

EXPANDDIFF: DYNAMIC RANGE EXPANDING DIFFUSION FOR SINGLE-IMAGE HDR RECONSTRUCTION

SUPPLEMENTARY MATERIAL

Mehmet Emre Andiran, Zhuoqian Yang, Liying Lu, Mathieu Salzmann, Sabine Süssstrunk*

School of Computer and Communication Sciences, EPFL, Switzerland
firstname.lastname@epfl.ch, *memreandiran@gmail.com

S1. EXTENDED RELATED WORK

This supplement expands the literature discussion, documents the prediction and evaluation conventions, and reports supporting analyses and visual comparisons. Condition and model names follow the main paper.

S1.1. Reconstructing Lost Dynamic Range

Learning-based inverse tone mapping combines preservation of visible image content with inference in regions affected by clipping. Eilertsen et al. [1] reconstruct HDR from a single exposure with a CNN, emphasizing saturated highlights. ExpandNet [2] combines features at multiple spatial scales to predict HDR, while Santos et al. [3] use feature masking and perceptual supervision to reduce artifacts and synthesize missing detail. Liu et al. [4] explicitly decompose reconstruction into stages that reverse the camera pipeline. An alternative is to predict exposure brackets and merge them: Endo et al. [5] synthesize virtual exposures, and Lee et al. [6] use adversarial learning for recursive exposure generation.

These methods establish synthetic degradation and reconstruction as a training strategy for recovering information lost to limited dynamic range. Randomized clipping levels are themselves part of that strategy: Eilertsen et al. [1] set the exposure of each training crop so that clipping removes a uniformly drawn 5–15% of the image information, using a percentile of the image histogram to fix the scaling. Dynamic Clipping Synthesis (DCS) samples clipping levels at both ends independently, over ranges spanning near-unclipped to heavily clipped inputs, with thresholds taken from percentiles of the per-pixel channel minima and maxima, and rescales the surviving interval to the full range. It couples the resulting guidance with spatial conditioning and a perceptual prediction domain. The comparison between ExpandDiff-P and ExpandDiff-B studies how the training degradation relates to performance under different test conditions.

S1.2. Diffusion-Based Inverse Tone Mapping

Dalal et al. [7] formulate LDR-to-HDR conversion with conditional diffusion. Goswami et al. [8] use semantic guidance for diffusion-based inverse tone mapping. LEDiff [9] preserves a pretrained LDR latent space, trains dedicated highlight and shadow denoisers to generate exposure brackets, and combines the resulting codes with learned fusion and an HDR decoder. Its formulation addresses both directions of dynamic range expansion separately. Our evaluation reports the released highlight and shadow configurations as LEDiff-HL and LEDiff-SH; each treats the input as one end of the exposure bracket and synthesizes the rest with its own denoiser. ExpandDiff uses one conditional network to predict an HDR RGB

target directly, with both clipping directions handled within the same sampling trajectory.

ExpoCM [10] uses exposure-aware consistency trajectories to treat overexposed, underexposed, and well-exposed regions within one inference step. It is related in its goal of handling different exposure regimes jointly, but differs in its generative formulation and exposure-dependent supervision. ExpandDiff uses Gaussian forward diffusion, clean-target prediction, and 24-step DDIM sampling. ExpoCM is discussed as related work but is not included in the reported quantitative comparison, as no implementation was available at the time of writing.

Perceptual prediction domains also have precedents. Ke et al. [11] study encodings for learning on RAW and HDR images. X2HDR [12] uses perceptually uniform encodings to adapt pretrained diffusion models to HDR generation and RAW-to-HDR reconstruction. Refusion-HDR likewise predicts PU21-encoded HDR [13].

S1.3. Diffusion as a Restoration Framework

Diffusion models can serve as a general framework for conditional image reconstruction. Refusion [14] demonstrates this on four restoration tasks: shadow removal, dehazing, stereo super-resolution, and bokeh transformation. The AIM 2025 Refusion-HDR entry [13] uses a mean-reverting diffusion process with a modified NAFNet backbone. We distinguish this HDR adaptation from the original general-restoration method throughout our comparison.

RAW-Diffusion [15] provides the backbone adopted by ExpandDiff: RGB guidance, spatially adaptive normalization, clean-image prediction, and a bounded output head. We repurpose it for three-channel HDR reconstruction and train it on HDR-derived clipping pairs. The diffusion machinery and conditioning modules are inherited; we change the target task, the training distribution, and the prediction domain.

S2. METHOD DETAILS

S2.1. Target Scale and Dynamic Clipping Synthesis

The HDR target is display-referred. After rescaling each training image to a median luminance of 20 cd/m^2 , values above 1000 cd/m^2 are clipped and the image is divided by 1000. A normalized target value y thus represents an absolute linear intensity of $1000y \text{ cd/m}^2$ on the chosen display scale.

Dynamic Clipping Synthesis (DCS) draws new lower and upper clipping percentiles for every training pair, from ranges that are

fixed throughout training; each HDR target is paired with three independently clipped guidance images. The lower and upper thresholds are percentiles of the per-pixel channel minima and maxima, respectively, and apply to every channel; the surviving interval is then rescaled affinely. A pixel therefore counts as clipped if *any* channel reaches a boundary; it need not be uniformly white or black. Measured clipping fractions can differ from the nominal percentiles after quantization or when many pixels share the same intensity.

S2.2. Guidance and Prediction Domain

The EDSR [16] guidance encoder has four residual blocks with 64 channels and no upsampling head. At each conditioned U-Net block, its feature map is resampled to the block resolution. A shared 3×3 convolution with 128 hidden channels precedes the prediction of the spatial γ and β maps [17]. These maps modulate group-normalized activations without learned affine normalization parameters. The conditioning is spatial at the bottleneck and decoder resolutions; self-attention acts at the two coarsest resolutions and the bottleneck.

The PU21 curve [18] is calibrated in absolute luminance, so it is applied to $L = 1000Y$ rather than to the normalized target Y itself:

$$\tilde{U}(L) = p_7 \left[\left(\frac{p_1 + p_2 L^{p_4}}{1 + p_3 L^{p_4}} \right)^{p_5} - p_6 \right], \quad (\text{S1})$$

where $p_1, \dots, p_7$ are the published `banding_glare` coefficients. The result is then renormalized by its value at the peak, giving $x_0 = 2U(1000Y)/U(1000) - 1$ with $U(L) = \tilde{U}(\max\{L, 0.005\})$, which is zero at black. The display scale of Sec. S2.1 is therefore undone before the curve and reimposed after it, and both endpoints are preserved: $Y = 0$ maps to $x_0 = -1$ and $Y = 1$ to $x_0 = 1$. Applying the transform per channel retains an image-space RGB representation. The inverse maps the prediction back to normalized linear HDR, which is multiplied by 1000 when expressed in absolute display units.

The bounded-head argument concerns local output sensitivity. At a desired output $a = \tanh(z)$, the derivative with respect to the preactivation is $1 - a^2$. A linearly encoded intensity of 0.5 cd/m^2 corresponds to $a = -0.999$ and a derivative of approximately 0.002; under PU21 the same intensity maps to $a = -0.894$, where the derivative is 0.20, a hundredfold increase. Encoding also changes the relative weighting of intensity errors in the reconstruction loss. The ablation measures their combined effect.

S2.3. Training Configurations and Sampling

ExpandDiff-P and ExpandDiff-B use the same four HDR sources, batch size 32, PU21 target, and SNR-weighted $L_1 + L_2$ loss. Clipping thresholds are computed on each full 512×512 image, and training draws random 256×256 crops with random rotations and flips. ExpandDiff-P samples $q_{lo} \sim \mathcal{U}[0, 10]$ and $q_{hi} \sim \mathcal{U}[0, 30]$. ExpandDiff-B samples $q_{lo} = 0$ and $q_{hi} \sim \mathcal{U}[3, 7]$, applies the randomized two-parameter camera response of Eilertsen et al. [1], also used by LEDiff [9], then performs an 8-bit round trip. The reported ExpandDiff-P and ExpandDiff-B checkpoints use 150k and 75k training steps at learning rates 2×10^{-4} and 10^{-4} , respectively. The four target/head ablation configurations each use the ExpandDiff-P setup for 150k steps. All checkpoints use AdamW without weight decay, cosine learning-rate decay to zero, and $T = 1000$ diffusion steps with a linear noise schedule.

Inference runs one 24-step DDIM [19] trajectory, i.e., 24 denoiser evaluations, for both clipping directions. At 512×512 , the measured runtime is 3.19 s on an NVIDIA A100 80 GB. Reflection

padding to multiples of 64 permits other input dimensions within memory limits. For ExpandDiff-P on C_{95} , increasing the step count to 48 and 100 changes gain-aligned PU21-PSNR by a paired -0.17 and -0.23 dB (95% CIs $[-0.38, +0.03]$ and $[-0.54, +0.08]$; measured on the floating-point rather than the 8-bit input), so additional steps give no evidence of improved accuracy.

S2.4. Linearizing the Inference Input

The guidance I of Eq. (1) in the main paper is linear, so an 8-bit input must be linearized before it can be used. We apply the inverse sRGB transfer function [20] to $v/255$ for each 8-bit value v ,

$$I = \begin{cases} \frac{v/255}{12.92}, & v/255 \leq 0.04045, \\ \left(\frac{v/255 + 0.055}{1.055} \right)^{2.4}, & \text{otherwise,} \end{cases} \quad (\text{S2})$$

which is piecewise rather than a single power law; the linear segment near black mirrors the encoding function, whose pure-gamma form would have unbounded slope at the origin. The same 8-bit input is used for every method and every condition. Our pipeline linearizes that 8-bit image with Eq. (S2), regardless of the original response that produced it.

This is an approximation whenever the true camera response differs from sRGB, and a single 8-bit image does not determine that response. ExpandDiff-B is trained through this mismatch rather than around it: its LDR guidance is formed with a randomized two-parameter response and an 8-bit round trip, and is then linearized with Eq. (S2) rather than with the response that produced it.

S3. SUPPORTING EXPERIMENTS

S3.1. Evaluation Conditions and Metric Interpretation

All conditions use the same 181 SI-HDR references and the same 512×512 center crop and box-filter resize, applied to the benchmark-supplied C_{95} images as well. The clipping fractions below are measured on the resulting 8-bit inputs, after that resampling and after quantization, so they need not match the nominal percentiles. A pixel counts as clipped when any channel reaches a boundary, as in Sec. S2.1. C_{95} averages 8.17% highlight clipping and zero shadow clipping. The constructed C_p inputs use fixed percentiles $(q_{lo}, q_{hi}) = (5, 15)$, yielding averages of 5.87% shadow clipping and 15.40% highlight clipping, with every one of the 181 images clipped at both ends.

SI-HDR emphasizes the sensitivity of objective scores to tone and color differences [21, 22]. We fit the brightness gain using only pixels clipped at neither boundary, then apply the benchmark’s 20-parameter tone-and-color correction on top of it for full-reference scoring. Correcting after the gain favors the competing methods rather than ours (Sec. S3.5). PIQE and FID-R are reported on the gain-only basis. Both alignments use the HDR reference and are part of evaluation, not inference.

Competing methods are run from their released weights on the same 8-bit input at 512×512 . From AIM 2025 [13], we evaluate Refusion-HDR and DITM, the two highest-ranked entries with released code, which placed third and fifth.

LEDiff’s pipeline requires a text prompt, which their paper does not specify for inverse tone mapping; in correspondence, the authors confirmed that they generated one per input image with LLaVA, so we caption every image with LLaVA-1.5-7B [23] under a single deterministic strategy. Additionally, LEDiff carries its own published

post-process, a saturation-gated soft-mask blend of the generated content with the input under a fitted gamma and exposure. We retain it because it is part of their method, and we use the constants from their supplementary material rather than those in their released code, since these are the values their paper describes. The shared alignment is then applied on top, as it is for every method.

PU21-encoded metrics use the `banding_glare` coefficients on absolute luminance clamped to $[0.005, 10000]$ cd/m^2 , the convention of the reference implementation [18]; this differs from the target encoding of Sec. S2.2, which additionally normalizes by $U(1000)$. HDR-VDP-3 [24] (v3.0.7, quality task, `rgb-native`) assumes a 24-inch display at 1920×1080 viewed from 1 m with a 1000 cd/m^2 peak, identically for every method and condition.

For FID-R, we compute FID [25] on 60 crops of 128×128 from each Reinhard-tone-mapped [26] prediction, giving 10,860 prediction crops. The reference distribution contains 20,000 crops sampled at independent locations from the same target images. We average scores over seeds $\{1234, 7, 99\}$. This follows the use of tone-mapped FID in LEDiff [9].

S3.2. Training Distribution and the Earlier ExpandDiff-D Checkpoint

At the reported checkpoints, the gain-aligned PSNR of ExpandDiff-B relative to ExpandDiff-P changes from $+1.20 \text{ dB}$ on C_{95} to -8.08 dB on C_p , a 9.28 dB change in relative performance. Because the reported ExpandDiff-P checkpoint was trained twice as long and at twice the learning rate, we also compare at matched steps: a 75k-step snapshot of ExpandDiff-P’s 150k run gives 21.38 dB on C_{95} , 1.23 dB below ExpandDiff-B (paired; 95% CI $[0.50, 1.96]$). This supports the advantage of the benchmark-oriented training regime on that condition.

For completeness, Table S1 retains ExpandDiff-D, an earlier checkpoint trained on HDR+’s 8-bit JPEG data [27]. ExpandDiff-D samples its clipping thresholds in stops below white ($\mathcal{U}[-12, -6]$ for shadows, $\mathcal{U}[-4, 0]$ for highlights) rather than as percentiles, and uses a linear target, an additional logarithmic error term, batch size 8, and 150k training steps. It obtains 20.97 dB on C_{95} and 25.61 dB on C_p . Sec. S3.3 reports a further checkpoint trained on a different degradation, evaluated under two additional input conditions and on C_{95} and C_p .

The unprocessed input remains an informative baseline. On C_{95} , all three ExpandDiff checkpoints in Table S1 exceed its gain-aligned PU21-PSNR, whereas none of the six competing configurations do. ExpandDiff-P and ExpandDiff-D improve all eight reported scores over the input, and ExpandDiff-B improves seven, with a PIQE 0.07 higher than the input’s. Under C_p , two competing models, MaskHDR and ExpandNet, also exceed the input in gain-aligned PU21-PSNR.

S3.3. Additional Input Conditions

C_0 applies LEDiff’s capture model [9]: per-scene HDR-FLIP exposures [28] with their randomized camera response, taken at the middle exposure E0. It blows 13.43% of pixels and crushes 5.33% , close to C_p on average, but only 27 of the 181 images exceed 1% at both ends, whereas every C_p image is clipped at both; similar average fractions do not imply equivalent reconstruction tasks. C_p^{hard} applies Eq. (1) of the main paper at $q_{\text{lo}} = 10$ and $q_{\text{hi}} = 30$, the upper end of ExpandDiff-P’s training ranges rather than their midpoints, blowing 27.99% and crushing 10.42% , with every image clipped at both ends.

Table S1. Table 1 of the main paper with ExpandDiff-D added: results on 181 SI-HDR references under two input conditions. PSNR, VSI, and PIQE use PU21; VDP-3 is in JOD. All scores include brightness alignment; CRF additionally applies the benchmark’s tone-and-color correction. FID denotes FID-R. Best values per condition are bold and second-best values underlined.

Method	PSNR $\uparrow$		VSI $\uparrow$		VDP-3 $\uparrow$		PIQE	FID
	gain	CRF	gain	CRF	gain	CRF	$\downarrow$	$\downarrow$
C_{95}								
ExpandDiff-B	22.61	30.15	.9866	.9897	9.06	9.46	28.22	8.76
ExpandDiff-P	21.41	28.65	.9801	.9867	<u>8.77</u>	9.35	27.32	9.99
ExpandDiff-D	20.97	<u>28.88</u>	<u>.9802</u>	.9872	8.73	9.34	<u>27.33</u>	9.90
LDR input	20.47	27.93	.9756	.9826	8.44	9.15	28.15	10.00
ExpandNet	19.17	28.09	.9743	.9840	8.37	9.12	28.60	10.65
DITM	19.06	27.64	.9767	.9857	8.52	9.00	30.43	10.70
LEDiff-HL	18.95	26.45	.9722	.9803	8.33	8.97	29.17	12.98
MaskHDR	18.77	28.51	.9767	<u>.9874</u>	8.59	<u>9.37</u>	29.20	11.92
LEDiff-SH	18.75	26.16	.9708	.9781	8.33	8.84	27.76	12.19
Refusion-HDR	16.85	27.77	.9555	.9804	8.10	9.08	45.24	<u>9.58</u>
C_p								
ExpandDiff-P	31.21	33.52	.9913	.9917	9.62	9.68	39.31	4.78
ExpandDiff-D	25.61	29.58	.9835	.9860	9.20	9.48	37.92	7.40
MaskHDR	23.88	28.72	<u>.9806</u>	<u>.9884</u>	<u>9.28</u>	<u>9.65</u>	38.78	7.88
ExpandDiff-B	23.13	28.27	.9754	.9845	8.98	9.42	39.06	8.99
ExpandNet	22.74	28.34	.9781	.9828	8.76	9.20	40.17	6.89
LDR input	22.23	27.37	.9699	.9754	8.49	9.00	40.11	<u>6.75</u>
LEDiff-HL	20.99	25.18	.9633	.9771	8.29	8.90	<u>37.81</u>	12.41
DITM	20.50	25.98	.9727	.9799	8.36	8.79	36.83	11.66
Refusion-HDR	19.46	25.82	.9629	.9757	8.34	9.04	44.24	11.89
LEDiff-SH	19.14	25.37	.9659	.9733	8.45	8.80	37.90	9.73

ExpandDiff-S is a fourth checkpoint, reported only here. It matches ExpandDiff-P in target, objective, batch size, training steps, learning rate and data sources, and differs only in its training degradation: LEDiff’s capture model, with the E−, E0 and E+ exposures emitted as separate training pairs. ExpandDiff-S is therefore in distribution at C_0 and out of distribution at C_p^{hard} .

At C_0 (Table S2), ExpandDiff-S leads every column except gain-aligned PSNR and PIQE. In gain-aligned PSNR, ExpandDiff-B is ahead of ExpandDiff-S by 0.62 dB and exceeds the strongest competing model, DITM, by a paired 2.44 dB ; in corrected PSNR, ExpandDiff-S exceeds the strongest competing model there, MaskHDR, by 2.32 dB . At C_p^{hard} , ExpandDiff-P leads by a paired 4.76 dB PU21-PSNR and takes every column except PIQE. An ExpandDiff checkpoint is best in every column of both conditions except PIQE, and the checkpoint trained on the matching degradation takes most columns. Every ExpandDiff variant, including those evaluated outside their training distribution, exceeds the unprocessed input and at least four of the six competing configurations in gain-aligned PSNR under both conditions.

On C_{95} and C_p , ExpandDiff-S reaches 18.99 and 21.47 dB gain-aligned PU21-PSNR, below both ExpandDiff-B (22.61 and 23.13 dB) and ExpandDiff-P (21.41 and 31.21 dB). ExpandDiff-B outperforms it in every column on C_{95} , as does ExpandDiff-P on C_p .

S3.4. Paired Reconstruction Results

Table S3 collects the paired mean PU21-PSNR differences and confidence intervals for the comparisons reported above and in the main

Table S2. Results under the two additional input conditions, 181 images each against the same references. Columns and emphasis follow Table S1; the two blocks are separate comparisons.

Method	PSNR $\uparrow$		VSI $\uparrow$		VDP-3 $\uparrow$		PIQE	FID
	gain	CRF	gain	CRF	gain	CRF	$\downarrow$	$\downarrow$
C_0								
ExpandDiff-B	24.53	30.07	.9835	.9865	8.79	9.04	39.66	7.61
ExpandDiff-S	23.91	30.57	.9860	.9891	8.93	9.28	39.85	6.97
DITM	22.10	26.86	.9803	.9834	8.49	8.67	38.24	10.81
Refusion-HDR	19.14	26.93	.9675	.9810	8.29	8.98	45.65	10.35
ExpandDiff-P	18.21	28.04	.9771	.9857	8.32	9.14	40.22	7.89
ExpandDiff-D	18.11	27.24	.9765	.9838	8.34	9.00	38.45	11.60
LEDiff-HL	17.70	26.41	.9689	.9786	8.13	8.87	40.43	8.33
MaskHDR	17.38	28.25	.9751	.9864	8.24	9.19	40.82	7.59
<i>LDR input</i>	16.96	27.87	.9700	.9800	7.83	8.77	42.24	7.75
LEDiff-SH	16.18	26.32	.9682	.9785	8.19	8.87	41.42	9.42
ExpandNet	15.93	26.33	.9595	.9793	7.54	8.55	45.02	9.72
C_p^{hard}								
ExpandDiff-P	26.22	27.85	.9781	.9788	8.73	8.83	39.99	8.09
ExpandDiff-S	21.69	25.98	.9650	.9756	8.41	8.79	40.41	12.10
ExpandDiff-D	21.61	25.37	.9668	.9716	8.18	8.59	38.82	11.58
MaskHDR	21.48	25.30	.9647	.9751	8.40	8.71	38.61	11.24
ExpandDiff-B	20.89	24.38	.9546	.9657	7.99	8.32	39.79	13.69
ExpandNet	20.67	24.78	.9596	.9659	7.74	8.11	40.14	11.62
DITM	19.06	23.18	.9559	.9646	7.67	7.96	37.99	14.98
LEDiff-HL	18.61	22.75	.9485	.9641	7.64	8.17	37.45	17.53
Refusion-HDR	17.87	22.61	.9434	.9540	7.48	7.94	44.84	17.45
<i>LDR input</i>	17.57	22.64	.9383	.9485	6.79	7.47	41.36	11.46
LEDiff-SH	17.00	22.23	.9440	.9544	7.57	7.86	37.50	16.40

paper. Each difference is computed per image against the competing method with the highest mean PU21-PSNR for that condition, then averaged over the 181 images. The confidence intervals quantify variation across the evaluated images, not across independently trained models. The corresponding full metric columns appear in Tables S1 and S2.

Table S3. Paired gain-aligned PU21-PSNR improvements over the strongest competing model on each condition. All comparisons use 181 images. CI denotes a 95% confidence interval; wins count the images on which the ExpandDiff checkpoint scores higher.

Condition	Model	Against	Δ dB	95% CI	Wins
C_{95}	ExpandDiff-B	ExpandNet	3.43	[2.54, 4.33]	130
C_{95}	ExpandDiff-P	ExpandNet	2.24	[1.50, 2.98]	121
C_{95}	ExpandDiff-D	ExpandNet	1.80	[1.08, 2.52]	106
C_p	ExpandDiff-P	MaskHDR	7.34	[6.52, 8.17]	175
C_0	ExpandDiff-B	DITM	2.44	[1.46, 3.42]	107
C_p^{hard}	ExpandDiff-P	MaskHDR	4.76	[4.21, 5.30]	173

S3.5. Effect of Brightness and CRF Alignment

Table S4 summarizes the gain fitted to predictions on C_{95} . The median is computed over $|\log_2 s|$ and measures deviation from unit gain; the interquartile range is computed over signed $\log_2 s$. ExpandDiff-B and ExpandDiff-P have the smallest values for both statistics, indicating that their predicted brightness scale is closer to the reference under this protocol.

Table S4. Brightness alignment on C_{95} , in stops. The median uses absolute gain offsets; the IQR uses signed offsets.

Method	Median $ \log_2 s $	IQR of $\log_2 s$
ExpandDiff-B	0.41	0.72
ExpandDiff-P	0.48	0.72
ExpandNet	0.75	1.71
ExpandDiff-D	0.96	1.96
Refusion-HDR	1.03	1.29
MaskHDR	1.62	2.00
DITM	2.27	2.43
LEDiff-HL	2.60	1.65
LEDiff-SH	2.86	1.70

CRF correction improves the mean PU21-PSNR of ExpandDiff-B and ExpandDiff-P by 7.54 and 7.24 dB, respectively, compared with 9.74 dB for MaskHDR and 10.92 dB for Refusion-HDR, the two largest among competing methods, and ExpandDiff-B retains the highest corrected PSNR. The gain-only advantage is thus partly reduced by a more flexible tone-and-color fit, but does not disappear.

The order of the two alignments matters, and the order used here is the less favorable one for ExpandDiff. On C_{95} , fitting the gain before the CRF correction raises the corrected PU21-PSNR of every method relative to correcting its unaligned prediction, but unevenly: by 0.09 and 0.15 dB for ExpandDiff-P and ExpandDiff-B and a similar 0.13 dB for ExpandNet, against 0.22 to 1.40 dB for DITM, Refusion-HDR and MaskHDR, and 4.15 and 8.00 dB for LEDiff-HL and LEDiff-SH (paired means over the 181 images; Table S5). Correcting unaligned predictions would therefore widen ExpandDiff-B's lead in mean corrected PU21-PSNR over the strongest competing method on C_{95} from 1.64 to 2.04 dB, with ExpandNet replacing MaskHDR in that position.

Table S5. Effect of the alignment order on corrected PU21-PSNR (dB) on C_{95} . The benchmark's 20-parameter correction is fitted either to the unaligned prediction or, as in all reported CRF scores, after gain alignment. Δ is the paired gain from aligning first, over 181 images, with its 95% confidence interval; it can differ slightly from subtracting the rounded columns.

Method	Unaligned	After gain	Δ	95% CI
ExpandDiff-P	28.56	28.65	+0.09	[+0.06, +0.12]
ExpandNet	27.96	28.09	+0.13	[+0.09, +0.17]
ExpandDiff-B	30.00	30.15	+0.15	[+0.10, +0.19]
DITM	27.42	27.64	+0.22	[+0.15, +0.28]
Refusion-HDR	27.50	27.77	+0.27	[+0.20, +0.34]
MaskHDR	27.10	28.51	+1.40	[+1.20, +1.61]
LEDiff-HL	22.30	26.45	+4.15	[+3.70, +4.61]
LEDiff-SH	18.16	26.16	+8.00	[+7.38, +8.61]

The benchmark's CRF correction has 20 parameters: 4 for luminance, a cubic $f(P)$ in PQ-encoded luminance P with basis $[P^3, P^2, P, 1]$, and 16 for chroma, a cubic $g(u', v')$ in the CIE 1976 chromaticity coordinates (u', v') with basis $[u'^3, v'^3, u'^2, v'^2, u'v', u', v', 1]$ for each of the two output channels. The luminance fit is unregularized, whereas the chroma fit is regularized toward the identity mapping. To measure how much the CRF correction must change each method's output, Table S6 reports the RMS deviation of the fitted correction from identity, evaluated on each prediction's own pixels: for luminance, the change $f(P) - P$; for chroma, the change $g(u', v') - (u', v')$. Because the correction is applied

after gain alignment, this measures the tone and color mismatch that a single gain cannot absorb. ExpandDiff-B requires the smallest correction on C_{95} in both luminance and chroma, and on C_p . ExpandDiff-P needs less than a third of the luminance correction of any competing method. The three competing methods that need the largest luminance correction on C_{95} , Refusion-HDR, ExpandNet and MaskHDR, are also the three whose PSNR gains most from it.

Table S6. Size of the benchmark’s tone-and-color correction: RMS deviation from identity of the fitted correction over each prediction’s pixels, median over 181 images, multiplied by 100 for readability. Luminance is in PQ-encoded units, chroma in CIE 1976 (u' , v') chromaticity. Lower means the prediction already matches the reference’s tone and color more closely. Best values per column are bold and second-best underlined.

Method	C_{95}		C_p	
	Luma	Chroma	Luma	Chroma
ExpandDiff-B	2.66	0.41	2.19	0.72
ExpandDiff-P	<u>2.97</u>	0.52	0.55	0.18
ExpandDiff-D	3.36	0.50	<u>1.41</u>	<u>0.51</u>
<i>LDR input</i>	3.58	0.63	2.60	0.95
ExpandNet	4.11	0.74	2.31	0.66
DITM	3.80	<u>0.43</u>	2.66	0.82
LEDiff-HL	3.92	0.71	2.50	1.17
MaskHDR	4.01	0.93	1.77	1.26
LEDiff-SH	3.45	1.20	3.46	1.65
Refusion-HDR	5.75	0.49	3.12	1.04

S3.6. Additional Qualitative Examples

Figs. S1–S4 show additional SI-HDR examples of highlight-only and two-sided clipping. Each example occupies two rows: the input and clipping mask, five competing configurations (all except LEDiff-SH), ExpandDiff-B, ExpandDiff-P, and the HDR reference. The C_0 examples in Fig. S3 add ExpandDiff-S, which is in distribution there. The masks mark pixels with a channel at a clipping boundary and are not inputs to ExpandDiff.

To show the methods on subjects and lighting outside SI-HDR’s range, Fig. S5 adds two hand-picked HDR+ captures [27]. Each capture’s final JPEG is linearized with the inverse sRGB transfer function, center-cropped and box-filtered to 512×512 as for SI-HDR, clipped with Eq. (1) of the main paper at the C_p percentiles, and passed through the same 8-bit round trip. All methods are run as in the benchmark evaluation. An HDR+ image is already tone-mapped by its capture pipeline, so the rightmost panel is the unclipped image the input was derived from rather than an HDR reference.

S4. LIMITATIONS AND SCOPE

All four scored conditions degrade the same 181 SI-HDR references, so they measure transfer across degradations over shared scenes rather than generalization to independent scene collections; the HDR+ captures of Fig. S5 are shown qualitatively only. C_p lies within ExpandDiff-P’s training percentile ranges and C_p^{hard} at their upper edge, while the benchmark-supplied C_{95} and LEDiff’s C_0 test degradations it was not trained on.

ExpandDiff predicts display-referred HDR up to 1000 cd/m^2 , not unrestricted scene radiance, and, like other methods, cannot know a scene’s absolute brightness. DCS varies clipping-related

information loss but omits other camera effects, including signal-dependent sensor noise and spatially varying tone processing, and inverse sRGB linearization only approximates an unknown camera response.

Reconstruction in clipped regions is inherently ambiguous: a plausible prediction need not reproduce the original scene detail, and reconstructions weaken over large contiguous clipped regions. The target/head ablation shows that the best full-reference scores need not coincide with the best PIQE or FID-R. Reducing sampling cost and evaluating these tradeoffs in HDR viewing studies are directions for further work.

S5. REFERENCES

- [1] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafal K. Mantiuk, and Jonas Unger, “HDR image reconstruction from a single exposure using deep CNNs,” *ACM Trans. Graph.*, vol. 36, no. 6, Nov. 2017.
- [2] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debatista, “ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content,” *Computer Graphics Forum*, vol. 37, no. 2, pp. 37–49, 2018.
- [3] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari, “Single image HDR reconstruction using a CNN with masked features and perceptual loss,” *ACM Trans. Graph.*, vol. 39, no. 4, Aug. 2020.
- [4] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1648–1657.
- [5] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani, “Deep reverse tone mapping,” *ACM Trans. Graph.*, vol. 36, no. 6, Nov. 2017.
- [6] Siyeon Lee, Gwon Hwan An, and Suk-Ju Kang, “Deep recursive HDRI: Inverse tone mapping using generative adversarial networks,” in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part II*, Berlin, Heidelberg, 2018, p. 613–628, Springer-Verlag.
- [7] Dwip Dalal, Gautam Vashishtha, Prajwal Singh, and Shanmuganathan Raman, “Single image LDR to HDR conversion using conditional diffusion,” in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 3533–3537.
- [8] Abhishek Goswami, Aru Ranjan Singh, Francesco Banterle, Kurt Debattista, and Thomas Bashford-Rogers, “Semantic aware diffusion inverse tone mapping,” *Journal of Physics: Conference Series*, vol. 3128, no. 1, pp. 012009, Oct. 2025.
- [9] Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang, “LEDiff: Latent exposure diffusion for HDR generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 453–464.
- [10] Aoyu Liu, Zhen Liu, Ziyi Wang, Dian Chen, Bing Zeng, and Shuaicheng Liu, “ExpoCM: Exposure-aware one-step generative single-image HDR reconstruction,” arXiv:2605.02464, 2026.

- [11] Andrew Yanzhe Ke, Lei Luo, Xiaoyu Xiang, Yuchen Fan, Rakesh Ranjan, Alexandre Chapiro, and Rafał Mantiuk, “Training neural networks on RAW and HDR images for restoration tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2025, pp. 810–819.
- [12] Ronghuan Wu, Wanchao Su, Kede Ma, Jing Liao, and Rafał K. Mantiuk, “X2HDR: HDR image generation in a perceptually uniform space,” arXiv:2602.04814, 2026.
- [13] Chao Wang et al., “AIM 2025 challenge on inverse tone mapping report: Methods and results,” arXiv:2508.13479, 2025.
- [14] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön, “Refusion: Enabling large-size realistic image restoration with latent-space diffusion models,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 1680–1691.
- [15] Christoph Reinders, Radu Berdan, Beril Besbinar, Junji Otsuka, and Daisuke Iso, “RAW-Diffusion: RGB-guided diffusion models for high-fidelity RAW image generation,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 8431–8443.
- [16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 2332–2341.
- [18] Rafał K. Mantiuk and Maryam Azimi, “PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR,” in *2021 Picture Coding Symposium (PCS)*, June 2021, pp. 1–5.
- [19] Jiaming Song, Chenlin Meng, and Stefano Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations*, 2021.
- [20] IEC, “Multimedia systems and equipment – colour measurement and management – part 2-1: Colour management – default RGB colour space – sRGB,” Standard IEC 61966-2-1:1999, International Electrotechnical Commission, Oct. 1999.
- [21] Param Hanji, Rafał Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger, “Comparison of single image HDR reconstruction methods — the caveats of quality assessment,” in *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022, SIGGRAPH ’22, Association for Computing Machinery.
- [22] Gabriel Eilertsen, Saghi Hajisharif, Param Hanji, Apostolia Tsirikoglou, Rafał K. Mantiuk, and Jonas Unger, “How to cheat with metrics in single-image HDR reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 3998–4007.
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, “Improved baselines with visual instruction tuning,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26286–26296.
- [24] Rafał K. Mantiuk, Dounia Hammou, and Param Hanji, “HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content,” arXiv:2304.13625, 2023.
- [25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.
- [26] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda, “Photographic tone reproduction for digital images,” *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, July 2002.
- [27] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy, “Burst photography for high dynamic range and low-light imaging on mobile cameras,” *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 35, no. 6, 2016.
- [28] Pontus Andersson, Jim Nilsson, Peter Shirley, and Tomas Akenine-Möller, “Visualizing Errors in Rendered High Dynamic Range Images,” in *Eurographics 2021 - Short Papers*, Holger Theisel and Michael Wimmer, Eds. 2021, pp. 25–28, The Eurographics Association.

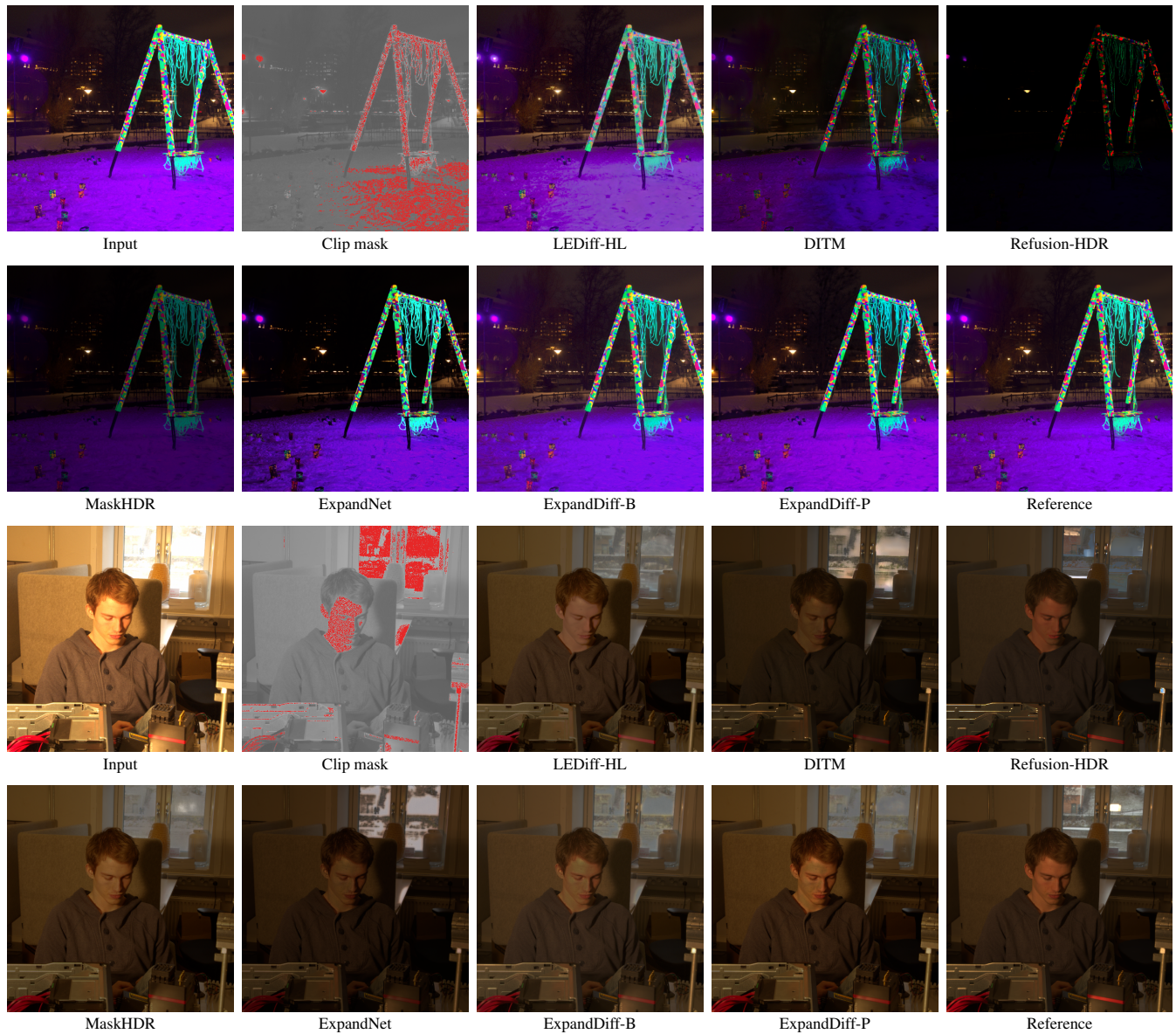


Fig. S1. Additional C_{95} examples with 9.2% and 8.3% clipped highlights. Red marks highlight-clipped pixels. Each example occupies two rows, with the same ordering of methods. Reconstructions use gain-only alignment.

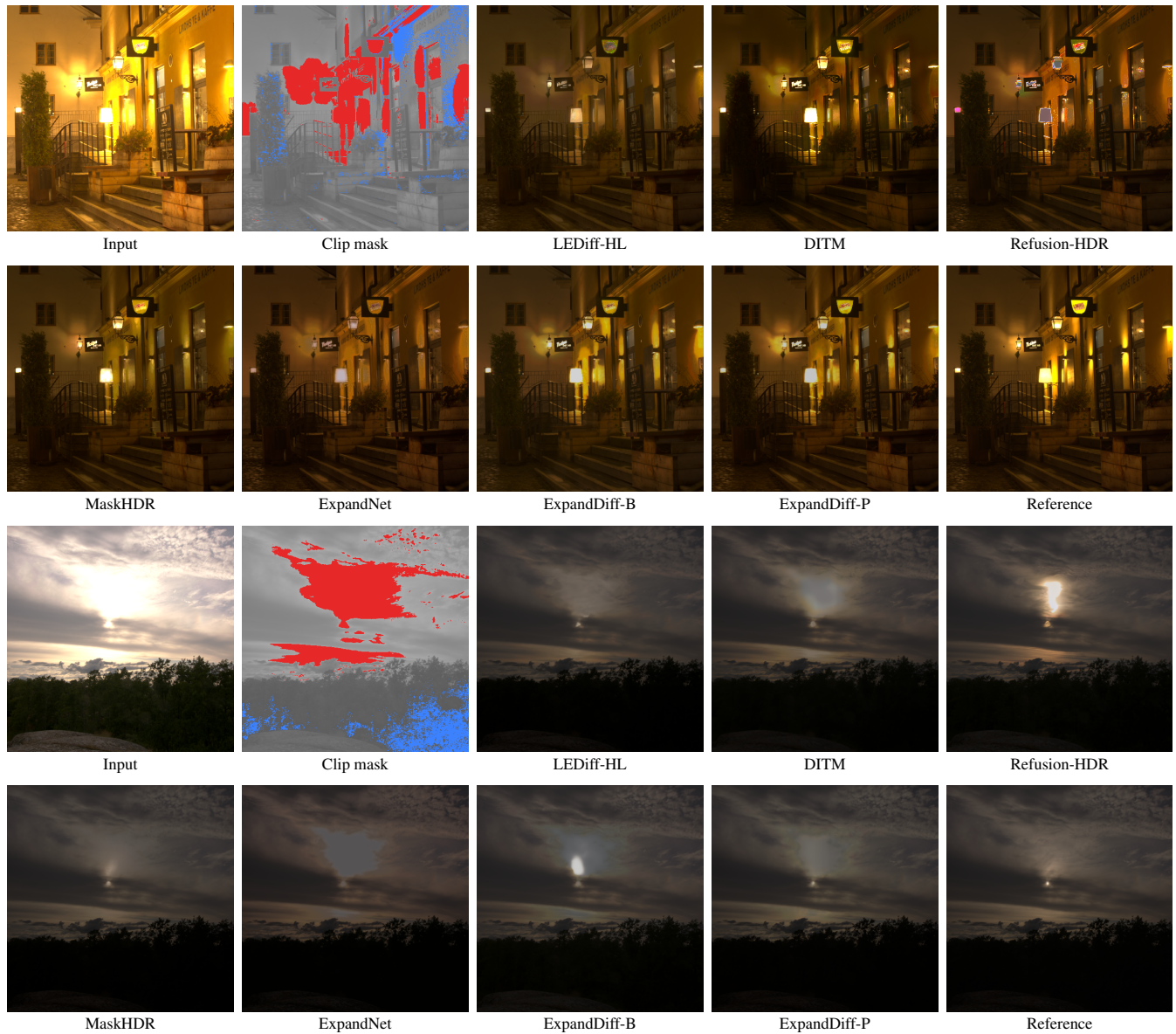


Fig. S2. Additional C_p examples, clipped over 20.6% and 21.4% of the frame. Red and blue mark highlight and shadow clipping. Each example occupies two rows, with the same ordering of methods. Reconstructions use gain-only alignment.

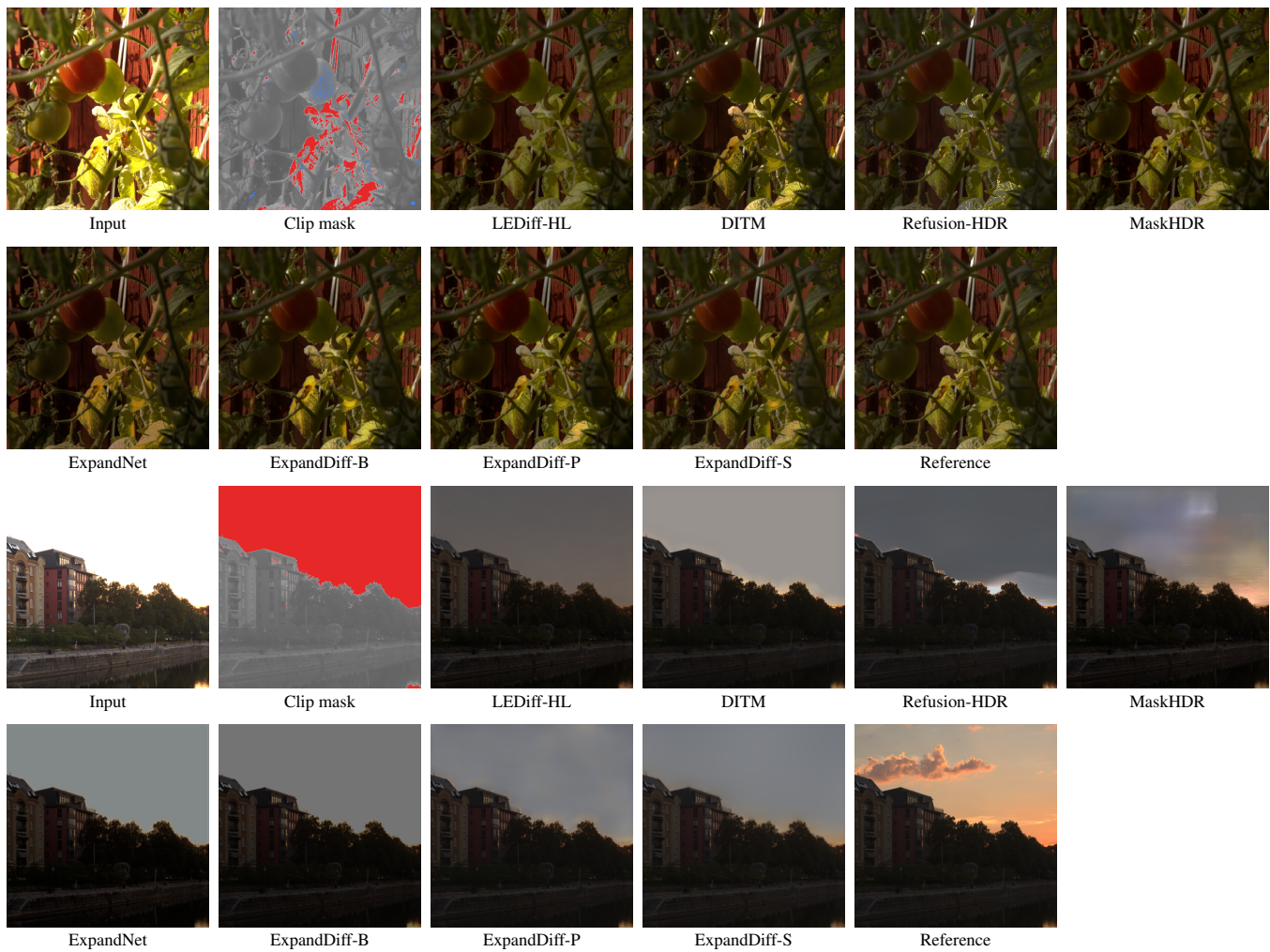


Fig. S3. C_0 examples, which additionally include ExpandDiff-S, clipped over 6.1% and 43.4% of the frame. Red and blue mark highlight and shadow clipping. Each example occupies two rows, with the same ordering of methods. Reconstructions use gain-only alignment.

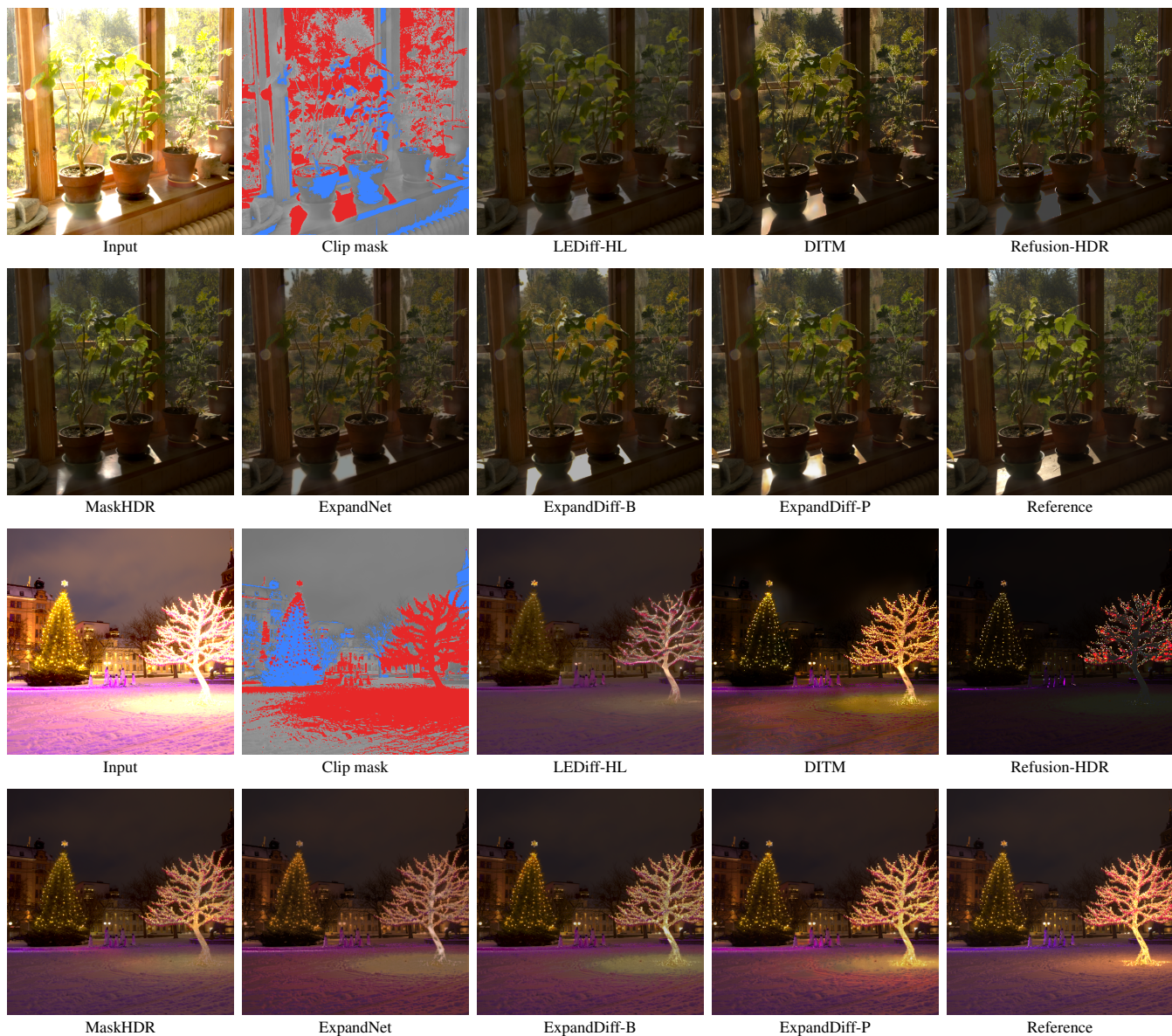


Fig. S4. Examples under C_p^{hard} , with doubled clipping percentiles (10, 30), clipped over 35.1% and 39.6% of the frame. Red and blue mark highlight and shadow clipping. Each example occupies two rows, with the same ordering of methods. Reconstructions use gain-only alignment.

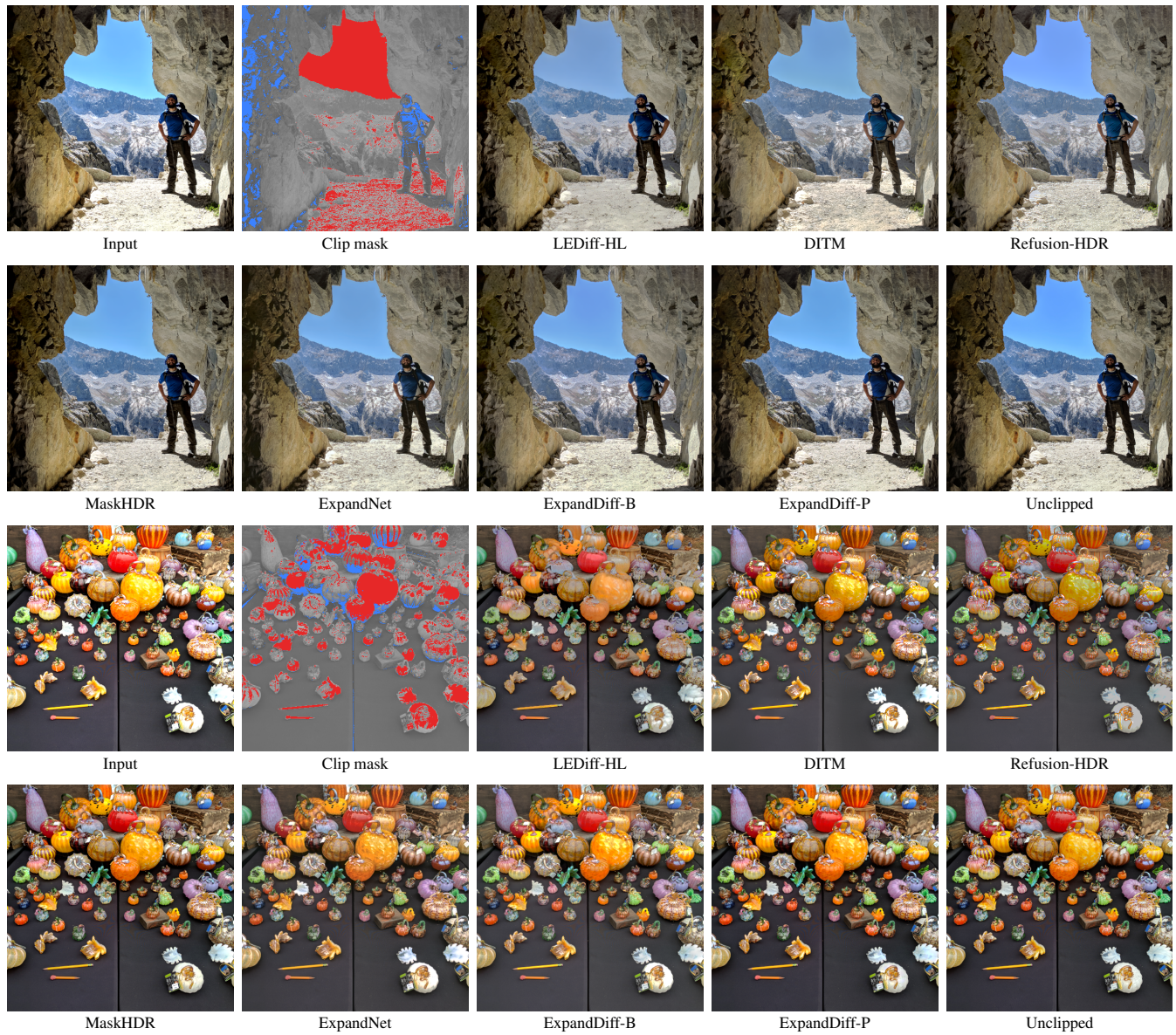


Fig. S5. HDR+ examples under C_p , clipped over 20.4% and 18.1% of the frame. Red and blue mark highlight and shadow clipping. Each example occupies two rows, with the same ordering of methods. Reconstructions use gain-only alignment.