

# Lighting in Motion: Spatiotemporal HDR Lighting Estimation

## Supplementary Material

### 6. Data generation details

We use Blender, paired with BlenderKit assets to procedurally generate indoor and outdoor renders. For indoor scenes, we use the full indoor scenes provided by BlenderKit, generating more cameras based on the original ones. Since scenes are not always completely modeled, we leverage existing camera assuming they point toward points of interest. We randomly sample a direction from the original camera frustum to obtain a target lookat point using ray-casting. Then we sample 3D points in the scene bounding box and check the visibility of the selected lookat point from them. If the point is visible from the sampled position, we consider the location to be a valid camera location and create a new camera pointed at the lookat.

For outdoor scenes, we select a central model from the building, vehicle or nature categories of BlenderKit. We add a ground plane, a random ground material, and add surrounding buildings, objects and vegetation using particle systems. We use HDRis from Polyhaven for lighting. We then derive cameras pointed at the central object from which it is visible. In total we use around 500 indoor scenes, reusing them for different motion for a total of 4400 scenes, and generate 1200 outdoor scenes. For each scene we render 4 viewpoints.

### 7. HDRI map optimization details

The predicted images from the network  $\hat{I}$  are cropped around the inpainted spheres. The same is done with the sphere mask, normals and position maps. The equirectangular HDRI is a Laplacian pyramid at a fixed resolution of 512x256 with 8 levels. We employ circular padding to leverage the cyclic nature of equirectangular maps. For faster convergence and better conditioning, the HDRI is defined in  $\log_2$  space. We optimize with Adam using a learning rate of  $5e-3$  for 1000 iterations per frame, for a total of 21 000 steps.

At every step, we randomly select a frame  $t$ , sphere type  $m$  (mirror or diffuse) and EV  $e$  from the predictions. The Laplacian pyramid is recomposed and is transformed back to linear space to obtain the HDRI map  $L_t$ . Then, the renderer  $\mathcal{R}$  is used to produce the image of corresponding sphere (mirror or diffuse). This rendered image is exposed according to the randomly selected EV and converted to sRGB color space to match the network’s prediction’s colors:

$$\hat{I}_t = sRGB(2^e \mathcal{R}(L_t, m)). \quad (7)$$

The loss function to optimize the HDRI representation is defined as:

$$\ell = M_{\text{sat}}(\ell_2(\hat{I}_t, I_t) + \frac{\lambda}{2}(\ell_1(\hat{I}_t, \hat{I}_{t-1}) + \ell_1(\hat{I}_t, \hat{I}_{t+1}))), \quad (8)$$

with  $\lambda = 0.1$  in all our experiments. The  $\ell_2$  loss enforces the rendered image to closely match the predicted image, and the two following  $\ell_1$  losses insure that the rendered image be similar to the neighboring frames, allowing for temporal smoothing. To prevent the saturated part of the image from lowering the overall intensity of the optimized HDRI, we define a saturation mask

$$M_{\text{sat}} = \begin{cases} 0, & \hat{I}_t > \tau \text{ and } I_t > \tau, \\ 1, & \text{otherwise.} \end{cases} \quad (9)$$

The renderer  $\mathcal{R}$  is a two modes differentiable Monte Carlo renderer for perfect mirror and perfect diffuse materials. The perfect reflection is implementing the reflection equation

$$\mathbf{v}_i - 2(\mathbf{v}_i \cdot \mathbf{n}_i) \mathbf{n}_i, \quad (10)$$

with  $\mathbf{v}_i$  computed from the sphere’s position map.

For the diffuse rendering, we first compute the luminance of the HDRI to use as importance weight:

$$L = 0.2126R + 0.7152G + 0.0722B \quad (11)$$

The importance map for each pixel of the HDRI map is computed as a multi-importance weighting of cosine and luminance:

$$w_i = (n_i \cdot r_i) L_i \sin(r_i), \quad (12)$$

where  $r_i$  is the ray direction corresponding to pixel  $i$  of the HDRI map. It is then normalized:

$$\mathbf{w} = \frac{w}{\sum_i w_i}. \quad (13)$$

The corresponding probability distribution function is computed by dividing the normalized importance map by the solid angle of the equirectangular map

$$PDF = \frac{\mathbf{w}}{\partial\omega}. \quad (14)$$

Samples  $s$  are drawn from the importance map  $\mathbf{w}$  and the final rendered colors  $R_i$  is

$$R_i = \frac{1}{S} \sum_{s \in S} \frac{L_s(n_i \cdot r_i)}{PDF}. \quad (15)$$

We use 64 samples with sub-pixel sampling in all our experiments.

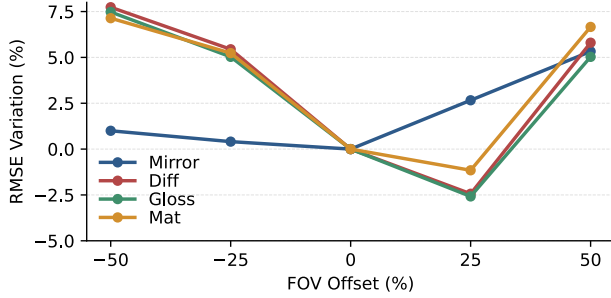


Figure 7. Effect of varying field of view on RMSE on the Infinigen dataset.

## 8. Ablation of field of view stability

As LiMO assumes known FOV, which is estimated from an off-the-shelf network in practice [44], we ablate the effect of perturbing the estimation in Fig. 7. We observe that LiMO is quite robust to varying FOVs. Interestingly, we observe a slight improvement (2%) in RMSE when the FOV is over-estimated by 25%. When the FOV is severely over- or under-estimated (by up to 50%), the performance hit is at most 8%.

## 9. Ablation on Laval Indoor SV

Ablations of our modules for the Laval Indoor SV Dataset [17] are presented in Tab. 6. Interestingly, while still improving colors (Angular error), the use of the diffuse prediction for the HDRi optimization results in less accurate overall intensity (RMSE). We hypothesize that this is because we predict spheres larger than those in the ground truth in order to increase the prediction resolution, biasing the prediction.

## 10. Additional results

In complement to Tab. 3, Tab. 5 reports metrics on our sequences test dataset for glossy and matte spheres.

Sample predictions from The Laval Indoor Spatially Varying HDR dataset [17] are presented in Fig. 8.

Sample predictions from the Laval Outdoor HDR Dataset [21] are shown in Fig. 10.

More in-the-wild results are presented in Fig. 9. We make use of the predicted pointcloud from the FOV and depthmap as shadow catcher when inserting objects in the scene.



Dataset	Method	RMSE $\downarrow$		SI-RMSE $\downarrow$		SSIM $\uparrow$		Ang. Err. $\downarrow$		T-LPIPS $\downarrow$		T-LPIPS-Diff $\downarrow$		Warped Err $\downarrow$	
		Gloss	Mat	Gloss	Mat	Gloss	Mat	Gloss	Mat	Gloss	Mat	Gloss	Mat	Gloss	Mat
Dynamic object	4D Lighting	0.30	0.33	0.21	0.34	0.89	0.88	4.6	4.9	0.0009	0.0006	0.0064	0.0016	0.0125	0.0099
	LiMo (image)	0.15	0.16	0.12	0.17	0.96	0.97	1.4	1.7	0.0224	0.0138	0.0166	0.0117	0.0489	0.0575
	LiMo (video)	0.18	0.21	0.13	0.21	0.96	0.96	2.9	3.0	0.0025	0.0020	0.0053	0.0017	0.0200	0.0156
Dynamic camera	4D Lighting	0.38	0.37	0.18	0.31	0.89	0.88	3.8	4.3	0.0011	0.0010	0.0047	0.0007	0.0163	0.0142
	LiMo (image)	0.16	0.17	0.12	0.18	0.96	0.97	1.4	1.8	0.0179	0.0114	0.0125	0.0102	0.0499	0.0541
	LiMo (video)	0.21	0.23	0.13	0.21	0.96	0.96	2.4	3.0	0.0019	0.0015	0.0042	0.0012	0.0178	0.0150
Dynamic lighting	4D Lighting	0.34	0.35	0.70	0.80	0.86	0.85	10.1	10.1	0.0008	0.0007	0.0011	0.0005	0.0095	0.0096
	LiMo (image)	0.17	0.18	0.13	0.19	0.95	0.96	1.8	2.2	0.0032	0.0022	0.0018	0.0015	0.0203	0.0217
	LiMo (video)	0.22	0.25	0.15	0.24	0.94	0.95	2.8	3.2	0.0009	0.0008	0.0010	0.0005	0.0077	0.0077
Combination	4D Lighting	0.32	0.33	0.20	0.33	0.89	0.88	4.0	4.3	0.0017	0.0013	0.0103	0.0018	0.0217	0.0167
	LiMo (image)	0.16	0.18	0.12	0.19	0.96	0.97	1.8	2.2	0.0249	0.0170	0.0137	0.0140	0.0615	0.0723
	LiMo (video)	0.23	0.24	0.16	0.24	0.94	0.94	2.7	3.0	0.0048	0.0042	0.0072	0.0033	0.0337	0.0289

Table 5. Quantitative evaluation of lighting estimation on dynamic scenes for “Gloss” (glossy) and “Mat” (matte) spheres in complement to Tab. 3. We compare LiMo with “4D Lighting” [42]. Results are color coded by best, second best.



Figure 8. Additional sample predictions from the Laval Indoor Spatially Varying test set [17] for, from left to right: DiffusionLight [31], 4D Lighting [42], and the image and video versions of the proposed LiMo. We visualize predictions by rendering the same four test spheres used for the quantitative metrics (see Tab. 2): mirror (top left), diffuse (top right), matte (bottom left) and glossy (bottom right).

Method	RMSE $\downarrow$				Si-RMSE $\downarrow$				SSIM $\uparrow$				Ang Err $\downarrow$			
	Mirr	Diff	Gloss	Mat	Mirr	Diff	Gloss	Mat	Mirr	Diff	Gloss	Mat	Mirr	Diff	Gloss	Mat
w/o Diffuse, Geo	0.328	0.231	0.246	0.269	0.651	0.181	0.201	0.310	0.796	0.960	0.943	0.941	5.22	3.26	3.31	3.46
w/o Diffuse	0.299	0.197	0.211	0.233	0.627	0.166	0.181	0.289	0.806	0.969	0.953	0.952	4.81	2.95	2.97	3.14
w/o Geo	0.333	0.238	0.253	0.277	0.614	0.182	0.200	0.300	0.800	0.958	0.942	0.939	4.95	3.05	3.09	3.35
L1Mo (full)	0.297	0.201	0.215	0.235	0.599	0.168	0.180	0.283	0.813	0.967	0.953	0.951	4.55	2.61	2.63	2.88

Table 6. Ablation of the use of the added geometric maps  $I_{\text{dir}}$  and  $I_{\text{dist}}$  for predictions (see Sec. 3.2) and diffuse sphere for HDRI optimization (see Sec. 3.1) on Laval Indoor SV with our image model. “Mirr” (mirror), “Diff” (diffuse), “Gloss” (glossy) and “Mat” (matte) refer to the different test spheres (see Sec. 4.2). Results are color coded by best, second and third best.

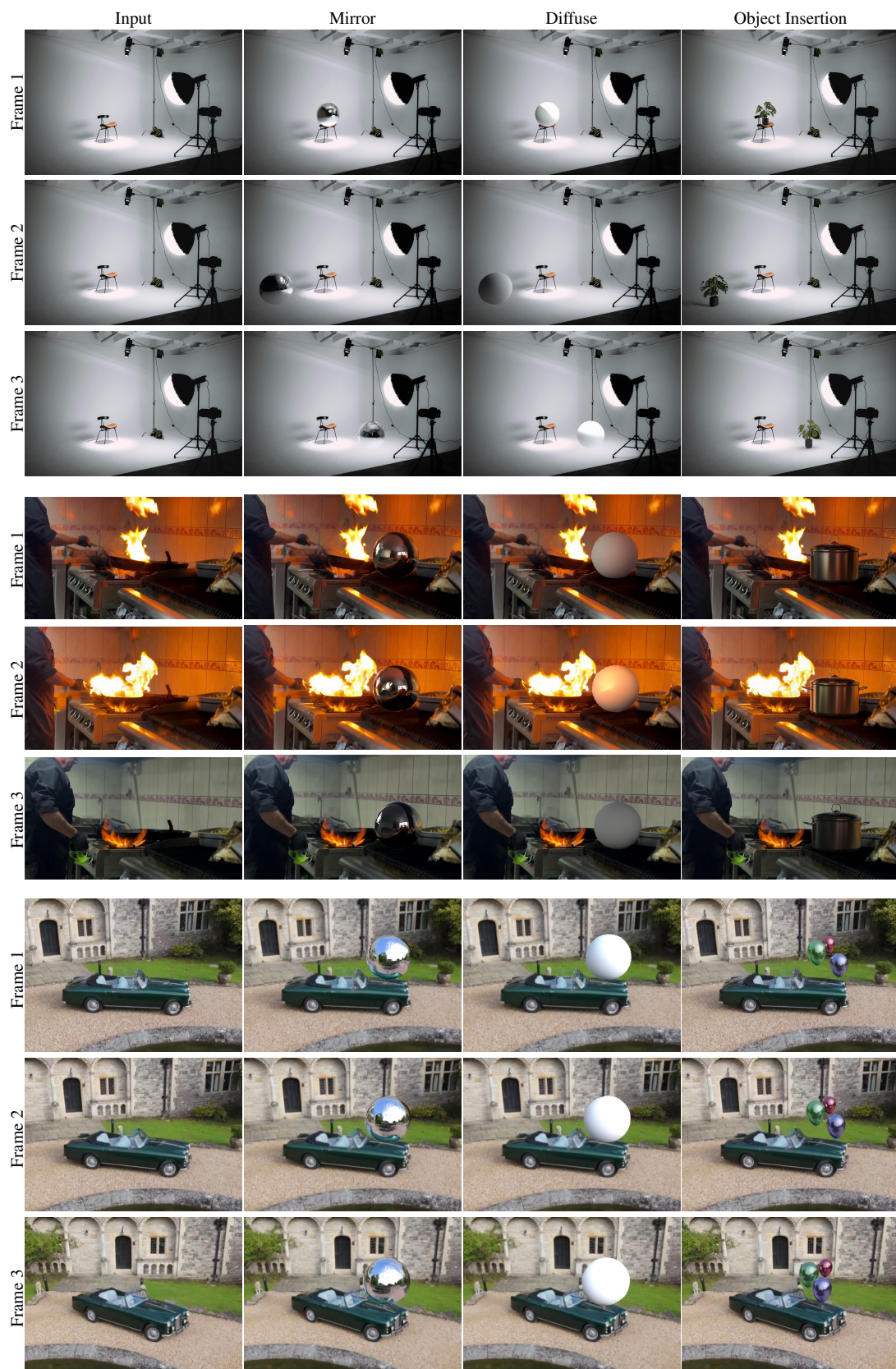


Figure 9. Additional examples of our method on in-the-wild images and videos, with from left to right: the input frame, the predicted mirror sphere at EV0, the predicted diffuse sphere at EV0 and the inserted object. The predicted pointcloud is used as shadow catcher.





Figure 10. Example qualitative prediction results on the Laval Outdoor HDR Dataset [21] with near-camera lighting estimations. 4D Lighting [42], trained exclusively on indoor scenes, cannot appropriately deal with outdoor scenes.