ALL-IN-ONE: JOINT MULTIMODAL IMAGE COMPRESSION VIA SHARED 2D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligned visible and thermal infrared images observe related boundaries but differ in texture and radiometric response. Independent compression repeats spatial support, whereas rigid sharing can suppress modality-specific structure. We study this trade-off with All-in-One, an explicit RGB–IR codec that stores shared 2D Gaussian anchors, a jointly coded appearance dictionary, and sparse infrared geometry corrections. Top-4 appearance routes are active during fitting, and correction support is fixed before its values are learned. Reconstruction uses normalized 2D splatting without a neural synthesis network. A ten-pair LLVIP component study combines recorded checkpoint quality with separately audited complete-stream sizes. Sparse corrections improve mean RGB/IR PSNR by 0.129/0.299 dB and SSIM by 0.00239/0.00072 over hard sharing, at an additional 0.0189 pair-bpp. Eight pairs improve all four quality metrics. Packet decomposition and paired-image analysis quantify the quality–rate trade-off of bounded thermal adaptation within a shared representation.

1 INTRODUCTION

Visible and thermal infrared cameras provide complementary observations in low-light conditions. RGB retains color and fine texture, while thermal IR can preserve salient structure when visible contrast is weak. Storing both images independently is straightforward, but it ignores their shared spatial organization. Conversely, treating the pair as four interchangeable channels assumes that the same local representation is equally suitable for both sensors. We consider aligned RGB–IR image compression as a problem of *shared support with a limited modality-specific correction budget*.

The object being compressed is the pair itself, not a fused image or a scene model. The decoder must reconstruct both the three-channel RGB image and the one-channel IR image at their original evaluation resolution. Neither modality is available to the decoder as free side information. This distinguishes the task from guided reconstruction, image fusion, and compression of only one sensor conditioned on an uncoded partner. It also makes transmitted geometry, routing information, and correction masks part of the rate, rather than auxiliary model parameters.

Explicit 2D Gaussian representations offer a useful design space: support location and extent are separate from appearance, and reconstruction can be performed directly from a compact parameter set (Zhang et al., 2024; 2025). For a pair of aligned observations, the location of a useful boundary may be shared even when its colors are not. The appropriate constraint is therefore not equality of RGB and IR appearance. It is reuse of localized spatial support, with independent dictionary columns and a bounded mechanism for geometry disagreement.

Our representation shares an anchor inventory and sparse appearance routes. A subset of anchors receives an IR position and scale correction. This design poses two coupled questions: whether shared support releases enough bits to justify its cross-modal constraints, and whether the parameters optimized by the encoder remain useful after serialization. We address the second question by retaining the sparse routing topology during fitting and fixing correction support before learning its values. We address the first through complete packet accounting and separate RGB/IR quality comparisons, rather than a pooled four-channel score.

Our empirical focus is the effect of relaxing hard geometry sharing through a small, coded correction layer. On ten LLVIP pairs, we measure the change in reconstruction quality, account for every

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.

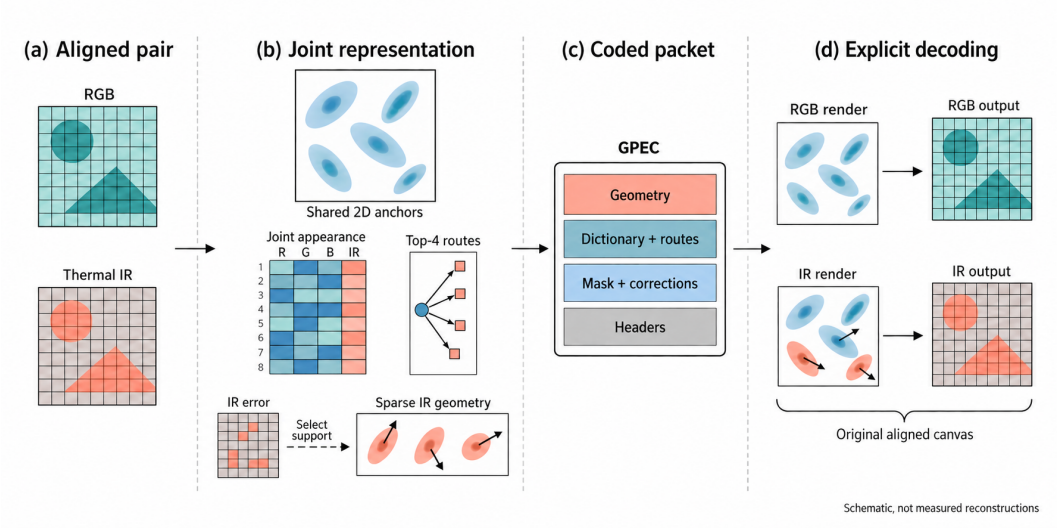

Figure 1: **RGB–IR compression pipeline.** Shared 2D anchors and an eight-entry, four-channel appearance table encode the aligned pair. Top-4 routes determine appearance; error-selected IR offsets and scale corrections adapt a sparse subset of geometry. GPEC transmits the required fields and headers. Explicit rendering reconstructs either sensor on its original aligned canvas. The grids are conceptual illustrations, not measured reconstructions.

transmitted byte, and examine paired outcomes rather than relying only on averages. This component comparison identifies the price and benefit of thermal adaptation; its operating points differ in rate and are not interpreted as an equal-rate comparison with independent compression.

The contributions of this RGB–IR study are:

- **Explicit shared support with bounded thermal adaptation.** A joint 2D Gaussian field separates reusable geometry, four-channel dictionary appearance, and sparse IR geometry corrections. It reconstructs both sensor images without a learned synthesis network or a 3D scene.
- **Compact sparse parameterization.** Top-4 routes and fixed correction support constrain the fitted inventory. Complete GPEC accounting includes support, dictionaries, masks, quantization metadata, and entropy payloads.
- **Measured thermal-adaptation trade-offs.** A paired RGB–IR component study quantifies reconstruction changes and complete packet costs, including per-image variability and the distinction between coded values and metadata.

These contributions concern a particular coding formulation and its evaluation; generic Gaussian splatting, quantization-aware training, or error-based selection are not claimed as inventions.

2 Related Work

Explicit image representations. GaussianImage demonstrates image representation and compression using 2D Gaussians (Zhang et al., 2024). Image-GS studies content-adaptive support (Zhang et al., 2025), while GaussianImage++ incorporates densification and quantization improvements (Li et al., 2026). These methods motivate direct rendering but also establish the appropriate explicit-codec references. Our distinction is the coupling of geometry and sparse appearance routes across a visible–thermal pair. Unlike single-image representations, the shared inventory couples support across two sensor appearances.

Learned image and multimodal compression. Hyperprior and autoregressive codecs exploit latent dependencies through learned probability models (Ballé et al., 2018; Minnen et al., 2018). CCEM

explicitly targets RGB–IR compression with a channel-wise cross-modality entropy model (Wang et al., 2026); it is a directly relevant comparator for this task. RGB-D coding also investigates cross-modality redundancy (Zheng & Gao, 2024). Our design shares explicit support instead of predicting cross-modal latent entropy. These are different mechanisms, not evidence that one is intrinsically superior. Comparisons require identical source images, bit depth, color domain, and complete coded rates.

Cross-view and thermal representations. Distributed multi-view coding, such as LDMIC (Zhang et al., 2023), uses a different encoder/decoder access pattern. Here both modalities are available at the encoder and neither is free at the decoder. ThermalGaussian (Lu et al., 2025) addresses thermal scene representation, whereas we encode an observed 2D pair without novel-view synthesis. The alignment assumption eliminates the need to estimate camera intrinsics or a depth-dependent projection. Small sensor-specific differences are handled as image-plane corrections, not as recovered physical motion.

Selection and sensitivity. Reconstruction-sensitive Gaussian pruning already exists, including PUP 3D-GS (Hanson et al., 2025). Sampling hard-sharing error is a practical support-selection heuristic, not a new general pruning principle or an exact deletion score. Our study instead asks how much coded correction to allocate, whether sparse topology survives serialization, and whether shared coding outperforms a fairly optimized independent reference.

3 RGB–IR Representation

3.1 Task and Rate Convention

Let $I_r \in [0,1]^{3\times H\times W}$ and $I_t \in [0,1]^{1\times H\times W}$ be the aligned RGB and thermal IR images. Each reconstruction is compared with its own GT on the dataset-provided canvas. We do not fuse the outputs or rescale intensities independently per image. The transmitted packet contains all information required to reconstruct both images, apart from the fixed public decoder implementation.

For a packet containing B bytes, the primary convention is

$$R_{\text{pair}} = \frac{8B}{HW}. \tag{1}$$

The denominator counts one aligned image grid; all bits for both modalities are included. A rate normalized by the sum of the two canvases is exactly $R_{\text{pair}}/2$ and is not mixed into our tables. RGB PSNR uses MSE over its three channels; IR PSNR uses its single channel. PSNR and SSIM are always reported separately. Joint-error PSNR would implicitly privilege the three RGB channels and is not used for model selection.

The conceptual optimization is

$$\min_{\theta,N,\mathcal{S}} \; w_r D_r(I_r,\hat{I}_r) + w_t D_t(I_t,\hat{I}_t) \quad \text{s.t.} \quad B(\theta,N,\mathcal{S}) \le B_{\max}, \tag{2}$$

where N is the number of shared anchors and $\mathcal{S}$ is the active correction support. This formulation distinguishes reconstruction distortion from complete coded cost. The component study evaluates two fixed inventories under this formulation, measuring their byte costs rather than solving a budget-constrained search.

3.2 Shared Anchors and Sparse Thermal Geometry

Figure 2 contrasts three geometry policies. Independent coding permits unrelated inventories; hard sharing ties their parameters; sparse correction retains most ties while allowing a transmitted subset of IR deviations. This is a constraint on parameter ownership, not a guarantee of a particular quality ordering.

Each anchor has center $p_i \in [0,1]^2$, positive scales $s_i \in \mathbb{R}^2_+$, and rotation θ_i. RGB uses this geometry directly. Thermal geometry is

$$p_i^t = p_i + m_i \,\mathrm{diag}(W^{-1},H^{-1})\delta_i, \qquad s_i^t = s_i \exp(m_i a_i), \qquad \theta_i^t = \theta_i, \tag{3}$$

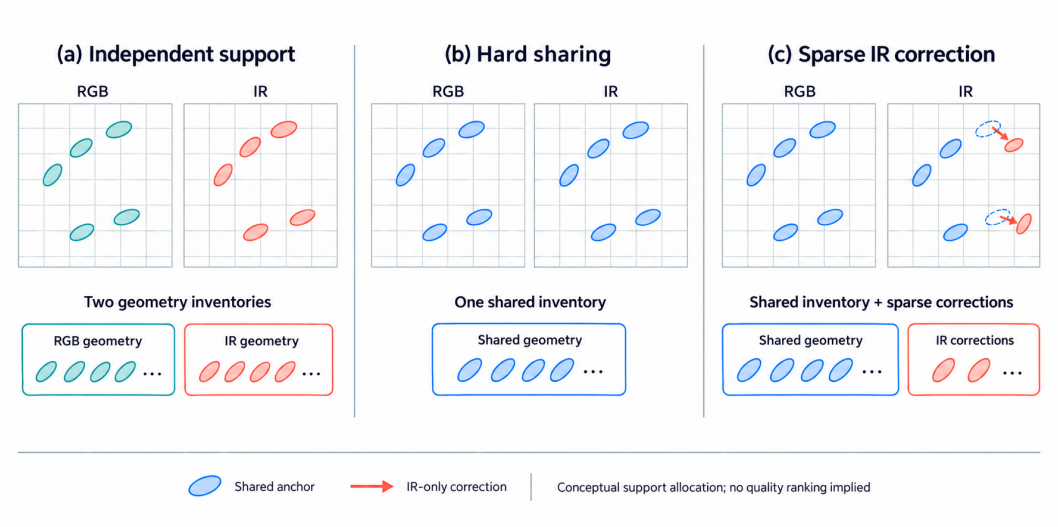

Figure 2: **Support ownership.** Independent fields store two geometry inventories. Hard sharing reuses one inventory. Sparse correction adds IR-only displacement and scale parameters on selected anchors. Dashed outlines indicate pre-correction support. The number of drawn ellipses is illustrative, not the experimental correction fraction.

where $m_i \in \{0, 1\}$ indicates support membership. The offset δ_i is stored in *image pixels*; a_i is a dimensionless isotropic log-scale correction. The development setting uses $|\delta_{i,x}|, |\delta_{i,y}| \leq 4$ pixels and $|a_i| \leq 0.25$, with $\sum_i m_i = \mathrm{round}(0.2N)$. These are operating settings, not claimed optima.

At the correction activation step, the hard-sharing IR reconstruction supplies a score

$$q_i = \left| I_t(p_i) - \hat{I}_t^{\mathrm{hard}}(p_i) \right|, \qquad \mathcal{S} = \mathrm{Top}_{\mathrm{round}(\rho N)}(q). \tag{4}$$

Bilinear sampling evaluates each image at the normalized center. The selected support is frozen before learning correction values, initialized to zero. Consequently, the final mask does not remove corrections that have already absorbed reconstruction responsibility. The score measures local error, not the exact value of deleting a Gaussian; normalized rendering can couple nearby points.

3.3 Sparse Joint Appearance

An eight-entry appearance dictionary $E \in \mathbb{R}^{8\times 4}$ represents RGB and IR in separate columns. For route logits z_i, let $\mathcal{T}_4(z_i)$ contain the four selected expert indices. The anchor appearance is

$$f_i = \sum_{e \in \mathcal{T}_4(z_i)} g_{i,e} E_e, \qquad g_{i,e} = \frac{\exp(z_{i,e}/\tau)}{\sum_{j \in \mathcal{T}_4(z_i)} \exp(z_{i,j}/\tau)}. \tag{5}$$

The first three components form RGB appearance and the last component forms IR appearance. Shared routes encourage a compact joint appearance organization without enforcing equal intensity across sensors. They can also constrain expressivity, which motivates separate-route and separate-dictionary controls.

Four selected entries among eight yield $\binom{8}{4} = 70$ subsets. The stream stores a subset symbol and three quantized weights; the remaining simplex weight is recovered by the codec. Expert values and their quantization ranges are also transmitted. Learning uses the sparse support rather than fitting dense routes and truncating them only at export. The codec uses sorted expert indices and integer simplex weights summing to 63. This export quantization is distinct from maintaining sparse topology during fitting.

3.4 Normalized Reconstruction

Let $w_{i,m}(x)$ denote the spatial weight of anchor i for modality m, with geometry from Equation 3. The top-10 renderer selects $T_{10,m}(x)$ within its spatial support and computes

$$\hat{I}_m(x) = \mathrm{clip}_{[0,1]}\left[\frac{\sum_{i\in T_{10,m}(x)} w_{i,m}(x)c_{i,m}}{\sum_{i\in T_{10,m}(x)} w_{i,m}(x) + \epsilon}\right]. \tag{6}$$

RGB and IR share appearance routing but may select different spatial neighborhoods after thermal correction. The correction is geometric: it is not a signed residual image added to RGB or to an independently rendered base. Removing or moving a point changes both the selected neighborhood and its normalization. Accordingly, exact candidate quality is established by re-rendering the full decoded parameter set.

The decoder applies the same explicit operations to either requested modality. RGB rendering does not require rendering an IR image first, and vice versa. This is output-selective computation, not proof that the complete entropy stream permits arbitrary byte-range access. The absence of a neural synthesis network describes the decoder architecture, not a measured speed advantage.

3.5 Why Local Correction Requires Global Verification

The benefit of a correction depends on the image already reconstructed by its neighbors. Write an unclipped pixel as $y = A/(Z + \epsilon)$, where A is the weighted color sum and Z the weight sum over the selected neighborhood. A change in geometry gives $A' = A + \Delta A$ and $Z' = Z + \Delta Z$, hence

$$y' - y = \frac{\Delta A - y\Delta Z}{Z + \Delta Z + \epsilon}. \tag{7}$$

This follows by bringing the two normalized sums to a common denominator. It holds channel-wise even when neighborhood membership changes, provided each sum uses its own selected points. Clipping is nonexpansive, so it cannot increase the magnitude of this pixel perturbation. In particular, the value of a correction cannot be estimated by simply adding or subtracting one Gaussian's unnormalized color.

For image error $e = \hat{I} - I$ and reconstruction change $d = \hat{I}' - \hat{I}$, the exact MSE difference over P scalar samples is

$$D'_{\mathrm{mse}} - D_{\mathrm{mse}} = \big(2\langle e, d\rangle + \|d\|_2^2\big)/P. \tag{8}$$

A useful correction must counteract existing error strongly enough to offset its own squared change. High local error therefore identifies a place to investigate, not a guaranteed gain. These elementary identities motivate full decoded-image checks; they are not a new rate–distortion bound or an SSIM guarantee. The practical separation is between cheap encoder-side support proposals and the complete measurements used to judge them.

4 Encoding and Complete-Stream Verification

4.1 Optimization Schedule

The encoder optimizes each pair for 12,000 updates. Top-4 appearance routing is active during fitting with temperature $\tau = 0.5$. Geometry grows from an initial inventory before the final correction stage; at step 6,000, hard-sharing error selects the correction support. The active corrections are initialized to zero and subsequently optimized with the shared parameters. The recorded configurations set the quantization warmup to 20,000 steps, beyond the fitting horizon. These experiments therefore use post-training export quantization, not in-training deployment-matched QAT.

The distortion combines robust pixel, squared-error, structural, and gradient terms,

$$D_m = \lambda_c D_{\mathrm{char}} + \lambda_2 D_{\mathrm{mse}} + \lambda_s(1 - \mathrm{SSIM}) + \lambda_g D_{\mathrm{grad}}. \tag{9}$$

Channels and spatial samples are averaged within each modality before modality weighting. The coefficients are $(\lambda_c, \lambda_2, \lambda_s, \lambda_g) = (0.4, 0.15, 0.45, 0.02)$, with RGB:IR weighting $1 : 1.5$ in both conditions. The correction changes geometric flexibility without changing this modality preference. The weights specify quality allocation rather than a pooled evaluation metric; all reported scores remain modality-specific.

Table 1: Measured mean packet composition on the ten development pairs, in bytes. Fractional values arise from averaging integer file sizes. These are actual packet costs, not the fixed-payload formula.

Configuration	Entropy payload	Headers / metadata	Total
Hard sharing	64,267.6	735.7	65,003.3
Sparse IR correction	67,200.1	896.8	68,096.9
Difference	2,932.5	161.1	3,093.6

4.2 Accounting for Support and Corrections

For the pilot's 16-bit center, scale, and rotation fields, geometry uses $80N$ fixed payload bits. Top-4 routes require $25N$ bits before entropy coding, and the 8×4 dictionary uses 384 bits at 12-bit precision. The mask and three 6-bit correction scalars add $N + 18\,\mathrm{round}(\rho N)$ bits. Thus

$$L_{\mathrm{fixed}} = 106N + 18\,\mathrm{round}(\rho N) + 384. \tag{10}$$

This expression is a field-level diagnostic, not the reported rate. The complete GPEC packet additionally contains field descriptors, quantization ranges, entropy-code metadata, alignment, and a container header. Entropy coding can make the packet smaller than the fixed payload estimate, even after adding metadata. We measure the final file and compute Equation 1 from its complete byte count.

For a no-correction stream, both correction values and their unused mask are omitted. This matters in the residual ablation: its added rate is the difference between two complete streams, not only the nominal offset payload. Two independently coded images likewise incur both containers and dictionaries. The codec comparison never counts only a selected subset of the files required for reconstruction.

When does shared support pay for itself? Let B_G, B_E, and B_U denote the coded geometry, appearance table, and route fields; let B_C contain the correction mask and active correction values. With header cost B_H, the shared stream is worthwhile in rate only if

$$B_G + B_E + B_U + B_C + B_H < B_r^{\mathrm{sep}} + B_t^{\mathrm{sep}}. \tag{11}$$

All terms refer to complete encoded components for a specified candidate. Geometry reuse can release bits, but joint routing may require a larger anchor inventory to preserve detail, and correction masks have a nonzero cost even when their values are small. Consequently, a smaller parameter count is not a sufficient acceptance criterion. Equation 11 is a transparent accounting test, not a prediction of the optimum sharing fraction. Per-modality distortion must be checked in addition to this rate condition.

Table 1 separates the measured overhead from the coded values. The difference includes any changes in the entropy coding of jointly optimized fields; it is not assumed to consist solely of offset symbols. This distinction lets the same reporting convention apply to an independent codec, a hard-sharing codec, and a corrected shared codec without favoring a particular file organization.

4.3 Fitting and Packet Audit

Table 2 separates image fitting from serialization and identifies the two measurement sources. The packet audit verifies stored symbols and their decoded parameters. It does not equate the checkpoint's forward path with the parsed packet's reconstruction: topology can agree while numerical weight quantizers differ.

The model's quantized route path rounds individual weights and renormalizes their sum; the packet encoder allocates integer weights whose sum is exactly 63. Round-trip equality concerns the latter serialized representation. Accordingly, the component tables explicitly identify checkpoint quality and do not present it as packet-decoded RD performance.

Table 2: Fitting and accounting sequence used for the component evidence. Checkpoint quality and packet costs are distinct measurements.

1	Read the aligned pair with fixed radiometric normalization and no per-image contrast transform.
2	Fit shared anchors and top-4 joint appearance; select correction support from hard-sharing IR error.
3	Optimize active corrections and shared parameters; record RGB/IR checkpoint PSNR and SSIM.
4	Quantize and encode geometry, appearance, routes, and active corrections into GPEC.
5	Parse the packet and check symbol and parameter round trips against the encoded reference.
6	Match pair identifiers to quality records; measure complete file bytes and separate payload from metadata.

Table 3: Ten-pair LLVIP development evidence at $N = 5100$ shared anchors. Quality comes from recorded checkpoint evaluations; rates come from complete GPEC files. The rows are not rate matched.

Method	Pair bpp	RGB PSNR	RGB SSIM	IR PSNR	IR SSIM
Shared, no IR correction	0.3967	29.270	0.83798	39.006	0.98102
Shared + sparse IR correction	0.4156	29.399	0.84037	39.305	0.98174

5 Experiments

5.1 Dataset and Evaluation Domain

We use ten aligned visible–infrared pairs from LLVIP (Jia et al., 2021), with identifiers 190001–190010. These images form a development cohort drawn from the official test split. The reported aggregates describe this cohort, rather than the full test distribution.

We evaluate the dataset-provided aligned images at their original 1280×1024 resolution. Stored 8-bit RGB and grayscale IR are normalized by 255, with no YUV420 conversion or contrast stretching. For each modality, PSNR is computed from its own image MSE. SSIM uses a fixed 11×11 Gaussian window, $\sigma = 1.5$, and data range one, averaging RGB channels within that modality. Dataset results are arithmetic means of per-image metrics, not PSNR calculated from pooled dataset errors.

Only PSNR and SSIM are primary quality metrics. Rate always counts the complete pair stream. Per-image variability is exposed through paired changes and the complete ten-pair coverage, including adverse outcomes.

5.2 Hard Sharing versus Sparse Correction

Both conditions use $N = 5100$ shared anchors, an eight-entry appearance dictionary, top-4 routes, and 12,000 optimization updates. Hard sharing ties RGB and IR geometry. Sparse correction activates 1,020 IR corrections with the position and scale bounds in Equation 3. Both conditions use the same recorded loss coefficients and modality weights. Here N counts the joint anchor inventory, not a per-modality independent budget.

This comparison asks whether a small relaxation of shared geometry improves fitted reconstruction and how much its packet costs. The appearance dictionary and route capacity remain the same; the correction values and mask add transmitted information. The comparison is therefore a component quality–rate trade-off, rather than a matched-rate ranking of codecs. We report mean quality, complete packet composition, and every paired quality change.

5.3 Measured Development Evidence

Table 3 is regenerated from the existing ten-pair checkpoint metric records and their corresponding GPEC v2 audits. The exporter matches pair identifiers, verifies actual file sizes, and requires the recorded symbol/parameter round-trip checks to pass. These are *checkpoint-quality measurements with separately audited stream sizes*. Symbol-level round trips establish packet integrity, not equality between checkpoint rendering and actual packet-decoded images.

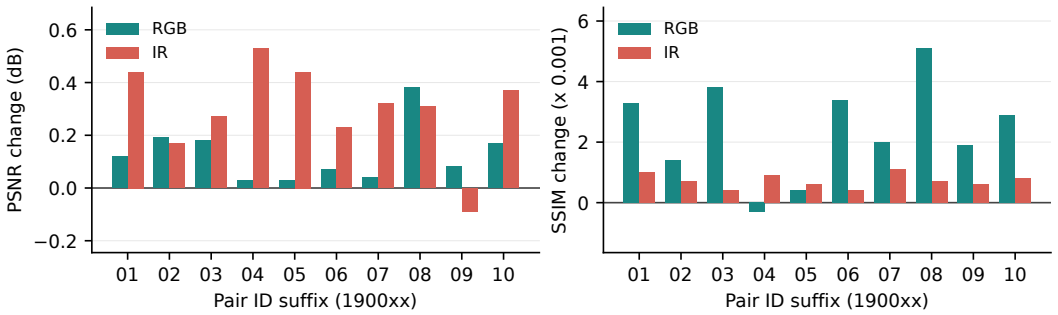

Figure 3: **Recorded ten-pair changes: sparse correction minus hard sharing.** Left: PSNR difference. Right: SSIM difference in units of 10^{-3}. Both negative changes remain visible. Each pair also incurs additional bytes; this is not an equal-rate comparison. Bars are generated directly from the historical metric CSV, not by an image-generation model.

Sparse correction changes mean RGB/IR PSNR from 29.270/39.006 to 29.399/39.305 dB and SSIM from 0.83798/0.98102 to 0.84037/0.98174. Complete rate changes from 0.39675 to 0.41563 pair-bpp. The corresponding mean packet sizes are 65,003.3 and 68,096.9 bytes. This supports a local quality–rate trade-off relative to hard sharing. It does not establish an equal-rate gain, an advantage over independent coding, or an optimal correction fraction.

The full corrected packet averages 67,200.1 entropy-payload bytes and 896.8 header/metadata bytes. This decomposition demonstrates why both field-level estimates and actual streams are useful: one explains where bits go, while the other establishes the rate that a decoder receives. No complete-stream saving is inferred from Gaussian counts alone.

Variation across pairs. Figure 3 shows the recorded paired changes. RGB PSNR improves on all ten pairs, whereas IR PSNR improves on nine and decreases by 0.09 dB on one. RGB SSIM improves on nine pairs and IR SSIM on all ten; eight pairs improve all four metrics simultaneously. The mean packet increase is 3,093.6 bytes, approximately 4.76% of the hard-sharing packet. These paired counts are descriptive, not significance tests. They show why the two mean PSNR gains should be interpreted alongside per-image behavior and the added rate, rather than as a universal Pareto improvement.

5.4 Paired Quality and Byte Analysis

Table 4 exposes the complete paired changes underlying the averages. No image is removed because one of its metrics decreases. The RGB PSNR gain ranges from 0.03 to 0.38 dB, while the IR gain ranges from -0.09 to 0.53 dB. The two exceptions to simultaneous four-metric improvement occur on different images: pair 190004 loses 0.0003 RGB SSIM despite a 0.53 dB IR PSNR gain, and pair 190009 loses 0.09 dB IR PSNR despite improved RGB quality and IR SSIM. Geometry flexibility thus affects the two sensors and the two quality criteria differently.

The added packet cost lies between 3,054 and 3,195 bytes per pair. This narrow interval contrasts with the variation in quality gains: equal correction counts do not imply equal reconstruction benefit. In particular, all pairs incur additional bytes, including the two with a metric regression. The aggregate gain is therefore best understood as the average response to bounded adaptation, not a per-image dominance guarantee.

Where do the additional bytes go? Mean entropy payload grows by 2,932.5 bytes and metadata by 161.1 bytes, together accounting for the 3,093.6-byte increment. Metadata contributes approximately 5.21% of the increment; the remaining 94.79% is coded payload. Total metadata occupies approximately 1.13% and 1.32% of the hard-sharing and corrected packets, respectively. Most of the observed cost is thus associated with coded values rather than container overhead.

At $N = 5100$, the uncompressed correction fields alone contain $5100 + 18 \times 1020 = 23{,}460$ bits, or 2,932.5 byte-equivalents before field packing. The numerical equality with the mean entropy-payload

Table 4: Per-pair changes from hard sharing to sparse correction. PSNR changes are in dB; SSIM changes are multiplied by 10^3. Quality is checkpoint-based; added bytes are measured from complete packets.

Pair	Δ RGB PSNR	Δ IR PSNR	Δ RGB SSIM	Δ IR SSIM	Added bytes
190001	+0.12	+0.44	+3.3	+1.0	3095
190002	+0.19	+0.17	+1.4	+0.7	3195
190003	+0.18	+0.27	+3.8	+0.4	3097
190004	+0.03	+0.53	-0.3	+0.9	3060
190005	+0.03	+0.44	+0.4	+0.6	3086
190006	+0.07	+0.23	+3.4	+0.4	3054
190007	+0.04	+0.32	+2.0	+1.1	3072
190008	+0.38	+0.31	+5.1	+0.7	3084
190009	+0.08	-0.09	+1.9	+0.6	3081
190010	+0.17	+0.37	+2.9	+0.8	3112
Mean	+0.129	+0.299	+2.39	+0.72	3093.6

increment in this cohort does not identify the correction fields' compressed length: the two fits also produce different geometry, route, and dictionary symbols. Only the complete packet difference measures the actual incremental transmission cost.

6 Discussion and Conclusion

Aligned RGB–IR compression offers a focused setting in which shared geometry is plausible but not automatically beneficial. All-in-One expresses this structure through shared 2D support, sparse joint appearance, and rate-limited thermal geometry. Its explicit decoder avoids a neural synthesis network, while complete packet accounting makes the cost of this design measurable. The development evidence quantifies a positive correction trade-off and establishes concrete byte costs.

The empirical conclusion is a measured thermal-adaptation trade-off within an aligned-pair shared representation. Its scope is the ten-pair component study with checkpoint quality and separately audited packet costs, not matched-rate superiority over independent or external codecs. Large misregistration, different sensor coverage, and thermal-specific structure fall outside the bounded-correction assumption. Within the evaluated setting, sparse geometry correction offers a compact way to improve average reconstruction while retaining shared support and explicit rendering.

AI USE STATEMENT

Generative AI tools were used only to edit the manuscript for readability and language. We have not used generative AI tools to generate synthetic data, develop theoretical models, formulate mathematical claims, assist with proofs, propose or refine hypotheses, design experiments, implement methods, translate text, clean or reformat data, analyze results, interpret findings, draft the paper, create figures, or write code. The authors reviewed the language edits and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies reconstruction of visible and thermal images from an existing dataset and does not introduce a person-identification or detection experiment. Visible and thermal imagery can contain sensitive human information; compression does not anonymize it. Dataset licensing, permitted redistribution, and privacy considerations must be respected when releasing examples or code. No new human-subject data are collected for this study.

REPRODUCIBILITY STATEMENT

Sections 3–4 define the representation, rate convention, and encoding sequence. Section 5 reports the paired component evidence and packet accounting. Appendix A records the configuration and evidence provenance. Tables and the paired-change plot are produced from per-pair measurement records by deterministic scripts; packet lengths are checked against the corresponding files. Table 4 exposes all paired changes used in the analysis.

REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *International Conference on Learning Representations (ICLR)*, 2018.

Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. PUP 3D-GS: Principled uncertainty pruning for 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5949–5958, 2025.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021.

Tiantian Li, Xinjie Zhang, Xingtong Ge, Tongda Xu, Dailan He, Jun Zhang, and Yan Wang. GaussianImage++: Boosted image representation and compression with 2D Gaussian splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2512.19108.

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 (ICLR)*, 2025. arXiv:2409.07200.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Haofeng Wang, Fangtao Zhou, Qi Zhang, Zeyuan Chen, Enci Zhang, Zhao Wang, Xiaofeng Huang, and Siwei Ma. End-to-end RGB-IR joint image compression with channel-wise cross-modality entropy model. In *Proceedings of the 2026 Data Compression Conference (DCC)*, 2026. doi: 10.1109/DCC66757.2026.00097. arXiv preprint arXiv:2506.21851.

Xinjie Zhang, Jiawei Shao, and Jun Zhang. LDMIC: Learning-based distributed multi-view image coding. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=ILQVw4cA5F9`.

Xinjie Zhang, Xingtong Ge, Tongda Xu, Dailan He, Yan Wang, Hongwei Qin, Guo Lu, Jing Geng, and Jun Zhang. GaussianImage: 1000 FPS image representation and compression by 2D Gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2403.08551.

Yunxiang Zhang, Alexandr Kuznetsov, Akshay Jindal, Kenneth Chen, Anton Sochenov, Anton Kaplanyan, and Qi Sun. Image-GS: Content-adaptive image representation via 2D gaussians. In *ACM SIGGRAPH Conference Proceedings*, 2025. URL https://arxiv.org/abs/2407.01866.

Huiming Zheng and Wei Gao. End-to-end RGB-D image compression via exploiting channel-modality redundancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

A Implementation and Evidence Scope

The study uses aligned RGB and thermal IR pairs 190001–190010 from LLVIP. All ten pairs are included in every aggregate, including the two pairs with a negative change in one quality metric.

Configuration. Both conditions use 5,100 final anchors, eight experts, top-4 appearance routes, temperature 0.5, and 12,000 fitting updates. Shared geometry has 16-bit fields, dictionary entries use 12 bits, and route weights use 6 bits. The corrected condition selects 20% of anchors at step 6,000; each active anchor has two 6-bit position offsets and one 6-bit isotropic log-scale correction. The displacement bound is four pixels per axis and the log-scale bound is 0.25. Correction support is fixed before its initially zero values are learned. RGB:IR loss weighting is 1:1.5; Charbonnier, MSE, structural, and gradient coefficients are 0.4, 0.15, 0.45, and 0.02. The quantization warmup exceeds the fitting horizon, so these runs are not deployment-matched QAT experiments.

Quality provenance. Quality measurements come from recorded checkpoint evaluations. Their source CSV stores PSNR to two decimal places and SSIM to four; additional digits in averages arise from averaging these records, not a higher-precision re-evaluation. The checkpoint and packet route quantizers have different rounding rules, so these values are not labeled as actual packet-decoded quality.

Packet provenance. The export script matches pair identifiers and metric rows, checks actual file size, and requires recorded symbol and parameter round-trip equality. These equalities compare serialized symbols and their decoded parameter reference, not the checkpoint renderer against packet rendering. Complete packet size includes geometry, appearance, routing, correction masks and values, quantization ranges, entropy metadata, and headers. Neither modality is supplied as free side information.

Rendering semantics. Equation 6 uses top-10 normalized blending and output clipping. Geometry corrections can change neighborhood membership and normalization, but they are not additive signed image residuals. Equations 7–8 describe this distinction. The renderer can compute either output without first rendering the other; this property is separate from byte-range access or comparative decoder latency.

Interpretation. The hard-sharing and correction conditions differ in transmitted bytes. Their comparison measures the effect and cost of additional geometric flexibility, not an equal-rate coding advantage. The development cohort is not a full-test aggregate. The scope of the numerical evidence is preserved consistently in the abstract, tables, and conclusion.