21.01	25.29	22.43	20.76	23.55	27.83	23.86	23.97	19.06	22.37
	DTG	26.45	33.69	33.58	31.54	33.91	27.55	30.90	32.28	34.94	35.40	38.57	33.38	32.68
	Ours	**28.18**	**34.64**	**33.94**	**32.00**	**35.97**	**33.25**	**33.17**	**36.26**	**37.43**	**36.72**	**41.35**	**36.06**	**34.92**
SSIM↑	D-3DGS	0.889	0.926	0.943	0.927	0.951	0.883	0.897	0.922	0.960	0.927	0.882	0.885	0.916
	4DGaussians	0.913	0.941	0.960	0.914	0.959	0.899	0.907	0.928	0.945	0.926	0.956	0.912	0.930
	E-D3DGS	0.923	0.941	0.963	0.941	0.954	0.870	0.918	0.943	0.934	0.931	0.965	0.921	0.934
	T-GS	0.867	0.814	0.935	0.848	0.912	0.859	0.850	0.869	0.916	0.874	0.889	0.852	0.874
	MMOne	0.881	0.831	0.939	0.876	0.923	0.873	0.858	0.879	0.921	0.888	0.907	0.861	0.886
	DTG	0.933	0.960	0.975	0.958	0.965	0.900	0.921	0.945	0.961	0.955	0.967	0.945	0.949
	Ours	**0.949**	**0.970**	**0.978**	**0.963**	**0.976**	**0.945**	**0.943**	**0.967**	**0.969**	**0.963**	**0.979**	**0.960**	**0.963**
LPIPS↓	D-3DGS	0.175	0.130	0.091	0.143	0.160	0.304	0.229	0.152	0.096	0.158	0.161	0.238	0.170
	4DGaussians	0.134	0.114	0.081	0.162	0.114	0.248	0.213	0.152	0.133	0.157	0.101	0.169	0.148
	E-D3DGS	0.113	0.097	0.073	0.122	0.135	0.284	0.205	0.127	0.165	0.160	0.085	0.161	0.144
	T-GS	0.216	0.263	0.134	0.279	0.263	0.317	0.315	0.255	0.215	0.270	0.224	0.305	0.255
	MMOne	0.213	0.261	0.131	0.259	0.258	0.318	0.323	0.259	0.220	0.273	0.217	0.306	0.253
	DTG	0.097	0.065	0.053	0.084	0.089	0.235	0.197	0.121	0.099	0.092	0.080	0.101	0.109
	Ours	**0.073**	**0.056**	**0.043**	**0.074**	**0.064**	**0.155**	**0.143**	**0.071**	**0.076**	**0.081**	**0.051**	**0.082**	**0.080**

Table 7: Complete per-scene reconstruction results in the RGB modality on DynamicRGBT-Scenes.

Metric	Method	Heating Table	Hot Bar	Heat Gun	Bacon	Ice Packs	Hair Dryer	Hair Dryer Dark	Ironing Cloth	Candles	Hot Water	Foam	Covers	Avg.
PSNR↑	D-3DGS	18.11	22.73	22.36	18.32	15.08	14.35	27.10	16.80	22.95	16.45	15.52	18.17	18.99
	4DGaussians	19.87	27.19	26.23	20.30	18.03	15.50	30.09	21.99	24.72	19.07	19.02	23.74	22.15
	E-D3DGS	21.55	27.68	28.00	23.25	18.60	15.66	31.73	23.75	26.84	22.39	22.45	25.38	23.94
	T-GS	15.96	16.61	18.81	11.89	11.43	13.20	21.64	13.49	14.26	11.70	12.58	15.57	14.76
	MMOne	16.14	16.96	18.42	12.45	11.63	12.69	21.34	13.35	14.43	11.79	12.65	15.29	14.76
	DTG	21.03	27.63	28.02	23.87	19.40	16.93	31.92	24.09	27.63	22.44	22.95	26.57	24.37
	Ours	**22.09**	**28.36**	**29.42**	**24.87**	**19.89**	**18.26**	**33.06**	**25.88**	**28.96**	**23.46**	**25.04**	**26.59**	**25.49**
SSIM↑	D-3DGS	0.576	0.804	0.776	0.697	0.555	0.632	0.816	0.633	0.854	0.587	0.665	0.688	0.690
	4DGaussians	0.613	0.869	0.845	0.704	0.629	0.618	0.848	0.755	0.868	0.651	0.691	0.839	0.744
	E-D3DGS	0.679	0.886	0.885	0.805	0.661	0.595	0.877	0.805	0.897	0.767	0.815	0.863	0.795
	T-GS	0.488	0.590	0.632	0.339	0.305	0.482	0.623	0.445	0.526	0.305	0.439	0.568	0.479
	MMOne	0.513	0.612	0.633	0.386	0.338	0.477	0.605	0.444	0.527	0.332	0.458	0.568	0.491
	DTG	0.659	0.887	0.879	0.827	0.690	0.667	0.886	0.816	0.909	0.775	0.823	**0.893**	0.809
	Ours	**0.698**	**0.902**	**0.906**	**0.848**	**0.709**	**0.718**	**0.901**	**0.860**	**0.929**	**0.799**	**0.864**	0.890	**0.835**
LPIPS↓	D-3DGS	0.459	0.274	0.303	0.336	0.458	0.369	0.333	0.410	0.158	0.446	0.336	0.333	0.351
	4DGaussians	0.402	0.186	0.232	0.365	0.374	0.386	0.308	0.311	0.156	0.387	0.352	0.171	0.303
	E-D3DGS	0.307	0.159	0.187	0.226	0.314	0.378	0.278	0.266	0.125	0.246	0.216	0.148	0.238
	T-GS	0.525	0.442	0.429	0.612	0.601	0.465	0.509	0.508	0.417	0.611	0.491	0.367	0.498
	MMOne	0.508	0.437	0.443	0.576	0.592	0.474	0.510	0.516	0.423	0.595	0.481	0.372	0.494
	DTG	0.349	0.162	0.204	0.212	0.297	0.336	0.255	0.265	0.119	0.247	0.217	0.121	0.232
	Ours	**0.266**	**0.139**	**0.159**	**0.183**	**0.270**	**0.298**	**0.233**	**0.208**	**0.090**	**0.221**	**0.173**	**0.120**	**0.196**