WHICH OUTPUT ABSORBS THE CONFLICT? GAUGE-AWARE ATTRIBUTION FOR FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D reconstruction models emit depth and cameras from a single pass, and we ask which of those outputs absorbs the disagreement when a scene's geometry conflicts with what the image looks like. The question is not well posed as it stands: for a model defined only up to scale a global depth rescale and a baseline rescale are observationally equivalent, so any "depth did $x\%$, cameras did $y\%$" split rests on a gauge convention the reader cannot see, and we fix one and compare only invariants. Nor is the natural statistic safe: the projection coefficient such decompositions invite is the product of directional alignment and magnitude, $\beta = M\cos\theta$, rewarding a component for being large rather than for being right. Our instrument is a ray-preserving warp: every scene point moves along its ray through the reference camera, so that image is byte-identical at every dose while the other views change. The monocular cue is therefore held fixed by construction rather than by assumption, the re-rendered views stay geometrically self-consistent with a different relief instead of being degraded, and the conflict has a dose. Across 224 measurements, 104 (46%) score $\beta > 0$ where explained variance is negative, worse than predicting no change, the worst at $\beta = +1.095$ against $R^2 = -41.9$; an earlier stage of this analysis relied on such a coefficient, and we correct it here. Under explained variance VGGT's outputs jointly reproduce 51–94% of the conflict across four doses, declining monotonically in γ, though the registered $|\log\gamma|$ form was not supported, while the relief it reports stays nearly inert (parallax-reliance slope $\rho = 0.012$, against 0.719 for RAFT triangulation with ground-truth poses at the same model-input resolution). DUSt3R and Fast3R reproduce none of it, in different ways, and the model whose outputs do reproduce it is the one with an explicit camera head, though it also differs in attention structure, so we state the architectural attribution as supported rather than established. Our identifiability claim is scoped to pointmap scale: a pre-committed rule declined to extend it to a joint disparity gauge over $f\,\|t\|$, and a normaliser-frozen control confirmed the decline was not an artifact.

1 INTRODUCTION

A feed-forward reconstruction model (Wang et al., 2024; 2025a;b; Keetha et al., 2026) is asked for a scene and a camera at once, and reports both. When those reports are mutually inconsistent with the scene, a natural question is which of them carries the inconsistency. That question is harder to pose than to ask, and two obstacles are structural rather than incidental.

Gauge. For a model whose output is defined only up to scale, the map $(d, t) \mapsto (\lambda d, \lambda t)$ changes nothing observable, so a baseline rescale and a global depth rescale are the same operation. Reporting that "the depth head did $x\%$ and the camera head did $y\%$" therefore depends on a convention the reader cannot see. We fix the gauge by normalising each solution's median depth and compare only invariants. The consequence reaches past our own measurements: the readouts conventionally used to assess these models—matching accuracy, epipolar consistency, Sampson error (Hartley & Zisserman, 2004)—are computed on correspondences or on the fundamental matrix, are invariant to this same gauge, and so cannot tell a depth rescale from a baseline rescale.

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.

Statistic. The natural way to attribute a measured change to a component—project that component's implied change onto the conflict—yields $\beta = M \cos\theta$, a product of alignment and magnitude, so a component can score $\beta \approx 1$ by being poorly aligned and large. We derive the explained-variance form $R^2 = \cos^2\theta - (M - \cos\theta)^2$, which penalises both errors, and identify the gate $M = 2\cos\theta$ where the two disagree in sign. That disagreement is not a corner case in our data: 104 of 224 component measurements fall on the wrong side of it (Section 5). An earlier stage of this analysis drew a result from β; we correct it here.

Both were caught by a protocol we recommend for any component-wise attribution: before trusting the estimator, run synthetic models with known answers through the whole of it. That self-test caught four defects, two of which would otherwise have produced clean-looking, publishable, wrong results.

With both problems settled the stimulus can do its work. Making two depth cues disagree cleanly is the remaining difficulty: perturbing an image changes the single-image channel and the parallax together, and changing the cameras changes what is visible. We instead move every scene point along its ray through the reference camera, so that image is unchanged *byte for byte* at every γ while the other views move. The monocular cue is not approximately held fixed; it is the same file. Nor are the re-rendered views degraded—they stay geometrically self-consistent with a different relief. γ gives the conflict a dose, and $\gamma = 1$ reproduces the real frames exactly, making the control a rendering identity rather than an assumption.

The empirical section applies the instrument to VGGT-1B (Wang et al., 2025a), DUSt3R (Wang et al., 2024) and Fast3R (Yang et al., 2025), chosen because they separate global cross-view attention from an explicit camera head. Only VGGT's outputs reproduce the conflict, and the architectural reading is supported against its alternative rather than established, with $n = 1$ per cell and one cell of the design empty (Section 6.2).

Finally, the paper reports a claim it failed to establish. We set out to show that an explicit camera head enlarges the gauge ambiguity into a joint disparity gauge over $f\,\|t\|$; a rule hashed before the quantities were extracted returned *undetermined* for the focal channel and *does not track* for the gauge, and a normaliser-frozen control left the null intact (Section 4.1).

Contributions. (i) A controlled cue-conflict instrument whose reference image is byte-identical across doses, so the monocular cue is held fixed by construction and the re-rendered views stay self-consistent rather than degraded. (ii) A gauge-identifiability argument: for scale-free models a depth rescale and a baseline rescale are observationally equivalent, so any component split rests on a convention. (iii) The projection-coefficient correction and the conclusion it overturned, with evidence that the hazard is routine rather than incidental, offered as a warning for component-wise attribution generally. (iv) Three-model output allocation under conflict, with pre-registered gates, one hypothesis (H_A) falsified and one cell of the design empty. (v) A pre-registered negative result on the disparity gauge, with its confound checked and the scope narrowed accordingly.

2 Related work

Feed-forward reconstruction. DUSt3R (Wang et al., 2024) regresses pairwise pointmaps and aligns them in a separate global step; Fast3R (Yang et al., 2025) removes that step, passing all views through all-to-all attention and recovering poses with a solver; VGGT (Wang et al., 2025a) predicts cameras, depth, pointmaps and tracks through separate heads, including an explicit camera head; π^3 (Wang et al., 2026) drops the reference view altogether; Depth Anything 3 (Lin et al., 2026) collapses the outputs to a single depth-ray target. What matters here is not accuracy but *arity*: these models emit several geometric quantities at once, and nothing requires those quantities to degrade together. Our three span the two axes public weights let us vary—pairwise versus global cross-view attention, and an explicit camera head versus a pose solver.

Geometric understanding versus learned priors. The closest work is Bratulić et al. (2025). Probes on intermediate features recover the fundamental matrix from DUSt3R, VGGT and DA3, emerging in middle layers behind correspondence-matching attention heads that attention knock-out shows to be causal. A second part perturbs VGGT's inputs and finds it robust where classical

pipelines are not, concluding that the model supplies missing correspondences from learned priors—robustness bought at the cost of geometric consistency (figures compared in Section 6.3).

That conclusion is our premise, not our target, but it is conditional on its measurement regime. The readouts establishing it—matching accuracy, Sampson error—are invariant to the scale gauge (Section 4), so they cannot tell a depth rescale from a baseline rescale.

Corruption is not conflict. Their manipulations remove or corrupt evidence: occluded patches carry no correspondence information, so "the evidence got worse" and "the prior took over" are observationally hard to separate. Our ray warp instead leaves the reference image byte-identical and supplies multi-view evidence fully self-consistent with a *different* relief—not degraded, but different. Robustness and allocation are therefore orthogonal axes: a model may be robust while misallocating, or fragile while allocating correctly. Their experiments measure robustness; ours measures allocation. Work that does pose conflicts scores whether depth goes wrong, under physically unsupported edge cues (Zhang et al., 2026) or 3D visual illusions (Yao et al., 2025), not which output absorbs the conflict.

Their prior-reliance half is also VGGT-only, and its manipulations are conditions, not doses: that priors dominate is established, but not *how much*, as a function of departure from the prior; our instrument returns that.

Probing internal representations. A broader line asks what geometric information a backbone carries (El Banani et al., 2024; Chen et al., 2025; An et al., 2025); closer to our models, probes read a DUSt3R variant's residual stream (Starý et al., 2025), one VGGT layer separates distractor views (Han et al., 2026), and another serves as a retrieval verifier (Maggio & Carlone, 2026). Together these establish that the geometric machinery is present, localised and exploitable.

Bratulić et al. (2025) separates work offering *observational evidence at the output* from mechanistic analysis inside the network, and on that axis this paper is plainly output-level. The axis does not order the two questions: internal probing establishes that a capability is present, but cannot establish how a disagreement is allocated, for the gauge reason above; our instrument, conversely, cannot see inside. The two compose at one point: since correspondence machinery has been shown causally load-bearing for these models' epipolar estimates, the inertness we measure is not attributable to multi-view evidence being absent from the input, because our manipulation keeps that evidence present and self-consistent.

Priors by design, and attributing a response. MP-SfM (Pataki et al., 2025) and MonST3R (Zhang et al., 2025) fuse monocular priors into geometry deliberately: they specify where the prior governs, whereas we measure where the trade falls in a model not built to expose it. On the deployment side, DriveVGGT (Jia et al., 2025) reports VGGT's absolute scale unreliable in driving and repairs it with calibration constraints, and Hillemann et al. (2026) find its raw confidences need thresholding; our results suggest that scale error has structure worth modelling, not only correcting.

3 METHOD

3.1 A CONFLICT WITH A BIT-IDENTICAL REFERENCE

For reference camera C_1 with ground-truth z-depth D, every scene point moves along its own ray through C_1:

$$g(z) = d_0 \left(\frac{z}{d_0}\right)^{\gamma}, \qquad X' = X\,\frac{g(z)}{z}, \qquad d_0 = \text{median valid depth}. \tag{1}$$

X' stays on its ray, so the reference image is the original file, byte for byte, at every γ. Only the other views change. Source views are re-rendered from each real frame's own RGB-D by forward splatting ($2\times$ supersampled z-buffer, background-favouring hole fill) with a deterministic z-tie break, so $\gamma = 1$ reproduces the real frames exactly: zero holes, mean $|\text{render} - \text{real}| = 0.012$ from the matched low-pass alone. Figure 1 shows the manipulation and its size. The manipulation is therefore a pure change of relief with the monocular cue held fixed, which is what separates it from occlusion, lighting or symmetry perturbations used elsewhere (Bratulić et al., 2025; Zhang et al., 2026).

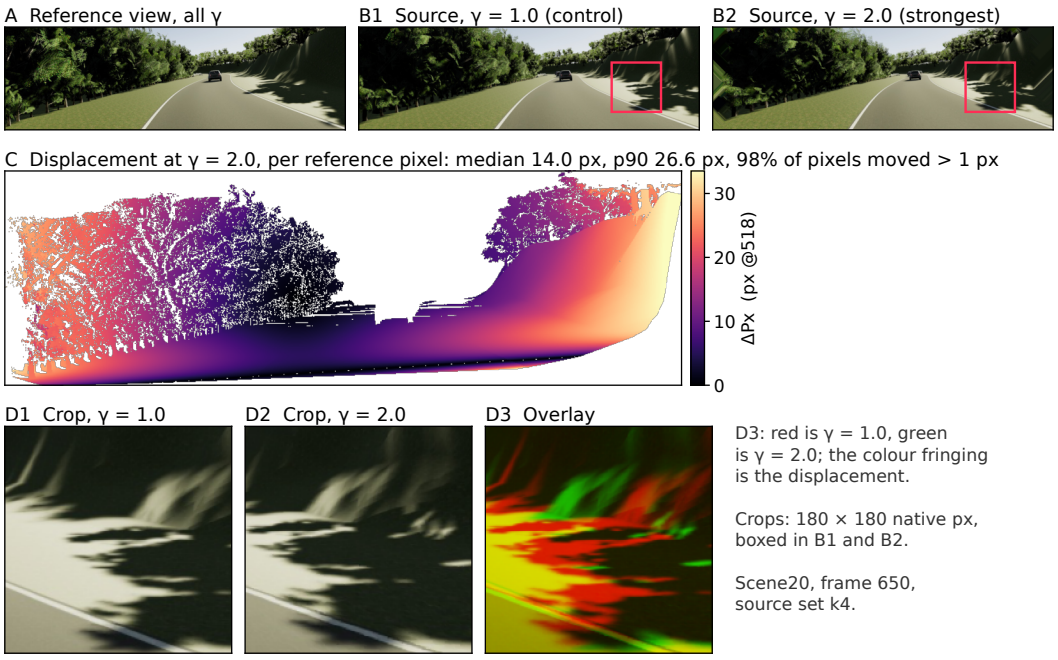

Figure 1: The cue-conflict stimulus. **A** the reference view, byte-identical at every γ. **B** the same source view at the control and the strongest dose. **C** the displacement the warp introduces per reference pixel: median 14.0 px, p90 26.6 px at the 518-px input, 98% of valid pixels moved over 1 px. **D** a native-resolution crop and red/green overlay. The manipulation sits near the perceptual threshold by design: a warp obvious in a side-by-side would push the source views out of distribution and break the $\gamma = 1$ identity.

3.2 Visibility, not clamping

Every projection discards points behind the source camera rather than clamping them. This is not a detail: a single reference pixel whose warped point falls behind a camera, with Z clamped to 10^{-6}, contributes $\sim 10^{12}$ to a pooled sum and silently swamps every other frame. The symptom is a decomposition that returns *exactly* 0.000 with a $[0, 0]$ confidence interval.

3.3 A validation protocol for component-wise attribution

The quantity of interest is how much of the conflict each output head reproduces. We recommend a general precaution for any such measurement: before trusting a component-wise attribution, run synthetic models with known answers through the entire estimator. A depth-follower, a pure baseline-rescaler and a no-change model have analytically known decompositions; an estimator that misreports them is not measuring what its name says, and no amount of plausible-looking output on real models will reveal it.

Applied here, the protocol caught four defects before any model result was believed—two of which would otherwise have produced clean-looking, publishable, wrong results. A *gauge leak*: removing the scale gauge by rescaling only the γ translation made a synthetic depth-follower with *unchanged cameras* report a spurious camera term of -1.03. *Non-additivity*: swapping components one at a time and summing is not swapping them jointly, so for a metric depth-follower the single swaps overshoot and cancel (shape $+1.47$, ratio -1.03, interaction $+0.55$) while the joint swap is exact, making a gate on the sum ill-posed. And the two described elsewhere—the unguarded projection above, and the statistic itself (Section 5).

Final form (Figure 3): put each condition's solution in the canonical gauge (median depth over the mask $= 1$) and compare only gauge-invariant components—normalised relief $\hat{s}$, baseline-to-depth

ratio $r = \|t\| \,/\, \mathrm{median}(d)$, rotation and translation direction, intrinsics. Report

$$R_x^2 = 1 - \frac{\sum \|\Delta_{\mathrm{true}} - \Delta_x\|^2}{\sum \|\Delta_{\mathrm{true}}\|^2} \tag{2}$$

with $\cos_x$ and mag_x alongside, always. Single swaps are attribution only and are explicitly non-additive.

3.4 The parallax-reliance slope, and the geometry reference

Two quantities carried over from this study's pilot are quoted in Section 6 as cross-checks, and are defined here for completeness.

The slope ρ. For each reference pixel p, let

$$\Delta\log(p) = \log \hat{D}_\gamma(p) - \log \hat{D}_1(p), \qquad x(p) = (\gamma - 1)\big(\log D(p) - \log d_0\big), \tag{3}$$

where $\hat{D}$ is the model's own reference depth and D the ground-truth depth. $\Delta\log$ is the model's change in reported depth between a dose and its own $\gamma = 1$ control, which cancels its per-pixel bias; x is the change the warp imposes. ρ is the OLS slope of $\Delta\log$ on x with a free intercept, the intercept absorbing global scale so that uncalibrated outputs are handled. $\rho = 1$ describes reported depth that tracks the manipulated parallax exactly; $\rho = 0$ describes reported depth unchanged by it. Values are computed per reference frame and aggregated, with 95% CIs by cluster bootstrap over scenes.

The geometry reference. Where Section 6 quotes ρ for *explicit triangulation*, the pipeline is RAFT-large optical flow (Teed & Deng, 2020) from the reference to each source view, combined with ground-truth relative pose and midpoint triangulation (Hartley & Zisserman, 2004). It is a validity check on the stimulus rather than a competing method: it establishes that the manipulated parallax is recoverable from the rendered views by image correspondence and known camera geometry alone, with no learned depth.

We quote $\rho = 0.719$ $[0.53, 0.87]$, measured at the 518-px input the models themselves receive, because the model values it is set against are measured there too; at native resolution the same pipeline gives 0.853. The gap is not cosmetic—at 518 px the lowest-parallax bin falls to 0.46—so 0.719 is both the like-for-like comparator and the weaker of the two.

4 Gauge identifiability: why "depth versus camera" is not a well-posed split

For a model whose output is defined only up to scale, the map $(d, t) \mapsto (\lambda d, \lambda t)$ leaves every projection unchanged (Triggs et al., 2000). A baseline rescale and a global depth rescale are the same operation. Any decomposition that reports "the depth head did $x\%$ and the camera head did $y\%$" therefore depends on a gauge convention, and different conventions move the split arbitrarily. Figure 4 draws the two solutions and lists what survives the convention.

This is not a corner case: in our earlier data the convention was nearly harmless only because the models barely changed their depth ($\lambda = 0.997$–1.002), which is why the defect went unnoticed until a self-test used a synthetic model that *did* change its depth.

Recommendation. Fix the gauge by normalising each solution's median depth to 1, then compare only invariants: relief shape, r, rotation, direction, intrinsics. Under the alternative convention—normalising by $\|t\|$ instead—a pure relief change is attributed to the depth term and a pure baseline change to the depth term as well, because the normaliser itself moves; the invariant comparison is stable under both.

4.1 Scope of this claim: pointmap scale only

The ambiguity established here is the pointmap scale gauge $(d, t) \mapsto (\lambda d, \lambda t)$. We do not claim that an explicit camera head enlarges it into a joint disparity gauge over $f\,\|t\|$, and the data do not support writing it.

Testing that required a pre-committed rule, since the question decides the section's scope. Writing $\Phi = f_0 \|t\| / \mathrm{median}(\hat{d})$ for the predicted disparity at the median depth, a broad claim needed the focal channel to move with the conflict *and* Φ to track what the warp requires. Measured on VGGT across 69 frames and four doses, the focal clears a 2% change with a CI excluding zero at one of four doses where the rule required two, and the tracking slope of $\Delta\log\Phi$ is -1.779, CI $[-3.293, -1.158]$—the opposite sign to the requirement, not merely absent (Figure 7).

So the focal verdict is *undetermined* and the gauge does not track; by the rule fixed beforehand, the identifiability claim stays at pointmap scale.

The null is clean, not confounded. Φ carries the model's own predicted depth in its denominator, and the depth channel is the one output that has moved throughout, so the negative slope could have been a normaliser artifact. Freezing the normaliser—so the constant cancels and $\Delta\log\Phi^* = \Delta\log f_0 + \Delta\log\|t\|$ is the camera channels' own change—leaves the slope essentially unchanged at -1.869, CI $[-3.377, -1.210]$. Decomposing it locates the effect entirely in the translation, while the focal's own slope includes zero; a gauge check rules out the mirror-image artifact, since $\Delta\log\|t\|$ co-moves with $\Delta\log\mathrm{median}(\hat{d})$ far from the $+1$ that global scale drift would produce (Appendix D).

Two observations are therefore recorded without being claimed: the disparity gauge moves *against* the requirement at three of four doses, with its confound checked rather than merely noted; and the change is carried by the translation channel, not the focal—consistent with Section 6, where the intrinsics term never carried the change either.

The broad claim is the result *we set out to establish*; a threshold hashed before the quantities were extracted killed it, and the narrow claim is what survived.

5 A PROJECTION COEFFICIENT IS NOT A FRACTION EXPLAINED

The natural-looking statistic

$$F_x = \frac{\langle \Delta_x, \Delta_{\mathrm{true}} \rangle}{\|\Delta_{\mathrm{true}}\|^2} = \cos(\Delta_x, \Delta_{\mathrm{true}}) \cdot \frac{\|\Delta_x\|}{\|\Delta_{\mathrm{true}}\|} \tag{4}$$

is a regression coefficient. It inflates whenever a large, weakly aligned field is projected onto a small one. We first read "the cameras explain 63–81% of the conflict" from it, and withdrew that reading: a direction probe showed the models' camera-induced fields are 2.5–$3\times$ larger than the conflict itself, with per-frame alignment ranging from $+0.21$ to -0.67.

This is not a property of the one cell we got wrong. Across the dose sweep—three models, sixteen dose $\times$ view-set cells each, five gauge-invariant terms per cell (four for Fast3R, whose intrinsics term is excluded), 224 measurements in all—104, or 46%, fall where $\beta > 0$ while $R^2 < 0$ (Figure 2). Every one of those would be reported under a projection coefficient as a positive contribution, and every one is in fact worse than predicting that the output did not move. The hazard is the common case in this data, not the exception, and we found it only because the two statistics were computed side by side.

The extreme bounds how far apart the two can be: in DUSt3R's `k2_g0.7` cell the joint swap scores $\beta = +1.095$ with $R^2 = -41.9$, from a field $6.6\times$ the size of the conflict at $\cos\theta = 0.165$—enormous, almost orthogonal, and scoring as a near-perfect explanation. No amount of care in *interpreting* such a β recovers the fact that the field it came from reproduces none of the conflict.

5.1 WHY β REWARDS SIZE

Write $M = \|\Delta_x\| / \|\Delta_{\mathrm{true}}\|$ for the magnitude ratio and θ for the angle between the candidate field and the conflict, so that $\beta = \langle \Delta_x, \Delta_{\mathrm{true}} \rangle / \|\Delta_{\mathrm{true}}\|^2 = M\cos\theta$. β is a *product* of alignment and magnitude, so any β can be reached by a badly aligned field that is large enough. It is not bounded above by 1 and carries no information about how much of the conflict is actually reproduced.

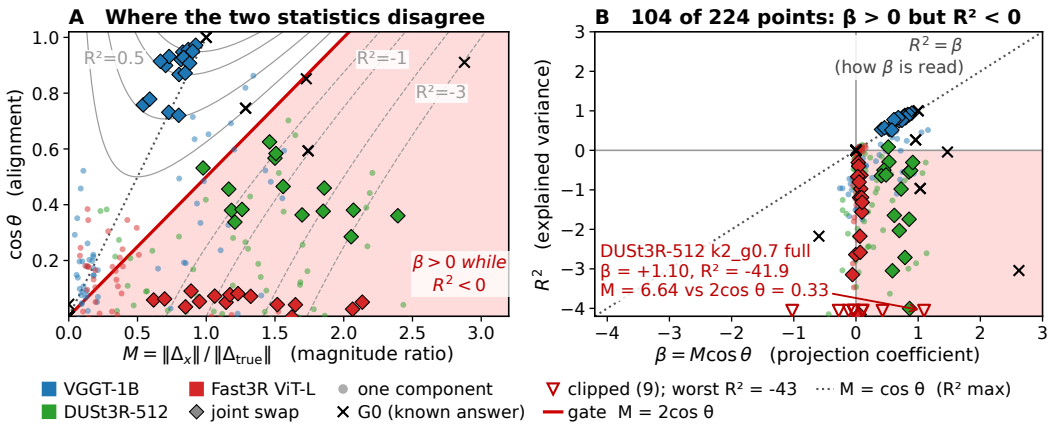

Figure 2: β rewards size, not fit. **A** the $(M, \cos\theta)$ plane with R^2 contours, the gate $M = 2\cos\theta$, and the wedge where $\beta > 0$ while $R^2 < 0$ shaded. **B** the same fields under both statistics; 104 of 224 measured points fall in the shaded region. Nine lie below the axis and are drawn as clipped triangles rather than dropped.

Explained variance does not factor that way. Expanding $\|\Delta_{\text{true}} - \Delta_x\|^2 = \|\Delta_{\text{true}}\|^2 (1 + M^2 - 2M\cos\theta)$,

$$R^2 = 1 - \frac{\|\Delta_{\text{true}} - \Delta_x\|^2}{\|\Delta_{\text{true}}\|^2} = 2M\cos\theta - M^2 = \cos^2\theta - (M - \cos\theta)^2. \tag{5}$$

The second form separates the two failure modes: $\cos^2\theta$ caps what alignment can buy, and $(M - \cos\theta)^2$ is the penalty for being the wrong size. A field can only score near 1 by being both well aligned *and* correctly scaled; β requires neither.

Where β inverts. From Equation 5, $R^2 > 0$ iff $M < 2\cos\theta$. Crossing that gate makes R^2 negative—worse than predicting no change—while β keeps rising with M. The two statistics disagree in sign wherever $\cos\theta > 0$ and $M > 2\cos\theta$.

The mechanism in isolation. The sweep above shows how often the two statistics part company; this shows why, on a case where the right answer is known in advance rather than measured. Our estimator self-test includes a model that changes only its baseline, which cannot reproduce a relief change. At $\gamma = 0.7$ it produced $M = 1.741$, $\cos\theta = 0.593$:

- $\beta = 1.741 \times 0.593 = +1.032$—indistinguishable from a perfect explanation;
- $R^2 = 2(1.741)(0.593) - 1.741^2 = -0.966$—worse than predicting nothing;
- the gate sits at $2\cos\theta = 1.186$, and this cell crosses it by $+0.555$.

Both figures come from the same field. Reported as β, a model that reproduces none of the conflict looks like one that reproduces all of it.

A note on the baseline. Our R^2 is *uncentred*: the null is "the outputs did not move" ($\Delta_x = 0$), not "the outputs moved by the average amount". We measured the alternative and rejected it—the conflict field has a systematic mean, since forward motion expands the image about the focus of expansion (mean displacement $(-0.52, -0.51)$ px against an RMS of 3.72 px, so $|\text{mean}|/\text{RMS} = 0.196$). Centring would credit a candidate for reproducing that mean, which is itself part of the signal under test. So Equation 5 is the coefficient against a no-change null, and it does not coincide with the centred form here.

Under R^2, one model's apparent "absorption" vanished entirely—no DUSt3R cell clears the 0.4 gate, 15 of 16 point estimates are negative, and the sixteenth $(+0.085)$ has a confidence interval spanning zero—while another's survived at a different magnitude (VGGT). The qualitative conclusion changed for one model and the quantitative claim changed for both.

General warning. A component-wise decomposition reported without (i) an explained-variance statistic, (ii) alignment and magnitude reported separately, and (iii) a test for non-additivity, can produce catastrophic false positives that look like clean results.

6 RESULTS

6.1 RELIEF IS INVARIANT; ONLY ONE MODEL'S OUTPUTS REPRODUCE THE CONFLICT

The multi-view evidence remains geometrically self-consistent under the manipulated scene, yet all three models' relief outputs remain largely invariant to the imposed parallax conflict: $\rho = 0.012$ (VGGT) and 0.047 (DUSt3R) against explicit triangulation at 0.719, measured at the same 518-px input resolution; shape $R^2 \leq +0.10$ (VGGT) and ± 0.09 (Fast3R).

We report this first because it is the least surprising thing here, not the most: that these models' outputs track appearance over parallax is established (Section 2), and what ρ adds is only that the same conclusion follows from a manipulation with no pixel change at all. It is a check on the instrument, not a finding of ours.

The substantive result is what the models' outputs *jointly* reproduce, and it differs sharply. Over the 16 dose $\times$ view-set cells, VGGT (global interleaved attention, explicit camera head) spans $+0.513$ to $+0.942$ with every interval clearing the pre-registered 0.4 gate; Fast3R (global all-to-all, PnP) spans -4.342 to -0.309 with every interval below zero; and DUSt3R (pairwise, aligner) spans -41.926 to $+0.085$ with no interval clearing 0.4 (Table 1 and the per-term breakdown in Figure 5).

The registered dose condition was not met, and we report that rather than the claim we expected. G1, fixed before any dose result was read, required R^2_{full} to decline monotonically in $|\log\gamma|$, on the rationale that a camera-carried explanation must degrade as the conflict stops resembling a global scale change. It does not. $|\log\gamma|$ pairs $\gamma = 0.5$ with $\gamma = 2.0$, and in all four view sets those cells differ with disjoint confidence intervals—by 0.36, 0.37, 0.18 and 0.31 R^2 respectively, the largest being $+0.907$ against $+0.535$ at `k2`. The compression member scores *higher*, everywhere. Pooled over the sixteen cells, R^2_{full} is essentially uncorrelated with $|\log\gamma|$ and only weakly related to the conflict's size in pixels; what it tracks is γ itself, declining monotonically from compression to expansion in 4/4 view sets—which we report as an unregistered observation, not as a confirmed prediction (Appendix C, Figure 6).

The other two G1 conditions were met: $R^2_{\text{full}} \geq 0.4$ at every dose (minimum 0.513, every CI clearing 0.4) and $R^2_{\text{shape}} < 0.2$ (maximum $+0.100$). We therefore record G1 as partially met—its two substantive conditions held, its dose sub-condition did not—rather than as a clean pass, and we state the sub-condition wherever the gate is cited. The failure mode the dose condition was written to detect—a flat response, indicating the decomposition tracks a degeneracy rather than a behaviour—is not what occurred, since R^2_{full} varies by 0.18–0.37 across doses with non-overlapping intervals; but the condition as registered was a magnitude claim, and a magnitude claim is what the data decline to support.

The response is strongly dose-dependent, then, and simply dependent on the *direction* of the relief change rather than its size—which we did not predict. Why compression should be reproduced better than expansion is open, and the obvious candidate—that expansion pushes depth further from the distribution the model was trained on—is untested here and would need its own registration (Section 7).

6.2 THE ARCHITECTURAL TEST

Fast3R was chosen because it fills the (global attention, no camera head) cell: verified in source before the prediction was registered—two DPT (Ranftl et al., 2021) pointmap heads and no pose regression anywhere, with all-to-all attention over concatenated view tokens.

Outcome: H_A falsified; the architectural reading is supported against its alternative, not established. Fast3R's outputs do not reproduce the conflict, so global cross-view attention alone is not sufficient for the reproduction we measure. The model whose outputs do reproduce it is the one that also has an explicit camera head—but that reading carries the caveat that the two cells differ

in attention structure as well as in camera head, and $n = 1$ per cell. H_A was not tested against a matched-attention control.

Two alternatives were fixed before the data: had Fast3R reproduced the conflict, H_C would have died and the search would have moved to training data, view count and model scale; had it landed in $0.2 \leq R^2_{\text{full}} \leq 0.4$, no architectural claim would have been made.

Limits stated with the claim. Three, and the third bounds the reading hardest. $n = 1$ model per cell. The (no attention, camera head) cell is empty, with no candidate model. And the axis is coarser than "global versus pairwise" suggests: VGGT interleaves global self-attention with frame-wise layers where Fast3R applies all-to-all attention throughout, so a reader is entitled to the alternative that interleaving, rather than the camera head, is what distinguishes VGGT. We cannot separate the two with public weights, and do not claim to. Separately, Fast3R's negative R^2 is carried entirely by the rotation-and-direction term, at alignment $\cos$ between 0.00 and 0.09—uncorrelated with the conflict rather than opposed to it—which is consistent with an unstable solver-derived pose rather than a stable output change of any kind.

6.3 Discussion: for VGGT, relief is the invariant channel

Across every probe we aimed at VGGT one channel stayed put and the others moved around it: the relief is inert under the conflict and stays fixed while the disparity gauge moves through the translation channel rather than the focal. For VGGT, relief is the invariant channel and the measured change appears in the others—what one would expect of a model that regresses cameras explicitly, and consistent with Section 6.2. This is a statement about VGGT, not about feed-forward reconstruction: the other two behave differently in kind, DUSt3R's pointmap destabilising (no cell clears the gate, 15 of 16 point estimates negative) while Fast3R's holds and its PnP-derived pose moves without tracking the conflict.

Stated against the prior result. The interpretability literature closes on a trade: learned priors buy robustness at the cost of geometric consistency, read off as a consistency score—for VGGT under occlusion a median Sampson error that barely moves, $0.61 \rightarrow 0.65$ px, where classical pipelines degrade several-fold (Bratulić et al., 2025). Our reading refines that in one respect: the cost is not spread across the outputs, and it is not absent because the score is small—it is *relocated*. Under a conflict that leaves the reference image untouched, the reported relief stays within ± 0.10 R^2 of no change while the baseline-to-depth channel moves enough for the outputs jointly to reproduce 51–94% of the conflict. A correspondence-space scalar cannot see this, being invariant to the gauge along which the relocation happens. The trade-off has an address, and it is the channel a downstream consumer is most likely to trust and least likely to check: camera and depth stability can diverge silently under geometric conflict, so systems relying on the camera estimate should monitor it independently of depth rather than through a scalar consistency check.

7 Limitations

Our evaluation uses one synthetic driving domain (Cabon et al., 2020), forward-motion sequences, three models, and 518-px inputs; the main empirical conclusion is centered on VGGT. The architectural comparison is not causal: the contrasted models differ in attention structure as well as in the presence of an explicit camera head, with $n = 1$ model per design cell, so the architectural reading is output-level rather than mechanistic. We likewise make no claim about internal computation beyond published evidence from prior work. One registered prediction was not met: R^2_{full} did not decline monotonically with conflict magnitude in $|\log \gamma|$; the observed directional asymmetry is reported as an unregistered observation without a mechanistic explanation. Finally, two ancillary probes were inconclusive—one downgraded to a no-kill-authority check before it ran, for five specification defects, and a blur probe that destabilized the camera estimate before conflict doses could be meaningfully tested; these do not support additional mechanism claims.

AI use statement

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools for research ideation (the cue-conflict design and the ray-preserving warp), for designing the methodology (the pre-registrations, their gates and kill criteria, and every dated amendment), for implementing the methods (all analysis code), for generating the synthetic stimuli (which that code renders deterministically from vKITTI2 RGB-D), for formulating the mathematical claims and writing the derivations in Sections 4 and 5.1, and for interpreting results. We have not used generative AI tools for collecting or cleaning the underlying dataset, which is pre-existing and unmodified, nor for training or fine-tuning any model; translating text and analysing qualitative data are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: writing code, creating all figures, drafting sections of the paper, summarising literature, brainstorming, and editing for readability.

We have reviewed all AI-assisted work. Before any model output was read, an estimator self-test ran synthetic models with analytically known decompositions through the entire pipeline; it caught four defects, two of which would otherwise have produced clean-looking, publishable, wrong results. Every numerical claim was re-verified against stored result files rather than against previously generated text, and every citation identifier was resolved against its listing before being written. Code, result files and figures are hashed with dated entries. The scientific judgements—including the decision to retain G1 rather than invoke its kill rule—are the authors', and are recorded with the arguments against them.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The stimulus generator, the estimator, the self-test, the analysis scripts and the scripts for every data figure are included in the supplementary material, with the result files they read. Every stimulus is rendered deterministically from vKITTI2 RGB-D with a fixed tie-break, so the cache is reproducible byte-for-byte; all bootstrap procedures are seeded. The pre-registrations and their amendments, including the corrections issued after results were seen, will be released after the review period; their SHA-256 digests are included in the supplementary material so the released documents can be checked against them.

References

Honggyu An, Jinhyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1103–1115, 2025. doi: 10.1109/CVPR52734.2025.00111.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3D reconstruction models. *arXiv preprint arXiv:2512.11508*, 2025. Presented at ECCV 2026.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2GS: Probing visual foundation models with Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6348–6361, 2025. doi: 10.1109/CVPR52734.2025.00595.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, 2024. doi: 10.1109/CVPR52733.2024.02059.

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–437, 2026.

Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

Markus Hillemann, Robert Langendörfer, Steven Landgraf, and Markus Ulrich. Uncertainty quality of VGGT: An analysis on the DTU benchmark dataset. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XI-2-2026:665–672, 2026. doi: 10.5194/isprs-annals-XI-2-2026-665-2026.

Xiaosong Jia, Yanhao Liu, Yu Hong, Renqiu Xia, Junqi You, Bin Sun, Zhihui Hao, and Junchi Yan. DriveVGGT: Calibration-constrained visual geometry transformers for multi-camera autonomous driving. *arXiv preprint arXiv:2511.22264*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509, 2026. doi: 10.1109/3DV69130.2026.00054.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.10647.

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. In *Proceedings of Robotics: Science and Systems (RSS)*, 2026. arXiv:2601.19887.

Zador Pataki, Paul-Edouard Sarlin, Johannes L. Schönberger, and Marc Pollefeys. MP-SfM: Monocular surface priors for robust structure-from-motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21891–21901, 2025. doi: 10.1109/CVPR52734.2025.02039.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168, 2021. doi: 10.1109/ICCV48922.2021.01196.

Michal Starý, Julien Gaubil, Ayush Tewari, and Vincent Sitzmann. Understanding multi-view transformers. In *ICCV 2025 End-to-End 3D Learning Workshop*, 2025. arXiv:2510.24907.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58536-5_24.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski (eds.), *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer Berlin Heidelberg, 2000. doi: 10.1007/3-540-44480-7_21.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. doi: 10.1109/CVPR52734.2025.00983.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, pp. 10481–10497, 2026.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025. doi: 10.1109/CVPR52734.2025.02042.

Chengtang Yao, Zhidan Liu, Jiaxi Zeng, Lidong Yu, Yuwei Wu, and Yunde Jia. 3D visual illusion depth estimation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 78267–78299. Curran Associates, Inc., 2025. doi: 10.52202/085713-2622.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, pp. 82863–82886, 2025. arXiv:2410.03825.

Wentao Zhang, Jinhu Qi, Weiqiang Jin, Yifei Zhang, Chan-Tong Lam, and Irwin King. Geometric collapse: When vision models fail to verify physical causality. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2607.06871.

A THE INSTRUMENT, END TO END

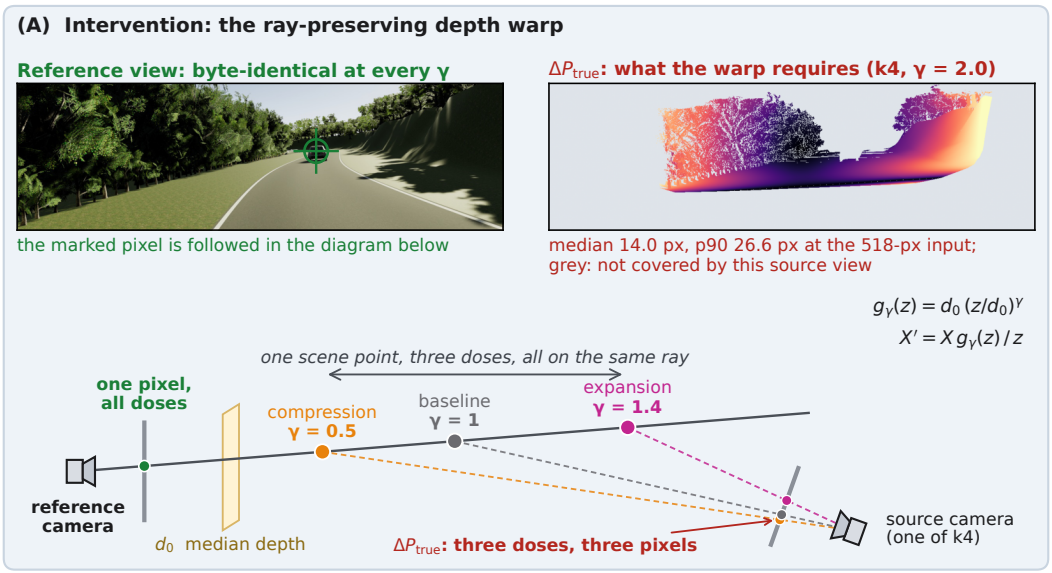

Figure 3: The instrument. **(A) The intervention.** Every scene point moves along its own ray through the reference camera, so the three doses of a single point project to *one* reference pixel and the reference image is the original file, byte for byte, at every γ; seen from a source camera the same three points land in three different places, and that separation is the conflict ΔP_{true} the estimator must account for. In the schematic the source crossings are computed by line intersection rather than placed by hand. The two insets are measured, not drawn: the reference frame as the models receive it, and the ΔPx field the warp requires, both for Scene20_00650—the frame of Figure 1—and computed by the same routine, so the quoted median and p90 agree with that figure by construction. **(B) The quantification.** The three models are used frozen, with their public weights, and treated as black boxes: only their depth and camera estimates enter the analysis. Each condition is put in the canonical gauge before anything is compared, because $(d, t) \mapsto (\lambda d, \lambda t)$ is unobservable; only gauge-invariant components are compared; and what is reported for every dose $\times$ view-set cell is explained variance, which is at most 1 and goes negative when a candidate field is worse than predicting no change at all. A single swap takes one component from the dose and every other component from that frame's own $\gamma = 1$ solution: these are attribution probes and are explicitly non-additive, while the joint swap is exact. Apart from the two insets, every quantity shown is a design constant.

B GAUGE IDENTIFIABILITY, IN FULL

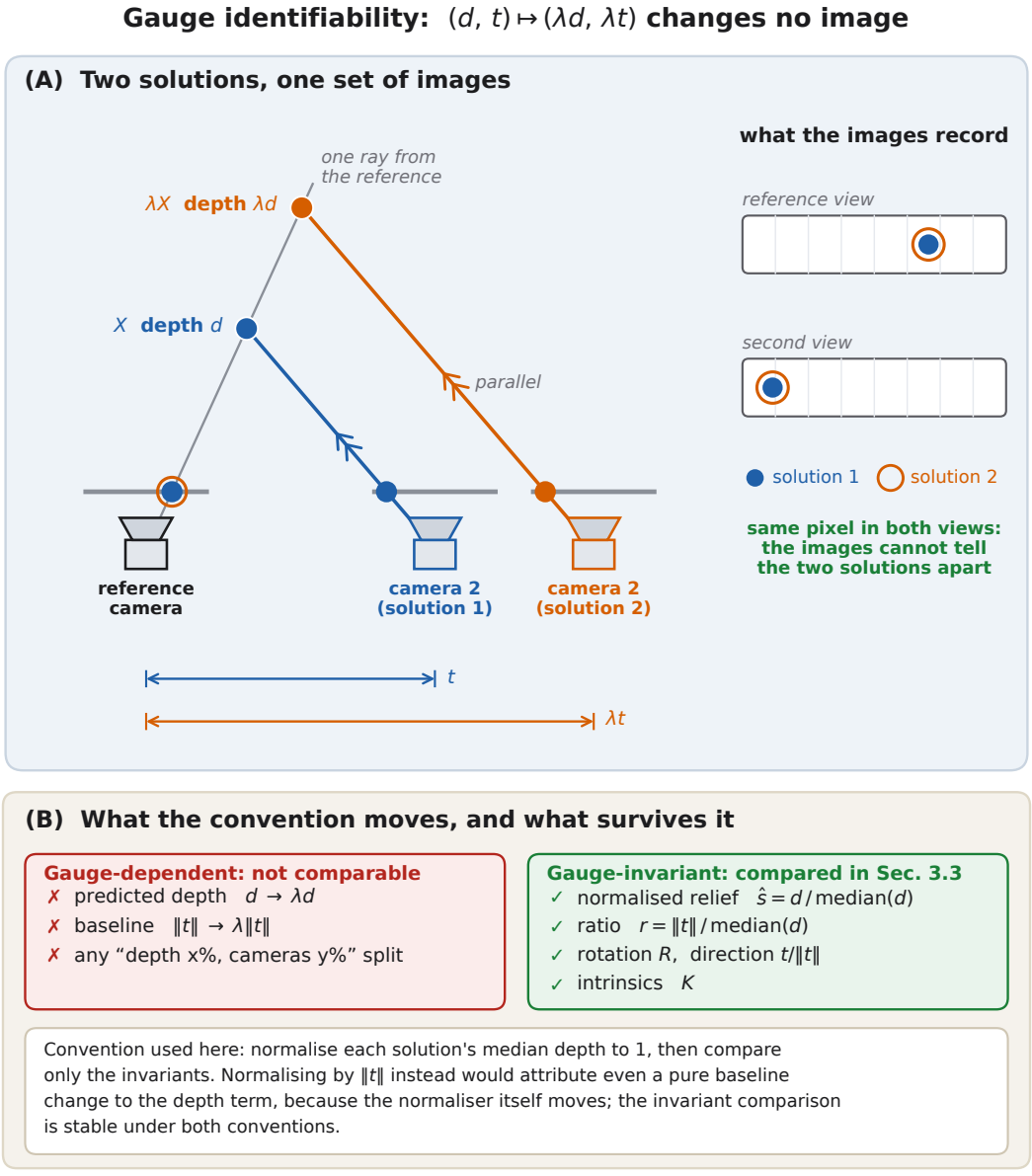

Figure 4: Gauge identifiability. **(A)** Two solutions related by $(d, t) \mapsto (\lambda d, \lambda t)$, drawn top-down with the shared rotation both must have. Both scene points lie on one ray from the reference camera, and camera 2's ray to each is parallel, so the two solutions put the point on the same pixel in both images (right); the pixels are computed from the geometry, not placed. **(B)** What the convention moves and what survives it: the predicted depth, the baseline and any "depth did $x\%$, cameras did $y\%$" split are gauge-dependent, while normalised relief, the baseline-to-depth ratio, rotation, translation direction and intrinsics are invariant. This figure contains no measurement; the claim is geometric and holds for any λ.

C THE ARCHITECTURAL TEST, PER CELL AND PER TERM

Table 1: How much of the conflict each model's outputs jointly reproduce, over 16 dose × view-set cells (95% cluster bootstrap over scenes). Every VGGT cell clears the pre-registered 0.4 gate; every Fast3R interval lies below 0; no DUSt3R interval clears 0.4.

MODEL	ATTENTION	CAM. HEAD	R^2_{full}
VGGT-1B	global, interleaved	yes	$+0.513 \ldots + 0.942$
Fast3R ViT-L	global, all-to-all	no (PnP)	$-4.342 \ldots - 0.309$
DUSt3R-512	pairwise	no (aligner)	$-41.926 \ldots + 0.085$

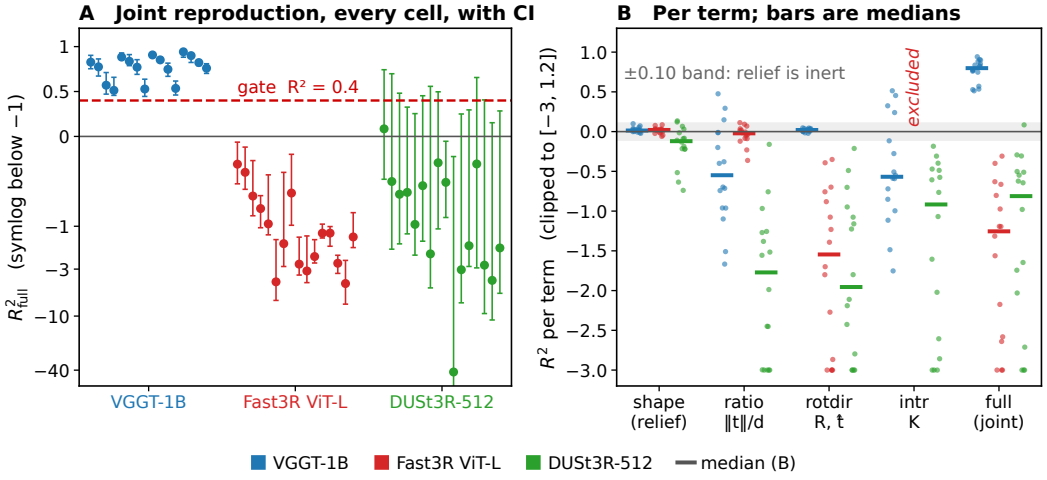

Figure 5: **A** R^2_{full} for every dose × view-set cell with its cluster-bootstrap interval, against the pre-registered 0.4 gate; the y-axis is symmetric-log below -1 so DUSt3R's -41.9 cell does not flatten the rest. **B** the per-term breakdown: the relief term sits at ~ 0 for all three models, so the response lives in the camera-side terms. Fast3R's intrinsics column is empty by amendment, shown rather than omitted.

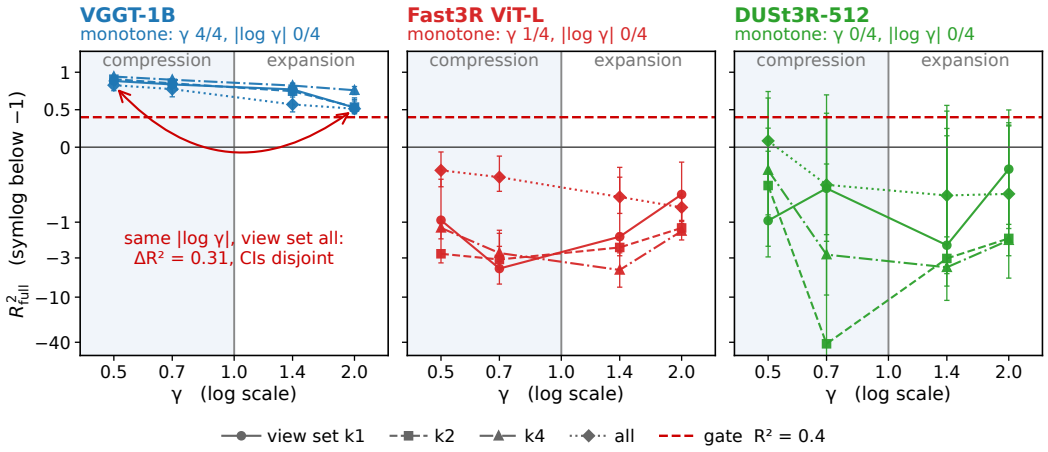

Figure 6: Dose response against $\log \gamma$, so equal-magnitude doses sit mirror-symmetric about $\gamma = 1$. The registered prediction required the members of each $|\log \gamma|$ pair to behave alike; they do not. VGGT is monotone in γ in 4/4 view sets and in $|\log \gamma|$ in 0/4. The annotated pair is the `all` view set, where $\gamma = 0.5$ and $\gamma = 2.0$ differ by $\Delta R^2 = 0.31$ with disjoint intervals; the largest such gap is 0.37 at `k2`.

D THE FOCAL CHANNEL AND THE DISPARITY GAUGE

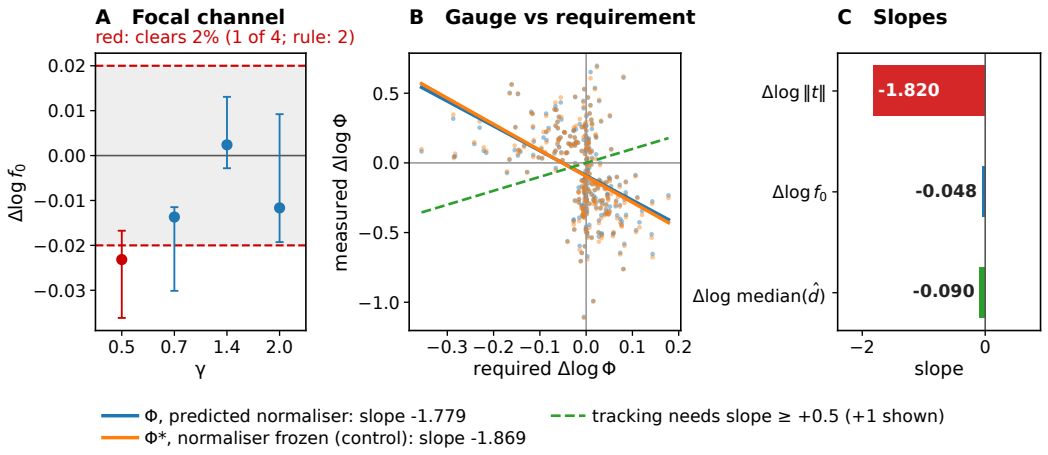

Figure 7: The focal channel is undetermined and the disparity gauge does not track. **A** $\Delta \log f_0$ per dose against the pre-registered 2% band; the rule required two doses to clear it with a CI excluding zero, and one does. **B** Φ against the change the warp requires, with the predicted-normaliser fit (-1.779) and the frozen-normaliser control (-1.869); tracking would have required a slope of at least $+0.5$. **C** the component slopes: the effect lives entirely in $\|t\|$, the focal's own slope includes zero, and $\|t\|$ co-moves with median depth at $+8.66$, far from the $+1$ a global scale drift would give.