GENERATIVE LIGHTING ESTIMATION FOR MULTI-VIEW CAPTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

An environment map is the standard representation of the distant illumination arriving at a point in a scene, and many inverse rendering and relighting pipelines on real captures depend on one. Generative methods predict a map from a single image with a learned prior. Everything outside that image's field of view is then inferred from indirect cues and never observed, so the estimate does not faithfully reproduce what is in the scene. Per-scene inverse rendering instead optimises against a whole multi-view capture, solving for illumination jointly with geometry and materials. It is constrained only where the capture has observations, and the light sources usually lie outside them. We recover a high dynamic range environment map from a sparse set of posed low dynamic range frames, so that the learned prior supplies only what the capture does not. The map is produced as a set of views that together cover all directions, generated jointly by a multi-view diffusion model so they agree with each other and with the input frames, and assembled into a panorama. The single-image setting is the smallest case of this. On it we improve on the state of the art on Poly Haven and both Laval benchmarks. Additional frames improve the estimate further, with diminishing returns. On captures with a moving camera, the frames need not be taken where the map is centred; the benefit falls off slowly with distance. Each frame enters with its pose, so unlike stitching and outpainting the model is not bound to a single viewpoint. Two controls attribute the gain to the camera positions and rule out the extra pixels: a checkpoint trained without displaced views and our own checkpoint with the translations zeroed at inference lose the same amount of it.

1 INTRODUCTION

Relighting an object in a photographed scene, or recovering the materials of one, requires the light arriving at a point in it. An environment map records that light as the radiance incident from every direction, on the sphere of directions around the point. It is used as though the scene around that point were infinitely far away, so it describes one location in the scene. Methods that recover materials and geometry from photographs either assume the map is given, which holds only for synthetic data, or estimate it from a single image. The captures they run on are not single images: photogrammetry, radiance field reconstruction (Mildenhall et al., 2020) and Gaussian splatting (Kerbl et al., 2023) all begin with tens or hundreds of posed views, and the lighting estimate is built from one of them while the rest are discarded.

Two bodies of work address lighting recovery and they do not meet. Per-scene inverse rendering optimises an environment map against multi-view observations (Zhang et al., 2021b;a; Boss et al., 2021; Jin et al., 2023; Liang et al., 2024; Gao et al., 2024): it uses every view, but explains observed radiance only, so the unobserved part of the sphere is unconstrained or smoothed away, and that is where the light usually is. Generative lighting estimation instead learns a prior over environment maps and completes what was never seen (Liang et al., 2025; Chen et al., 2022; Phongthawee et al., 2024), but its interface is one image, so the completion is a hallucination and the remaining $N - 1$ frames play no part. Neither line offers a generative prior that is constrained by multi-view evidence and improves as that evidence grows.

We estimate an environment map from $N \geq 1$ posed views with two learned stages and one assembly step. The first stage generates low dynamic range frames that cover the sphere, conditioned on

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.

the input views and trained with varying N, so the single-view case uses the same model. The second stage maps that generated content to high dynamic range, frame by frame. Assembly stitches the frames into an equirectangular map. The split follows the data. Generating content for directions no camera observed needs training views separated by real camera translation, and that is widely available as 360 degree low dynamic range video. The radiometric lift is the easier task, and it can be learned from the far scarcer high dynamic range panoramas, which are stills and carry no translation. No corpus holds both, so a model trained end to end on high dynamic range targets is confined to the stills, and we show what that costs.

More views bring more pixels whether or not their positions are used, so a score on our own model cannot show that the camera translations are what help. We therefore attribute the gain by difference, against two controls. One is a checkpoint trained with every conditioning frame routed back to the target location, so it never sees a non-zero baseline. The other changes no weights at all: it is our own checkpoint, handed the same images, with the presented translations collapsed to zero. Both cost the model the same multi-view gain at every source distance. Each is open to an objection the other closes, so what the displaced training data buys is the use of the camera-position channel, with extra pixels and a better image prior both ruled out.

Our contributions are the formulation and the evidence that a model exploits it.

- Lighting estimation from $N \geq 1$ posed views, trained for variable N, where existing generative methods accept one image and discard the rest of the capture. The split into generation, radiometric lift and assembly follows from the corpora the two halves require.
- At one view we improve on the state of the art on the sphere metrics of three benchmarks with high dynamic range ground truth. Going to four views improves every sphere metric further, and the improvement reaches the 61 percent of the sphere no input view covers.
- Causal evidence that the camera translations are what carry the gain, with the extra pixels ruled out: a checkpoint trained without displaced views and our own checkpoint with the translations zeroed at inference lose the same amount of it.

2 Related Work

Lighting estimation from a single image. A line of work learns a prior over environment maps and samples from it conditioned on one photograph. Chen et al. (2022) generate high dynamic range panoramas from text, Phongthawee et al. (2024) recover a light probe by inpainting a chrome ball into the input image, and Liang et al. (2025) condition a video diffusion transformer on a single view and set the current state of the art on the benchmark we adopt. They produce plausible illumination for the part of the sphere no camera observed, which is most of it; the interface is the limitation we address, since a capture with many posed views is reduced to one image and the evidence in the others is discarded.

Lighting from multiple views. Per-scene inverse rendering recovers illumination jointly with geometry and materials by optimising against all available views (Zhang et al., 2021b;a; Boss et al., 2021; Jin et al., 2023; Liang et al., 2024; Gao et al., 2024). Its objective is photometric consistency with the observations, so the recovered map is constrained only where the scene reflects light back into a camera, and directions no surface reveals stay at whatever the regulariser prefers; these methods also require dense capture and per-scene optimisation and carry no prior over the unobserved sphere. Closest to our setting, Srinivasan et al. (2020) predict a volumetric lighting representation from a stereo pair and Li et al. (2023) estimate lighting from video, but neither fills the directions never observed from a generative prior, relying instead on a narrow stereo baseline and on temporally adjacent frames. We take sparse views with wide baselines and combine them with such a prior.

High dynamic range reconstruction. Inverse tone mapping expands a low dynamic range image into high dynamic range (Liu et al., 2020; Marnerides et al., 2018), and recent work performs the expansion with diffusion models, either by denoising a consistent exposure stack (Bemana et al., 2025) or in a latent space (Fekri et al., 2026). Our second stage solves the same problem by adapting the same backbone the first stage uses, on cube map faces. Because the two stages are separate, that module could be replaced by any of these operators without retraining the first, an opening the decomposition leaves, outside the scope of this paper.

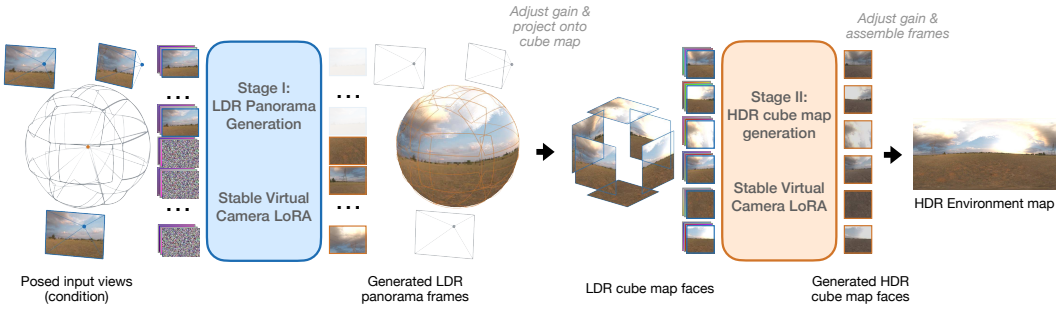

Figure 1: The pipeline. Stage 1 takes N posed low dynamic range views, each at its own position, and generates a fixed grid of overlapping perspective crops covering the sphere from one query location. Each crop comes at its own exposure, so we solve for the relative gains over the overlaps, place the crops on a common radiance scale, and reproject them onto six cube faces, again rendered wide enough to overlap. Stage 2 treats those faces as posed images and predicts a high dynamic range encoding at the same cameras. The same gain solve over the overlaps assembles the predictions into the equirectangular map.

Multi-view diffusion. Novel view synthesis models condition a diffusion backbone on camera geometry and generate unseen views (Sargent et al., 2024; Zhou et al., 2025). We build on Zhou et al. (2025), which encodes cameras as Plücker rays, attends across frames, and accepts a variable number of conditioning views, and Kalischek et al. (2025) report that overlapping cube map faces make cross-frame consistency easier to obtain, which informs the field of view in Section 3.2.

3 Method

We estimate an environment map with two learned stages and one assembly step (Figure 1). Stage 1 generates low dynamic range frames that cover the sphere, conditioned on the input views and their poses. Stage 2 maps that generated content to high dynamic range, frame by frame. Each frame is generated at its own exposure, so assembly solves for the relative gains and composites the frames into a single map (Section 3.4). Both stages adapt the same multi-view diffusion backbone under separate adapters, and assembly is not learned.

3.1 Problem Setup

Given $N \geq 1$ images $I_1, \ldots, I_N$ with known intrinsics and extrinsics $\pi_1, \ldots, \pi_N$ and a query location p, we produce a high dynamic range environment map E at p, in the orientation of the first input view. The query location need not coincide with any input camera: Section 4 queries at one, Section 5 away from one. Existing methods learn $I \mapsto E$ and are applied to one frame of a capture; we learn $\{(I_i, \pi_i)\}_{i=1}^{N} \mapsto E$, so the poses carry what distinguishes the formulations, the inter-camera translations a single image cannot express. Positions are centred and normalised by their RMS radius, so the model never observes an absolute position or scale. Each photograph also carries its own exposure, so at $N > 1$ the frames disagree in scale where they overlap. We keep that variation and recover the relative gains needed to combine the frames into one map, which also gives the map range no single frame holds.

3.2 Stage 1: Multi-View Generation

We build on a 1.3B-parameter multi-view diffusion model for novel view synthesis (Zhou et al., 2025). It denoises one sequence holding the conditioning views as clean latents and the targets as noisy ones, with attention running jointly over the patches of every frame. Each frame carries its own camera as a Plücker ray map, concatenated to its latents and injected through adaptive layer normalisation. We freeze the autoencoder and the image encoder and adapt the attention, feed-forward, projection and convolution layers with rank-128 low-rank updates (Hu et al., 2022). The number of conditioning frames is resampled between one and four at every training step, so one set of weights covers the range.

Output geometry. The targets are perspective crops whose union covers the sphere, which keeps the backbone in the mode it was trained for. We use two arrangements: twenty crops at 60 degrees of vertical field of view, six around the equator and fewer in each row towards the poles, and a cube map of six faces at 100 degrees (Appendix A). Both are rendered wider than tiling the sphere requires, so that neighbours overlap. Kalischek et al. (2025) widen cube faces from 90 to 95 degrees for the same reason and report the overlap as an aid to cross-frame consistency, then crop it away when assembling; we keep it, because it is also what assembly solves the relative gains against. The two arrangements trade quality against cost, twenty forward passes against six, and Section 6 measures the gap.

Training data. The corpus is hand-held 360 degree video posed by structure from motion, with high dynamic range panoramic stills mixed in at half rate for image quality and scene variety (Appendix E). The stills alone are not sufficient: crops from one panorama share an optical centre, so the translation between any pair is zero and the model never receives a gradient distinguishing placement by translation from placement by rotation. Every crop, from either corpus and at either stage, is auto-exposed to a fixed mean luminance, clipped and gamma encoded by one shared operator, which imitates the camera that would have taken it. Per-frame exposure is therefore a property of the targets throughout the pipeline, in both corpora.

3.3 Stage 2: High Dynamic Range Prediction

Stage 2 lifts the low dynamic range panorama that Stage 1 produces to high dynamic range. It works on six cube faces, so Stage 1's frames are first placed on a common scale and reprojected onto those faces (Appendix D). Each face is then auto-exposed to a fixed mean luminance, clipped and gamma encoded by the operator that produced every training crop, so it matches what Stage 2 was trained on. Stage 2 predicts high dynamic range at those same faces, reusing the backbone under a separate adapter and without architectural change. Its only job is to extend the dynamic range while preserving the content Stage 1 produced, and Appendix H reports how much range it recovers.

Keeping the two tasks apart also matters for the backbone. Adapting a pretrained generative model works best when its targets stay close to the distribution it already models, and high dynamic range targets are a large shift in colour statistics. Stage 1 therefore keeps the backbone's native output range, while Stage 2 is left a remapping at fixed content, which it learns without difficulty.

Encoding. To use unbounded radiance as a target, we encode it as an image in $[-1, 1]$ that the autoencoder can take: for each face we rescale its median luminance to a fixed value, take the logarithm, and squash the result through a hyperbolic tangent,

$$y = \tanh\left(\frac{\log(k\,x + \epsilon) - \log m}{s}\right), \qquad k = \frac{m}{\operatorname{median}(x)}, \tag{1}$$

with $m = 0.18$ and $s = 4$, which places the saturation point near 55 times the median. Rescaling by k removes the face's exposure, so the encoding does not depend on the gain Stage 1 happened to produce. It also discards the relative scale between faces, which we recover afterwards by matching radiance where the faces overlap. Decoding inverts Equation equation 1, clamping the argument first because the inverse hyperbolic tangent diverges at ± 1 (Appendix C).

Training data. Stage 2 trains on high dynamic range panoramas alone and needs no video, because every frame it sees comes from one location. Each batch takes a panorama, splits it into faces, and computes an exposure per face from that face's own statistics; the conditioning frame is the face at that exposure and the target is the same face under Equation equation 1. Neither the encoding nor the gain solve fixes an absolute scale, so our estimate is determined only up to a global factor.

3.4 Assembly

Both stages leave the same quantity undetermined, Stage 1 by returning each frame at its own exposure and Stage 2 by removing any common scale. Let g_i be the unknown gain of frame i and μ_{ij} the mean luminance of frame i over its intersection with frame j, so that each overlapping pair contributes one linear constraint in the log domain,

$$\log g_i - \log g_j = \log \mu_{ji} - \log \mu_{ij}. \tag{2}$$

Table 1: Single-view comparison with the baselines on three benchmarks. The last column is directional error in degrees. A dagger marks a tie: our paired per-scene difference from the best baseline is under two standard errors. Bold marks the best value in a column and any cell tied with it.

		Scale-invariant RMSE ↓			Angular Error (°) ↓			Normalised RMSE ↓			
Dataset	Method	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Dir. ↓
Poly Haven	DiffusionLight	0.113	0.290	0.516	1.779	2.458	3.463	0.222	0.352	0.455	47.59
	LuxDiT	0.085	0.194	0.403	**1.674**	2.680	3.519	0.160	**0.232**	0.348	**33.26**
	Ours	**0.066**	**0.167**	**0.363**	**1.690**†	**2.199**	**2.840**	**0.116**	**0.226**†	**0.326**	**35.55**†
Laval Indoor	DiffusionLight	0.124	0.329	0.606	**2.430**	3.362	5.644	0.214	0.359	0.427	45.50
	LuxDiT	0.114	**0.289**	0.577	2.723	3.777	5.899	0.203	0.358	0.511	41.48
	Ours	**0.103**	**0.293**†	**0.561**	**2.410**†	**3.186**	**5.060**	**0.168**	**0.313**	**0.392**	**35.57**
Laval Outdoor	DiffusionLight	0.097	0.259	0.445	2.315	3.268	5.763	0.174	0.329	0.448	49.11
	LuxDiT	0.106	0.238	0.413	2.339	3.685	6.051	0.198	0.297	0.423	**40.94**
	Ours	**0.080**	**0.207**	**0.375**	**1.735**	**2.579**	**4.761**	**0.149**	**0.276**	**0.388**	**42.29**†

Both halves of this step are standard: Brown & Lowe (2007) recover a per-image gain by least squares on the intensities of overlapping pixels, and Burt & Adelson (1983) blend the result across a Laplacian pyramid. We solve Equation equation 2 by weighted least squares in the log domain, where the constraint above is linear, and composite with pyramid blending across the overlap bands (Appendix B). The step is not learned and serves both outputs: Stage 1 with assembly gives a low dynamic range panorama, which is what we evaluate where no high dynamic range ground truth exists, and adding Stage 2 gives the high dynamic range map.

4 Experiments: Lighting Estimation Benchmarks

4.1 Setup

We follow the render-and-compare protocol of Wang et al. (2022): predicted and ground-truth maps light a mirror, a matte silver and a diffuse sphere in Blender Cycles, compared per material with scale-invariant and normalised RMSE and per-pixel angular error, to which we add the directional error of Somanath & Kurz (2021). We re-ran both baselines on all three splits from their public checkpoints, which is what the paired per-scene tests require; our LuxDiT figures reproduce the published ones up to small differences from the re-evaluation. Poly Haven (Poly Haven, 2025) and Laval Indoor (Gardner et al., 2017) follow the scene selections of Liang et al. (2025), 181 and 289 scenes; Laval Outdoor (Hold-Geoffroy et al., 2019) is the full 205-scene set. Our panoramas come from that work's Poly Haven training split, so the Poly Haven rows are held out and neither Laval set is seen in training at all. Mean tables invite comparisons the spread does not support, so every comparison below is a paired per-scene difference in standard errors, and differences under two are marked as ties.

4.2 Single-View Results

The picture is consistent across the three corpora. We are ahead of Liang et al. (2025) on seven of the nine sphere-metric cells on Poly Haven, on eight of nine on Laval Indoor, and on all nine on Laval Outdoor against both baselines, with the remaining cells ties. No baseline leads us anywhere at two standard errors, and the margins are largest on Laval Outdoor. Directional error is the exception: we improve on Liang et al. (2025) on Laval Indoor, but elsewhere the paired differences are under one standard error and we report them as ties.

4.3 Scaling with the Number of Views

The comparison above uses one view because that is what the baselines accept; here we add views of the same scene and measure what they buy. They are cut from each scene's own ground-truth panorama at yaws 180, then 90, then 270 degrees, at the benchmark crop's field of view, so they share its optical centre and the input carries angular coverage and no parallax. The sets are nested,

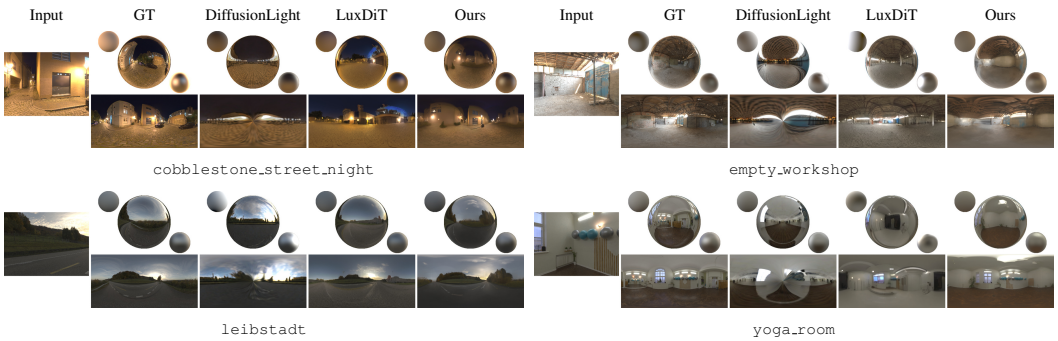

Figure 2: Qualitative comparison with prior work on Poly Haven from a single input view, two outdoor scenes and two indoor. Each method is drawn as mirror, matte silver and diffuse spheres rendered under its estimated map, above the map itself. Maps are median-aligned and tone mapped by the benchmark operator. Further scenes are in Appendix M.

Table 2: Two further studies of lighting estimation on Poly Haven. The upper block adds views at a shared optical centre: every metric is better at four views than at one. The lower block ablates Stage 1, each run at one view and read against the one-view row above it. With no fine-tuning at all the pipeline only reaches the level of DiffusionLight; the cube map is behind the grid; and training the targets on the grid used at inference beats random orientations, on a benchmark scored on that grid.

	Scale-invariant RMSE ↓			Angular Error (°) ↓			Normalised RMSE ↓		
	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror
Views at a shared optical centre									
1 view	0.0662	0.1665	0.3633	1.690	2.199	2.840	0.1160	0.2263	0.3260
2 views	0.0660	0.1515	0.3349	1.426	1.618	2.238	0.1174	0.2173	0.3091
3 views	0.0483	0.1278	0.3105	0.808	1.213	1.742	0.0906	0.1899	0.2809
4 views	**0.0436**	**0.1125**	**0.2934**	**0.766**	**1.039**	**1.487**	**0.0809**	**0.1738**	**0.2697**
Stage 1 variants, one view									
Training-free	0.1186	0.2332	0.4184	2.983	3.393	3.972	0.2075	0.3347	0.4211
Random target views	0.0744	0.1823	0.3805	1.824	2.444	3.092	0.1281	0.2331	0.3306
Cube map grid	0.0855	0.1983	0.4007	1.893	2.411	3.103	0.1551	0.2509	0.3548

four views is the training maximum so no run extrapolates, and coverage reaches 39.4 percent of the sphere. This is an oracle study, since the extra views are ground truth: it measures what a multi-view capture would buy, not a new capability of a single-image method, and conditioning pixels are never composited into the output.

Extra views improve the estimate, monotonically. Every cell of the upper block of Table 2 is better at four views than at one, and the mirror columns improve at better than nine standard errors paired per scene. The walk is not monotone everywhere: diffuse normalised RMSE rises slightly at two views before falling, and diffuse scale-invariant RMSE is flat over the same step. Two views already place us ahead of the single-view state of the art on every Poly Haven cell of Table 1. Directional error, the angle between the dominant light directions, falls until three views and then stops.

The gain is integration, not reproduction. The runs were handed real pixels for part of the sphere, so the sphere metrics alone cannot separate using the evidence from copying it. We therefore score the stitched panorama on the 61 percent of the sphere outside the four-view coverage, which is shown to no run. There scale-invariant PSNR rises by 2.25 dB from one view to four, and every step is significant. Evidence about one part of the sphere therefore improves the part the model still has to invent (Appendix G).

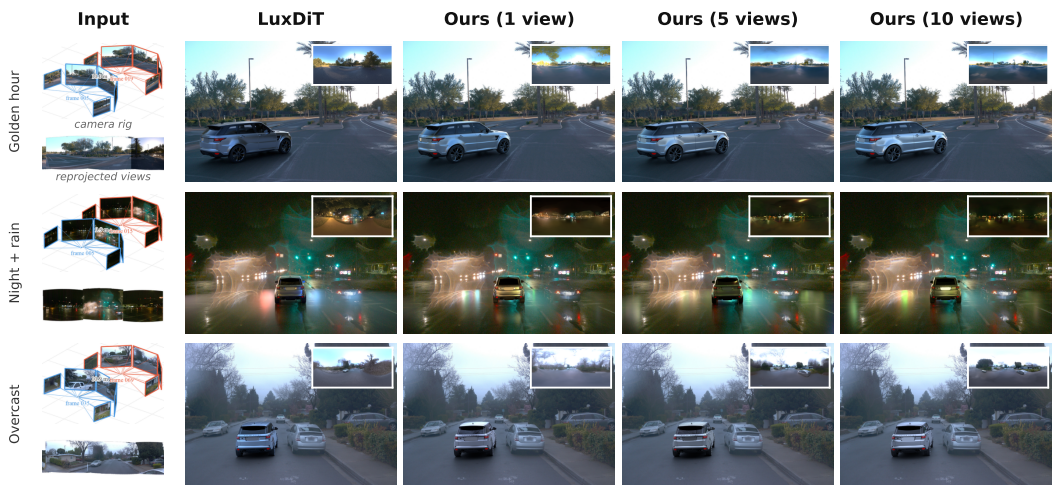

Figure 3: Lighting estimation on driving captures, one row per scene. Multi-view input is the normal case here: an autonomous vehicle rig carries several cameras at once, and the trajectory supplies more of the scene from further on. The input column shows both, five cameras at each of two positions about ten metres apart, blue at the earlier frame and red at the later one, with the same frames projected onto a sphere at infinity below. The remaining columns insert a car at a fixed point on the road and light it with the estimated map, shown inset: the single-image baseline, then ours from one camera, from the full rig, and from both positions. No ground truth exists for these captures, so the panels are qualitative; the measured comparisons are Tables 1 and 2 at a fixed location, and Figure 5 for displaced views.

5 Experiments: Spatial Variability

Section 4.3 added views at a shared optical centre. A real capture supplies them from wherever the camera happened to be, and panorama generation from a few such views is usually posed as stitching followed by outpainting: align the inputs by rotation onto a common sphere, composite, and generate the rest. That pipeline assumes every input camera shares one optical centre, which fails as soon as the capture involves walking, and its error is the angle between an object's true placement and its rotation-only placement, growing with baseline over scene depth. A posed multi-view model is not obliged to make that assumption, since the translations are part of its conditioning, but the question of whether a given model uses them is what this section settles. We measure what displaced views buy, as a function of their number and of how far from the target they were captured, and isolate how much of that is the camera translation and how much the extra pixels.

Figure 3 first shows what the setting looks like on real captures, where an autonomous vehicle rig supplies several views at once and the trajectory supplies more from ten metres further on.

Construction. We use 13 held-out clips of posed panoramic video from the corpus of Section 3.2, six target frames each for 78 targets per cell, with the n conditioning views at nested azimuths for $n = 1$ to 8 and view j cut from the frame whose distance to the target is j-th closest to d. Source distances d are in trajectory-normalised units, against training baselines with median 0.357, so $d = 1.4$ sits at the edge of the training distribution and $d = 2.8$ well beyond it. Structure from motion recovers each trajectory only up to scale, and the captures are hand-held walks, so an assumed walking pace is what gives a distance a metric reading: one unit is 5 to 15 metres (Appendix J). The RMS radius of the conditioning cameras about the target is pinned in every cell; otherwise it would shrink as n grows and be read as a view-count effect. The scored region is the sphere minus the eight-view coverage and minus the nadir where the operator sits, identical for every cell and seen by no conditioning set, so no cell can win on it by copying. All numbers are scale-invariant, solid-angle-weighted PSNR.

The three runs. Three runs share identical cells, targets, ground truth and seed. The first is our checkpoint with the real camera geometry. The *training control* is the same architecture, LoRA recipe, corpus, seed and step count, differing only in that every conditioning frame is routed back to



Figure 4: The inputs are eight views taken away from the query location; the panorama is generated at that location, and the ground truth for it is on the left. Nearby content sits in a different direction from each input position, so a model that ignores where the cameras stood cannot place it correctly. The remaining two columns do exactly that: our own checkpoint with the camera translations zeroed, and a checkpoint trained only on single-location data. Both put the near tree over the buildings in each scene, where ours leaves them clear. Badges are scale-invariant PSNR on the region no conditioning view covers. Panoramas are cropped above -45 degrees, the latitudes the metric scores. Six further scenes are in Appendix K.

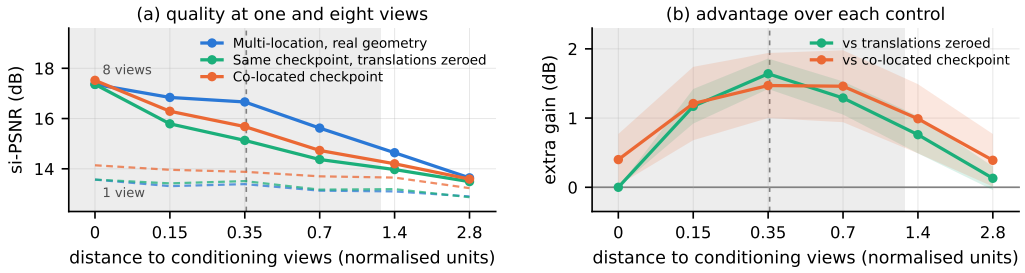

Figure 5: Panorama generation from conditioning views taken away from the query location, on held-out panoramic video posed by structure from motion. We compare our model with two controls: a checkpoint fine-tuned only on views from a single location, and our own checkpoint run with the camera translations zeroed. Every run is scored on identical pixels of the region no conditioning view covers. (a) Scale-invariant PSNR against that distance, solid at eight conditioning views and dashed at one, so what extra views are worth is the vertical gap within a colour. Levels fall with distance in every run because the task itself gets harder; the gap stays open for our model wherever the sources are displaced, and both controls close it. Levels carry no error band, for the reason Appendix J gives. (b) The same gain relative to each control, paired per target, with bands of one standard error clustered by clip. Both differences peak at the median camera separation seen in training, marked by the dashed vertical, and return to zero at 2.3 times the largest. Shading marks the range of separations the model was trained for.

the target panorama, so it trains with all cameras at one optical centre. The *pose control* changes no weights at all: it is our checkpoint handed the same images and view directions with the presented translations collapsed to zero. The pose control admits an exact validity check: at $d = 0$ the conditioning cameras already coincide, so collapsing the translations is a no-op and the two runs come out bit-identical, while at every displaced rung they differ. Figure 4 shows all three in pixels.

Extra views help, and the help decays with distance. Going from one view to eight improves the unseen region by 3.79 dB with co-located sources. From $d = 0.35$ outward the gain falls by about 0.84 dB per doubling of source distance and is nearly flat below that, keeping a fifth of its co-located value beyond the largest training baseline; the single-view level meanwhile barely moves, which is what one expects for a region invented at every distance. Every step from one view to four pays at every distance, and each is smaller than the one before. The fourth-to-eighth step does not pay: null with co-located sources and negative beyond the training range. The benefit therefore saturates at four views, the conditioning count used in training (Appendix I), and where it saturates depends on rendering the conditioning crops at the shape the checkpoint was trained on (Appendix J).

The advantage is the translation channel, and it lives where the model was trained. Inside the training range our model extracts one to two decibels more one-to-eight-view gain than either control, every displaced rung significant (Figure 5). A model denied its translations at inference therefore behaves like one that never had any, which rules out the more varied training corpus as the explanation. Both differences trace an inverted U peaking at the median training baseline: at $d = 0$ there is no parallax to resolve, and beyond the training range the model stops using the translations.

Neither control settles the question alone, and their objections are disjoint. The pose control presents genuinely displaced images as though they were co-located, so its deficit could come from sensitivity to inconsistent conditioning. The training control is a second checkpoint at one seed, so a checkpoint-level difference is not excluded by errors clustered over clips. The two agree within their errors at every rung, which is why we read the pair: the capability is real, causally attributed, and confined to the separations the model was trained for.

No single-image method is run on this corpus, so every comparison in this section is internal, our model against its own two controls. Appendix I adds the two model-free references, rotation-only reprojection of the same conditioning crops and a whole real source panorama resampled into the target frame. The second still beats generation on the unseen region at every displaced rung, by up to 1.8 dB. It receives a complete real panorama where the model receives crops covering a fraction of the sphere, so the comparison is not like for like. It is still the option to prefer wherever such a capture exists. Appendix J records the remaining caveats.

6 Ablations

All runs below are scored on Poly Haven at one input view, in the lower block of Table 2 and read against the one-view row above them, with the paired per-scene statistics in Appendix L. The comparison between single-location and multi-location training is the training control of Section 5 and we do not repeat it here.

The backbone alone. Both of our stages are fine-tunes, so dropping them asks whether the pipeline works with no training at all. The backbone generates the frames and assembly stitches them, and the only dynamic range the result carries comes from the gain solve across frames. Every one of the nine cells degrades by a wide margin paired per scene, so the benchmark result does not come from the pretrained model.

Target sampling. Training with target crops drawn at random instead of on the fixed grid costs accuracy on every scale-invariant and angular cell, and ties on normalised RMSE except on the diffuse sphere. The variant draws targets from the same distribution as the conditioning views, in orientation and in field of view, which is how the backbone was trained before we adapted it. Fixing the target grid is therefore worth a small but consistent margin when inference runs on that grid, and the plausible reason is that a fixed grid presents the same overlap pattern at every training step, where arbitrary target cameras do not.

Output geometry. We also trained Stage 1 to emit six cube map faces as its target grid instead of the twenty crops at narrower field of view, and ran both through the same Stage 2 and assembly. The grid is ahead on every sphere metric at roughly three times the generation cost, while directional error does not separate them. The cube map remains usable: against Table 1 it is ahead of DiffusionLight on most cells and splits with LuxDiT.

7 Discussion, Limitations, and Future Work

Stage 2 trains on frames rendered from real panoramas and runs on Stage 1's output, and we have not measured what that mismatch costs. The backbone also extrapolates poorly beyond the angular region the input views cover, so a stronger multi-view generator is the most direct route to better content on the parts of the sphere no camera observed. Separately, the standard protocols score what a panorama contains and compress the range a high dynamic range estimate exists to produce; Appendix F gives a corrected route that renders on the linear map and scores all three methods under it on Laval Indoor.

AI USE STATEMENT

Of the tasks ICLR asks us to report, we used generative AI tools for research methodology and experiment design, implementation, data cleaning, and results interpretation. Methodology and experiment design were discussed with the tools rather than delegated to them, and the method this paper proposes is the authors' own. Implementation was delegated: the tools wrote the training scaffolding and the data processing pipelines used for training and evaluation. We checked that code with automated tests and by inspecting its outputs on hand-built probes. Data cleaning was agentic, cutting the video tours into clips and running them through structure from motion, and clips the pipeline could not reconstruct were discarded. The discard criterion was whether reconstruction succeeded rather than a judgement about content. For results interpretation the tools wrote the scripts that run the evaluations and aggregate the statistics, and the authors drew the conclusions from those statistics.

We also used generative AI tools for figure creation, code editing, literature analysis, brainstorming, and paper editing. The related work was drafted from model recall, after discussion of the problem setup in which the authors supplied the reference points. Every bibliography entry was then checked against the primary record: that pass corrected two entries naming papers that do not exist as first described and six carrying wrong authors, years or venues. We did not use generative AI tools for synthetic data generation, hypothesis refinement, or qualitative analysis. Theoretical model development, mathematical claims, proof assistance, translation, and survey or interview questions do not arise here: the work has no theoretical results, no text drafted in another language, and no human subjects.

We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including text, claims and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

The corpus that trains Stage 1 (Section 3.2) was collected from publicly accessible panoramic video on YouTube. We did not establish the licence of the individual videos. It was used only to train the models reported here, and we release neither the corpus nor the clips, poses and panoramas derived from it, beyond the figures in this paper. That footage records people and property incidentally and without their consent, which is a further reason it stays unreleased. The driving captures of Figure 3 are from the Waymo Open Dataset. This publication was made using the Waymo Open Dataset, provided by Waymo LLC under the Waymo Dataset License Agreement for Non-Commercial Use, available at `waymo.com/open/terms`, and your access and use of such work are governed by the terms and conditions therein. The benchmark results of Sections 4 and 6 use Poly Haven (Poly Haven, 2025), which is CC0, and the Laval Indoor and Laval Outdoor datasets (Gardner et al., 2017; Hold-Geoffroy et al., 2019) under the licence their authors require.

REPRODUCIBILITY STATEMENT

Appendix E holds the training configuration for both stages of Section 3: corpus, sequence construction, crop sampling, the low-rank update, the optimiser schedule, and the sampler settings used at test time. It also names the high dynamic range panorama corpus both stages train on and the split it comes from, which is Liang et al. (2025)'s, so their runs and ours are scored on the same held-out scenes. Appendices A to D give the output geometry, the assembly operator and its gauge solve, the high dynamic range encoding with its constants, and the measurements that verify the handoff between the stages against deliberately corrupted controls. Appendix F defines the evaluation protocol and which terms the corrected variant removes, so either route can be recomputed. Every number comes from one seed, so the paired per-scene statistics of Appendix L are what carry significance rather than differences of means. The baseline rows of Table 1 are our own runs of the public checkpoints, not values copied from their papers. We do not release the video corpus, for the reasons in the ethics statement, so Stage 1 cannot be retrained from public data; we will release the scripts that build it from raw 360 degree video, and the implementation, after publication.

REFERENCES

Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Bracket diffusion: HDR image generation by consistent LDR denoising. *Computer Graphics Forum*, 44(2), 2025.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P. A. Lensch. NeRD: Neural reflectance decomposition from image collections. In *ICCV*, 2021.

Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 2(4):217–236, 1983.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2Light: Zero-shot text-driven HDR panorama generation. *ACM Transactions on Graphics*, 2022.

Pedram Fekri, WenChen Li, William Chen, and Peter Altamirano. LatentHDR: Decoupling exposure from diffusion via conditional latent-to-latent mapping for text/image-to-panoramic HDR. *arXiv preprint arXiv:2605.11115*, 2026.

Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with BRDF decomposition and ray tracing. In *ECCV*, 2024.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics*, 36(6):176:1–176:14, 2017.

Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *CVPR*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial inverse rendering. In *CVPR*, 2023.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. CubeDiff: Repurposing diffusion-based image models for panorama generation. In *ICLR*, 2025.

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

Zhengqin Li, Li Yu, Mikhail Okunev, Manmohan Chandraker, and Zhao Dong. Spatiotemporally consistent HDR indoor lighting estimation. *ACM Transactions on Graphics*, 2023.

Ruofan Liang, Kai He, Zan Gojcic, Igor Gilitschenski, Sanja Fidler, Nandita Vijaykumar, and Zian Wang. LuxDiT: Lighting estimation with video diffusion transformer. *arXiv preprint arXiv:2509.03680*, 2025.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. GS-IR: 3d gaussian splatting for inverse rendering. In *CVPR*, 2024.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *CVPR*, 2020.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum*, 2018.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. DiffusionLight: Light probes for free by painting a chrome ball. In *CVPR*, 2024.

Poly Haven. Poly haven: CC0 HDRIs, textures and models. `https://polyhaven.com`, 2025. Assets licensed CC0.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *CVPR*, 2024.

Gowri Somanath and Daniel Kurz. HDR environment map estimation for real-time augmented reality. In *CVPR*, pp. 11298–11306, 2021.

Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *CVPR*, 2020.

Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. StyleLight: HDR panorama generation for lighting estimation and editing. In *ECCV*, 2022.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021a.

Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, 2021b.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

A Output Geometry

The pitch-row grid places five rows of crops at latitudes -75, -37.5, 0, $+37.5$ and $+75$ degrees, with 2, 5, 6, 5 and 2 yaws in each and every row starting at yaw zero, twenty crops in all. Each crop is rendered at 768×576 pixels with a 60 degree vertical field of view, so its horizontal field of view is 75.2 degrees and neighbours overlap by at least 4.7 degrees everywhere on the sphere. Rendering the same grid at 576 pixels square, which is 60 degrees horizontally, instead leaves ten holes totalling 0.08 percent of the sphere at latitudes between 28 and 32 degrees, where the corners of the equator row fail to meet those of the rows above and below. The cube map arrangement uses six faces at 100 degrees, 576 pixels square, with the front face anchored to the first input frame. Both arrangements are deliberately wider than tiling the sphere requires, so that assembly has overlaps to solve the relative gains against.

B Assembly and Gauge Recovery

Equation equation 2 contributes one row per overlapping pair of frames. Rows are weighted by the square root of the overlap area, so a pair meeting over a wide band constrains the solution more than one meeting at a corner. The system determines the gains only up to a common factor, so the log-gain of the first frame is pinned to zero to fix the gauge. We add a ridge of 10^{-8} to the normal equations, solve in the log domain and exponentiate, then composite into an equirectangular map with Laplacian pyramid blending across the overlap bands.

One implementation note carries a real failure mode. The gain search range has to be widened when the frames being stitched are median-normalised high dynamic range faces; low dynamic range crops need no widening. At the default range, overlaps falling on bright sky sit outside it and are dropped from the solve, which removes exactly the constraints that matter most for a map whose dynamic range lives in the sky.

C High Dynamic Range Encoding

Decoding inverts Equation equation 1. The inverse hyperbolic tangent diverges as its argument approaches ± 1, so we clamp to 0.976 first and evaluate in single precision. Left unclamped, a handful of near-saturated pixels per scene run away past 10^6 times the median, which is enough to dominate the mean of any metric computed on the result. The clamp is tighter than numerical safety alone would require, and it removes that tail without visibly affecting the reconstructed highlight.

D Handoff Verification

The pipeline reprojects Stage 1's grid onto six cube faces before Stage 2 consumes them, which is an opportunity to lose content silently. We verified the step by stitching the Stage 1 grid and the handoff faces separately and comparing them per scene in linear light. Across all 675 scenes of the three corpora the worst case is a correlation of 0.9968, with median PSNR between 45.5 and 46.6 dB and no uncovered pixel on any face in any scene. Deliberately transposed, vertically flipped and face-order-rolled controls score between 0.007 and 0.33, so the check discriminates. The multi-view runs of Section 4.3 were checked the same way, with correlation at least 0.992 and median PSNR 42.4 to 43.9 dB on all 181 scenes.

This is a geometric check. It does not address the radiometric mismatch discussed in Section 3.3.

E Training Configuration

Corpus and sequences. Stage 1 trains on 148 captures of hand-held 360 degree video at two frames per second, $31{,}974$ frames in total, posed by structure from motion, with high dynamic range panoramic stills mixed in at half rate. Those stills are the Poly Haven training split of Liang et al. (2025), and they are also the whole of Stage 2's training data. We take the split from that work instead of drawing our own, which puts both methods on the same held-out scenes: the 181 Poly Haven scenes of Table 1 are its test split and enter no training run. No Laval panorama is used in training at either stage, so the Laval Indoor and Laval Outdoor rows measure transfer to corpora neither stage has seen. A training sequence holds the conditioning views together with a random subset of eight grid targets, at most twelve frames in all, so the sequences the model produces at test time are longer than any it saw in training. Conditioning frames are drawn from the distance-rank quantiles $[0.15, 0.95]$ among the twenty-four frames nearest the target, so they carry real non-zero baselines.

Crops. Conditioning crops are sampled with a vertical field of view between 45 and 90 degrees and a pitch within 15 degrees of the horizon, which is where hand-held footage puts its cameras. Target crops are fixed at 60 degrees and reach 75 degrees of pitch, so the model is asked for the poles during training even though it is never conditioned on them.

Optimisation and sampling. Both stages use low-rank updates of rank 128 and alpha 64, trained for thirty thousand steps at batch size four and learning rate 5×10^{-5} in bfloat16, with AdamW under linear warmup and cosine decay. Sampling uses 50 Euler steps at classifier-free guidance 3.0, and every number reported in this paper uses one fixed seed.

F Evaluation Details and the Corrected Protocol

Inputs and resolutions. Input crops follow the StyleLight preparation path, a 256×192 perspective view at 60 degrees vertical field of view cut from the panorama at yaw zero, and predicted and ground-truth maps are resized to 128×256 before rendering.

A calibration difference the sphere metrics do not see. The protocol fits its tone-mapping exposure per method, and on Laval Outdoor the fitted values differ by a factor of 24: 0.64 for Liang et al. (2025), 1.20 for Phongthawee et al. (2024) and 15.34 for ours. The tone map median-normalises every map, so this does not enter the scale-invariant results of Section 4.2. It is still a real difference in the absolute radiance we predict, and ours is the outlier.

The corrected route. The standard route tone maps the environment map and then renders. That tone map raises radiance to the power $1/2.4$, scales the median to one half and clips, so everything above roughly five times the median is white before the renderer sees it, and the range reported in Appendix H is invisible to the scores. A corrected route renders on the raw linear map and applies the tone curve to the rendered spheres afterwards, so nothing compresses the range before it reaches the renderer. We ran all three methods through it on Laval Indoor.

Table 3: Laval Indoor under the corrected route. Means with no paired test, so the ordering carries no significance claim, and the margins on mirror si-RMSE are thin. The two routes must not be mixed in one comparison.

Metric	Material	DiffusionLight	LuxDiT	Ours
si-RMSE ↓	Diffuse	0.00455	0.00402	**0.00359**
	Matte	0.01556	0.01477	**0.01413**
	Mirror	0.04757	0.04726	**0.04612**
Angular (°) ↓	Diffuse	6.158	5.706	**5.498**
	Matte	6.420	6.683	**5.894**
	Mirror	9.081	9.327	**8.046**
norm-RMSE ↓	Diffuse	0.4197	0.3705	**0.3107**
	Matte	0.4124	0.4114	**0.3623**
	Mirror	**2.2642**	6.8874	2.4548

The route changes the numbers substantially and the ordering very little. Angular error roughly doubles for every method, and the material ranking shifts: Phongthawee et al. (2024) reads 9.08 degrees on the mirror here against 5.64 under the standard route. Our method leads scale-invariant RMSE and angular error on all three materials and normalised RMSE on two of the three. The absolute-scale variant of RMSE under this route ranks the methods in the opposite order, which measures predicted absolute radiance and is the calibration difference reported above. We ran the route on Laval Indoor alone, so it establishes what the correction does to the numbers on one corpus and leaves the comparative question to the standard protocol of Table 1.

G Region Breakdown of the Benchmark View-Count Walk

Table 4 breaks the view-count walk of Section 4.3 down by region. The *unseen* row is the claim quoted in the main text. The *extra* row is the check that the conditioning reaches the output at all, and is reported here so that the two are not confused.

Table 4: The view-count walk of Section 4.3 broken down by region, scored against the ground-truth panorama and solid-angle weighted. Regions are fixed by the four-view geometry so that every run is scored on identical pixels: *extra* is the part the additional views can cover, *unseen* the complement of the widest run's coverage, which is shown to no run. Directional error is over the scenes where both sides detect a light.

Views	1	2	3	4
Scale-invariant PSNR (dB) ↑				
extra (29.5% of sphere)	15.05	17.03	19.79	**24.22**
unseen (60.6%)	17.29	17.78	18.85	**19.54**
full sphere	16.35	17.40	18.92	**20.42**
Directional error (°) ↓	35.55	32.62	**28.05**	28.56

H Dynamic Range by Corpus

The sphere metrics of Section 4 are computed after the protocol tone maps the map, which removes most of what Stage 2 produces, so Table 5 reports range directly as the per-scene ratio of brightest pixel to median. Stage 2 multiplies the Stage 1 range by roughly four and still falls two orders of

magnitude short of the ground truth: it recovers part of the range a low dynamic range output cannot represent, and it does not reconstruct the sun. The benchmark registers neither fact, for the reason given in Appendix F. The consequence is measurable: swapping in an earlier Stage 2 checkpoint moves every sphere metric by under one percent while the output range changes by a fifth. We report the standard protocol in the main text regardless, because it is what the baselines were scored on.

Table 5: Dynamic range, the per-scene ratio of maximum to median radiance. Ground truth was measured on Poly Haven only, and the Laval rows must not be compared against it.

Configuration	median	p90
Stage 1 alone, Poly Haven	18.6×	78.7×
Cascade, cube map Stage 1, Poly Haven	123.5×	6718×
Cascade, Poly Haven	71.6×	2915×
Cascade, Laval Indoor	446.3×	6960×
Cascade, Laval Outdoor	39.6×	551×
Poly Haven ground truth	15516×	1.45×10^6×

I Transport and Invention, Measured Separately

Section 5 scores one region, the part of the sphere that no conditioning set covers at any view count. Scoring its complement as well separates two things the model has to do at once. On the covered region the content is present in the conditioning and has to be carried to the target viewpoint, which we call transport. On the unseen region there is nothing to carry and the content has to be invented. Table 6 reports both, and Figure 6 gives the unseen region by view count in full.

Table 6: Scale-invariant, solid-angle-weighted PSNR in dB. *Transport* is each cell's own conditioning coverage, scored with one gain per crop region; *invention* is the region and the global-gain convention of Section 5. The model-free columns are rotation-only reprojection of the same eight crops, on transport, and a whole real source panorama resampled into the target frame, on invention. The two regions are different pixels, so levels compare only within a column group; what to read across them is the response to distance and to view count.

	Transport (covered) ↑				Invention (unseen) ↑				Model-free	
d	n=1	2	4	8	n=1	2	4	8	Reproj.	Source pan.
0	25.66	23.88	23.64	24.18	13.57	16.19	17.38	17.36	39.81	82.60
0.15	16.97	16.56	16.61	19.40	13.31	15.28	16.23	16.84	16.24	18.61
0.35	16.99	16.87	16.68	17.85	13.39	15.66	16.49	16.66	14.69	17.13
0.70	15.07	15.05	15.23	15.99	13.13	15.14	15.67	15.62	13.56	15.95
1.40	13.53	13.57	13.57	14.07	13.10	14.51	14.70	14.64	12.92	15.12
2.80	12.89	12.82	12.82	12.88	12.90	13.87	14.02	13.64	12.43	14.42

The two capabilities respond to distance in opposite ways. Transport loses almost everything at the first displacement step, falling 8.7 dB at one view between co-located sources and the nearest rung, because pixel-exact copying stops applying as soon as the view comes from anywhere other than the target position; it then declines slowly. Invention barely moves across the whole ladder, from 13.6 to 12.9 dB at one view, which is what one expects for a region that is invented at every distance and therefore cannot be made harder by moving the camera. By $d = 2.8$ the two have met: 12.89 against 12.90 dB. At 2.3 times the largest training baseline, content the conditioning actually contains is worth no more to the output than content the model has to make up.

They respond to view count in opposite ways too. Extra views buy nothing on the transport region with co-located sources, where the single view already contains the answer and the score in fact drifts down from 25.66 to 24.18 dB. On the invention region the same step is worth $+3.79$ dB. Where the sources are displaced the pattern softens but survives: at $d = 0.35$ transport gains 0.86 dB from one view to eight against 3.27 dB for invention. Additional views mainly buy a better guess about the part of the sphere nobody photographed; the part somebody did photograph gains far less.

Neither region reaches its own ceiling, and the two ceilings are different in kind. On the transport region a model-free reprojection of the same eight crops scores 39.81 dB with co-located sources where the model reaches 24.18, a gap of more than 15 dB set by the autoencoder round trip, resampling onto a different crop grid and per-crop auto-exposure. That ceiling is mechanical, and every generated number in Section 5 sits below it. On the invention region the reference is a whole real source panorama resampled into the target frame, which still beats generation at every displaced rung, by 1.8 dB at $d = 0.15$ and by between 0.3 and 0.8 dB beyond it. That comparison is not like-for-like: the reference receives a complete real panorama where the model receives crops covering 10 to 49 percent of the sphere. It is still the baseline a practitioner would weigh, and the method addresses the case where no such capture exists.

Where this leaves the pipeline. Read together, the two columns locate the weaker half. The model transports what the conditioning constrains better than it invents what the conditioning does not reach, and it is on the invention side that a real panorama still beats it. We report both scoring conventions because the choice biases the view-count sweep. Fitting one gain over the whole scored region charges a method for exposure disagreement between its own crops, and that charge grows with crop count. Fitting one gain per crop region removes the nuisance from the model and the model-free references symmetrically, and raises every level by 0.4 to 0.9 dB. It leaves the one-to-eight-view gains on the invention region at +3.38, +3.44, +3.26, +2.70, +2.01 and +1.43 dB across the six rungs, within 0.7 dB of the curve Section 5 reports.

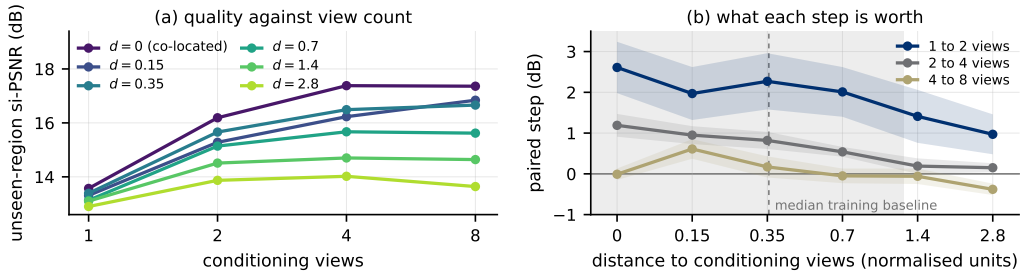

Figure 6: Our checkpoint on the invention region. (a) Scale-invariant PSNR against conditioning-view count, one curve per source distance, scored on the complement of the eight-view coverage so every view count is measured on identical pixels. (b) The paired step between successive view counts, against source distance, with bands of one standard error clustered by clip. Saturation is a statement about a step: the fourth-to-eighth curve sits on zero from $d = 0.35$ outward and turns negative beyond the training range. Shading as in Figure 5.

J The View-Count Sweep in Detail

Evaluation split. Eighteen clips, 3528 frames, are held out whole; the split is by clip because neighbouring frames are near-duplicates. Five frames across four clips sit far off their trajectory under structure from motion and are removed as both targets and sources before anything else. Every clip was scale-normalised at export to an RMS camera radius of 3.66 ± 7 percent, so a distance means the same fraction of trajectory extent on every clip. That normalisation carries no metric meaning on its own; a walking-pace estimate puts one unit at 5 to 15 metres, uncertain by about a factor of 1.6. Section 5 analyses the 13 clips in which the operator actually moves.

Crop shape determines which view count saturates. Conditioning crops in Section 5 are 768×576 at 60 degrees vertical field of view, the shape the checkpoint was trained on. An earlier pass rendered them square, at 60 degrees horizontal, where the checkpoint was trained at 75.2. It lowered the levels and also changed the step pattern: two-to-four measured as a coin flip and four-to-eight as the strong step, reversing what we report. Any claim about where the benefit saturates depends on rendering at the trained shape.

The excluded clips, and what they were doing. Five clips are excluded because the camera stays essentially in one place throughout them. The distance ladder selects, for each rung, the source frame

whose distance to the target is closest to that rung. On a clip with no displacement the selector returns nearly the same frames at every rung, so the curve is flat in d by construction; one such clip reaches a mean source distance of 2.29 units at both the 0.15 and the 0.7 rung. Those rows are therefore not measurements of displacement at all, which matters for an experiment whose independent variable is displacement. Their effect went beyond noise. They contributed a near-constant offset of roughly -1.5 dB to the gap between our model and the controls at every rung, manufacturing a monotone distance trend that is not there. They also made generation appear to lose to reprojection of its own crops at low view counts, since on a barely displaced source reprojection at infinity is close to exact. Restricted to clips that move, both readings change: the control comparison shows the inverted U of Figure 5, and the model beats the reprojection floor generally. Two limits survive the restriction. The rungs still widen with view count, and stalled cameras still place some sources at zero achieved distance at every nominal rung. The horizontal axis therefore gives a nominal rung, the achieved displacement is not what it plots, and re-binning by achieved distance remains open.

At one conditioning view the three runs agree. They sit within half a decibel on the unseen region, with the co-located checkpoint insignificantly ahead, which we report because with nothing to fuse there is no reason displaced training views should help. What separates the runs in Section 5 is therefore the multi-view gain alone.

Why the level panels carry no error band. A standard error on a level is the between-clip spread of scene difficulty, 0.6 to 1.1 dB, which is larger than most of the differences the section reports and is not the uncertainty any claim rests on. It cancels in the pairing, so panel (b) of Figure 5 and panel (b) of Figure 6 carry bands and the level panels do not. Drawing it on a level would invite the between-clip comparison when the within-target one is what matters.

Mechanism, and remaining limits. The unseen-region metric rewards a mean-colour guess, since that region is invented in every cell and PSNR favours blur there. Our numbers on it therefore bound how closely the model reproduces the reference, and they say nothing about how plausible its invention looks. A pilot on three clips pointed at directional coverage as the source of the gain, with the clip where geometry can contribute nothing gaining the most. The full run does not reproduce that ordering, and the pooled parallax measurement is unreliable across a corpus whose scene depth varies by roughly 300 times. The controls of Section 5 establish that the camera-translation channel carries the effect; which geometric quantity the model extracts from it is a different question and is unanswered. Seed variance was not measured on this grid; the paired design cancels target difficulty but not seed noise, and all runs share one seed and one checkpoint each.

K Further Controls in Pixels

Figure 4 shows the two controls failing at one operating point. Figure 7 extends it with six scenes spanning the source distances and view counts of the sweep, so the failure can be read across the grid instead of at a single cell. Both controls place nearby content as though every camera had stood in the same place, at two conditioning views as well as at eight, and at the edge of the training baseline distribution as well as inside it. The pose control and the training control fail on the same objects in the same scenes, which is the qualitative form of the agreement Figure 5 reports in decibels.

L Ablation Tables

Table 7 carries the paired per-scene statistics behind Section 6 on the nine sphere-metric cells, and the directional error of the same three variants is reported below it.

Directional error separates only the training-free run, which costs $+6.74 \pm 2.33$ degrees. The cube map at $+2.37 \pm 2.10$ is a null, which is what Section 6 reports. Random targets post the best directional mean of any run here, -3.37 ± 1.31 degrees, but on a 53 percent win rate, so it is carried by outliers and we do not read it as an effect.

M Additional Qualitative Results

Figure 8 shows the view-count walk of Section 4.3 on six scenes, which Table 2 reports quantitatively. Figures 9 to 11 extend Figure 2 with twelve further Poly Haven scenes, seven outdoor and



Figure 7: Six further scenes in the layout of Figure 4, ordered by source distance and then by view count and labelled with both. The first row is inside the training baseline distribution, the last is at its edge.

Table 7: The three Stage 1 variants of Table 2 minus the reference, paired per scene on Poly Haven. Each cell is the mean difference, its distance from zero in standard errors, and the share of scenes on which the variant is ahead. Positive means the variant is worse.

Variant	Material	si-RMSE	angular (°)	norm-RMSE
Training-free	Diffuse	+0.0524 (9.5, 18%)	+1.293 (9.1, 20%)	+0.0915 (9.6, 20%)
	Matte	+0.0667 (9.2, 21%)	+1.194 (9.2, 20%)	+0.1084 (11.1, 22%)
	Mirror	+0.0550 (8.3, 25%)	+1.132 (9.7, 16%)	+0.0951 (10.9, 17%)
Random targets	Diffuse	+0.0082 (6.3, 34%)	+0.134 (3.4, 38%)	+0.0121 (4.6, 39%)
	Matte	+0.0158 (7.6, 29%)	+0.245 (6.9, 30%)	+0.0068 (1.4, 41%)
	Mirror	+0.0172 (7.1, 24%)	+0.251 (7.6, 25%)	+0.0046 (1.0, 39%)
Cube map	Diffuse	+0.0193 (6.4, 28%)	+0.203 (2.3, 50%)	+0.0391 (7.8, 27%)
	Matte	+0.0318 (7.5, 27%)	+0.212 (3.6, 34%)	+0.0246 (3.8, 39%)
	Mirror	+0.0373 (8.0, 23%)	+0.263 (4.9, 30%)	+0.0288 (4.8, 33%)

five indoor, each from a single input view. The scenes were picked to span the lighting conditions of the split, with no selection for our own results: clear sun, fog, snow, closed canopy, urban dusk and urban night outdoors, and dark industrial interiors, window-lit halls and small rooms with artificial sources indoors.

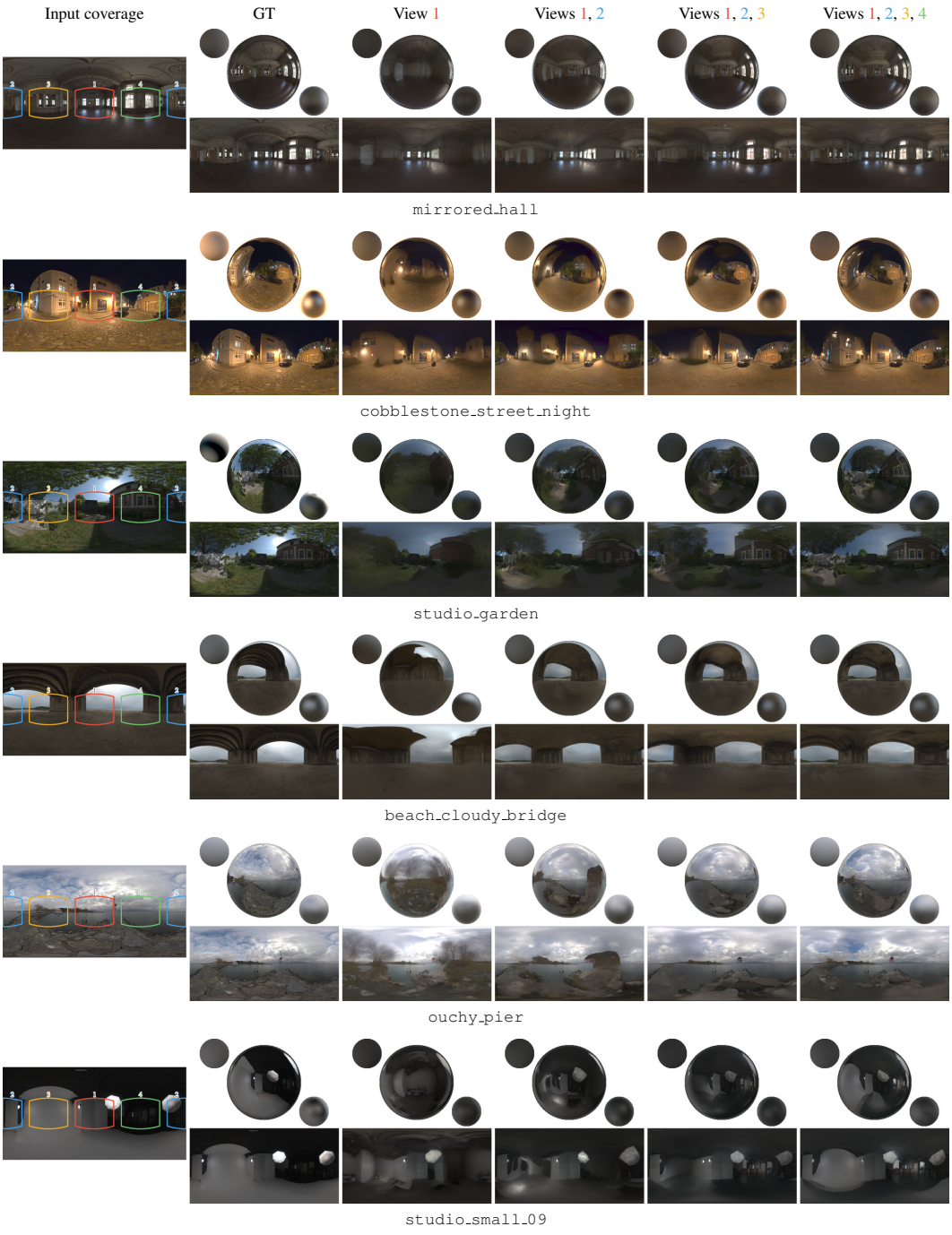

Figure 8: The view-count runs of Table 2 on six Poly Haven scenes, with blocks laid out as in Figure 2. The leftmost panel marks the four input views on the ground-truth panorama in the header colours; the blocks are the ground truth and the one, two, three and four view runs. The view sets are nested and share an optical centre, so each block adds angular coverage to the one on its left. Conditioning pixels are never composited into the output: every map is the model's own generated grid, stitched.

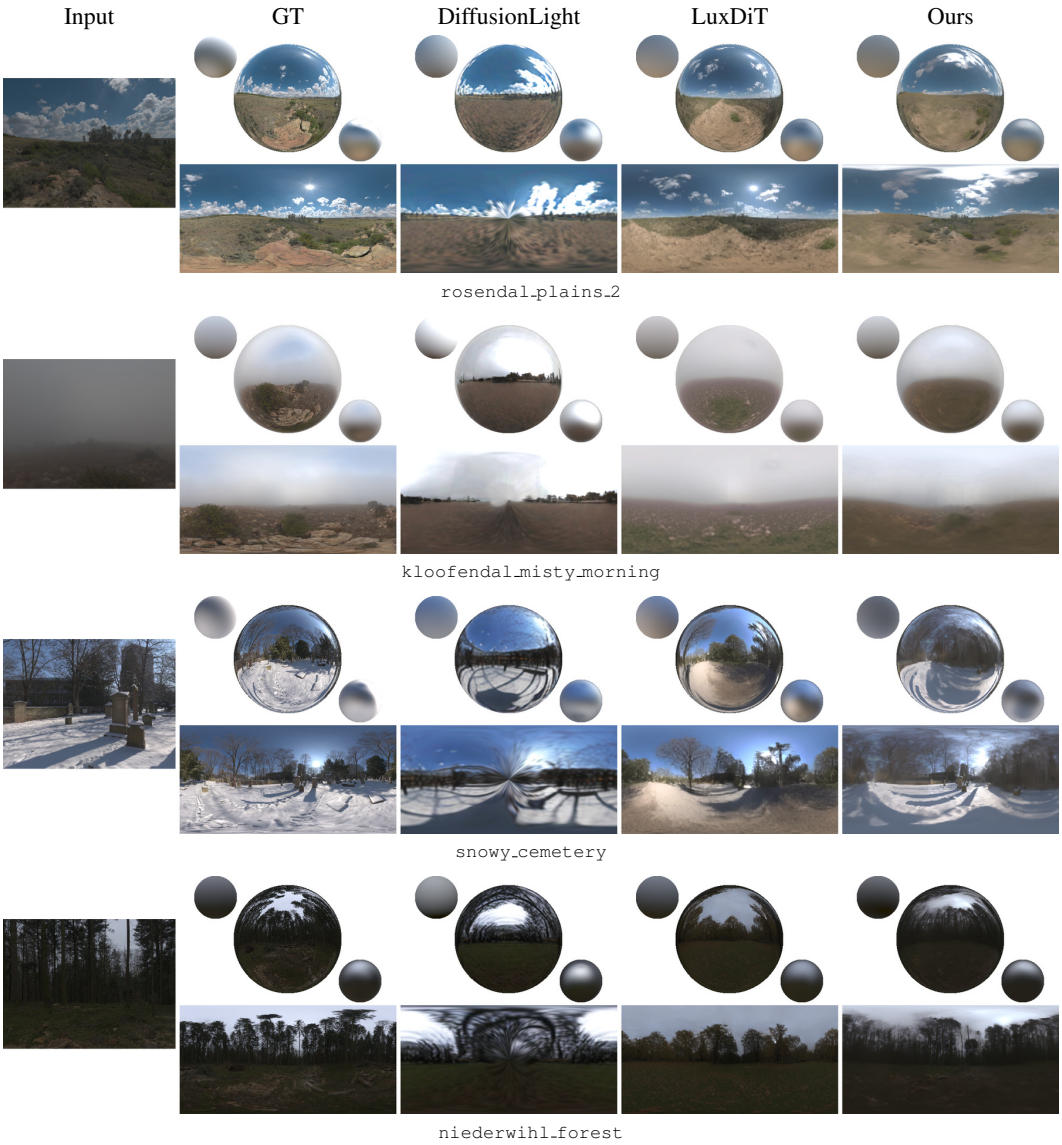

Figure 9: Further outdoor Poly Haven scenes from a single input view, in the layout of Figure 2. Continued in Figures 10 and 11.

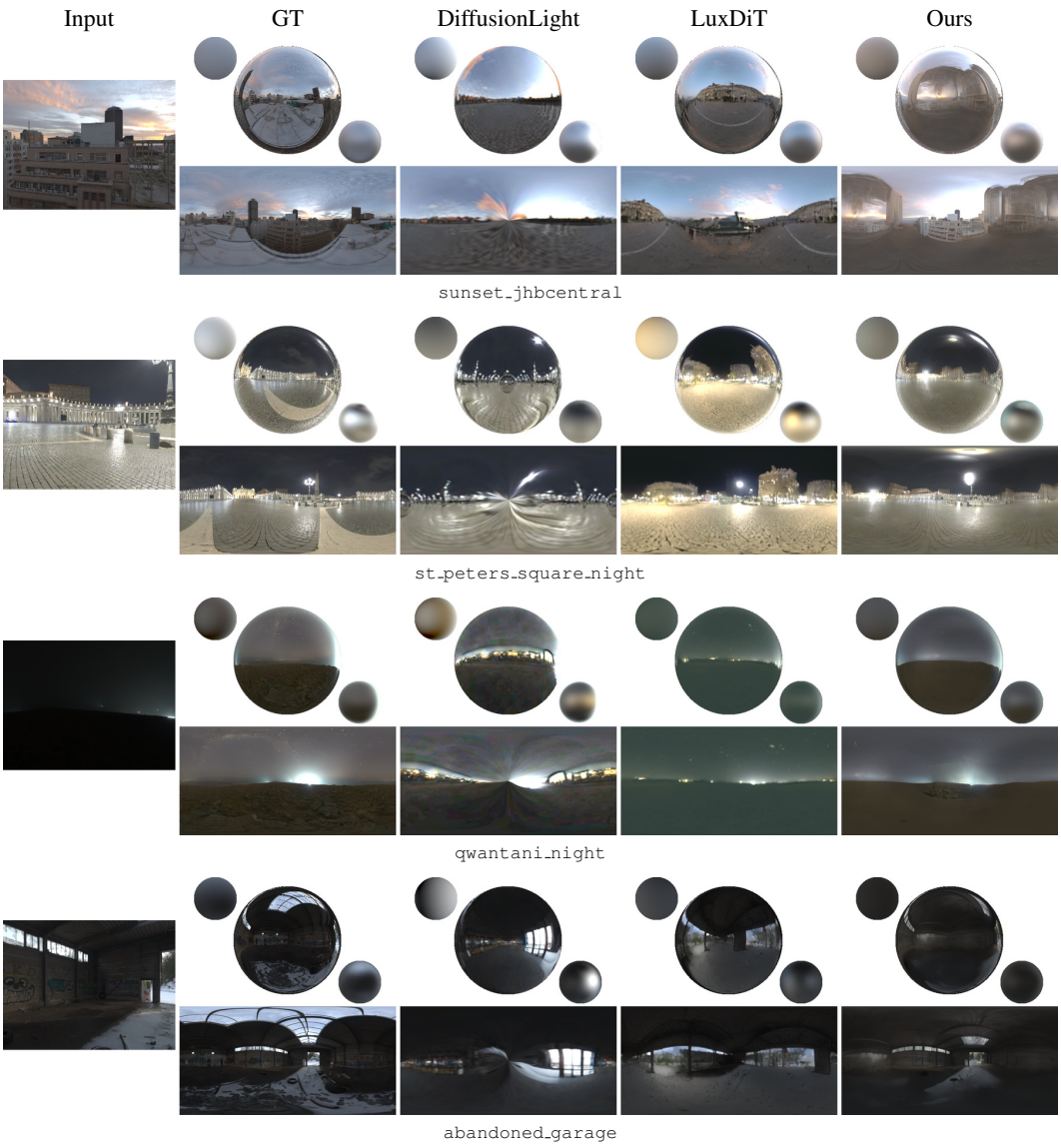

Figure 10: Continued from Figure 9. The first three rows are outdoor scenes and the last is indoor.



Figure 11: Further indoor Poly Haven scenes, continued from Figure 10.