EVENTGAUGE: WHEN CAN 3D EVALUATION VALIDATE METRIC SCALE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned 3D reconstruction increasingly claims metric depth and absolute scene scale, yet standard evaluation can remain unchanged when the recovered global scale is wrong. Similarity alignment can absorb global scale, while co-centred rendering can preserve projections under joint rescaling, removing scale evidence from both common reporting channels. We introduce **EventGauge**, a physical audit that withholds a calibrated displaced camera until reconstruction is frozen. The audit asks whether the finished reconstruction is best supported at the true physical viewpoint and whether that support is sharp enough to justify a metric conclusion. Controlled rescaling verifies that the response moves with scale as predicted by geometry. Across positive and failure cases, the audit can support physical consistency, expose a residual inconsistency, or remain unresolved when the evidence is too broad. EventGauge therefore turns a metric-scale claim from something standard reporting can silently absorb into a falsifiable physical test. Selecting a scale is a property of reconstruction; validating that scale is a property of evaluation.

1 INTRODUCTION

Recovering metric 3D is a stronger scientific claim than recovering a scene only up to an unknown similarity transform. Evaluation should become stronger accordingly. Modern monocular systems increasingly predict metric depth, camera geometry, or absolute scene scale. Some explicitly seek a consistent global scale Ye et al. (2025; 2026), while others target an absolute metric representation Mescheder et al. (2026). Once scale enters the scientific claim, a low reconstruction error is no longer sufficient. The reported evaluation must contain evidence that can distinguish the correct physical scale from an incorrect one.

In common monocular evaluation, scale evidence can disappear at both major stages of reporting. Similarity alignment explicitly fits global scale away before trajectory error is reported. Co-centred rendering adds no independent physical translation, so jointly rescaling scene geometry and reconstruction-camera translation leaves the image projection unchanged. Neither invariance is mathematically surprising. Their combination is consequential because it can remove scale evidence from both trajectory and rendering scores. A traceable audit of eleven recent event-based protocols finds no independent physical translation in the quantitative rendering references we could trace. Standard reporting can therefore certify trajectory shape and image quality while providing no independent evidence that the recovered global scale is physically correct (Figure 1).

Scale sensitivity is not metric-scale validation. Recent generative novel-view synthesis work measures scale inconsistency across generated views Forghani et al. (2025), but a system can be perfectly consistent and still be wrong by a fixed factor in physical units. Absolute validation additionally requires a physical reference that does not inherit the reconstruction's scale.

Rendering itself is not the problem. Metric-scale evidence requires an independently fixed physical reference. A metric camera displacement is one such reference. Metric depth, calibrated 3D geometry, and native metric camera motion can play the same role. SHARP is a positive example because it evaluates a metric representation under metric camera motion Mescheder et al. (2026).

EventGauge uses a displaced camera because its images can remain hidden until reconstruction is fixed. This makes the physical reference evaluation-only.

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.

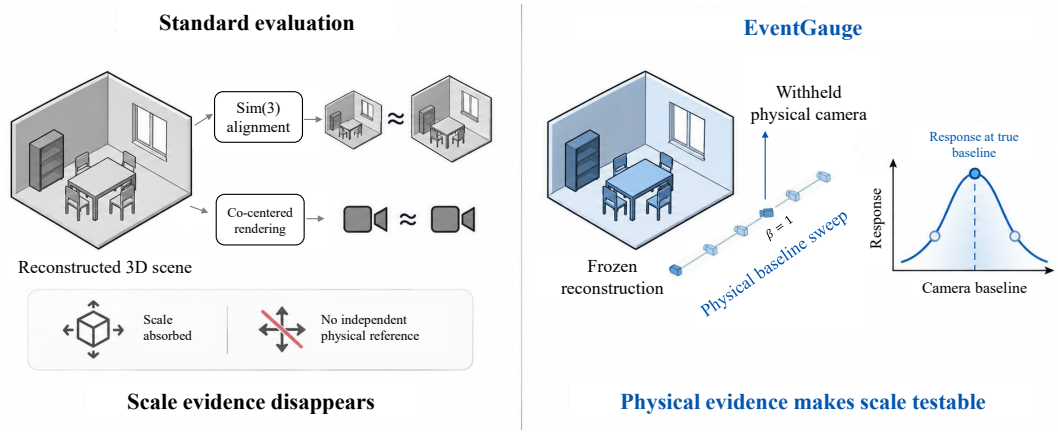


Figure 1: **Why metric scale requires physical evidence.** (a) Standard evaluation can hide metric scale because Sim(3) alignment absorbs global scale and co-centered rendering provides no independent physical reference. (b) EventGauge restores scale evidence by evaluating a frozen reconstruction from a withheld physical camera and measuring the response across a physical baseline sweep. The resulting response provides a physical basis for testing whether the selected scale is supported.

We build **EventGauge** around two complementary checks. The trajectory check keeps the reconstruction-to-metric conversion absorbed by similarity alignment and reports the residual when rescaling is forbidden. The image check then tests the frozen scene under that conversion using a displaced camera withheld from reconstruction and scale fitting. Its baseline response separates true-view quality, the best-matching physical baseline, and the sharpness of that location. EventGauge is designed as an audit. A sharp response can support a metric conclusion; a broad or censored response withholds unsupported precision.

Contributions.

1. **An evaluation capability test for metric scale.** We show how similarity alignment and co-centred rendering can jointly erase scale evidence from the two main reporting channels. A traceable protocol audit shows that this capability gap occurs in current event-based evaluation practice.
2. **EventGauge, a calibrated physical audit for metric claims.** It keeps trajectory-scale information visible and tests the frozen scene with an evaluation-only physical camera. Known-answer calibration establishes the response law, while a native-metric RGB case shows that the audit can positively support physical consistency.
3. **Evidence that independent physical evaluation changes what can be concluded.** Standard reporting can hide materially different metric conversions. A held-out physical view gives a coherent judgment of a controlled geometric correction, while broad second-rig responses remain unresolved rather than being promoted to a precise metric estimate.

2 Related Work

2.1 Learned 3D Reconstruction and Scale

Global scale ambiguity is fundamental to monocular geometry. Learned systems can still choose a global scale through data, priors, camera conditioning, or normalization. Pose-free radiance-field methods jointly optimize scene and cameras Lin et al. (2021); Bian et al. (2023). NoPoSplat uses camera intrinsics to address scale ambiguity Ye et al. (2025). YoNoSplat uses camera-distance normalization to make reconstruction scale more consistent Ye et al. (2026). SHARP makes the stronger claim of an absolute-scale metric representation Mescheder et al. (2026). These goals are

not equivalent: resolving scale ambiguity, producing a repeatable scale, representing metric scale, and independently validating metric scale.

2.2 Scale-sensitive evaluation versus physical scale validation

Forghani et al. introduce SFC and SS-TSED to measure scale inconsistency in generative novel-view synthesis Forghani et al. (2025). ScaleMaster studies intra-session scale drift and inter-session scale ambiguity in deep monocular SLAM Ju et al. (2026). It complements trajectory metrics with direct map-to-map evaluation against 3D ground truth. Dense metric depth or 3D ground truth is a stronger direct anchor when it is available. EventGauge asks how evaluation can retain physical scale evidence when dense scene ground truth is absent but a calibrated displaced view is available. A consistency score alone cannot establish physical correctness without an independent metric anchor.

2.3 Evaluation validity

Prior work asks whether benchmarks retain the information needed for their conclusions. Examples include audits of single-view 3D reconstruction, weakly supervised localization, texture bias, and generative-image metrics Tatarchenko et al. (2019); Choe et al. (2020); Kalischek et al. (2023); Jayasumana et al. (2024). Monocular depth provides a close geometric analogy. Scale alignment is appropriate when absolute scale is unavailable, whereas an external metric reference can preserve it Godard et al. (2019). Our focus is more specific: does the reported evaluation retain at least one independent observation that changes when the physical quantity named in the claim changes? This is a necessary observability condition for evaluation identifiability.

2.4 Event-based 3D reconstruction as a testbed

Event NeRF and Gaussian methods span pose-given and pose-free reconstruction Rudnev et al. (2023); Klenk et al. (2023); Low and Lee (2023); Han et al. (2024); Huang et al. (2025); Yura et al. (2025); Wu et al. (2025); Kim et al. (2026); Wang et al. (2024). Their photometric pipelines differ. Across the eleven protocols we trace, quantitative rendering references introduce no independent physical translation. E-3DGS explicitly discusses the viewpoint discrepancy introduced by a physically separate reference camera Zahid et al. (2025). We use such a displaced camera only after reconstruction is frozen, so it contributes no information to reconstruction or scale fitting. Our question is whether this withheld physical view can provide evaluation evidence about metric scale.

3 When Does Evaluation See Metric Scale?

3.1 Similarity alignment removes scale evidence

Let $\mathbf{x}_r \in \mathbb{R}^3$ denote an estimated camera centre in reconstruction coordinates. Let $\mathbf{x}_m \in \mathbb{R}^3$ denote the corresponding centre in the metric reference frame. A Sim(3) trajectory alignment fits a positive global scale s, a rotation $\mathbf{R} \in \mathrm{SO}(3)$, and a translation $\mathbf{t} \in \mathbb{R}^3$ such that

$$\mathbf{x}_m = s\mathbf{R}\mathbf{x}_r + \mathbf{t}. \tag{1}$$

The scalar $s > 0$ is the reconstruction-to-metric conversion factor. Reporting the residual after this fit is useful for trajectory shape, but it removes s from the residual by construction. Two reconstructions can therefore have similarly small Sim(3)-aligned absolute trajectory error (ATE) while requiring substantially different metric conversions. If metric scale is part of the scientific claim, the fitted s is evidence and should remain visible in reporting.

3.2 Co-centred rendering also cannot see scale

For rendering to provide metric-scale evidence, the viewing geometry must supply an independently fixed physical reference. If the rendering camera shares the same optical centre as the reconstruction camera, jointly rescaling the scene and all reconstruction-camera translations leaves projection rays unchanged. Photometric quality at that view can still test appearance and view consistency, but it cannot distinguish a family of globally rescaled reconstructions.

This geometry arises naturally in event vision. DAVIS produces events and frames from the same pixel array Brandli et al. (2014), and CeleX-V combines event and full-frame imaging in one sensor Chen and Guo (2019). Beam-splitter rigs such as EDS are likewise designed to approximate co-located reference geometry Hidalgo-Carrió et al. (2022). These references are valuable for image comparison, but without independent translation they do not add metric-scale evidence.

3.3 Metric camera motion breaks the ambiguity

Consider a target camera displaced by a calibrated physical baseline $B > 0$ that does not rescale with the reconstruction. Let f be the focal length in pixels, Z the scene depth measured in the same physical units as B, and d the resulting first-order image disparity in pixels. For small viewpoint changes,

$$d \approx \frac{fB}{Z}. \tag{2}$$

The physical displacement B is fixed by calibration. If the reconstruction uses the wrong scene scale, the predicted depth-to-baseline ratio is wrong and so is the predicted parallax. A metric displaced-view evaluation can therefore carry information about global scale.

Equation (2) provides only local geometric intuition near the physical baseline; all EventGauge sweeps render the exact calibrated rigid poses in Section 4.2. Under the local sweep $B \mapsto \beta B$, $d(\beta) \approx f\beta B/Z$, so fB/Z sets the first-order geometric sensitivity to β. This sensitivity does not determine localization precision. Precision also depends on reconstruction quality, visibility, sensor differences, and the comparison score. EventGauge therefore measures response sharpness rather than inferring precision from geometry alone.

This is why metric novel-view synthesis and co-centred rendering should not be grouped together. SHARP, for example, evaluates at the true metric camera displacement Mescheder et al. (2026), so its rendering score is already sensitive to scene scale. EventGauge uses the same geometric fact but evaluates multiple displacements along the known physical baseline. The scan reveals not only the score at the true displacement, but also where the best image match occurs and whether that location is well localized.

3.4 A simple evaluation capability test

The two blind spots above are instances of the same invariance. Let $\mathcal{R}$ denote a reconstruction and its reconstruction camera trajectory. Let $G_\gamma\mathcal{R}$ denote the reconstruction obtained by multiplying scene coordinates and reconstruction-camera translations by a global scale factor $\gamma > 0$. Let E denote whatever quantity the benchmark ultimately reports. If

$$E(G_\gamma\mathcal{R}) = E(\mathcal{R}) \qquad \text{for all } \gamma > 0, \tag{3}$$

then E cannot distinguish reconstructions that differ only by that global scale. The reported quantity may still measure other properties, but it cannot support a conclusion about recovered metric scale. More generally, a benchmark can validate a physical degree of freedom only if at least one reported observation changes when that degree of freedom changes.

3.5 An audit of eleven traced protocols

We trace eleven recent event-based 3D reconstruction papers whose rendering-reference geometry can be established from the paper or released protocol. None of the reported quantitative references contributes a physical translation independent of the event reconstruction. The audit is limited to protocols whose quantitative geometry can be traced, and it does not imply that the cited methods claim metric scale. Its result is an evidence ceiling: these rendering scores can assess reconstruction or image quality, but cannot independently certify absolute physical scale. Per-paper evidence locators are in Appendix Table 2.

4 EventGauge: A Calibrated Physical Scale Audit

EventGauge is an evaluation protocol, not a reconstruction method. It leaves the estimated trajectory and scene unchanged and asks which scale information survives into reporting. It combines two

Readout	Question
$fitted\ s$ + SE(3) ATE	How are reconstruction distances converted to metric units?
true-view quality	Does the reconstruction predict the real camera image?
best-matching baseline	Which tested baseline gives the best image match?
response sharpness	How sharply is the best-matching baseline localized?

Table 1: **EventGauge separates four evaluation questions that are easily conflated.** The first row uses trajectory alignment. The remaining rows use images from the held-out physical camera.

complementary checks. The trajectory check exposes the metric conversion absorbed by similarity alignment. The image check then tests the frozen scene with a withheld physical camera at a known metric displacement and across nearby displacements.

4.1 Two complementary checks

The first level reuses the trajectory alignment but keeps all of its scale information. We report conventional Sim(3)-aligned ATE, the fitted scale s, and rigid SE(3)-aligned ATE. Here SE(3) denotes a 3D rigid transform: it allows rotation and translation, but no global rescaling. The three quantities answer different questions. Sim(3)-aligned ATE measures trajectory agreement after similarity alignment. The fitted scale records the scale needed to convert reconstruction coordinates into metric coordinates. SE(3)-aligned ATE shows how much error remains when rescaling is forbidden.

For a reconstruction that is natively metric, this conversion is simply $s = 1$. In our pose-free testbed, the fitted s metrizes the estimated camera centres but does not test the final reconstructed scene. The second level asks whether the frozen scene is consistent with that conversion when viewed from a separate physical camera. Its frames are withheld from training, pose estimation, model selection, and fitting s, and are opened only after the reconstruction and s have been frozen.

4.2 Physical baseline sweep

We write T_{ab} for the rigid transform that maps coordinates from frame b to frame a. The symbols w, e, and f denote the reconstruction world, event-camera, and held-out frame-camera coordinate frames. The calibrated event-to-frame transform T_{ef} has rotation R_{ef} and physical translation $\mathbf{t}_{ef}$. Because $s > 0$ converts reconstruction distances into metric units, $\mathbf{t}_{ef}/s$ expresses the physical translation in reconstruction units. We introduce a dimensionless baseline multiplier $\beta > 0$. We also write $\mathsf{T}(R, \mathbf{t})$ for a rigid transform with rotation R and translation $\mathbf{t}$. With these definitions, EventGauge tests the held-out camera pose

$$T_{wf}(\beta) = T_{we} \circ \mathsf{T}\left(R_{ef}, \frac{\beta \mathbf{t}_{ef}}{s}\right), \tag{4}$$

where T_{we} is the event-camera pose in the reconstruction world frame and $\circ$ denotes rigid-transform composition. At $\beta = 1$, the translation equals the calibrated physical baseline expressed in reconstruction units. Values $\beta < 1$ and $\beta > 1$ respectively shorten and lengthen that baseline while preserving its calibrated direction and rotation. Both the reconstruction and s remain fixed throughout the sweep. Re-fitting s inside the sweep would absorb the scale difference under test.

A different fitted s across separately aligned reconstructions is not itself expected to move $\hat{\beta}_Q$. A pure change of reconstruction coordinate gauge is absorbed by the corresponding fitted s, leaving the true physical camera at $\beta = 1$. The sweep instead tests whether the frozen scene remains physically consistent after that trajectory-to-metric conversion is fixed. The known-answer intervention in Section 5.1 deliberately breaks this compensation by changing reconstruction scale while holding s fixed.

4.3 True-view quality, best-matching baseline, and response sharpness

Let $Q(\beta)$ denote the image-comparison score between the view rendered at baseline multiplier β and the corresponding held-out frame image. Larger values mean a better image match. We use the same scoring rule and the same valid image region for every β. This ensures that changes in Q reflect the tested camera position rather than a changing evaluation mask. The particular score used in our experiments is specified in Section 5 and the appendix.

True-view quality. $Q(1)$ is the image-comparison score at the true calibrated camera position. It answers a simple question: can the fixed reconstruction predict the image seen by the real camera? Because this score compares final images, a poor value can be caused by geometry, appearance, occlusion, sensor-modality differences, or calibration. We therefore use $Q(1)$ as an overall true-view quality measure, not as a scale estimate by itself.

Best-matching baseline. The response maximum is

$$\hat{\beta}_Q = \arg\max_{\beta} Q(\beta), \tag{5}$$

where $\hat{\beta}_Q$ is the baseline multiplier that maximizes $Q(\beta)$. In other words, it is the tested camera position that gives the best image match. If $\hat{\beta}_Q$ is near one, the best match occurs near the true calibrated camera position. If it is far from one, the reconstruction matches a shifted camera position better than the true one. This reveals an inconsistency somewhere in the reconstruction-to-rig geometry, but it does not by itself identify the cause.

Response sharpness. A peak location is meaningful only when the response has a clear peak. If many nearby baseline values give almost the same score, the exact maximum is not reliable. We therefore report the near-peak set

$$\mathcal{W}_\delta = \{\beta : Q(\beta) \geq Q_{\max} - \delta\}, \tag{6}$$

where $Q_{\max} = \max_\beta Q(\beta)$ is the peak score and $\delta > 0$ is a fixed tolerance in the same units as Q. The set $\mathcal{W}_\delta$ contains all baseline multipliers whose scores lie within δ of the best score. We summarize its span by

$$W_\delta = \sup \mathcal{W}_\delta - \inf \mathcal{W}_\delta, \tag{7}$$

where W_δ is the near-peak width in baseline-multiplier units. A narrow W_δ means the best-matching baseline is sharply localized; a broad value means that many physically different baselines are nearly indistinguishable under the reported score. Whether $\mathcal{W}_\delta$ reaches the lower or upper sweep boundary is reported separately as censoring. A bounded interval can still be too broad to support a precise scale conclusion, whereas a censored interval gives only a lower bound on W_δ. Thus $Q(1)$ measures how well the true camera is predicted. The value $\hat{\beta}_Q$ gives the best-matching location. The width W_δ and censoring describe how precisely that location can be interpreted.

4.4 From metric NVS to a response curve

A standard metric novel-view evaluation renders at one physical displacement: the source-to-target camera motion. In EventGauge notation, its score is $Q(1)$. EventGauge instead evaluates several displacements and obtains the curve $Q(\beta)$. The curve shows which physical baseline gives the best image match, and its width shows how precisely that baseline is localized. EventGauge therefore extends metric novel-view evaluation rather than replacing it.

4.5 Known-answer scale calibration

Before using the response curve to diagnose unknown reconstruction errors, we first test it with a scale change whose effect is known in advance. Let G_γ multiply scene positions, reconstruction-camera translations, and spatial primitive sizes by a global factor $\gamma > 0$. Because the scene and reconstruction-camera motion scale together, the original monocular projections are unchanged. During this test, the previously fitted s and the physical rig baseline remain fixed.

Let $Q_\gamma(\beta)$ denote the response after applying G_γ, and let $Q_1(\beta)$ denote the response without an injected scale change. Their maximizing baseline multipliers are $\hat{\beta}_{Q,\gamma}$ and $\hat{\beta}_{Q,1}$, respectively. Geometry predicts

$$Q_\gamma(\beta) \simeq Q_1(\beta/\gamma), \qquad \hat{\beta}_{Q,\gamma} = \gamma \hat{\beta}_{Q,1}. \tag{8}$$

This gives a direct calibration test. Before inspecting any unknown failure, we know how the entire EventGauge curve should move under a controlled global rescaling. Agreement with this prediction shows that the evaluation responds to scale in the intended way. The detailed injection procedure, fixed scan range, and a control that intentionally breaks the joint rescaling are in the appendix.

5 What Becomes Visible Once Evaluation Can See Scale?

We evaluate EventGauge as an instrument before using it diagnostically. We first verify how its response moves under a known scale change and whether it can support a clean native-metric case. We then ask what standard reporting hides, whether independent physical evidence changes a controlled judgment, and when the available evidence is too broad for precise localization. IncEventGS is the primary event-only pipeline. Pooled PSNR is the primary score; Pearson re-audits the same fixed pixels and frames. The appendix reports complete values and robustness checks.

5.1 Calibrating the audit before diagnosis

Response-law calibration. We apply the global scale intervention from Equation (8) to a fixed corrected TUM-VIE `desk` reconstruction. Scene geometry and reconstruction-camera translations are scaled jointly, leaving the event-camera projections used by reconstruction effectively unchanged. Geometry predicts a horizontal shift of the EventGauge response along the β axis.

The measured response follows that prediction under two image scores. Across five injected scales, both PSNR and Pearson maxima lie within one 0.01 grid step of their score-specific predicted locations. After the predicted β-axis warp, PSNR curves agree to at most 0.0031 dB RMSE, while Pearson curves agree to below 6×10^{-5} RMSE. An intentionally broken joint-rescaling control violates both relations. A geometry-only control anchors the sweep itself: on TUM-VIE `desk`, it peaks at $\beta = 1$, with $W_{0.2} = 0.0030$ at the physical 4.6 cm baseline. Thus the control fixes the physical endpoint, while known-scale injection calibrates how the response moves under scale.

Native-metric positive control. We next ask whether the calibrated audit can positively support a reconstruction whose camera geometry is already metric. We freeze a vanilla RGB 3DGS trained only on the right TUM-VIE `mocap-desk` camera with metric poses, so $s = 1$. The left RGB camera is completely withheld and supplies a 10.94 cm evaluation-only baseline. We scan the predeclared range $\beta \in [0.5, 1.5]$. Pearson peaks at 1.00. PSNR peaks at 0.99 under an 87.1% fixed mask and at 1.00 under an independent fixed mask. All near-peak sets are bounded and contain $\beta = 1$. The 0.2-dB response width is 0.044–0.071 across the two masks, giving a sharp positive reference. Thus the audit gives a clean positive metric result without events or Sim(3) fitting.

5.2 Sim(3) reporting hides materially different selected scales

Three IncEventGS TUM-VIE `desk` runs differ only by random seed, yet the fitted reconstruction-to-metric scale spans $s = 1.51$–1.95. After Sim(3) alignment, all three trajectories still appear excellent at 1.5–2.2 mm ATE. Once rescaling is forbidden, rigid SE(3) alignment exposes 3.1–4.5 cm of residual error (Figure 2). The standard aligned score therefore assigns nearly the same quality to reconstructions that require substantially different metric conversions.

5.3 Independent physical evidence changes the judgment of a geometric correction

The released TUM-VIE path contains a concrete camera-domain mismatch. The accumulated event target is formed in an undistorted pinhole domain, while Gaussian rasterization uses the raw fisheye-calibration intrinsics. We change only the rasterizer to the same undistorted camera model as the target. Event data, random seed, training schedule, and held-out evaluation remain fixed. This creates a controlled test of whether a real displaced camera prefers a specific geometric correction that the optimized event loss does not reliably rank.

The sequence-level physical-view evidence is coherent. Median true-view PSNR improves on all five sequences. The median resolvable baseline distance moves toward the calibrated baseline on four sequences and is unchanged on `3d-trans`. Median event loss, by contrast, improves on three sequences and worsens on two. Pearson gives the same sequence-level judgment: median Corr@1 improves on all five sequences, and median $|\hat{\beta}_{\mathrm{corr}} - 1|$ decreases on all five. On `desk`, corrected peaks remain at $\beta = 1.15$–1.20 under both scores. The correction is strongly supported, but a residual physical inconsistency remains. Run-level values, shared-control bookkeeping, and the single score-sensitive case are reported in the appendix.

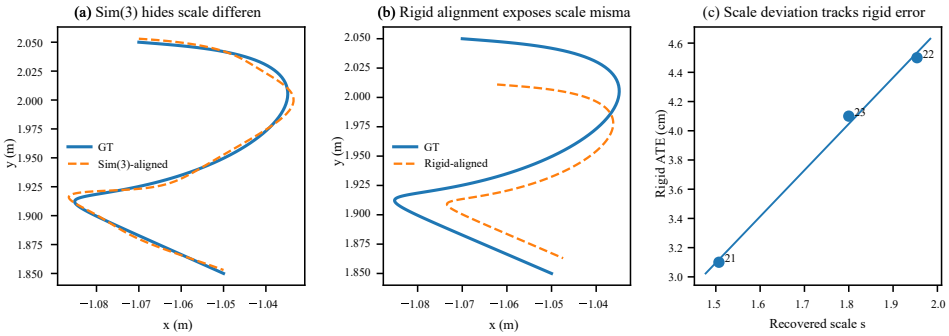


Figure 2: **Similarity alignment can hide physically different scales.** Released-path `desk` runs span $s = 1.51$–1.95. (a) Sim(3) alignment absorbs global scale differences and preserves low aligned ATE. (b) Removing scale freedom with rigid alignment exposes the resulting geometric mismatch. (c) The rigid residual tracks the scale deviation absorbed by Sim(3), showing that the hidden scale difference is physically meaningful.

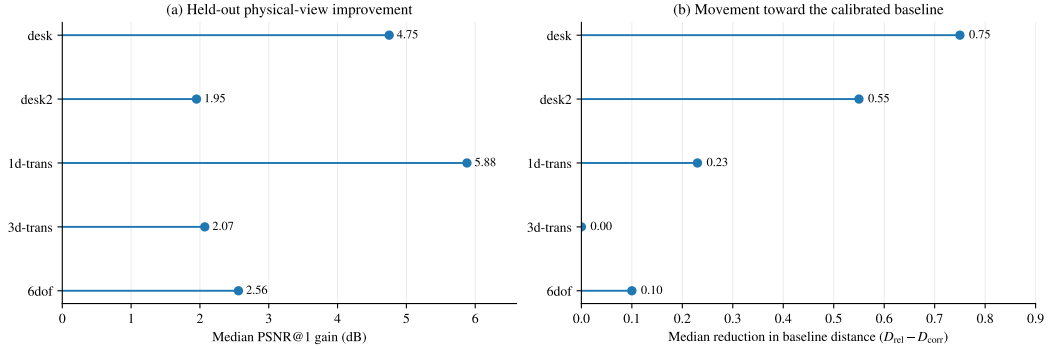


Figure 3: **Independent physical evidence changes the judgment of geometric corrections.** EventGauge evaluates corrections from a withheld physical viewpoint rather than relying only on reconstruction-space metrics. (a) Median held-out physical-view PSNR@1 improves on all five sequences. (b) The median reduction in baseline distance, $D_{\text{rel}} - D_{\text{corr}}$, is positive when the correction moves toward the calibrated baseline. Baseline-distance statistics are computed on resolved corrected pairs only: four sequences move toward the calibrated baseline and `3d-trans` is unchanged. Median optimized event loss, by contrast, improves on three sequences and worsens on two.

5.4 Geometric sensitivity does not guarantee precise localization

A known physical baseline breaks the exact scale invariance, but it does not guarantee a sharp end-to-end response. M3ED provides a second rig for this distinction. The event-only reconstruction is frozen first. Metric poses are then opened for trajectory alignment, while OVC frames remain withheld until physical-view evaluation. The physical displacement is not a sub-pixel perturbation. On masked metric-depth samples, the median first-order disparity $f_x B/Z$ is 34.1 px on `stairs` and 12.3 px on `flight_2`.

On `flight_2`, all three reconstructions attain their highest sampled PSNR at $\hat{\beta}_Q = 0.85$, but the gain over the true baseline is only 0.010–0.015 dB. The PSNR response is broad: even at $\delta = 0.1$ dB, the three widths are about 1.08–1.20. Pearson independently reproduces the same refusal case. All three Pearson curves also peak at $\beta = 0.85$ and improve over $\beta = 1$ by only 0.007–0.010. Their responses remain much broader than corrected TUM-VIE `desk` under the same Pearson width definitions. Under PSNR, the contrast is similarly large: the native-metric RGB positive control has $W_{0.2} = 0.044$–0.071, whereas the M3ED end-to-end responses span 1.25–2.45.

The geometry-only PSNR control has a 0.2-dB width of about 0.007–0.009, showing that the calibrated baseline can support a sharp preference under idealized geometric evidence. The end-to-end responses are much broader. Their width reflects the full reconstruction-and-comparison chain, so we do not assign it to reconstruction geometry alone. The operational conclusion is stronger and score-robust: the available physical image evidence does not support precise metric localization. A scale audit should expose that limitation rather than convert a repeatable argmax into a metric estimate. Threshold, censoring, disparity, and score-robustness checks are in the appendix.

6 Implications for Learned 3D Evaluation

EventGauge treats unresolved as a valid scientific outcome. A sharp, uncensored response can support precise localization; a broad or censored response cannot. Making that judgment requires keeping trajectory and image evidence separate. Fitted Sim(3) scale metrizes reconstruction-camera distances, while $Q(1)$ and $Q(\beta)$ test the frozen scene from withheld physical viewpoints.

Future evaluations should keep these channels separate. With metric trajectories, retain fitted scale and rigid error beside similarity-aligned error. With an independent metric baseline, report $Q(1)$, best-matching baseline, response width, and censoring. Broad or censored responses should remain unresolved.

7 Conclusion

A scale-invariant score cannot validate metric scale. EventGauge is a calibrated physical audit that tests a frozen scene with a withheld physical camera while keeping trajectory-scale information visible. Known-answer changes calibrate the PSNR/Pearson response, and a native-metric RGB control localizes the true physical baseline without fitted scale conversion. Across five event-reconstruction sequences, both scores support the same controlled correction, while a second rig yields a repeatable but unresolved optimum. EventGauge lets physical evidence support or withhold a metric conclusion. Reconstruction chooses a scale; evaluation determines whether physical evidence supports that choice.

Reproducibility statement

The main text gives the scientific logic, reported quantities, interventions, and headline evidence. The appendix documents the fixed evaluation protocol, protocol tracing, known-answer calibration, native-metric RGB control, camera-domain provenance, complete paired results, PSNR/Pearson robustness, and second-rig disparity, response-width, and metric-depth diagnostics. We will release the EventGauge evaluation code, frozen evaluation-frame lists, masks, baseline-sweep scripts, and protocol-audit artifact.

AI use statement

Generative AI tools assisted manuscript editing and internal translation. The authors manually verified all AI-assisted prose, citations, equations, analyses, interpretations, and claims. The authors take responsibility for the final submission.

References

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

Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 dB 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. doi: 10.1109/JSSC.2014.2342715.

Shoushun Chen and Menghan Guo. Live demonstration: CeleX-V: A 1m pixel multi-mode event-based sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3133–3142, 2020.

Fereshteh Forghani, Jason J. Yu, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, and Marcus A. Brubaker. Learn your scales: Towards scale-consistent generative novel view synthesis. *arXiv preprint arXiv:2503.15412*, 2025.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Haiqian Han, Jianing Li, Henglu Wei, and Xiangyang Ji. Event-3dgs: Event-based 3d reconstruction using 3d gaussian splatting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5781–5790, 2022.

Jian Huang, Chengrui Dong, Xuanhua Chen, and Peidong Liu. Inceventgs: Pose-free gaussian splatting from a single event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26933–26942, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315, 2024.

Hyoseok Ju, Bokeon Suh, and Giseop Kim. Have we mastered scale in deep monocular visual slam? the scalemaster dataset and benchmark. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Nikolai Kalischek, Rodrigo Caye Daudt, Torben Peters, Reinhard Furrer, Jan D. Wegner, and Konrad Schindler. Biasbed – rigorous texture bias evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22221–22230, 2023.

Yunsoo Kim, Changki Sung, Dasol Hong, and Hyun Myung. E2egs: Event-to-edge gaussian splatting for pose-free 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4922–4931, 2026.

Simon Klenk, Lukas Koestler, Davide Scaramuzza, and Daniel Cremers. E-nerf: Neural radiance fields from a moving event camera. *IEEE Robotics and Automation Letters*, 8(3):1587–1594, 2023.

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

Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18335–18346, 2023.

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amael Delaunoy, Tian Fang, Yanghai Tsin, Stephan Richter, and Vladlen Koltun. Sharp monocular view synthesis in less than a second. In *International Conference on Learning Representations (ICLR)*, 2026.

Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single colour event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4992–5002, 2023.

Maxim Tatarchenko, Stephan R. Richter, Rene Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3405–3414, 2019.

Jiaxu Wang, Junhao He, Ziyi Zhang, Mingyuan Sun, Jingkai Sun, and Renjing Xu. Evggs: A collaborative learning framework for event-based generalizable gaussian splatting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Jingqian Wu, Shuo Zhu, Chutian Wang, Boxin Shi, and Edmund Y. Lam. Sweepevgs: Event-based 3d gaussian splatting for macro and micro radiance field rendering from a single sweep. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(12):12734–12746, 2025. doi: 10.1109/TCSVT.2025.3583735.

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 (ICLR)*, 2025.

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 (ICLR)*, 2026.

Toshiya Yura, Ashkan Mirzaei, and Igor Gilitschenski. Eventsplat: 3d gaussian splatting from moving event cameras for real-time rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26876–26886, 2025.

Sohaib Zahid, Viktor Rudnev, Eddy Ilg, and Vladislav Golyanik. E-3dgs: Event-based novel view rendering of large-scale scenes using 3d gaussian splatting. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 926–934, 2025.

Zixin Zhang, Kanghao Chen, and Lin Wang. Elite-evgs: Learning event-based 3d gaussian splatting by distilling event-to-video priors. *arXiv preprint arXiv:2409.13392*, 2024.

A PROTOCOL-AUDIT TRACEABILITY

The main-paper audit asks one geometric question for each reported quantitative rendering reference: does it introduce a physical camera translation that is independent of the event reconstruction? We include recent event-based 3D reconstruction papers with quantitative rendering or novel-view evaluation when the reference-camera geometry can be traced from the paper or released protocol. Table 2 gives a compact evidence locator for each classification.

Work	quantitative reference	geometry / independent metric translation	evidence locator
IncEventGS	event-view renders; Replica GT	co-centered / same virtual camera; No	Tables 1–2; Sec. 4 (Huang et al., 2025)
EventNeRF	DAVIS frames	same DAVIS pixel array; No	real DAVIS evaluation; same sensor (Rudnev et al., 2023)
E-NeRF	DAVIS frames	same DAVIS pixel array; No	real DAVIS evaluation; same sensor (Klenk et al., 2023)
Robust e-NeRF	DAVIS frames / synthetic GT	same DAVIS / same virtual camera; No	real DAVIS + synthetic evaluation (Low and Lee, 2023)
Event-3DGS	simulated intensity	same virtual camera; No	Table 1; DeepVoxels synthetic evaluation (Han et al., 2024)
E-3DGS	synthetic frames	synthetic same camera; real cameras pose-only; No	Sec. 5.2 (Zahid et al., 2025)
EventSplat	EDS frames	beam-splitter / optically co-located; No	Secs. 5.1.2, 5.4.2; EDS supp. Secs. 4–5 (Yura et al., 2025)
SweepEvGS	DAVIS frames (GT sweep)	fast/slow sweeps, same DAVIS346C; No	Sec. IV-A; Fig. 3; Secs. IV-D–E (Wu et al., 2025)
Elite-EvGS	E2VID unseen-view image	pseudo-reference from same event stream; No	Sec. IV-A; Table II; Sec. IV-C (Zhang et al., 2024)
E2EGS	Replica synthetic frames	inherits IncEventGS Replica geometry; No	Table 1; Secs. 4.1–4.2 (Kim et al., 2026)
EvGGS	synthetic frames	same simulated camera; No	Ev3D-S synthetic evaluation (Wang et al., 2024)

Table 2: Traceability for the eleven-protocol audit in the main paper. "No" means that the quantitative rendering reference does not introduce an independently fixed metric camera translation relative to the event reconstruction. The classification concerns the geometry used by the reported quantitative metric. It does not imply that the cited work itself claims absolute metric scale.

B FIXED EVALUATION PROTOCOL

All TUM-VIE headline results use event data as the only reconstruction input. The frame stream is withheld from training, pose estimation, model selection, and fitting the reconstruction-to-metric scale s. Only after the reconstruction and s are frozen are held-out frame-camera images opened for EventGauge evaluation.

For the baseline scan, β changes only the calibrated physical baseline defined in the main paper. A single `covis_mask.npy` is computed once and reused for every β on the same evaluation frames. Headline PSNR pools masked luminance squared error over all valid pixels before converting to PSNR. Pearson correlation is pooled over the same valid pixels across all frames. It is not averaged per frame. Both scores therefore use the same fixed mask, frame set, luminance conversion, and camera sweep.

C KNOWN-ANSWER SCALE CALIBRATION

The calibration uses the fixed corrected TUM-VIE `desk` seed-21 reconstruction. No model is retrained or re-aligned. Before injection we freeze $s_0 = 1.53238$ and the physical baseline $B = 45.983$ mm. We also freeze 38 evaluation frames, held-out intrinsics and extrinsics, the fixed evaluation mask, and the scan $\beta \in [0.5, 1.8]$ with step 0.01. The injected global scale is $\gamma \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$.

The intervention multiplies Gaussian positions, reconstruction-camera translations, and Gaussian spatial sizes by γ. It leaves rotations, appearance, intrinsics, the physical camera extrinsic, and s_0 fixed. Joint scene-camera scaling preserves the reconstruction-camera projections, while Event-

γ	PSNR pred.	PSNR peak	corr. pred.	corr. peak	PSNR RMSE	corr. RMSE
0.80	0.920	0.92	0.912	0.91	0.00309	5.8×10^{-5}
0.90	1.035	1.03	1.026	1.03	0.00148	4.6×10^{-5}
1.00	1.150	1.14	1.140	1.14	0	0
1.10	1.265	1.26	1.254	1.26	0.00058	4.2×10^{-5}
1.20	1.380	1.37	1.368	1.37	0.00057	4.6×10^{-5}

Table 3: Known-answer response calibration under PSNR and pooled Pearson correlation. The PSNR predictions were frozen from a pre-injection 0.05-step scan, whose reference peak was $\hat{\beta}_{\mathrm{PSNR},1} = 1.15$; the formal calibration then used the finer fixed 0.01 grid, where the unscaled maximum is 1.14. Pearson predictions use its 0.01-grid reference peak of 1.14. Thus measured peak deviations are at most one 0.01-grid step from their pre-registered predictions. Curve RMSE is measured after the predicted β-axis warp; PSNR RMSE is in dB and correlation RMSE is unitless.

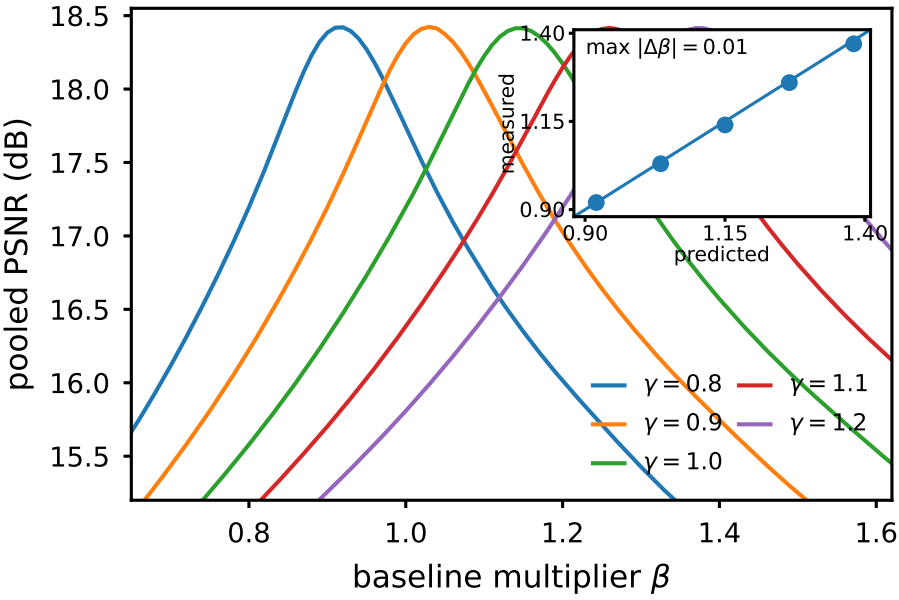


Figure 4: Complete known-answer scale calibration. The five valid global rescalings shift the Event-Gauge response by the predicted factor.

Gauge predicts

$$Q_\gamma(\beta) \simeq Q_1(\beta/\gamma), \qquad \hat{\beta}_{Q,\gamma} = \gamma\hat{\beta}_{Q,1}. \tag{9}$$

After the predicted β-axis warp, the five valid PSNR curves have correlation at least 0.999999 and RMSE at most 0.0031 dB. The five Pearson curves have warped-curve correlation 1.0 and RMSE below 6×10^{-5}. In the intentionally broken $\gamma = 1.2$ control, PSNR warped-curve correlation falls to 0.861 and RMSE rises to 1.76 dB. Pearson also breaks: warped-curve correlation falls to 0.895 and RMSE rises to 0.077. The calibration therefore tests both the expected response and a failure of the joint-scaling assumption under two image scores.

D NATIVE-METRIC RGB POSITIVE CONTROL

To test the audit outside event reconstruction and without fitted scale conversion, we use a frozen vanilla RGB 3DGS control on TUM-VIE `mocap-desk`. The reconstruction was trained only on right-camera RGB frames 90–129 with metric camera poses. The left RGB stream is withheld from training and model selection. Hence $s = 1$, and the calibrated right-to-left translation of 10.942 cm supplies the evaluation-only physical baseline. Across the 38 evaluation frames, the $\beta = 1$

virtual camera pose agrees with the calibrated left-camera pose to numerical precision. The sweep is predeclared as $\beta \in [0.5, 1.5]$ with step 0.01 and is not expanded after inspection.

We evaluate two fixed, β-invariant masks. The primary alpha-pair mask covers 87.1% of pixels; an independent frozen co-visibility mask covers 31.7%. Under the alpha-pair mask, PSNR peaks at $\hat{\beta} = 0.99$ with only 0.0045 dB gain over $\beta = 1$, while Pearson peaks exactly at 1.00. Under the co-visibility mask, both scores peak at 1.00. All near-peak sets are bounded and contain the physical baseline; $W_{0.2}$ is 0.071 and 0.044 for the two masks. A geometry-only endpoint control also peaks exactly at $\beta = 1$ with $W_{0.2} = 0.00075$. This provides a clean positive native-metric case without event reconstruction or Sim(3) fitting.

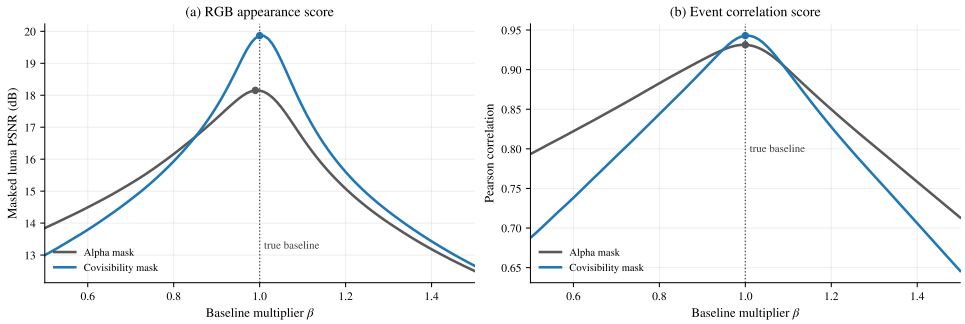


Figure 5: **Non-event native-metric positive control.** A vanilla RGB 3DGS is trained only on the right TUM-VIE camera with metric poses ($s = 1$) and evaluated on the completely withheld left camera. PSNR (left) and Pearson correlation (right) both localize the calibrated physical baseline near $\beta = 1$ under two fixed masks. The dotted line marks the true right-to-left baseline.

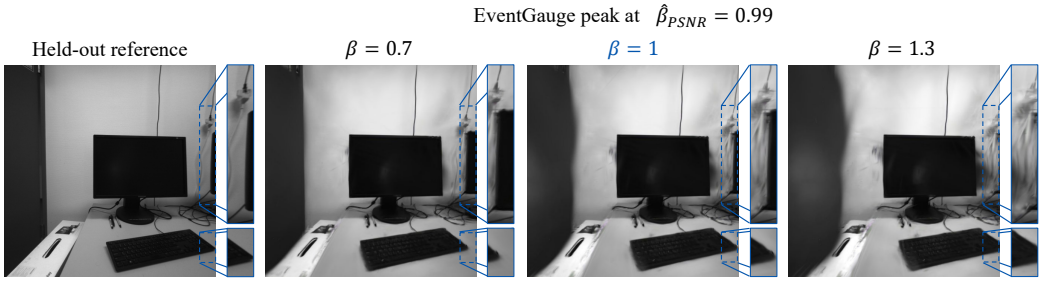


Figure 6: **Qualitative view of the native-metric RGB control.** The first panel is the completely held-out left-camera reference. The remaining panels render the frozen right-camera reconstruction at shorter, calibrated, best-matching, and longer physical baselines. The calibrated and best-matching views are visually nearly identical, while larger baseline changes produce visible parallax mismatch. This figure is illustrative; all reported response values pool the fixed evaluation frames.

E FIVE-SEQUENCE CAMERA-DOMAIN INTERVENTION

The released TUM-VIE path forms the event target in an undistorted pinhole domain while Gaussian rasterization uses the raw fisheye-calibration intrinsics. Our matched correction changes only the event-camera matrix consumed by the rasterizer from $\mathbf{K}_{\text{raw}}$ to $\mathbf{K}_{\text{und}}$. Event data, random seed, training schedule, and held-out evaluation are otherwise unchanged.

We freeze five sequences (`desk`, `desk2`, `1d-trans`, `3d-trans`, and `6dof`) and seeds 21–23. For an interior near-peak interval, let $D = |\hat{\beta}_{\text{PSNR}} - 1|$. Table 4 gives every released-to-corrected comparison. The three `3d-trans` corrected seeds share one released run, so run-level counts are descriptive rather than independent. Event-loss change is corrected minus released; negative is

lower. Main summaries are sequence-level. Baseline location uses the median paired change in D over interior corrected responses.

sequence	seed	released $\hat{\beta}$	corrected $\hat{\beta}$	D_{raw}	D_{corr}	PSNR@1 change	event-loss change	censoring
desk	21	1.90	1.15	0.90	0.15	+4.96	+0.0017	ok/ok
desk	22	2.00	1.20	1.00	0.20	+4.75	+0.0021	ok/ok
desk	23	1.95	1.20	0.95	0.20	+4.58	−0.0091	ok/ok
desk2	21	1.40	1.05	0.40	0.05	+2.20	−0.0233	ok/ok
desk2	22	1.55	1.00	0.55	0.00	+1.87	−0.0097	ok/ok
desk2	23	1.55	1.00	0.55	0.00	+1.95	+0.0173	ok/ok
1d-trans	21	0.80	1.05*	0.20	–	−4.73	+0.0995	ok/both
1d-trans	22	0.45	1.15	0.55	0.15	+7.80	+0.2617	ok/ok
1d-trans	23	1.20	1.15	0.20	0.15	+5.88	+1.2249	ok/ok
3d-trans	21	1.25	1.25	0.25	0.25	+1.71	−0.0193	ok/ok
3d-trans	22	1.25	1.25	0.25	0.25	+2.07	−0.0001	ok/ok
3d-trans	23	1.25	1.25	0.25	0.25	+2.27	−0.0263	ok/ok
6dof	21	1.20	1.05	0.20	0.05	+2.68	−0.0041	ok/ok
6dof	22	1.10	1.05	0.10	0.05	+2.49	+0.0048	ok/ok
6dof	23	1.10	1.00	0.10	0.00	+2.56	−0.0009	ok/ok

Table 4: Complete five-sequence paired results. "ok" means the 0.2-dB near-peak interval is interior to the frozen sweep. "both" means both boundaries are reached. *The discrete maximum is reported only as a locator because the corrected interval is boundary-censored.

Across the five sequences, the median physical-view PSNR change is positive for every sequence. The sequence-median baseline change moves toward $\beta = 1$ on desk, desk2, 1d-trans, and 6dof, and is unchanged on 3d-trans. One 1d-trans corrected response is boundary-censored and is not used in this location summary. The median event-loss change is negative on three sequences and positive on two, so the training-loss direction is mixed. At the individual-run level, PSNR@1 improves in 14/15 comparisons. Among the 14 interior corrected responses, 11 best-matching baselines move closer to $\beta = 1$ and three remain unchanged.

The same fixed pixels also support a pooled Pearson audit. Median Corr@1 increases on all five sequences. The sequence-median reduction in $|\hat{\beta}_{\mathrm{corr}} - 1|$ is positive on every sequence. It is 0.75 on desk, 0.55 on desk2, 0.25 on both 1d-trans and 3d-trans, and 0.10 on 6dof. At the run level, Corr@1 improves in 15/15 comparisons; 14 correlation maxima move closer to $\beta = 1$ and one is unchanged. One run is score-sensitive. On 1d-trans seed 21, PSNR is boundary-censored and decreases at the true view. Pearson instead rises from 0.445 to 0.748 and moves from $\hat{\beta} = 0.70$ to 1.05. The sequence-level physical judgment is unchanged.

F SECOND-RIG RESPONSE WIDTH ON M3ED

M3ED provides a second calibrated sensor rig with synchronized event and OVC grayscale cameras. We evaluate stairs and flight_2 with three seeds each. During reconstruction, only event-derived inputs are exposed. After reconstruction is frozen, metric poses are opened for trajectory alignment and fitting the trajectory-to-metric conversion. OVC frames remain withheld until the EventGauge image evaluation. The physical event-to-OVC baseline is 70.342 mm on stairs and 70.260 mm on flight_2.

We quantify the available first-order parallax on the same formal evaluation frames. Sparse LiDAR metric depth is projected into the OVC image and intersected with the fixed evaluation mask. For each valid sample i, $d_i = f_x B / Z_i$ is the first-order disparity magnitude induced by the physical baseline.

scene	B (mm)	d p10 / median / p90 (px)	mask coverage
`stairs`	70.342	18.40 / 34.09 / 65.27	90.01%
`flight_2`	70.260	9.49 / 12.28 / 20.06	87.28%

Table 5: M3ED physical-baseline observability. The disparity summary uses $d_i = f_x B / Z_i$ on masked LiDAR-derived metric-depth samples from the formal evaluation frames. The baseline is therefore not a sub-pixel perturbation at the observed scene depths.

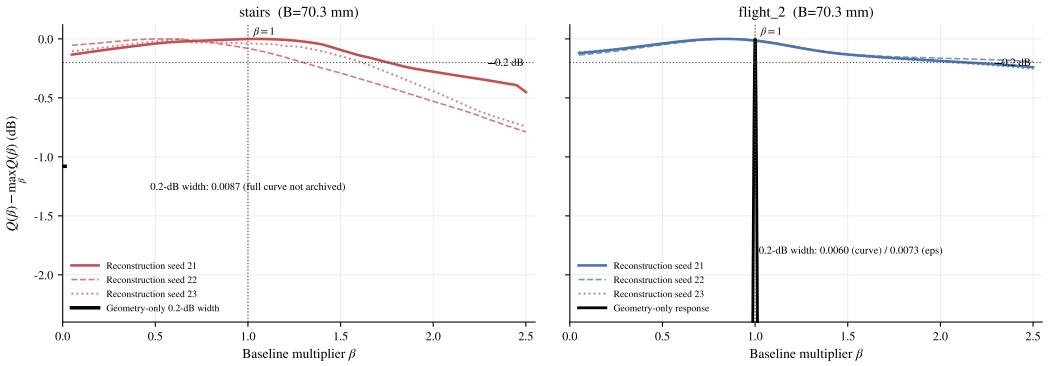


Figure 7: M3ED response sharpness under end-to-end evaluation. Reconstruction curves show $Q(\beta) - \max_\beta Q(\beta)$ for three seeds. The geometry-only response is sharply concentrated near the physical baseline, whereas the end-to-end responses can remain broad. For `stairs`, only the independently archived 0.2-dB width is available.

On `flight_2`, all three PSNR curves attain their discrete maximum at $\hat{\beta}_Q = 0.85$. The gain over $\beta = 1$ is only 0.010–0.015 dB. The response width, rather than the repeated argmax, determines whether that location supports a precise conclusion.

run	$\hat{\beta}_Q$	$W_{0.1}$	$W_{0.2}$	$W_{0.5}$	geometry-only $W_{0.2}$
TUM-VIE `desk` s21	1.15	0.047	0.093	0.160	0.0030
TUM-VIE `desk` s22	1.20	0.062	0.105	0.178	0.0030
TUM-VIE `desk` s23	1.20	0.076	0.104	0.177	0.0030
M3ED `stairs` s21	1.05	1.316	1.706^L	2.450^B	$0.0087^\dagger$
M3ED `stairs` s22	0.55	1.011^L	1.248^L	1.891^L	$0.0087^\dagger$
M3ED `stairs` s23	0.65	1.309	1.565^L	2.034^L	$0.0087^\dagger$
M3ED `flight_2` s21	0.85	1.145	2.095^L	2.450^B	$0.0073^\ddagger$
M3ED `flight_2` s22	0.85	1.078	2.450^B	2.450^B	$0.0073^\ddagger$
M3ED `flight_2` s23	0.85	1.196	2.015^L	2.450^B	$0.0073^\ddagger$

Table 6: Near-peak widths at $\delta = 0.1, 0.2, 0.5$ dB. L and B denote lower- and both-side censoring; unmarked values are bounded. At 0.1 dB, M3ED widths remain above 1.0 while TUM-VIE `desk` controls stay below 0.08. †The archived `stairs` geometry-only width uses baseline-perturbation interpolation. ‡For `flight_2`, the same interpolation gives 0.0073; direct pooled-curve interpolation gives 0.0060.

The geometry-only controls are also anchored at the physical endpoint. Every archived full TUM-VIE `desk` geometry-only curve peaks at $\beta = 1$; the physical 4.598 cm row has $W_{0.2} = 0.00303$. The archived M3ED `flight_2` full curve likewise peaks at $\beta = 1$, with $W_{0.2} = 0.00601$ at the physical 70.260 mm baseline. The absolute `stairs` peak is not reported because its full curve was not archived. At 0.2 dB, M3ED end-to-end responses are 1.25–2.45 wide and often censored, versus 0.006–0.009 for geometry-only controls. At 0.1 dB, most become bounded but remain 1.01–1.32 wide. Thus the conclusion is not tied to one threshold or censoring. Geometry-only controls show that the baseline can support a sharp preference under idealized evidence. The broader end-to-end

width can reflect reconstruction fidelity, sensor-domain appearance, or the comparison score, so we do not assign it to reconstruction geometry alone.

Post-hoc metric-depth diagnostic. As an independent geometric check, we compare the same frozen M3ED reconstructions with sparse LiDAR metric depth on the formal evaluation frames. On `flight_2`, the median ratio $Z_{\mathrm{rec}}/Z_{\mathrm{LiDAR}}$ is 0.53–0.56 across the three seeds, so the matched reconstructed depths are systematically shallower. Under a uniform depth-scale model this ratio would suggest a smaller baseline multiplier than the sampled image optimum at 0.85. Crucially, those LiDAR-implied locations are not excluded by the image evidence: they lie inside every 0.2-dB PSNR near-peak interval, whose lower boundary is 0.05 and upper boundary is 2.05–2.50. The same pattern holds on `stairs`: median depth ratios are 0.40–0.43, while the broad 0.2-dB image supports also contain those locations even though the sampled optima straddle $\beta = 1$. Thus the independent depth evidence cross-validates the refusal decision rather than the discrete argmax. The per-pixel ratios are broad, and LiDAR-valid coverage is sparse (about 4.3% on `flight_2` and 0.9% on `stairs`). We therefore treat $\hat{\beta}$ as a best-matching physical viewpoint, not a direct global depth-scale estimate.

G Score robustness across PSNR and Pearson

The primary paper uses pooled masked PSNR because its near-peak width has a direct tolerance in dB. We independently repeat the key response checks with pooled Pearson correlation on exactly the same fixed pixels and frames. Pearson is not used to select runs, masks, or thresholds. The known-answer calibration, five-sequence correction judgment, and TUM-VIE-sharp versus M3ED-broad distinction all survive the score change.

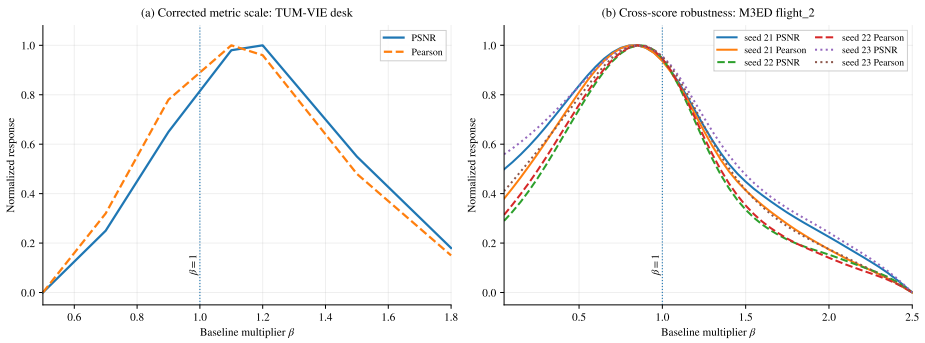


Figure 8: **Response conclusions are not specific to PSNR.** Left: on corrected TUM-VIE `desk`, normalized PSNR and Pearson responses both form a localized interior peak. Right: on M3ED `flight_2`, all three seeds remain broad under both scores, and both scores peak near $\beta = 0.85$. Curves are independently normalized only to compare shape; all peak locations and width summaries are computed from raw scores. Dotted lines mark $\beta = 1$ and the common sampled optimum where applicable.

For Pearson, we define the absolute near-peak set $\{\beta : \mathrm{Corr}(\beta) \geq \mathrm{Corr}_{\max} - \tau\}$. The table below reports a predeclared $\tau = 0.01$ summary and a score-normalized 90%-of-range width as a second scale-free view. We also checked $\tau \in \{0.005, 0.01, 0.02, 0.05\}$; the sharp-vs-broad ordering is unchanged.

case	Pearson $\hat{\beta}$	peak $-$ Corr@1	$W_{0.01}^{\mathrm{corr}}$	normalized $R \geq 0.90$ width
TUM-VIE `desk` corrected	1.15–1.20	0.032–0.042	0.136–0.154	0.297–0.325
M3ED `stairs`	0.75–1.25	0.003–0.008	0.559–0.733	0.485–0.869
M3ED `flight_2`	0.85 (3/3)	0.007–0.010	0.318–0.353	0.404–0.442

Table 7: Pearson response sharpness. The TUM-VIE corrected controls remain substantially sharper than either M3ED scene under both an absolute correlation tolerance and a normalized response-width definition. All intervals are interior for the quantities shown.