ANYTHERMALSPLAT: SCALING POSE-FREE FEED-FORWARD THERMAL SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

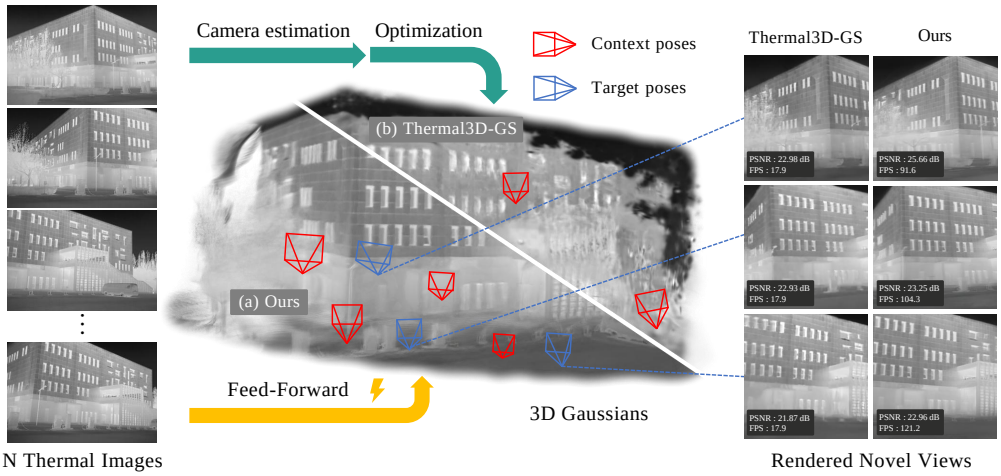


Figure 1: Optimization-based methods rely on externally estimated cameras and per-scene fitting, whereas **AnyThermalSplat** predicts a renderable thermal Gaussian scene from *unposed* and *uncalibrated* thermal images in a single forward pass. We report novel-view PSNR and rendering speed (FPS), excluding scene reconstruction time.

ABSTRACT

Thermal 3D reconstruction typically relies on known cameras and per-scene optimization, while the scarcity of multi-view thermal data hinders scalable feed-forward reconstruction. We present AnyThermalSplat, a pose-free feed-forward framework that predicts renderable thermal Gaussian scenes without external camera parameters or test-time scene optimization. It adapts RGB-pretrained geometric priors through low-rank updates and residual feature bridges, with a frozen RGB teacher providing geometric supervision alongside thermal reconstruction and held-out novel-view objectives; inference requires thermal images only. To expand supervision, we introduce **ThermalVerse**, a multi-domain RGB–thermal dataset, and a geometry-grounded thermalization pipeline that fits shared thermal appearance on fixed RGB-reconstructed geometry. Together with curated public data, these resources form a 500-scene training pool. On in-domain scenes, AnyThermalSplat achieves 26.76 and 28.48 dB PSNR with 6 and 12 context views, respectively, outperforming our matched thermally adapted **AnySplat-IR** baseline by 2.84 and 1.77 dB. This baseline is trained on the same thermal data under the same adaptation protocol, providing a controlled comparison beyond directly applying the original RGB model. We also conduct a controlled scaling study on nested 100–500-scene subsets under a fixed update budget and find substantial gains in both in-domain and out-of-domain reconstruction as the training set grows from 100 to 300 scenes. Beyond this regime, out-of-domain performance largely saturates, while in-domain reconstruction retains an overall benefit from larger training sets. We will release the data and code, along with model checkpoints.

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.

1 INTRODUCTION

Thermal infrared (TIR) imaging remains informative when darkness or changing illumination degrades visible-light sensing. A renderable thermal 3D representation integrates observations across viewpoints and enables novel-view synthesis. Existing thermal radiance-field and Gaussian Splatting methods typically require externally estimated cameras and per-scene optimization (Ye et al., 2024; Lu et al., 2025; Chen et al., 2024a), making it difficult to reconstruct new scenes directly from unposed thermal images in a feed-forward manner.

Recent geometric foundation models and feed-forward Gaussian predictors have enabled camera estimation and renderable scene reconstruction from image collections (Wang et al., 2024; 2025; 2026a; Jiang et al., 2025b; Ye et al., 2026). Extending these advances to thermal imagery, however, presents two challenges. First, RGB-pretrained priors face a modality gap: thermal appearance depends on surface temperature, emissivity, reflected radiation, and sensor response, so thermal contrast and boundaries need not align with visible-light appearance. Cross-spectral methods have explored geometric transfer to thermal imagery (Skorokhodov et al., 2026; Kweon et al., 2026), but do not directly address pose-free prediction of renderable thermal Gaussian scenes. Second, diverse multi-view thermal supervision remains scarce. Acquiring real RGB–thermal data at scale is costly, while independently translating RGB views into thermal images can assign inconsistent appearances to the same 3D surface across viewpoints. We therefore jointly address cross-modal geometric transfer and scalable multi-view thermal supervision.

We present AnyThermalSplat, a pose-free feed-forward framework that predicts cameras, depth, and a thermal Gaussian scene from unposed and uncalibrated thermal context images. Low-rank adaptation transfers RGB-pretrained geometric priors to thermal imagery, while residual feature bridges adapt intermediate features for a pretrained Gaussian decoder. A frozen RGB teacher provides geometric supervision during training, and held-out thermal views enforce novel-view consistency of the context-derived scene. At inference, AnyThermalSplat requires only thermal images, without externally supplied camera parameters or per-scene optimization.

To expand multi-view supervision, we introduce ThermalVerse, a self-collected RGB–thermal dataset spanning 209 scenes, together with a geometry-grounded thermalization pipeline that fits shared thermal appearance on fixed RGB-reconstructed geometry. Combined with curated public data, these resources form a 500-scene real-and-synthetic training pool. Nested 100–500-scene subsets further support a controlled data-scaling study under a fixed update budget, allowing us to examine how increasing thermal supervision affects in-domain reconstruction and out-of-domain generalization.

We evaluate AnyThermalSplat on both in-domain scenes and out-of-domain TI-NSD (Chen et al., 2024a) scenes under sparse and dense context settings. Compared with our matched AnySplat-IR baseline, thermally adapted using the same training data and protocol, AnyThermalSplat improves in-domain PSNR by 2.84 and 1.77 dB with 6 and 12 context views, respectively. Scaling the training set from 100 to 300 scenes yields clear gains in both in-domain and out-of-domain reconstruction. Beyond this regime, out-of-domain performance largely saturates, whereas in-domain reconstruction retains an overall benefit from larger training sets.

Our contributions are:

- A pose-free feed-forward thermal Gaussian reconstruction framework that jointly predicts cameras and renderable thermal scenes from unposed and uncalibrated thermal images without per-scene optimization.
- A self-collected multi-domain RGB–thermal dataset, ThermalVerse, comprising 209 scenes across 30 content categories and 26K RGB–thermal image pairs, together with a geometry-grounded thermalization pipeline for cross-view-consistent synthetic supervision.
- A controlled 100–500-scene data-scaling study that validates the benefits of expanding thermal supervision under a fixed update budget, showing clear reconstruction gains as the training set scales up.

2 Related Work

Feed-forward and pose-free reconstruction. Feed-forward reconstruction methods predict explicit Gaussian scenes without per-scene optimization. Early approaches such as Splatter Image, pixelSplat, and MVSplat focus on single-view or sparse multi-view reconstruction (Szymanowicz et al., 2024; Charatan et al., 2024; Chen et al., 2024b), while DepthSplat incorporates depth estimation and GS-LRM scales model capacity (Xu et al., 2025; Zhang et al., 2024). More recent work progressively relaxes camera assumptions: NoPoSplat accepts unposed images while retaining known intrinsics (Ye et al., 2025); PF3plat removes supplied poses, PreF3R supports variable-length image sequences, and AnySplat and YoNoSplat reconstruct from unposed and uncalibrated image collections with varying numbers of views (Hong et al., 2025; Chen et al., 2024c; Jiang et al., 2025b; Ye et al., 2026). TokenSplat further studies token alignment, while ForeSplat introduces optimization-aware foresight (Li et al., 2026a;b). These methods establish strong feed-forward reconstruction priors for RGB imagery, whereas extending such priors to thermal inputs requires adaptation to the substantial cross-modal gap.

Thermal reconstruction and cross-modal transfer. ThermalNeRF, ThermoNeRF, Thermal-Gaussian, Thermal3D-GS, and PhysIR-Splat reconstruct thermal appearance through per-scene optimization (Lin et al., 2024; Hassan et al., 2025; Lu et al., 2025; Chen et al., 2024a; Gao et al., 2026). PhysIR-Splat introduces a learned thermal geometry initializer, but still requires subsequent Gaussian optimization. Cross-modal methods such as SEAR adapt geometric transformers for RGB–thermal pose and geometry estimation, while DarkVGGT leverages thermal cues for geometric reasoning under degraded illumination (Skorokhodov et al., 2026; Kweon et al., 2026). RoRE explores feed-forward multimodal view synthesis using ray-conditioned camera embeddings (Griffiths & Dansereau, 2026). In contrast, our setting targets pose-free feed-forward reconstruction of an explicit thermal Gaussian scene from thermal images, with RGB supervision used only during training.

Data scaling for 3D reconstruction. MegaSynth provides 700K procedurally generated scenes for reconstruction training (Jiang et al., 2025a), while VGGT-Ω studies the effects of scaling RGB training data and model capacity (Wang et al., 2026a). In contrast, large-scale multi-view thermal resources remain scarce. We study this data-constrained regime through a controlled 100–500-scene scaling study under a fixed architecture and update budget, examining how increasing thermal supervision affects in-domain reconstruction and out-of-domain generalization.

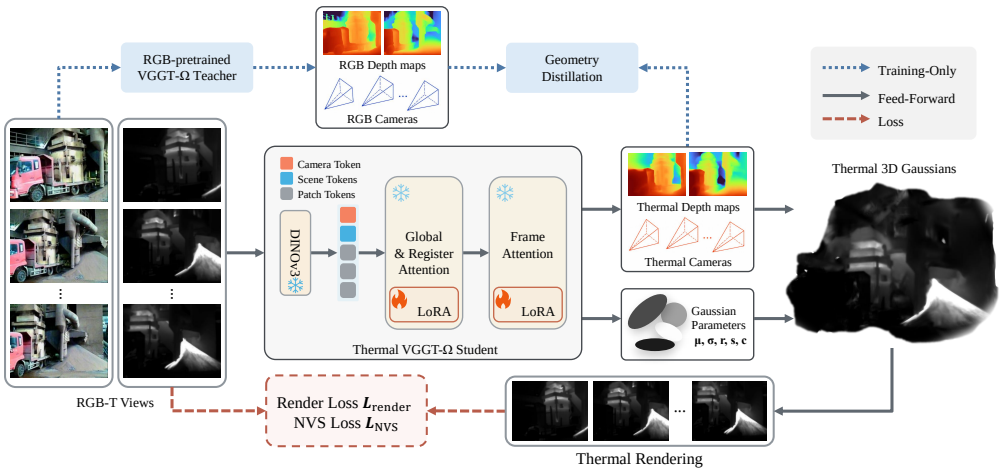


Figure 2: **Overview of AnyThermalSplat.** A thermal student adapts an RGB-pretrained geometric backbone with LoRA and predicts cameras, depth, and a renderable Gaussian scene from the context images. During training, a frozen teacher processes paired RGB images and provides camera and depth pseudo-supervision. At inference, reconstruction uses only thermal context images.

3 METHOD

Figure 2 summarizes the architecture of AnyThermalSplat, including thermal scene prediction, training-time RGB guidance, and held-out novel-view supervision.

Problem formulation. Given uncalibrated thermal context images $\mathcal{I}_C^T$, AnyThermalSplat predicts cameras, depth, and a renderable Gaussian scene :

$$\mathcal{G}_C^T = f_{\mathrm{GS}}(\mathcal{I}_C^T) = \{(\boldsymbol{\mu}_j, \sigma_j, \mathbf{r}_j, \mathbf{s}_j, \mathbf{c}_j)\}_{j=1}^N . \tag{1}$$

The attributes are position, opacity, rotation, scale, and thermal spherical harmonic coefficients. At camera $\widehat{C}_q$, differentiable rendering produces $\widehat{I}_q^T = \mathcal{R}(\mathcal{G}_C^T, \widehat{C}_q)$. The predicted scene depends only on the thermal context images.

3.1 THERMAL GEOMETRY ADAPTATION

To transfer RGB-pretrained geometric priors to thermal input, we combine a VGGT-Ω backbone (Wang et al., 2026a) with an AnySplat-initialized Gaussian decoder (Jiang et al., 2025b). Pretrained geometric weights remain frozen, while rank-8 LoRA (Hu et al., 2022) adapts attention and feed-forward network (FFN) projections. Four residual bridges accommodate the adapted features at the decoder input levels:

$$\widetilde{F}_\ell^T = F_\ell^T + g_\ell^{\uparrow}\Big(\mathrm{GELU}\Big(g_\ell^{\downarrow}\big(\mathrm{LN}(F_\ell^T)\big)\Big)\Big) . \tag{2}$$

The learned projections $g_\ell^{\downarrow}$ and $g_\ell^{\uparrow}$ reduce and restore feature dimensions. Zero initialization of the final projection preserves the initial input features. We unproject the predicted depths using the estimated cameras to obtain 3D Gaussian candidates. A trainable head predicts their attributes and fusion weights, and confidence-weighted voxel fusion merges redundant primitives.

3.2 TRAINING-TIME GEOMETRIC GUIDANCE

To guide thermal reconstruction with geometric priors pretrained with RGB, we use a frozen VGGT-Ω teacher (Wang et al., 2026a) that processes paired RGB context images to predict camera encodings $\mathbf{p}_C^R$ and depth D_C^R. Let $\mathbf{p}_C^T, D_C^T$ be the student predictions and $\widehat{D}_C^{\mathrm{rend}}$ the Gaussian-rendered depth. Following AnySplat, we combine camera distillation and depth supervision:

$$\mathcal{L}_{\mathrm{geo}} = 10\,\mathcal{L}_{\mathrm{cam}} + 0.1\ \mathrm{MSE}_M\big(\widehat{D}_C^{\mathrm{rend}}, D_C^R\big) + 0.1\ \mathrm{MSE}_M\big(\widehat{D}_C^{\mathrm{rend}}, D_C^T\big). \tag{3}$$

Here, $\mathcal{L}_{\mathrm{cam}}$ distills camera encodings, and MSE_M averages depth error over valid pixels selected by teacher confidence. The depth terms transfer teacher geometry and enforce rendered–predicted depth consistency, respectively. Teacher outputs are detached.

3.3 NOVEL-VIEW SUPERVISION AND INFERENCE

For novel-view supervision, each training episode reserves a set of thermal target images $\mathcal{I}_Q^T$ that is disjoint from the context set $\mathcal{I}_C^T$. A separate joint camera pass estimates $\widehat{\mathcal{C}}_{C\cup Q} = \mathrm{sg}[f_{\mathrm{cam}}(\mathcal{I}_C^T \cup \mathcal{I}_Q^T)]$, where sg denotes stop-gradient. No target features are passed to the Gaussian decoder. The context-only and joint camera passes use the same first context image as reference. A detached scale alignment then maps the target camera translations to the coordinate scale of the context scene (Appendix A.3). Consequently, gradients from target-view rendering update only the context-to-Gaussian reconstruction path.

With $\mathcal{L}_{\mathrm{render}} = \mathcal{L}_{\mathrm{MSE}} + 0.05\,\mathcal{L}_{\mathrm{LPIPS}}$, the objective over B scenes is

$$\mathcal{L} = \mathcal{L}_{\mathrm{geo}} + \frac{1}{B}\sum_{b=1}^{B}\Big[w_b\,\mathcal{L}_{\mathrm{render}}\big(\widehat{\mathcal{I}}_{C,b}^T, \mathcal{I}_{C,b}^T\big) + \lambda_{\mathrm{NVS}}(t)\,\mathcal{L}_{\mathrm{render}}\big(\widehat{\mathcal{I}}_{Q,b}^T, \mathcal{I}_{Q,b}^T\big)\Big] . \tag{4}$$

The source weight w_b applies only to context reconstruction; $\lambda_{\mathrm{NVS}}(t)$ enables held-out novel-view supervision after warm-up. Appendix A details the model, losses, and training settings. Inference constructs the scene in one thermal-only forward pass. Rendering at a specified camera needs no target image; pose-free evaluation uses the separate camera pass without supplied calibration or pose refinement.

4 Real and Synthetic Multi-View Thermal Data

We combine real RGB–thermal observations with geometry-grounded pseudo-thermal scenes to build diverse multi-view supervision for cross-scene learning and controlled data scaling.

4.1 ThermalVerse and Real Observations

ThermalVerse comprises 209 self-collected multi-view scenes with 26,071 RGB–thermal image pairs across 30 content categories, with 67–291 RGB–thermal pairs per scene. The captures span diverse scene types and spatial scales, including low-light environments and localized thermal contrast. Figure 3 presents representative paired views and the distribution of categories across the collection. We use a FLIR E6 Pro and retain its coarse cross-modal registration without additional calibration-board-based pixel alignment.

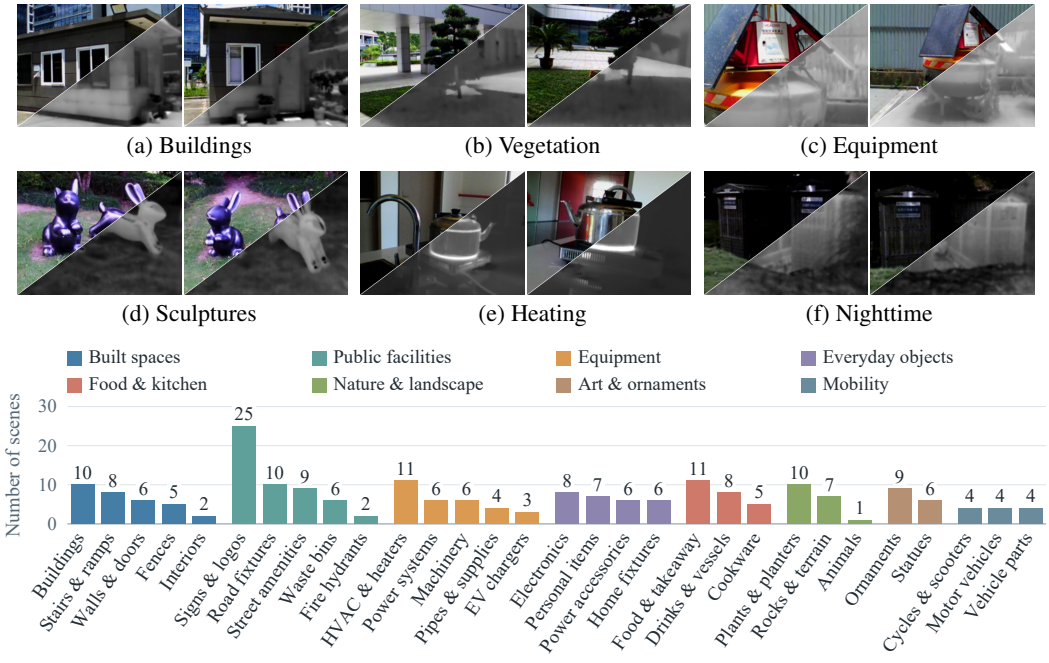


Figure 3: **Scene diversity in ThermalVerse.** Top: two viewpoints per scene, with RGB and thermal observations shown in the upper-left and lower-right portions, respectively. Bottom: distribution of the complete 209-scene collection across 30 primary content categories, colored by eight broader groups.

After curation and exclusion of evaluation scenes, we use 194 ThermalVerse scenes with 24,200 image pairs for training, supplemented by 46 public real-world scenes (Özer et al., 2024; Lin et al., 2024; Hassan et al., 2025; Lu et al., 2025; Wang et al., 2026b; Skorokhodov et al., 2026). Evaluation scenes are excluded from both training and cluster fitting. Appendix B provides source-level statistics and split details.

4.2 Geometry-Grounded Synthetic Thermalization

We construct pseudo-thermal observations from CO3D, MVImgNet, DL3DV, RealEstate10K, MipNeRF 360, and ACID (Reizenstein et al., 2021; Yu et al., 2023; Ling et al., 2024; Zhou et al., 2018; Barron et al., 2022; Liu et al., 2021). Independently translating RGB views into thermal images does not enforce cross-view consistency and can assign inconsistent appearances to the same 3D surface across viewpoints. To mitigate this issue, we reconstruct shared RGB Gaussian geometry with 3DGS-MCMC (Kheradmand et al., 2024) and use F-ViTA translations (Paranjape et al., 2026) as appearance targets. With geometry fixed, we fit a shared thermal appearance field and re-render it from the source cameras to obtain multi-view pseudo-thermal observations. This offline construction couples thermal appearance across views through a common 3D representation, while retaining

possible reconstruction and translation errors. The resulting observations represent pseudo-thermal appearance rather than calibrated temperature. Additional multi-view examples of the translation and shared-field rendering stages are shown in Appendix B.

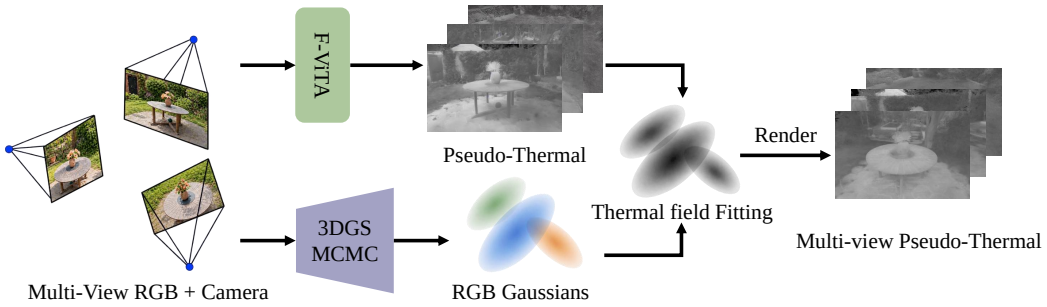


Figure 4: **Geometry-grounded synthetic thermalization.** F-ViTA targets supervise a shared thermal appearance field on fixed RGB geometry reconstructed with 3DGS-MCMC. Re-rendering produces multi-view pseudo-thermal training observations.

4.3 Training-Pool Curation and Controlled Scaling

Combining real and synthetic observations yields a curated training pool of 500 scenes (240 real and 260 synthetic) with 75,859 image pairs. We retain source identities and explicit frame pairs while filtering unusable observations. Frozen DINOv3 ViT-L/16 descriptors (Siméoni et al., 2025), extracted from up to 16 uniformly selected paired RGB frames per scene, are averaged and normalized. We fit two feature clusterings, distinct from the descriptive categories in Figure 3: $K = 8$ for stratification and sampling, and $K = 24$ for diagnostics.

Nested 100–500-scene subsets are stratified by source category and coarse cluster to approximately preserve the 48%/52% real/synthetic composition across scales. Training samples real and synthetic sources equally, with cluster-aware sampling used to maintain coverage across the training pool. Descriptors, cluster assignments, and full-pool reference statistics are shared across scales. Appendix D provides the complete sampling procedure and implementation details.

5 Experiments

5.1 Experimental Setup

Training setup. We train the main AnyThermalSplat model on the 500-scene pool for 7,000 optimization steps using rank-8 LoRA in both attention and FFN projections. The pretrained DINOv3 encoder, geometry backbone, camera head, and depth head remain frozen, while the LoRA adapters, residual bridges, and Gaussian prediction modules are optimized. Training uses AdamW on four NVIDIA A800 GPUs with BF16 mixed precision and a base learning rate of 1×10^{-4}, followed by cosine decay after a 500-step linear warm-up. Each training episode samples 2–16 thermal context views and four disjoint target views at 448×448 resolution. Held-out novel-view supervision is enabled after the first 200 steps. For the controlled scaling study and component ablations, we use the attention-only rank-8 configuration with the same 7,000 optimization steps update budget. Additional optimization, sampling, and parameter-group details are provided in Appendix A.

Evaluation protocol. We evaluate on ten held-out in-domain scenes and twenty out-of-domain TI-NSD (Chen et al., 2024a) scenes, using eight fixed target views and nested context sets of 6/12 (sparse) and 24/32 (dense) images per scene. Target images are used only by the camera branch and never enter Gaussian construction. We report PSNR, SSIM, and VGG-based LPIPS (Zhang et al., 2018) for novel-view synthesis. Additional input processing and method-specific camera protocols are provided in Appendix C.

Baselines. Sparse-view comparisons include zero-shot AnySplat, DepthSplat, and YoNoSplat (Jiang et al., 2025b; Xu et al., 2025; Ye et al., 2026), plus AnySplat-IR, our thermal counterpart

retaining the original AnySplat backbone under controlled training and evaluation settings (Appendix C.2). Dense-view comparisons additionally include 3DGS (Kerbl et al., 2023) and Thermal-3DGS (Chen et al., 2024a), each optimized per scene for 7,000 iterations. Camera conditions are summarized in Table 1.

5.2 Novel-View Synthesis

In-domain results. Table 1 shows that AnyThermalSplat achieves the best in-domain performance among all evaluated methods across every reported metric and context size. Compared with the matched AnySplat-IR baseline, the largest PSNR gains occur under sparse inputs, reaching 2.84 and 1.77 dB with 6 and 12 context views, respectively. The advantage remains at 24 and 32 views, where AnyThermalSplat also outperforms the per-scene optimized baselines under their reported evaluation conditions.

Out-of-domain results. On TI-NSD, AnyThermalSplat matches the AnySplat-IR baseline in PSNR while consistently improving SSIM and LPIPS across all context sizes. PSNR is slightly higher at 12, 24, and 32 views and only 0.07 dB lower at 6 views. These results show that AnyThermalSplat maintains competitive out-of-domain reconstruction quality while improving structural and perceptual fidelity.

Table 1: **Thermal novel-view synthesis under different input settings.** AnySplat-IR is our matched thermally adapted baseline. p/k indicate supplied reference poses/intrinsics; Opt denotes target-pose refinement for YoNoSplat or per-scene fitting for 3DGS variants. Gray rules separate the baselines from AnySplat-IR and ours. Bold and underlined values mark the best and second-best results within each setting.

				In-domain						Out-of-domain					
(a) Sparse views				6 views			12 views			6 views			12 views		
Method	p	k	Opt	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AnySplat	–	–	–	17.10	0.8644	0.4342	17.34	0.8540	0.4451	14.46	0.6077	0.5281	17.98	0.6575	0.4716
DepthSplat	✓	✓	–	22.85	0.9301	0.3663	23.34	0.9374	0.3593	17.95	**0.7164**	0.5037	18.38	0.7339	0.4770
YoNoSplat	✓	✓	✓	24.40	0.9395	0.3472	24.35	0.9399	0.3500	17.84	0.6763	0.5316	18.23	0.6797	0.5198
AnySplat-IR	–	–	–	23.92	0.9277	0.3516	26.71	0.9486	0.3030	**18.21**	0.6723	0.4928	20.56	0.7155	0.4297
Ours	–	–	–	**26.76**	**0.9542**	**0.3003**	**28.48**	**0.9620**	**0.2740**	18.14	0.6820	**0.4883**	**20.57**	**0.7391**	**0.4266**
(b) Dense views				24 views			32 views			24 views			32 views		
Method	p	k	Opt	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	–	–	✓	28.16	0.9452	0.3099	28.39	0.9506	0.3035	**23.90**	0.7773	**0.3621**	**24.83**	0.7959	**0.3501**
Thermal3D-GS	–	–	✓	27.29	0.9531	0.3068	27.83	0.9588	0.2974	21.80	**0.7805**	0.4041	22.76	**0.7979**	0.3975
AnySplat-IR	–	–	–	29.06	0.9588	0.2767	29.32	0.9621	0.2697	21.72	0.7456	0.4067	22.01	0.7587	0.3989
Ours	–	–	–	**29.33**	**0.9654**	**0.2611**	**29.93**	**0.9664**	**0.2580**	21.90	0.7787	0.3952	22.05	0.7865	0.3915

Qualitative results. Figure 5 shows that AnyThermalSplat better preserves object boundaries, structural edges, and thermal contrast in several scenes. Additional scene comparisons and failure cases are provided in Appendix I.

5.3 Scaling the Training Data

We independently train the attention-only reference on nested subsets ranging from 100 to 500 scenes, using the same initialization, 7,000-step update budget, and view-sampling policy at every scale. Figure 6 reports PSNR averaged over the four context-view settings within each domain. Complete metrics are provided in Appendix D.

Scaling the training set from 100 to 300 scenes yields clear gains in both domains: in-domain PSNR increases from 26.63 to 28.01 dB, while out-of-domain PSNR on Out-of-domain improves from 19.77 to 20.67 dB. Beyond this regime, out-of-domain performance largely saturates, reaching 20.57 dB at 500 scenes, whereas in-domain reconstruction retains an overall benefit from larger training sets despite a dip at 400 scenes. These results demonstrate the value of increasing thermal

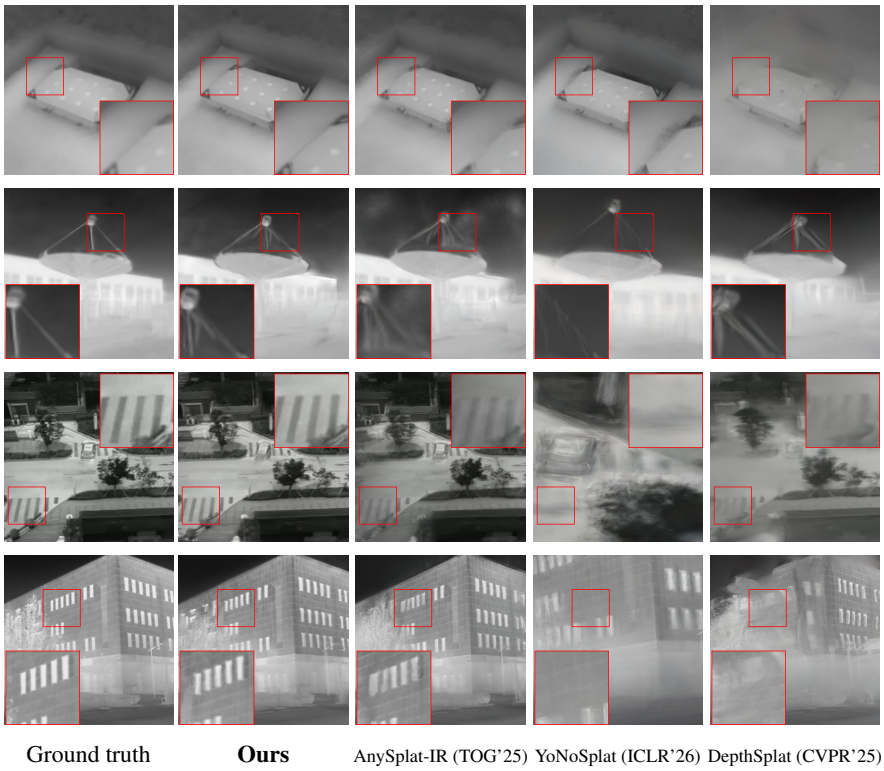


Figure 5: **Qualitative comparison of thermal novel-view synthesis.** Each row shows a scene, with ground truth and predictions from the compared methods. Enlarged insets highlight local differences in boundaries, thermal contrast, and rendering artifacts. AnySplat-IR denotes our thermally adapted AnySplat baseline.

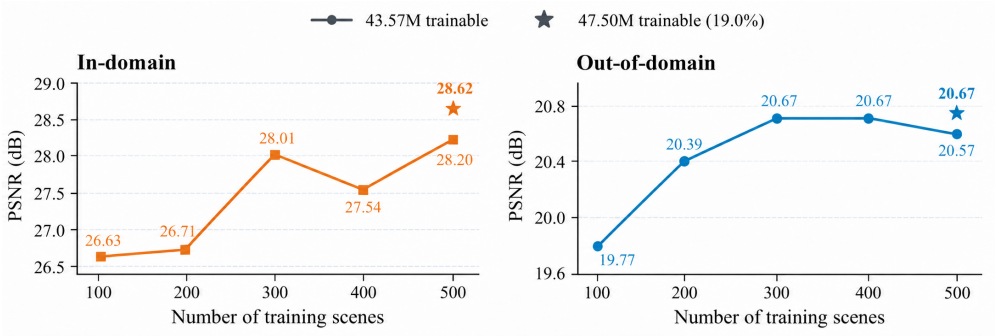


Figure 6: **Data scaling under a fixed training budget.** Curves show attention-only rank-8 models with 43.57M trainable parameters. Stars denote the 500-scene attention-plus-FFN model with 47.50M trainable parameters (+9.0%).

supervision under a fixed update budget, while also revealing saturation in out-of-domain generalization at larger scales.

5.4 Geometry and Component Analysis

Camera pose and efficiency. We evaluate camera poses on out-of-domain using ten-image groups over five fixed repeats and COLMAP reference poses (Appendix G). The attention-only reference improves AUC@5 and AUC@10 over VGGT-Ω by 7.61 and 8.05 percentage points, respectively (Table 2). We also compare reconstruction and rendering efficiency on one A800 (Table 3). Our main model reconstructs a scene in 1.15 s without per-scene optimization; 3DGS renders faster per view. Thermal-3DGS timing includes implementation-specific numerical checks (Appendix H).

Table 2: **Camera-pose estimation on out-of-domain.** Joint rotation/translation AUC (Edstedt et al., 2024); Ours is the attention-only reference.

Method	AUC ↑			
	5°	10°	20°	30°
VGGT	18.92	34.01	48.71	56.50
AnySplat	14.85	32.22	48.86	56.88
VGGT-Ω	36.76	53.13	64.40	69.44
Ours	**44.37**	**61.18**	**71.91**	**76.55**

Table 3: **Efficiency on one A800.** Mean over 30 scenes (24 context views, 448×448); reconstruction excludes target-camera estimation. Both 3DGS baselines use per-scene fitting.

Method	Recon. (s)↓	Render (ms/view)↓	Gaussians (M)
3DGS	83.71	**3.34**	**0.443**
Thermal3D-GS	757.46	48.99	0.489
Ours	**1.15**	5.52	2.267

Ablation study. Table 4 evaluates the contributions of architectural components and synthetic thermal data construction. Removing LoRA leads to the largest in-domain degradation, while removing novel-view supervision has the strongest impact on out-of-domain, highlighting the importance of thermal adaptation and held-out view consistency. Replacing the RGB teacher with thermal inputs further degrades performance, supporting the use of RGB geometric guidance. Replacing geometry-grounded thermalization with independent per-view translation also degrades performance, supporting the value of cross-view-consistent synthetic thermal supervision. Further component ablations and LoRA capacity analyses are provided in Appendices F and E.

Table 4: **Component ablations of the attention-only reference.** Scores are averaged over the 6/12/24/32-view settings within each benchmark. Bold and underlined values indicate the best and second-best results.

Configuration	In-domain			Out-of-domain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o LoRA	24.318	0.9387	0.3289	19.596	0.7295	0.4519
w/o NVS loss	25.564	0.9424	0.3133	19.443	0.7251	0.4386
Thermal teacher	25.575	0.9465	0.3185	<u>20.137</u>	0.7374	<u>0.4382</u>
Per-view translation	<u>27.088</u>	<u>0.9553</u>	<u>0.2929</u>	20.070	<u>0.7406</u>	0.4418
Full reference	**28.196**	**0.9593**	**0.2809**	**20.566**	**0.7488**	**0.4245**

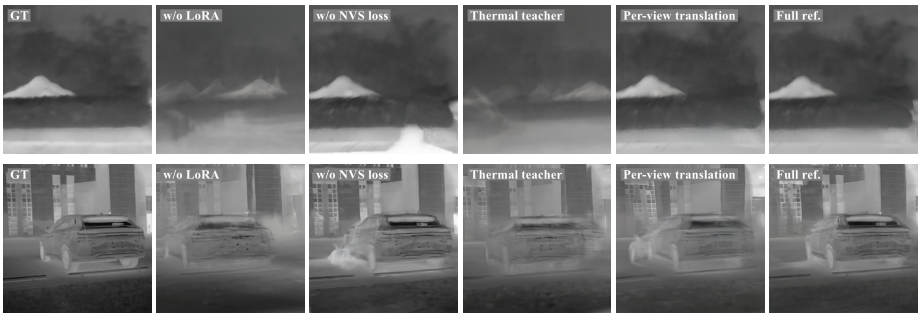


Figure 7: **Qualitative ablations with 12 context views.** Top: pond-stones (in-domain); bottom: car (out-of-domain). Each row uses the same target view, and the variants follow Table 4.

6 Conclusion and Limitations

AnyThermalSplat enables pose-free feed-forward thermal Gaussian reconstruction through parameter-efficient geometric adaptation, RGB-teacher guidance, and held-out novel-view supervision. Together with ThermalVerse and geometry-grounded thermalization, our 500-scene training pool enables controlled data scaling, showing clear in-domain and out-of-domain gains from 100 to 300 scenes, followed by saturation in out-of-domain performance at larger scales.

Limitations. Coarse RGB–thermal registration and pseudo-thermal synthesis may introduce supervision errors, with synthetic observations modeling thermal appearance rather than calibrated temperature. Novel-view evaluation also depends on target-camera estimation, and fixed-budget scaling does not establish a universal scaling law.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and language editing, code debugging, preliminary figure design, and table formatting. They also assisted with interpreting experimental results provided by the authors and drafting the corresponding discussion. These tools were not used to generate experimental measurements or select the results to report. The authors reviewed all AI-assisted text, code, and visual drafts and checked the experimental claims against the corresponding configurations, logs, and evaluation outputs. The authors take full responsibility for the final submission, including all AI-assisted content.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.12077.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024.

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*, pp. 253–269, 2024a. doi: 10.1007/978-3-031-73383-3_15.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386, 2024b. doi: 10.1007/978-3-031-72664-4_21.

Zequn Chen, Jiezhi Yang, and Heng Yang. PreF3R: Pose-free feed-forward 3d gaussian splatting from variable-length image sequence. *arXiv preprint arXiv:2411.16877*, 2024c. URL https://arxiv.org/abs/2411.16877.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

Ryan Griffiths and Donald G. Dansereau. RoRE: Rotary ray embedding for generalised multimodal scene understanding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BR2ItBcqOo.

Mariam Hassan, Florent Forest, Olga Fink, and Malcolm Mielle. ThermoNeRF: A multimodal neural radiance field for joint RGB-thermal novel view synthesis of building facades. *Advanced Engineering Informatics*, 65:103345, 2025. doi: 10.1016/j.aei.2025.103345.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. PF3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23662–23681. PMLR, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, Georgios Pavlakos, and Hao Tan. MegaSynth: Scaling up 3d scene reconstruction with synthesized data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16441–16452, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_MegaSynth_Scaling_Up_3D_Scene_Reconstruction_with_Synthesized_Data_CVPR_2025_paper.html.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3D gaussian splatting from unconstrained views. *ACM Transactions on Graphics*, 44(6), 2025b. doi: 10.1145/3763326. URL https://arxiv.org/abs/2505.23716.

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.

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3D gaussian splatting as Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.09591.

Minseong Kweon, Wenyuan Zhao, Nuo Chen, Lulin Liu, Huiwen Han, Zihao Zhu, Srinivas Shakkottai, Chao Tian, and Zhiwen Fan. DarkVGGT: Seeing through darkness using thermal geometry without daylight tax. *arXiv preprint arXiv:2606.11326*, 2026. URL https://arxiv.org/abs/2606.11326.

Yihui Li, Chengxin Lv, Zichen Tang, Hongyu Yang, and Di Huang. TokenSplat: Token-aligned 3d gaussian splatting for feed-forward pose-free reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40886–40895, 2026a.

Yuke Li, Weihang Liu, Cheng Zhang, Yuefeng Zhang, Jiadi Cui, Zixuan Wang, Junran Ding, Haoyu Wu, Yujiao Shi, Jingyi Yu, and Xin Lou. ForeSplat: Optimization-aware foresight for feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2605.22020*, 2026b. URL https://arxiv.org/abs/2605.22020.

Yvette Y. Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. ThermalNeRF: Thermal radiance fields. *arXiv preprint arXiv:2407.15337*, 2024. URL https://arxiv.org/abs/2407.15337.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.16256.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://infinite-nature.github.io/.

Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. ThermalGaussian: Thermal 3d gaussian splatting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/03bdba50e3741ac5e3eaa0e55423587e-Abstract-Conference.html.

Mert Özer, Maximilian Weiherer, Martin Hundhausen, and Bernhard Egger. Exploring multimodal neural scene representations with applications on thermal imaging. *arXiv preprint arXiv:2403.11865*, 2024. URL https://arxiv.org/abs/2403.11865.

Jay Nitin Paranjape, Celso M. De Melo, and Vishal M. Patel. F-ViTA: Foundation model guided visible-to-infrared translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5633–5642, 2026.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2109.00512.

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. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Vsevolod Skorokhodov, Chenghao Xu, Shuo Sun, Olga Fink, and Malcolm Mielle. SEAR: Simple and efficient adaptation of visual geometric transformers for unpaired RGB+Thermal 3D reconstruction. *arXiv preprint arXiv:2603.18774*, 2026. URL https://arxiv.org/abs/2603.18774.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10208–10217, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21499, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_VGGT-ohm_CVPR_2026_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

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, 2026b. doi: 10.1109/WACV61042.2026.00429.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting gaussian splatting and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16453–16463, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_DepthSplat_Connecting_Gaussian_Splatting_and_Depth_CVPR_2025_paper.html.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/857b34d81f0a8bfe3e18879dee3b5086-Abstract-Conference.html.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. YoNoSplat: You only need one model for feedforward 3D gaussian splatting. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.07321.

Tianxiang Ye, Qi Wu, Junyuan Deng, Guoqing Liu, Liu Liu, Songpengcheng Xia, Liang Pang, Wenxian Yu, and Ling Pei. Thermal-NeRF: Neural radiance fields from an infrared camera. *arXiv preprint arXiv:2403.10340*, 2024. URL https://arxiv.org/abs/2403.10340.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. MVImgNet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.06042.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19, 2024. doi: 10.1007/978-3-031-72670-5_1.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. URL https://tinghuiz.github.io/projects/mpi/.

A Implementation Details

A.1 Model variants and trainable modules

The main model uses rank-8 LoRA in both attention and feed-forward network (FFN) projections. The controlled data-scaling models and component-ablation reference instead use attention-only rank-8 LoRA. These configurations are trained independently; the attention+FFN model at 500 scenes is therefore not the endpoint of the attention-only scaling curve.

The student is initialized from the RGB-pretrained VGGT-Ω geometric model and an AnySplat Gaussian decoder. The pretrained DINOv3 image encoder, geometry backbone, camera head, and depth head remain frozen. In the frame-attention and inter-frame/global-attention blocks, LoRA is applied to the `qkv` and output `proj` matrices. The main model additionally adapts the FFN `fc1` and `fc2` projections. For a frozen matrix W, the adapter computes

$$Wx + \frac{\alpha}{r} BAx, \tag{5}$$

with $r = 8$, $\alpha = 16$, and dropout 0.05. The attention-only rank study uses $\alpha = 2r$, keeping the scale factor α/r fixed.

Features from layers $\{4, 11, 17, 23\}$ are passed through four residual bridges with widths $2048 \to 512 \to 2048$, layer normalization, and GELU activation. The final projection of each bridge is initialized to zero. The residual bridges and the Gaussian prediction modules initialized from AnySplat remain trainable. We retain degree-four spherical harmonics, a Gaussian-adapter scale multiplier of 0.005, and confidence-weighted voxel fusion with a scene-relative voxel size of 0.002. Single-channel thermal inputs are replicated across three channels.

A.2 Optimization and view sampling

Optimization. All configurations are trained independently for 7,000 optimization steps in a single mixed-source training stage and are evaluated using the final checkpoint. Training uses four NVIDIA A800 GPUs, BF16 mixed precision, gradient checkpointing, gradient clipping at 0.5, and no gradient accumulation. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.05.

The base learning rate is 10^{-4}. Multipliers of 3.0, 1.0, and 0.3 are applied to the LoRA, residual-bridge, and Gaussian-head-body parameter groups, respectively. A 500-step linear learning-rate warm-up is followed by cosine decay. This learning-rate warm-up is independent of the 200-step delay before held-out novel-view supervision is enabled.

View sampling. Each training episode contains C thermal context views and four thermal target views that are disjoint from the context set. All training images are processed at 448×448 resolution. We sample $C \in \{2, \dots, 16\}$ according to

$$P(C = c) = \frac{c^2}{\sum_{j=2}^{16} j^2}, \qquad B_{\mathrm{GPU}}(c) = \left\lfloor \frac{48}{c+4} \right\rfloor. \tag{6}$$

This distribution favors larger context sets while reducing the number of scenes that fit in each GPU batch. Scene sampling follows the source- and cluster-aware policy described in Appendix D.

A.3 Supervision and gradient paths

A frozen, unadapted RGB teacher processes the RGB observations paired with context views from both real and synthetic scenes. For each image, pixels whose teacher confidence exceeds the 30th percentile form the valid depth mask. Camera encodings are supervised with a Huber loss of threshold 1 and weight 10. The MSE terms comparing rendered depth with teacher depth and student depth each have weight 0.1. Teacher outputs are detached; in the latter consistency term, gradients propagate through both student-side quantities. Dataset camera and depth annotations are not used as regression targets in this geometric objective.

For depth maps A and B, the confidence-masked mean squared error is

$$\mathrm{MSE}_M(A, B) = \frac{1}{|\mathcal{V}_M|} \sum_{u \in \mathcal{V}_M} \big(A(u) - B(u)\big)^2, \tag{7}$$

where $\mathcal{V}_M$ is the nonempty set of valid pixels and u indexes the context view and pixel location.

Context and target image losses combine MSE and VGG-based LPIPS with weights 1 and 0.05, respectively. Context reconstruction uses a source weight of 1 for real scenes and 0.1 for synthetic scenes. These weights are applied to the per-scene context loss before averaging over the batch; the denominator is the batch size rather than the sum of source weights. The synthetic context weight does not apply to the geometric terms or the held-out novel-view loss. The held-out novel-view weight, denoted $\lambda_{\mathrm{NVB}}(t)$ in the objective, is 0 before step 200 and 1 thereafter; the loss includes all four target views. This training configuration uses neither SSIM nor surface-normal loss.

Translation-scale transfer. Let $\mathbf{c}_i^C$ and $\mathbf{c}_i^J$ denote the context-camera centers predicted by the context-only and joint passes, respectively. The detached scale conversion is

$$s = \frac{\frac{1}{V}\sum_{i=1}^{V} \|\mathbf{c}_i^C\|_2}{\max\left(\frac{1}{V}\sum_{i=1}^{V} \|\mathbf{c}_i^J\|_2, 10^{-8}\right)}. \tag{8}$$

The scene is constructed exclusively from context features. Target images enter only a separate camera-estimation pass with the context images, and gradients through this pass are stopped. Both predicted camera systems are anchored by the same first context image. Their translation scales are reconciled using the ratio of mean context-camera-center norms (Eq. 8), while target rotations and intrinsics are retained. The scale conversion is also detached. Target reconstruction gradients therefore update the context-to-Gaussian path rather than the target-camera estimation path. The confidence mask does not make coarse RGB–thermal registration pixel-accurate.

B Data Sources and Construction

B.1 Training data composition.

Table 5 summarizes the training data. The real-data component contains 30,306 image pairs, including 24,200 from ThermalVerse. Synthetic thermalization adds 260 scenes and 45,553 image pairs from six RGB datasets. Together, these sources expand multi-view thermal supervision beyond the available real captures and form the 500-scene training pool used in our experiments.

Table 5: **Composition of the curated training pool.** The pool combines 240 real and 260 synthetic scenes, totaling 75,859 RGB–thermal image pairs. ThermalVerse contributes 194 real scenes, while public real-world datasets and thermalized RGB datasets supply the remaining training scenes.

Type	Source	Scenes	Image pairs
Real	PARID	7	1,586
Real	RGBT-Scenes family	12	1,425
Real	SEAR	1	130
Real	ThermalMix	4	287
Real	ThermalNeRF	6	624
Real	ThermoScenes	16	2,054
Real	ThermalVerse	194	24,200
	Subtotal	240	30,306
Synthetic	ACID	15	480
Synthetic	CO3D	121	20,722
Synthetic	DL3DV	52	17,465
Synthetic	MVImgNet	42	1,303
Synthetic	Mip-NeRF 360	4	996
Synthetic	RealEstate10K	26	4,587
	Subtotal	260	45,553
	Training total	**500**	**75,859**

Self-collected scene accounting. The complete ThermalVerse collection contains 209 captured scenes and 26,071 RGB–thermal pairs. Of these, 194 scenes (24,200 matched pairs) are used for

training, six scenes are reserved for evaluation, and nine scenes are excluded during quality curation. The 46 public real scenes listed in Table 5 are separate from this 209-scene self-collected collection. The in-domain evaluation consists of ten scenes in total, including six self-collected scenes and four scenes from public real-world datasets. None of the evaluation or excluded scenes are used for training or DINOv3 feature-cluster fitting. TI-NSD is likewise excluded from both training and feature-cluster fitting. Scene counts refer to canonical scene records, while frame counts refer to matched RGB–thermal observations rather than all files contained in each source directory.

B.2 Real observations and pairing

ThermalVerse was captured with a FLIR E6 Pro. We retain the coarse registration provided by the device without additional calibration-board-based pixel alignment. The captures cover built and natural environments, public infrastructure, industrial equipment, everyday objects, and heat-related devices, ranging from close-up objects to larger spatial layouts. Multiple viewpoints support cross-view reconstruction, but coarse pairing does not imply coincident optical centers or pixel-accurate correspondence.

Canonical scene records preserve source identities and explicit frame pairs. We pair images by normalized filename stems and filter unusable files, rather than assuming that directory listings have matching orders. Unpaired observations are excluded from paired geometric supervision. Thermal images retain their stored pixel-value encoding; normalization for network input does not convert these values to temperature. RGB images provide teacher inputs during training, while reconstruction at inference uses thermal images.

B.3 Geometry-grounded pseudo-thermal observations

The RGB sources include CO3D, MVImgNet, DL3DV, RealEstate10K, Mip-NeRF 360, and ACID. For each source scene, F-ViTA produces translated appearance targets for individual views, and 3DGS-MCMC reconstructs shared RGB geometry. We keep this geometry fixed, fit a shared thermal appearance field, and re-render it at the source cameras to obtain paired training observations. The translated targets supervise offline fitting, whereas the re-rendered observations form the multi-view pseudo-thermal training data.

The geometric constraint couples the views but cannot eliminate errors in RGB reconstruction or translation. Re-rendering neither calibrates the translated appearance radiometrically nor physically simulates heat transfer. Offline field fitting is part of dataset construction, not test-time optimization of AnyThermalSplat.

What is optimized offline. The RGB reconstruction supplies the scene geometry and source cameras; both remain fixed while the thermal appearance is fitted to translated view targets. The resulting field is rendered at those same cameras, yielding a shared-view training signal rather than a collection of independently translated frames. This fitting stage is separate from the 7,000-step training of AnyThermalSplat and has no test-time counterpart. The surviving 260-scene image archive verifies the translated targets, fixed-camera re-renderings, and RGB pairings. It does not retain one uniform, scene-level record of appearance-fitting iterations and loss weights, so we do not assign unverified numerical settings to this offline stage. In particular, the MSE/LPIPS weights in Appendix A.3 describe AnyThermalSplat training, not the appearance-field fitting objective.

Multi-view examples. Figures 8 and 9 compare paired source RGB images, independent F-ViTA translations, and the final shared-field renderings at two source cameras in each of four scenes. The rendered training images are stored as 16-bit grayscale values; for these figures only, we map them linearly to 8-bit display values by dividing by 257, without per-image contrast adjustment. The examples illustrate the effect of a common scene representation while also exposing its residual blur and reconstruction artifacts.

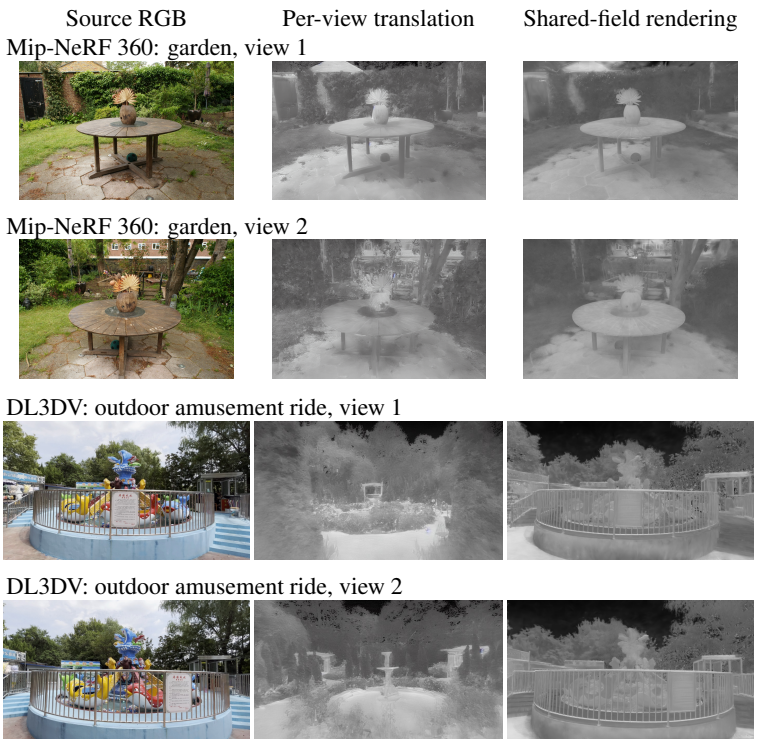


Figure 8: **Outdoor multi-view pseudo-thermal examples.** Two distinct source-camera views per scene show how independent translations are consolidated into shared-field renderings on fixed RGB geometry. These are pseudo-thermal appearances, not measured temperatures.

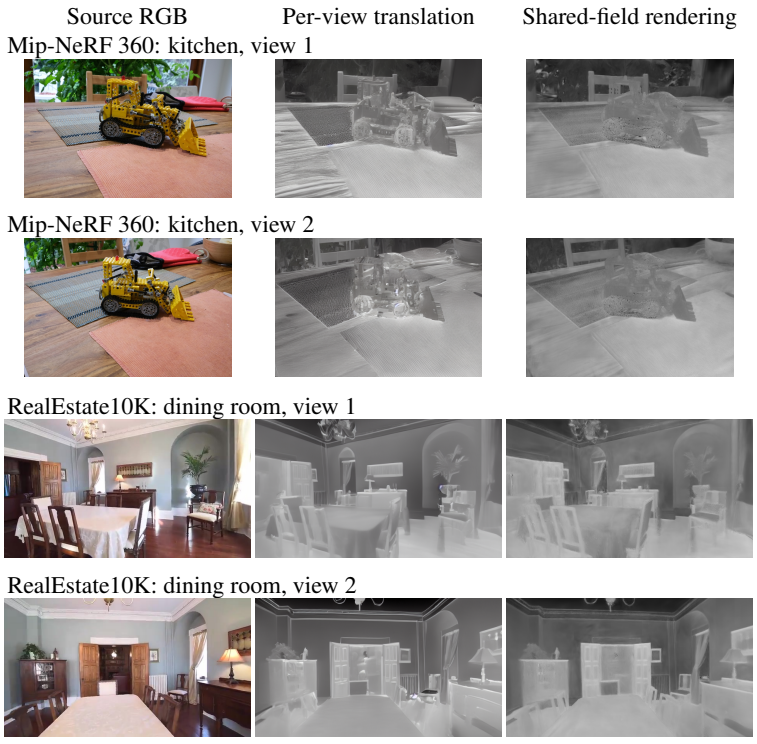


Figure 9: **Indoor multi-view pseudo-thermal examples.** View pairs expose both the improved cross-view structural coupling and remaining limitations at thin structures, reflective surfaces, and regions poorly reconstructed in RGB. Display mapping matches Figure 8.

C Novel-View Evaluation Protocols

C.1 Fixed views, preprocessing, and reduction

Novel-view synthesis is evaluated on ten in-domain scenes and twenty TI-NSD scenes. Each scene has eight fixed target views and context sets satisfying $C_6 \subset C_{12} \subset C_{24} \subset C_{32}$. Context and target frames are disjoint. Smaller context sets are ordered prefixes of the larger selection, preserving the first context image and target ordering. All methods use the same target selection.

Thermal images are converted to three-channel network inputs, and all final renderings and image-quality metrics use 448×448 resolution. The AnyThermalSplat evaluation loader resizes the shorter image dimension and applies a center crop to obtain square inputs. Model-specific preprocessing and camera conventions are retained where necessary. A common evaluation resolution therefore does not imply identical encoder input resolutions or camera access.

For a metric m, we first average over the eight target views in each scene, then average equally across scenes within each domain:

$$\overline{m}_{D,V} = \frac{1}{|D|} \sum_{s \in D} \frac{1}{8} \sum_{q=1}^{8} m_{s,V,q}, \qquad \overline{m}_{D} = \frac{1}{4} \sum_{V \in \{6,12,24,32\}} \overline{m}_{D,V}. \tag{9}$$

The second expression applies only to results explicitly averaged over context sizes, such as the scaling and ablation summaries. PSNR is averaged in dB after evaluation, rather than recomputed from pooled pixel MSE. SSIM and VGG-based LPIPS use the same scene-balanced aggregation. In-domain and TI-NSD results are reported separately. When efficiency results are pooled over all thirty scenes, the two domains receive weights in the ratio $1 : 2$, rather than equal weights.

C.2 Baseline-specific camera access

AnySplat and AnySplat-IR. AnySplat denotes the original RGB-pretrained model evaluated without thermal fine-tuning. We construct AnySplat-IR as a controlled thermal counterpart that retains the original AnySplat geometric backbone and pretrained initialization, rather than replacing them with VGGT-Ω. It follows the same thermal-training framework as AnyThermalSplat: low-rank adaptation of pretrained geometric features, frozen RGB-teacher geometric supervision, thermal context reconstruction, and held-out novel-view supervision. Table 6 distinguishes the shared training and evaluation settings from the architectural differences. Here, "matched" refers to the controlled training and evaluation protocols, not an identical backbone or a parameter-matched architecture.

Both trained models use the same 500-scene pool of 240 real and 260 shared-field synthetic scenes, four NVIDIA A800 GPUs, 7,000 optimization steps, BF16 precision, and the final checkpoint for evaluation. They share the 448×448 training resolution, four held-out targets, context-count distribution and image-budget policy in Appendix A.2, and source- and cluster-aware scene sampling in Appendix D. Both use AdamW with a base learning rate of 10^{-4}, a 500-step warm-up, and the same context, geometry, and novel-view loss weights. The synthetic context-loss weight is 0.1, while the held-out novel-view objective is enabled after the initial 200-step delay. Thus, AnySplat-IR receives both thermal training data and RGB geometric supervision; it is not the zero-shot AnySplat row.

The native baseline uses VGGT with DINOv2 and patch size 14. Its rank-8 LoRA updates the `qkv` and output `proj` matrices in frame and global attention blocks, and its frozen RGB teacher is initialized from the same original AnySplat checkpoint. It retains the native Gaussian prediction head and does not add the residual bridges used to connect VGGT-Ω features in AnyThermalSplat. The main model instead uses DINOv3 with patch size 16, adds FFN adapters, and uses four zero-initialized residual bridges. The attention-only models used for scaling and component ablations omit the FFN adapters but retain the VGGT-Ω backbone and bridges. Consequently, the comparison with the main model is not a backbone-only intervention: image encoding, teacher initialization, feature interfacing, and adapter placement also differ. It measures the complete thermal framework built on each geometric initialization under controlled data and training protocols.

At evaluation, both AnySplat-IR and AnyThermalSplat construct the scene exclusively from thermal context images. Targets enter only a separate context-plus-target camera-estimation pass, without

Table 6: **AnySplat-IR and AnyThermalSplat: controlled settings and architectural differences.** The AnyThermalSplat column describes the attention-plus-FFN main model. Spanning entries apply to both models; shared protocols do not imply identical feature interfaces, teacher weights, or adaptation capacity.

Setting	AnySplat-IR	AnyThermalSplat
Geometric backbone	Native AnySplat VGGT	VGGT-Ω
Image encoder / patch	DINOv2 / 14	DINOv3 / 16
Geometric initialization	Original AnySplat checkpoint	RGB-pretrained VGGT-Ω
LoRA placement	Attention `qkv/proj`	Attention `qkv/proj` and FFN `fc1/fc2`
LoRA hyperparameters	$r = 8$, $\alpha = 16$, dropout 0.05	
Frozen RGB teacher	Native AnySplat geometry	VGGT-Ω geometry
Feature interface	Native intermediate features	Four zero-initialized residual bridges
Gaussian prediction	Native AnySplat Gaussian head	AnySplat-initialized Gaussian head with bridged features
Thermal training pool	Same 240 real and 260 shared-field synthetic scenes	
Training objectives	Thermal context reconstruction, RGB-teacher geometry, and held-out novel-view supervision	
Training budget	7,000 steps; four A800 GPUs; BF16	
Optimizer / warm-up	AdamW; base learning rate 10^{-4}; 500-step warm-up	
Training view policy	Same context distribution, four targets, and per-GPU image budget	
Evaluation views	Same nested 6/12/24/32 contexts and eight fixed targets per scene; 448×448 renderings	
Camera protocol	Context-only scene construction; joint target-camera estimation; no reference cameras or target-pose refinement	

reference cameras, scene fitting, or target-pose refinement. Both use the norm-based translation-scale conversion in Appendix A.3, the same fixed target selections, and the scene-balanced metric reduction in Appendix C. Zero-shot AnySplat also uses the corrected norm-based scale conversion rather than signed coordinate means. Native patch conventions are retained; common rendering resolution and camera access do not imply identical internal tokenization. For memory management, the recorded AnySplat-IR run computes LPIPS in two-image chunks with activation recomputation, retaining all images and the same loss reduction rather than changing the training objective.

DepthSplat. The RGB-pretrained DepthSplat model is evaluated without thermal fine-tuning and receives reference context and target camera parameters. The evaluated adapter uses 448×448 inputs and renderings, with intrinsics adjusted for the evaluation crop. DepthSplat therefore provides a camera-assisted comparison rather than a strictly pose-free baseline.

YoNoSplat. YoNoSplat uses pretrained weights, native 224×224 processing, and 100 steps of target-pose refinement. Refinement starts from reference target cameras and keeps the predicted Gaussians fixed. The refined cameras are transferred to the 448×448 rendering crop with the corresponding intrinsic adjustment. This procedure optimizes target poses but not the scene representation. The reported scores are kept separate from those of the independently tested setting without refinement.

3DGS and Thermal-3DGS. Dense-view comparisons use 24 or 32 context views and 7,000 optimization iterations per scene. VGGT-Ω provides the initial geometry and predicted cameras, while a separate joint camera pass localizes target views. Only context images are used for scene fitting. These methods retain per-scene optimization and are therefore not feed-forward baselines.

Scope of pose-free evaluation. For AnyThermalSplat, evaluation uses neither reference cameras nor test-time scene or pose optimization. Target images nevertheless enter the detached camera-estimation branch. The reconstructed scene is independent of target features, but rendering scores reflect both reconstruction and target-camera localization. Alternatively, a known rendering camera can be supplied to render the predicted scene without a target image. The baseline comparison

reports the camera information available to each method rather than assuming identical camera access.

D Controlled Data-Scaling Details

D.1 Scene descriptors and nested subsets

We extract frozen DINOv3 ViT-L/16 descriptors from up to sixteen uniformly selected paired RGB frames per scene, average them, and apply ℓ_2 normalization. Clustering is fitted exclusively on the 500-scene training pool. We use two independent clusterings: $K = 8$ for subset stratification and scene sampling, and $K = 24$ for diagnostics. The finer clustering is not treated as a hierarchical subdivision of the coarser clustering. Both are feature-based clusters, not the manually assigned content categories shown in Figure 3.

The nested subsets contain 100, 200, 300, 400, and 500 scenes, with real/synthetic counts of $(48, 52)$, $(96, 104)$, $(144, 156)$, $(192, 208)$, and $(240, 260)$, respectively. Subsets are stratified by source category and coarse feature cluster.

D.2 Scene sampling and fixed-update interpretation

Real and synthetic sources each have sampling probability $1/2$ during training, although their proportions in the training pool are 48%/52%. Let $N_{a,k}$ denote the full-pool count for source category a and cluster k, and let $\mathcal{K}_a(T)$ contain the clusters represented in subset T. The cluster distribution is

$$P(k \mid a, T) = \frac{\sqrt{N_{a,k}}}{\sum_{j \in \mathcal{K}_a(T)} \sqrt{N_{a,j}}}, \quad k \in \mathcal{K}_a(T). \tag{10}$$

Clusters absent from a subset receive zero probability. Within the selected available cluster, scenes are sampled uniformly. Descriptors, cluster labels, and full-pool reference counts are shared across scales. However, renormalization over available clusters can change the realized distribution when smaller subsets omit clusters.

All scaling models use attention-only rank-8 LoRA and are trained independently from the same pretrained initialization for 7,000 steps. We control the architecture, update budget, view-count distribution, and source and cluster sampling policy. This comparison matches update counts rather than convergence or per-scene exposure. A larger pool generally reduces repeated exposure to individual scenes. Equal update counts do not guarantee equal wall-clock time or identical numbers of images successfully processed during training.

Table 7: **Effect of training-set size on thermal reconstruction.** All configurations use attention-only rank-8 LoRA and are trained for 7,000 updates. Expanding the training pool from 100 to 500 scenes improves all three metrics on both evaluation sets, although intermediate changes are non-monotonic. Results are averaged over the 6-, 12-, 24-, and 32-view settings. Bold and underlined values indicate the best and second-best results, including ties at the displayed precision.

Scenes	100	200	300	400	500
In-domain					
PSNR↑	26.63	26.71	<u>28.01</u>	27.54	**28.20**
SSIM↑	0.9540	0.9504	**0.9600**	0.9571	<u>0.9593</u>
LPIPS↓	0.2974	0.2977	**0.2797**	0.2856	<u>0.2809</u>
Out-of-domain					
PSNR↑	19.77	20.39	**20.67**	**20.67**	<u>20.57</u>
SSIM↑	0.7118	0.7394	**0.7488**	<u>0.7470</u>	**0.7488**
LPIPS↓	0.4478	0.4312	<u>0.4252</u>	**0.4245**	**0.4245**

In-domain reconstruction improves overall as the training pool grows from 100 to 500 scenes, with non-monotonic changes at intermediate scales. Cross-dataset gains are smaller and also non-monotonic at larger scales. These results do not establish an intrinsic saturation point or a universal scaling law. Stars in the main scaling figure denote the attention-plus-FFN model trained on 500

scenes, with 47.50M rather than 43.57M trainable parameters. They represent a separate capacity configuration rather than points on the fixed-architecture scaling curve.

E LoRA Rank and Adaptation Placement

The capacity study compares attention-only ranks $r \in \{4, 8, 16, 32, 64\}$ with attention-plus-FFN rank 8 on the 500-scene pool. Each configuration is trained independently for 7,000 steps with $\alpha = 2r$. Tables 8 and 9 report parameter counts and PSNR for each context size, respectively. Table 10 reports all three metrics averaged over the four context sizes.

Table 8: **Adaptation capacity.** Parameter counts are reported in millions (M). Total trainable counts include adapters, residual bridges, and Gaussian prediction modules.

Configuration	LoRA (M)	Trainable (M)
Attention, $r = 4$	1.180	42.390
Attention, $r = 8$	2.359	43.569
Attention, $r = 16$	4.719	45.929
Attention, $r = 32$	9.437	50.647
Attention, $r = 64$	18.874	60.085
Attention+FFN, $r = 8$	6.291	47.502

Table 9: **Capacity comparison across context sizes.** Each cell reports in-domain / TI-NSD PSNR in dB. Bold and underlined values indicate the best and second-best results for each domain and context size, respectively.

Configuration	6 views	12 views	24 views	32 views
Attention, $r = 4$	26.09 / 18.27	27.88 / <u>20.51</u>	29.25 / 21.55	29.63 / 21.73
Attention, $r = 8$	25.57 / <u>18.60</u>	<u>28.02</u> / 20.37	<u>29.40</u> / 21.54	<u>29.79</u> / 21.76
Attention, $r = 16$	26.12 / **18.66**	28.01 / 20.41	28.78 / <u>21.63</u>	29.22 / <u>21.97</u>
Attention, $r = 32$	<u>26.19</u> / 18.40	27.51 / 20.38	29.15 / 21.47	29.26 / 21.66
Attention, $r = 64$	26.04 / 18.06	27.99 / 20.25	**29.43** / 21.47	29.50 / 21.72
Attention+FFN, $r = 8$	**26.76** / 18.14	**28.48 / 20.57**	29.33 / **21.90**	**29.93 / 22.05**

Table 10: **Capacity comparison averaged over context sizes.** Metrics are averaged over targets and scenes, then equally over the 6/12/24/32-view settings. Bold and underlined values indicate the best and second-best results in each column, respectively, including ties at the displayed precision.

	In-domain			Out-of-domain		
Configuration	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Attention, $r = 4$	28.21	0.9587	0.2824	20.51	<u>0.7472</u>	0.4303
Attention, $r = 8$	28.20	<u>0.9593</u>	0.2809	<u>20.57</u>	**0.7488**	**0.4245**
Attention, $r = 16$	28.03	0.9589	<u>0.2781</u>	**20.67**	0.7436	0.4255
Attention, $r = 32$	28.03	0.9587	0.2789	20.48	0.7434	0.4278
Attention, $r = 64$	<u>28.24</u>	0.9585	0.2804	20.38	0.7300	0.4351
Attention+FFN, $r = 8$	**28.62**	**0.9620**	**0.2734**	**20.67**	0.7466	<u>0.4254</u>

Increasing attention-only rank does not consistently improve reconstruction quality, and the best configuration varies across domains and context sizes. Attention-plus-FFN rank 8 achieves the highest PSNR in six of the eight domain/view settings. However, attention-only rank 16 performs better on six-view TI-NSD, and rank 64 performs better on 24-view in-domain reconstruction. The joint rank-8 model uses 6.29M LoRA parameters and 47.50M total trainable parameters, compared with 18.87M and 60.08M for attention-only rank 64. Because adding FFN adapters changes both placement and parameter count, this comparison does not isolate adapter placement at a matched parameter budget.

F Component Ablation Details

Component ablations use the attention-only rank-8 reference trained on 500 scenes with the same 7,000-step budget. The no-LoRA variant disables the adapters while retaining trainable Gaussian prediction and bridge modules. The thermal-input-teacher variant keeps the teacher frozen but replaces paired RGB inputs with the thermal inputs used by the student. The variant without novel-view supervision sets the held-out rendering-loss weight to zero while retaining context reconstruction and geometric losses. Each variant is trained separately rather than applied as a test-time modification to a shared checkpoint.

Table 11: **Component ablations across context sizes.** Higher PSNR and SSIM and lower LPIPS indicate better reconstruction. All variants are trained separately and compared with the attention-only reference. Bold and underlined values indicate the best and second-best results for each domain, metric, and context size, respectively.

		In-domain			Out-of-domain		
Configuration	Views	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o LoRA	6	22.04	0.9237	0.3566	17.33	<u>0.6847</u>	0.5032
	12	23.37	0.9331	0.3328	19.53	0.7273	0.4509
	24	25.50	0.9472	0.3156	20.63	0.7492	0.4293
	32	26.36	0.9507	0.3104	20.89	0.7568	0.4241
Thermal-input teacher	6	<u>23.06</u>	<u>0.9321</u>	<u>0.3546</u>	<u>17.73</u>	0.6800	<u>0.4956</u>
	12	24.51	<u>0.9421</u>	0.3247	<u>20.07</u>	<u>0.7340</u>	<u>0.4385</u>
	24	27.06	<u>0.9549</u>	0.3017	21.20	0.7632	0.4128
	32	27.67	0.9570	0.2930	21.55	0.7722	0.4057
w/o novel-view loss	6	22.04	0.9196	0.3563	14.97	0.6389	0.5014
	12	<u>24.89</u>	0.9376	<u>0.3213</u>	19.20	0.7194	0.4403
	24	<u>27.38</u>	0.9541	<u>0.2919</u>	**21.67**	<u>0.7687</u>	<u>0.4080</u>
	32	<u>27.94</u>	<u>0.9584</u>	<u>0.2836</u>	**21.94**	<u>0.7733</u>	<u>0.4048</u>
Full attention reference	6	**25.57**	**0.9485**	**0.3131**	**18.60**	**0.7022**	**0.4733**
	12	**28.02**	**0.9594**	**0.2841**	**20.37**	**0.7381**	**0.4282**
	24	**29.40**	**0.9642**	**0.2656**	<u>21.54</u>	**0.7732**	**0.4010**
	32	**29.79**	**0.9653**	**0.2608**	<u>21.76</u>	**0.7814**	**0.3956**

The main ablation table averages these results over 6/12/24/32 views according to Eq. 9. Its full reference is the attention-only model rather than the larger attention-plus-FFN main model. Removing LoRA reduces mean in-domain PSNR by 3.88 dB, while removing held-out supervision reduces it by 2.63 dB. The frozen teacher also performs worse with thermal inputs than with RGB inputs. These comparisons support the combined use of adaptation, cross-modal guidance, and held-out-view supervision in the reference configuration; they do not isolate the effect of each component in the larger FFN-adapted model.

F.1 Per-view translation versus shared-field training data

This ablation replaces independent F-ViTA translations with shared-field re-renderings on fixed RGB geometry and source cameras. Both variants retain the same 240 real and 260 synthetic scene identities, paired RGB observations, view-sampling policy, attention-only rank-8 architecture, optimizer, and 7,000-step budget. Evaluation uses identical nested 6/12/24/32-view context sets and eight fixed targets per scene. Only the synthetic thermal supervision changes.

Table 12 shows improvements in all three average metrics on both benchmarks. Shared-field supervision increases PSNR by 1.108 dB in-domain and 0.496 dB on TI-NSD. SSIM increases by 0.0040 and 0.0082, while LPIPS decreases by 0.0120 and 0.0173, respectively. These results support geometry-grounded shared-field construction as a useful source of thermal supervision, without increasing training-scene count or model capacity.

Figure 10 compares downstream model predictions at matched held-out views, illustrating changes in reconstructed thermal structure and boundaries rather than the synthetic training images themselves.

Table 12: **Data-construction ablation.** Scene-balanced metrics are averaged equally over 6/12/24/32 context views. Both variants use matched scenes, architectures, and training budgets. Bold indicates the better result.

	In-domain			TI-NSD		
Synthetic supervision	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Per-view translation	27.088	0.9553	0.2929	20.070	0.7406	0.4418
Shared-field rendering	**28.196**	**0.9593**	**0.2809**	**20.566**	**0.7488**	**0.4245**

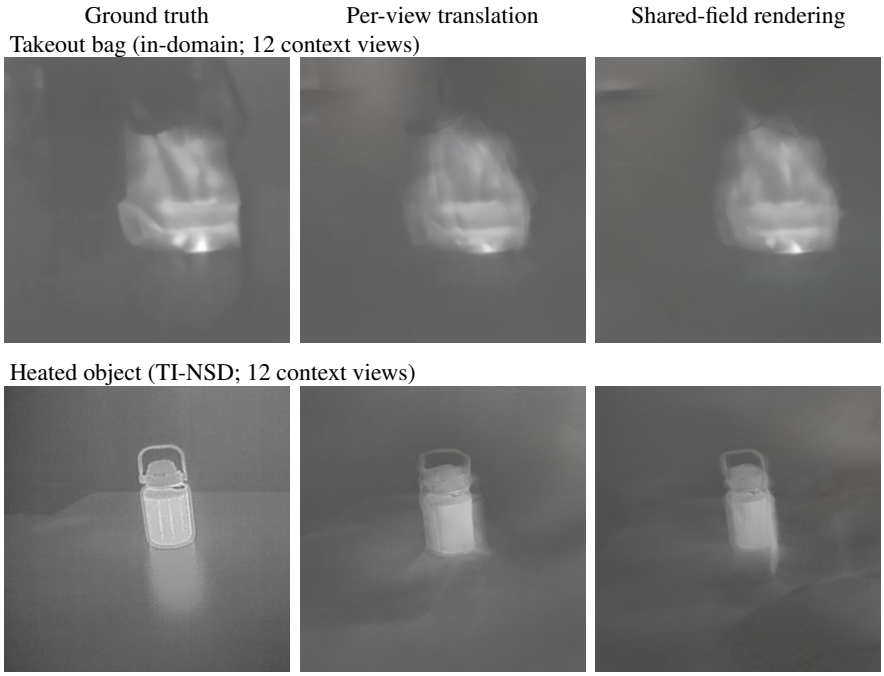


Figure 10: **Qualitative comparison of synthetic-supervision variants.** Models trained with independent translations and shared-field renderings use the same 12 context images and held-out target in each row. Panels are original evaluation outputs without in-image labels or contrast adjustment.

G Independent Camera-Pose Evaluation

G.1 Sampling and reference cameras

Pose evaluation uses a separate sampling protocol from the context/target partition for novel-view synthesis. Each of the twenty TI-NSD scenes contributes five fixed groups of ten thermal images, sampled uniformly without replacement within each group. All models receive images in the same order. Groups can overlap, producing 100 groups and 4,500 nominal unordered image pairs per model; the pairs are therefore not 4,500 independent test samples.

Inputs are resized and center-cropped to 448×448. Models receive no RGB images or reference calibration and perform no target-pose refinement. Reference poses are taken from the original TI-NSD COLMAP registrations rather than model-initialized camera files. These poses are reconstruction references, not independently tracked metric ground truth. A COLMAP world-to-camera pose $[R, t]$ is converted to the camera-to-world pose $[R^\top, -R^\top t]$.

G.2 Relative-pose error and AUC

For each unordered pair $i < j$, the relative transform is $T_{ij} = C_i^{-1} C_j$. Rotation error is the geodesic angle between predicted and reference relative rotations. Translation-direction error is defined as

$$e_t = \arccos\left(\left|\frac{\widehat{t}_{ij}^{\top} t_{ij}}{\|\widehat{t}_{ij}\|_2 \|t_{ij}\|_2}\right|\right), \qquad e = \max(e_R, e_t), \tag{11}$$

with angles measured in degrees. Taking the absolute inner product makes the translation metric insensitive to sign, consistent with the AnySplat evaluation helper; a signed diagnostic is retained separately. AUC@5/10/20/30 uses the same one-degree histogram/CDF discretization and is reported as a percentage, rather than computed by continuous trapezoidal integration.

Reference pairs with translation baselines no larger than 10^{-6} times the largest reference baseline in the group are excluded from translation and joint AUC. This reference-only mask is shared across models. Collapsed predicted baselines are retained and assigned a translation error of 90°. Scores are averaged over the five groups within each scene and then equally across the twenty scenes. The evaluation uses no fitted similarity transform, PnP, bundle adjustment, or Gaussian rendering.

Table 13: **Pose evaluation on TI-NSD.** AUC is reported as a percentage over 100 groups of ten images. Both thermal variants use rank-8 LoRA. Bold and underlined values indicate the best and second-best results at each threshold, respectively.

Method	AUC@5°	AUC@10°	AUC@20°	AUC@30°
AnySplat	14.85	32.22	48.86	56.88
VGGT	18.92	34.01	48.71	56.50
VGGT-Ω	36.76	53.13	64.40	69.44
Ours (attention-only)	**44.37**	**61.18**	**71.91**	**76.55**
Ours (attention+FFN)	<u>43.20</u>	<u>60.26</u>	<u>71.51</u>	<u>76.26</u>

The attention-plus-FFN model improves AUC at all four thresholds relative to the RGB-pretrained VGGT-Ω reference, but slightly trails the attention-only thermal model. Its better rendering performance therefore does not imply uniformly better pose accuracy. The main pose table explicitly identifies the attention-only configuration; the additional FFN row here reports the independently evaluated main model.

H Runtime, Memory, and Representation Size

H.1 Measurement scope

Efficiency is measured on a single A800 GPU using 24 context images and eight targets for each of the thirty evaluation scenes. Final renderings have resolution 448×448. Methods are measured sequentially. Checkpoint loading, image loading and saving, and quality-metric computation are excluded, so the results report stage timings rather than end-to-end application latency.

We time scene reconstruction, target-camera handling, and per-view rendering separately. Forward and rendering stages use one warm-up run followed by three timed repetitions, with CUDA synchronization at the start and end of each measurement. We take the median for each scene and then average equally across scenes. Each 7,000-step per-scene optimization is measured once rather than repeated three times. For 3DGS and Thermal-3DGS, reconstruction time includes VGGT-Ω initialization and scene optimization; target-camera handling is timed separately. For YoNoSplat, camera time includes 100 steps of target-pose refinement.

Peak memory denotes the peak allocated CUDA memory reported by the framework, rather than total device memory usage reported by a system monitor. It is aggregated from per-scene peaks across the measured stages, including initialization for the optimization-based baselines. For stages measured in separate processes, we use the largest stage peak rather than summing allocations from processes that do not run concurrently.

The measured Thermal-3DGS runtime includes numerical-validity checks in the evaluated training implementation. The timing gap is therefore implementation-specific and cannot be attributed en-

Table 14: **Stage timings and peak allocated memory.** Results are averaged over thirty scenes with 24 context views and eight targets each. Camera time includes refinement where applicable; – denotes supplied cameras.

Method	Recon. (s)	Cameras (s)	Render (ms/view)	Peak (GiB)
AnySplat	1.414	1.228	5.693	9.301
DepthSplat	0.812	–	23.209	21.131
YoNoSplat (pose opt.)	0.243	20.143	12.569	6.395
AnySplat-IR	1.247	1.399	6.123	10.836
Ours (attention+FFN)	1.150	0.931	5.522	11.333
3DGS	83.708	1.280	3.341	8.325
Thermal-3DGS	757.463	1.280	48.985	9.506

tirely to an intrinsic difference in algorithmic speed. Similarly, the absence of a target-camera stage for DepthSplat reflects the use of supplied reference cameras, rather than cost-free pose estimation. Because camera access differs across methods, comparisons within each column do not use identical camera conditions.

H.2 Gaussian counts and trade-offs

Gaussian counts refer to the final scene representation before view-dependent rasterization culling. For AnyThermalSplat, we count the encoder output after voxel fusion; for optimized methods, we count the point cloud after 7,000 iterations. No additional opacity threshold is applied. These primitive counts do not include parameters of auxiliary appearance networks.

Table 15: **Final Gaussian counts (millions).** Counts are measured after encoder fusion or per-scene optimization and before per-view culling.

Evaluation set	3DGS	Thermal-3DGS	Ours (attention+FFN)
In-domain	0.213	0.204	2.127
Out-of-domain	0.558	0.632	2.336
All 30 scenes	0.443	0.489	2.267

Under this protocol, the attention-plus-FFN model reconstructs a scene in approximately 1.15 seconds on average, but produces more Gaussians and requires more peak allocated memory than the two optimized references. Once reconstruction is complete, 3DGS renders faster than AnyThermalSplat in this experiment. The primary efficiency benefit is avoiding per-scene fitting, rather than simultaneously reducing representation size, memory usage, and rendering time.

I Qualitative Comparisons

Unlike the four-scene comparison in the main paper, the examples below use different evaluation scenes. All panels use 12 context views and the same fixed held-out target within each row; predictions are from the attention-plus-FFN main model and the matched AnySplat-IR baseline. The scene-level PSNR values average all eight targets, not just the single displayed frame. Camera conditions follow Appendix C.

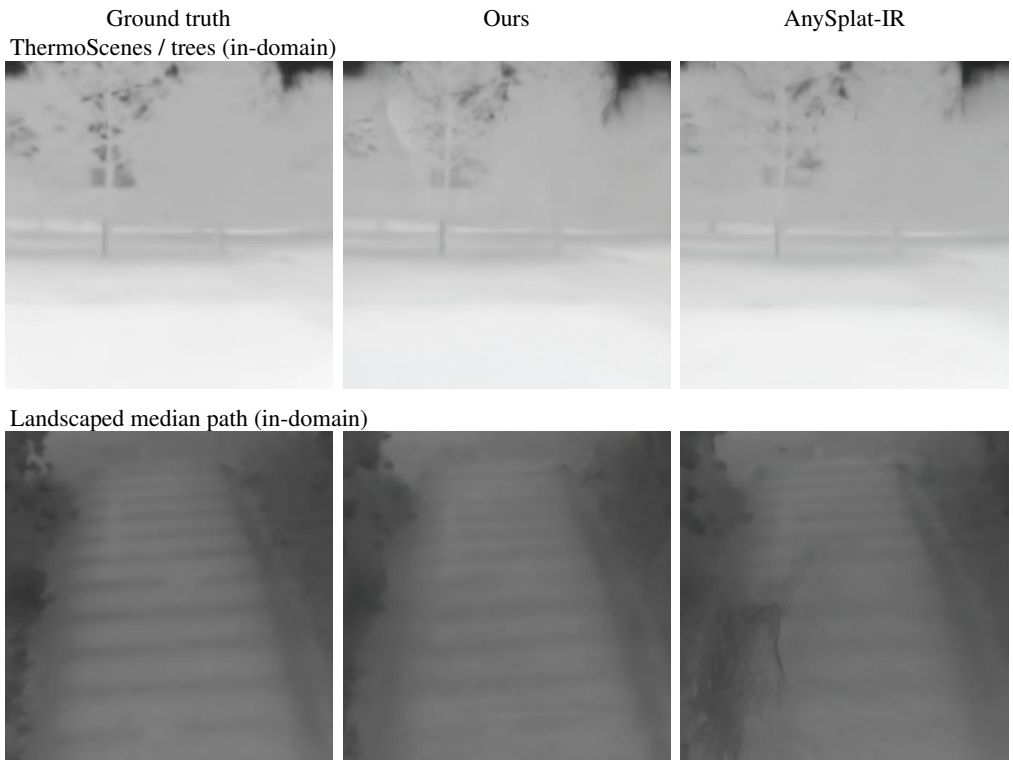


Figure 11: **Additional in-domain novel-view comparisons.** The main model attains 29.86 versus 27.80 dB scene-level PSNR on *trees* and 33.23 versus 30.03 dB on the *landscaped median path*. All images are original evaluation outputs, without contrast adjustment.

Failure cases. The gains are not universal. In TI-NSD *corridor*, *standing*, *heated*, and *bridge*, the main model scores 22.06 versus 24.15 dB, 19.20 versus 21.79 dB, 18.40 versus 20.21 dB, and 19.86 versus 20.98 dB, respectively, against AnySplat-IR at 12 views. Figure 12 shows matched target views. The failure modes include smoothed door-frame and bridge details, distorted thermal structure around the subject, and loss of a heated object's sharp boundary. Since target cameras are also estimated from thermal images, these panels alone do not separate pose error from scene-representation error.

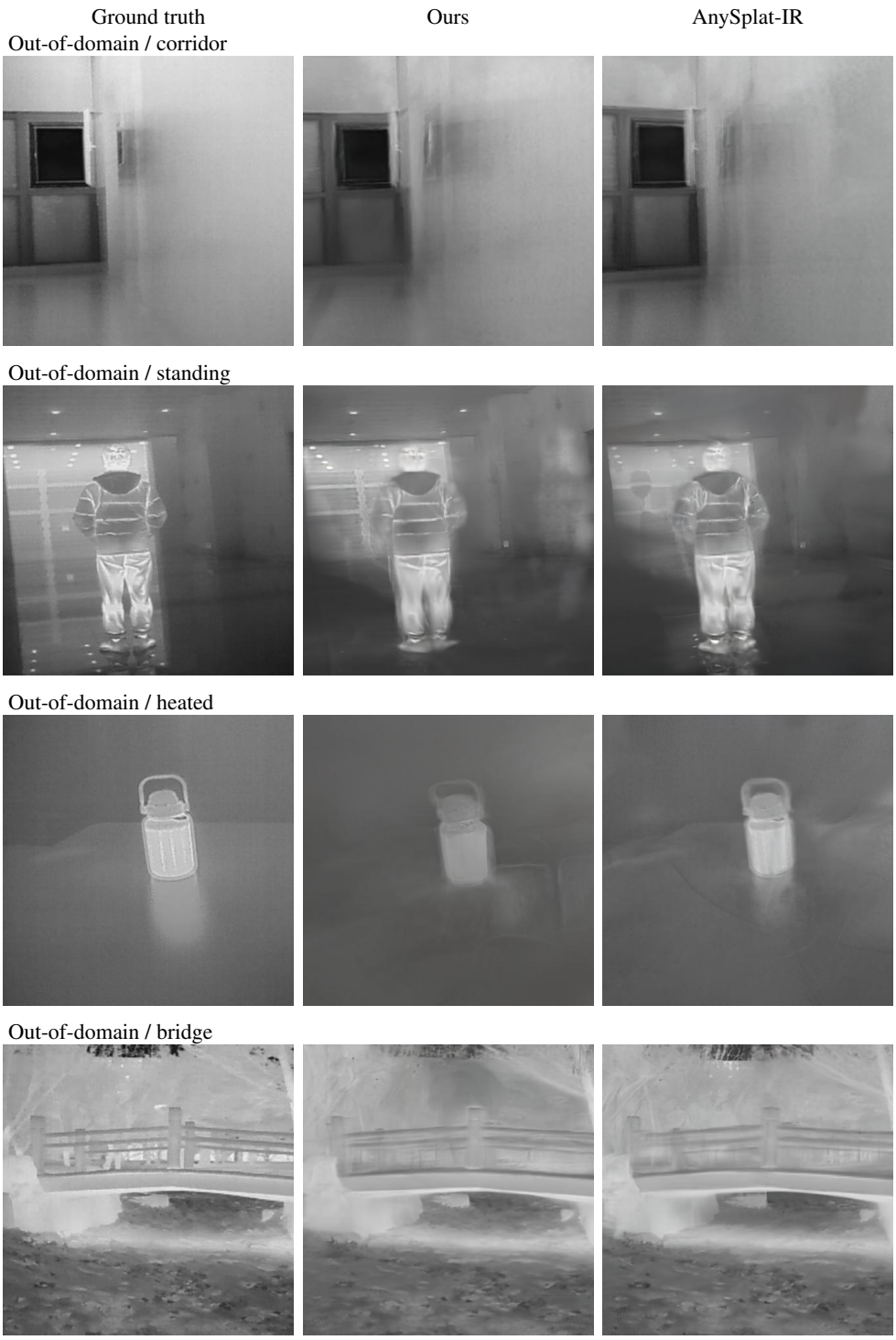


Figure 12: **Out-of-domain failure cases.** These scenes are selected to expose limitations rather than to represent average TI-NSD performance. Both methods use the same target selection; the displayed frame is one of eight target views used in each scene score.