CoRest: Multi-View Image Restoration with Geometric Guidance

Supplementary Material

A. Implementation Details

Training samples. We concatenate the Hypersim and TartanGround training datasets. On TartanGround, we sample a target and draw four distinct references from a surrounding window of up to 300 frames, with Gaussian probabilities favoring nearby frames. Half of these reference sets are ordered from farther to nearer frame indices relative to the target; the remaining sets retain their sampled order. On Hypersim, the target and four references are sampled uniformly without replacement from the scene. Each view is encoded independently by the VAE. The target is always degraded, and each reference is degraded independently with probability 0.25; all supervision uses the corresponding original clean target.

Degradation and resizing. The training motion-blur transform samples random trajectory kernels. Kernel width is obtained by rounding a uniform sample between 10 and 100 pixels, and the trajectory intensity is sampled uniformly between 0.1 and 1. Images are resized with BOX interpolation to a 480-pixel short side, with the other dimension rounded down to a multiple of 16. This approximately preserves aspect ratio without a square crop. The 25%, 50%, and 75% test conditions specify degraded-view proportions, with fixed targets shared by all methods; they do not specify degradation strength. Original clean target images are used only for training supervision or evaluation, never as inference references for those targets.

Optimization. We use eight GPUs with one sample per GPU and four gradient accumulation steps, giving an effective batch size of 32. LoRA, the retrieval adapter, and the timestep gate use a learning rate of 10^{-4} with cosine decay and 50 warmup steps. Training lasts 1,000 optimizer steps. Rank-16 LoRA adapts self-attention query, key, value, and output projections and the feed-forward projections in both the video backbone and VACE blocks. All other pretrained parameters, including the conditioning patch embedding, remain frozen. The flow-matching loss scores only the target view; the reference-view loss weight is zero. The confidence-weighted retrieval loss uses Smooth L1 with transition parameter 0.1 and overall weight 0.1. We use seed 42.

Geometric retrieval. VGGT inputs have a 336-pixel short side, adjusted for patch-grid alignment. Descriptors are extracted from aggregator index 17 using the difference between global and frame features. They are bilinearly resized to the native VACE conditioning grid and then ℓ_2-normalized. Each target token retrieves four tokens across the reference views with softmax temperature 0.07. The retrieved values are raw VACE conditioning features, not VGGT descriptors. The adapter has hidden width 256. Match confidence is computed with a sigmoid using similarity threshold 0.2 and temperature 0.05. The timestep gate is initialized to 0.1, and training uses retrieval-branch dropout of 0.1. The residual is added at target positions in every VACE control hint.

Inference and reference selection. OmniGlue [6] selects the reference views used at inference. The scene manifest identifies the degraded views to restore; reference selection uses the observed images, without clean target images or ground-truth cameras. For matching, we resize images to a 320-pixel short side with Lanczos interpolation, preserving aspect ratio. The score for an image pair is the number of OmniGlue correspondences with confidence ≥ 0.01. We initially score all pairs involving at least one degraded view. If both views remain degraded, we subtract 100 from the count, favoring cleaner references without excluding degraded candidates.

At each iteration, we sum the positive pairwise scores for each remaining degraded target and choose the target with the largest sum. We select its four highest-scoring reference views from the scene, order them by increasing score, and place the degraded target last in the conditioning sequence. Scores below zero do not contribute to target priority, but remain in the ranking used to select references. Thus, target priority depends on its matching support across the scene, while reference selection retains the four strongest individual pairs.

We restore the selected target using 30 UniPC sampling steps, flow shift 3, guidance scale 1, and seed 42. The generated target replaces its degraded version in both the restoration and matching image pools. We mark it as processed and recompute its OmniGlue scores against the other views; pairs involving this restored view no longer receive the degraded–degraded penalty. Original clean and already restored views remain eligible as references but are not selected as restoration targets. We repeat this procedure until every listed degraded view has been restored once. Every output is reused, without an output-quality acceptance threshold.

OmniGlue remains active for view selection when the geometric retrieval branch (VGGT, retrieval adapter, and timestep gate) is omitted at *CoRest* inference. The jointly trained LoRA and VACE reference conditioning use the selected views to restore the target. This inference configuration differs from a model trained without geometric retrieval.

B. Restoration Evaluation

Targets and aggregation. We evaluate only the degraded targets listed in each fixed scene manifest. The three fractions contain 312/599/916 targets on DL3DV, 195/386/597 on ScanNet++, and 19/49/78 on ETH3D, respectively. ETH3D excludes the easy split. A physical view can occur in more than one fraction, so these are condition–target cases rather than independent images. We average targets within each dataset and fraction and report the fractions separately.

Full-reference metrics. Ground-truth RGB images retain their original resolution and pixel values: 960×540 for DL3DV, 1920×1440 for ScanNet++, and 768×512 for ETH3D. Predictions are resized directly to this grid with PIL bicubic interpolation for PSNR, SSIM, and LPIPS, with no intermediate resize, cropping, mask, or color alignment. PSNR is computed per image before averaging. SSIM uses Gaussian weighting and an 11×11 window. LPIPS uses VGG version 0.1 in FP32. DreamSim uses the DINOv2 ViT-B/14 patch model; original predictions and ground-truth images enter its own preprocessing independently. Its score is $100(1 - \text{distance})$.

No-reference metrics and baseline handling. NIQE uses `niqe_matlab`, and MUSIQ uses the KonIQ-10k weights in pyiqa 0.1.16, with each implementation's default preprocessing on native outputs. BSSTNet falls back to the degraded input in 85 condition–target cases: 24 on DL3DV, 56 on ScanNet++, and five on ETH3D. These cases remain in all averages. For R2R, we use the five-degradation fine-tuned checkpoint with automatic prompt retrieval, supplying one image and no degradation label. Our SIR-Diff reproduction uses checkpoint 30100, 50 denoising steps, processing resolution 576, FP16 autocasting, and seed 42. Views are grouped consecutively into non-overlapping groups of four. GARD restores each half-scene independently, using 12 views per group on DL3DV and eight on the other datasets. *CoRest* can select four references from the full scene and reuse earlier outputs. Shared evaluation targets and metrics do not equalize these context policies.

C. Downstream Evaluation

Tables 7 and 8 report complete downstream results at 25%, 50%, and 75% degraded views on DL3DV and ETH3D. The main paper shows the 50% condition. *CoRest* improves all depth metrics over every baseline across the six conditions. It also exceeds SIR-Diff in both matching counts throughout.

Depth consistency. DA3-GIANT-1.1 receives full sampled scenes with 24 DL3DV views or 16 ETH3D views. We replace only the degraded targets with each method's output, retaining the original clean context. Outputs are resized to the sampled RGB grid before DA3's standard 504-pixel upper-bound resizing. We supply no ground-truth cameras. One clean-input prediction per scene is reused as the reference for all methods and fractions.

Table 7. **Downstream depth consistency on DL3DV and ETH3D.** For motion blur, DA3 predictions are aligned by one positive scale per scene and compared with clean-input predictions, not measured depth. Each degraded-view fraction is evaluated separately, with equal scene weighting. nRMSE is divided by the scene median reference depth. Colors indicate first and second place per dataset and fraction. * denotes our SIR-Diff reproduction.

Fraction	Method	DL3DV			ETH3D		
		AbsRel↓	nRMSE↓	δ_1 ↑	AbsRel↓	nRMSE↓	δ_1 ↑
25%	Degraded input	0.123	0.519	0.821	0.093	0.258	0.889
	NAFNet	0.115	0.484	0.834	0.107	0.280	0.850
	MoCE-IR	0.115	0.485	0.834	0.107	0.278	0.844
	R2R	0.112	0.471	0.839	0.094	0.268	0.876
	VRT	0.126	0.540	0.816	0.134	0.347	0.789
	BSSTNet	0.118	0.521	0.829	0.097	0.266	0.877
	SIR-Diff*	0.109	0.467	0.847	0.086	0.274	0.892
	GARD	0.086	0.284	0.914	0.069	0.189	0.936
	CoRest	0.062	0.219	0.940	0.053	0.160	0.959
50%	Degraded input	0.122	0.505	0.827	0.114	0.347	0.842
	NAFNet	0.122	0.493	0.827	0.118	0.340	0.838
	MoCE-IR	0.118	0.488	0.832	0.110	0.338	0.845
	R2R	0.112	0.474	0.840	0.101	0.317	0.862
	VRT	0.124	0.523	0.820	0.145	0.397	0.783
	BSSTNet	0.118	0.497	0.833	0.119	0.345	0.840
	SIR-Diff*	0.113	0.467	0.846	0.106	0.313	0.849
	GARD	0.086	0.284	0.913	0.087	0.256	0.907
	CoRest	0.071	0.236	0.936	0.075	0.211	0.917
75%	Degraded input	0.121	0.499	0.829	0.109	0.358	0.853
	NAFNet	0.119	0.484	0.832	0.112	0.341	0.848
	MoCE-IR	0.116	0.477	0.838	0.102	0.335	0.866
	R2R	0.111	0.461	0.844	0.100	0.320	0.863
	VRT	0.125	0.517	0.824	0.151	0.453	0.761
	BSSTNet	0.114	0.480	0.839	0.120	0.359	0.824
	SIR-Diff*	0.113	0.472	0.842	0.096	0.322	0.869
	GARD	0.089	0.303	0.907	0.093	0.271	0.892
	CoRest	0.077	0.259	0.925	0.073	0.214	0.926

For each scene, condition, and method, we fit a single positive scale as the median reference-to-prediction depth ratio over all scored target pixels. We fit neither a shift nor separate per-frame scales. We score finite positive reference depths without confidence filtering. AbsRel and δ_1 use the aligned predictions, with δ_1 counting pixels whose symmetric depth ratio is below 1.25. Normalized RMSE (nRMSE) is RMSE divided by the scene's reference median depth and is dimensionless. We average frame scores within each scene, then weight scenes equally. These measurements assess agreement with the clean-input estimator, not accuracy against independently measured geometry.

Figs. 4 and 5 show four selected targets per dataset at 50% degraded views, comparing all seven restoration baselines with *CoRest*. The examples illustrate recovery of object boundaries and scene layout, including display objects, archways, and vegetation, that are less distinct in the degraded-input predictions. We retain the scene-level scale alignment

Table 8. **Downstream feature matching with OmniGlue.** Mean match counts on fixed pairs of motion-blur targets, reported separately for each degraded-view fraction. *Matches* reports the number of correspondences at the native threshold of 0.001; *Conf.* reports the number with confidence ≥ 0.01. Colors indicate first and second place per dataset and fraction, excluding the clean-input reference. * denotes our SIR-Diff reproduction.

Fraction	Method	DL3DV		ETH3D	
		Matches↑	Conf.↑	Matches↑	Conf.↑
25%	Clean input (reference)	287.7	237.5	117.0	84.60
	Degraded input	57.10	36.10	28.60	17.20
	NAFNet	102.1	71.33	36.40	20.00
	MoCE-IR	91.53	63.24	33.00	21.80
	R2R	92.86	62.55	38.40	19.20
	VRT	82.73	51.80	30.60	15.20
	BSSTNet	99.08	63.78	35.80	21.60
	SIR-Diff*	87.37	64.92	30.00	21.20
	GARD	184.6	136.9	101.2	58.40
	CoRest	280.2	229.5	99.60	65.20
50%	Clean input (reference)	282.1	227.8	151.0	116.0
	Degraded input	59.38	37.66	24.86	9.286
	NAFNet	118.6	82.96	43.57	22.29
	MoCE-IR	98.82	69.94	29.00	11.43
	R2R	101.1	71.36	38.57	23.00
	VRT	101.4	69.86	41.29	19.14
	BSSTNet	111.7	76.44	38.86	19.14
	SIR-Diff*	87.48	62.96	41.71	24.71
	GARD	184.9	132.7	94.86	60.29
	CoRest	262.2	206.8	131.7	92.43
75%	Clean input (reference)	296.6	240.9	161.3	133.4
	Degraded input	45.56	26.96	21.43	14.00
	NAFNet	99.72	64.22	44.57	29.43
	MoCE-IR	85.98	53.38	38.29	23.29
	R2R	86.52	53.78	37.43	24.57
	VRT	86.58	51.74	31.14	16.43
	BSSTNet	98.56	60.26	38.14	21.43
	SIR-Diff*	75.94	48.74	32.57	23.29
	GARD	192.0	139.0	91.57	58.29
	CoRest	269.8	208.1	107.9	83.14

used for evaluation and use a shared color range across methods for each target; the accompanying AbsRel is computed for that target.

Feature matching. For each scene and fraction, we uniformly sample one unordered pair of degraded targets using a deterministic SHA256-based seed derived from seed 42, the dataset, the scene identifier, and the condition. Every method uses the same pair. Scenes with fewer than two targets are omitted, giving 49/50/50 pairs on DL3DV and 5/7/7 pairs on ETH3D at 25%/50%/75%. Each output is first resized bicubically to its original ground-truth dimensions and then downscaled, if needed, to a maximum long side of 640 pixels. There is no upsampling at this second stage and no cropping.

We use the pretrained OmniGlue matcher with at most 1,024 keypoints, its native acceptance threshold of 0.001, and an additional count at confidence ≥ 0.01. Means include zero-match pairs. One pair per scene gives equal scene weighting. The selected pair can differ across fractions, so clean-input counts also vary across fractions. No geometric ground-truth verification is applied. Because OmniGlue also drives reference selection, this experiment measures downstream matching support under a shared matcher and should be interpreted together with the independent depth-consistency evaluation.

Figs. 6 and 7 show one selected image pair from each dataset at 50% degraded views, with all seven baselines and *CoRest* evaluated on the same pair. Each panel draws every accepted correspondence and reports its count. In these examples, *CoRest* recovers matches across scene regions that have sparse matching support in the degraded inputs, bringing the counts closer to those of the clean inputs. Line colors distinguish correspondences by image location.

Table 9. **Training and inference ablations on DL3DV.** For motion blur, unguided LoRA adapts self-attention alone (Attention-only) or also feed-forward layers (Attention+FFN). *CoRest* adds geometric training to Attention+FFN, without inference-time VGGT; *CoRest* + VGGT activates retrieval in the same checkpoint. Colors indicate first and second within each fraction.

Fraction	Configuration	PSNR↑	SSIM↑	LPIPS↓
25%	Attention-only	18.71	0.527	0.370
	Attention+FFN	19.26	0.543	0.351
	CoRest + VGGT	19.44	0.548	0.346
	CoRest	19.45	0.549	0.347
50%	Attention-only	18.37	0.514	0.391
	Attention+FFN	18.87	0.527	0.373
	CoRest + VGGT	19.02	0.531	0.368
	CoRest	19.04	0.533	0.369
75%	Attention-only	18.17	0.509	0.420
	Attention+FFN	18.71	0.524	0.401
	CoRest + VGGT	18.88	0.528	0.396
	CoRest	18.90	0.529	0.397

D. Additional Motion-Blur Ablations

Tables 9 and 10 extend the main-paper ablations to all three degraded-view fractions on DL3DV. Extending unguided LoRA from attention to attention and feed-forward layers reduces LPIPS by 0.0187–0.0195. Geometrically guided training further reduces LPIPS by 0.0033–0.0041 when evaluated without inference-time VGGT. The *CoRest* + VGGT rows enable the retrieval branch in the same checkpoint, with retrieval dropout disabled and the reference-selection policy unchanged. On DL3DV, this changes PSNR by less than

Table 10. **Reference selection, order, and reuse.** *CoRest* on DL3DV motion blur. All rows use the same checkpoint without VGGT. Reuse marks whether restored outputs are eligible as references (✓: yes; ×: no; –: not applicable to wrong-scene references). Colors indicate first and second place within each fraction.

Fraction	Reference input	Reuse	PSNR↑	SSIM↑	LPIPS↓
25%	Wrong scene	–	17.26	0.474	0.558
	Random selection	✓	18.72	0.520	0.404
	Shuffled order	✓	19.42	0.548	0.347
	Matching-based selection	×	19.42	0.548	0.345
	Matching-based selection (ours)	✓	19.45	0.549	0.347
50%	Wrong scene	–	17.08	0.467	0.565
	Random selection	✓	18.39	0.507	0.429
	Shuffled order	✓	19.01	0.531	0.370
	Matching-based selection	×	18.93	0.528	0.367
	Matching-based selection (ours)	✓	19.04	0.533	0.369
75%	Wrong scene	–	17.08	0.465	0.566
	Random selection	✓	18.41	0.509	0.451
	Shuffled order	✓	18.90	0.529	0.397
	Matching-based selection	×	18.70	0.521	0.389
	Matching-based selection (ours)	✓	18.90	0.529	0.397

0.022 dB, SSIM by less than 0.0015, and LPIPS by less than 0.0016. Across all three datasets and fractions, the corresponding absolute changes remain below 0.044 dB, 0.0018, and 0.0027. Their direction varies by dataset and metric, indicating limited dependence on inference-time retrieval.

For reference selection, random references increase LPIPS by 0.0537–0.0600 relative to matching-based selection, while wrong-scene references increase it by 0.1692–0.2116. Shuffling the selected references changes LPIPS by at most 0.0013. Across fractions, reference relevance matters more than the order of a fixed reference set. Reuse is evaluated with the same VGGT-trained checkpoint without inference-time VGGT. The no-reuse policy keeps the reference pool at its original input values and does not recompute matches using generated outputs. Already processed views remain eligible in their original degraded form; only their generated restorations are excluded. Enabling reuse improves PSNR by 0.027–0.201 dB and SSIM by 0.0012–0.0086 across the three fractions, while increasing LPIPS by 0.0016–0.0079. The PSNR and SSIM gains grow as more views are degraded, but the LPIPS tradeoff persists. Wrong-scene inputs use fixed foreign clean references, so restored-reference reuse is not applicable to that row.

E. Brightness Restoration

We evaluate brightness restoration for both synthetic underexposure (darkening) and overexposure (brightening). This experiment uses a separately trained brightness checkpoint; it does not test the motion-blur checkpoint on an unseen degradation. The brightness checkpoint is trained for 1,000 steps on Hypersim and TartanGround with four references and rank-16 LoRA, replacing the target motion-blur transform with brightness augmentation. Training and test generation use `ColorJitter` with brightness parameter 2, sampling a multiplicative factor uniformly from $[0, 3]$. Factors below one darken the image, whereas factors above one brighten it and can clip highlights. This models brightness-induced exposure degradation in RGB images. *CoRest* uses the brightness-trained LoRA and reference conditioning without VGGT at inference.

Table 11 reports the completed results on all three datasets, keeping the degraded-view fractions separate. We use the same original-GT-resolution evaluation as in the main experiments and average only the manifest-selected degraded targets. We compare against Exposure-slot [9], evaluating its six MSEC/SICE/LCDP level-2/3 checkpoints separately, without selecting a checkpoint per test image using ground truth. Exposure-slot receives a single image, whereas *CoRest* uses four same-scene references; the comparison shares targets and metrics. *CoRest* exceeds all six Exposure-slot checkpoints in PSNR and SSIM across the nine conditions, and achieves lower LPIPS throughout DL3DV and ETH3D. Figs. 15 to 18 show eight selected brightness-restoration targets per dataset.

F. Mixed-Degradation Restoration

The mix setting combines motion blur with brightness degradation, including both underexposure and overexposure. For each degraded target, `MixTransform` randomly selects one or both transformations; when both are selected, their order is random. Thus, the test set includes motion blur alone, brightness alteration alone, and their composition. We use a separate checkpoint trained with this mixed-degradation transform and report the three degraded-view fractions separately; the fractions specify how many views are degraded, not the mixture of degradation types. Table 12 compares MoCE-IR [37], R2R [29], and *CoRest*. MoCE-IR uses its five-degradation checkpoint, and R2R uses its five-degradation finetuned checkpoint with automatic retrieval and no oracle degradation label. Both baselines restore one image at a time, whereas *CoRest* conditions on four same-scene references. All methods share the manifest-selected targets and the original-GT-resolution evaluation protocol, but differ in their input context. We report PSNR, SSIM, and LPIPS on all three datasets.

CoRest improves PSNR and SSIM over MoCE-IR and R2R in all nine dataset–fraction conditions. It also reduces LPIPS on DL3DV and ETH3D.

F.1. Novel-View Synthesis

We evaluate whether restoring mixed-degradation observations improves novel-view synthesis (NVS) with DA3-GIANT-1.1 [15] on 50 DL3DV scenes. For each method, we replace only the degraded context views with restored out-

Table 11. **Brightness restoration and Exposure-slot comparison.** Targets include synthetic underexposure and overexposure. Six Exposure-slot checkpoints are evaluated separately with single-image inputs. *CoRest* uses a separately trained brightness checkpoint with four references. *CoRest* omits VGGT at inference. All methods share scored targets. Fractions denote degraded-view proportions. Colors indicate first and second place per dataset and fraction.

Fraction	Configuration	DL3DV			ScanNet++			ETH3D		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
25%	Exposure-slot (MSEC L2)	18.45	0.783	0.254	17.33	0.847	0.352	16.26	0.701	0.264
	Exposure-slot (MSEC L3)	18.05	0.779	0.256	16.51	0.840	0.356	15.92	0.694	0.265
	Exposure-slot (SICE L2)	16.02	0.731	0.258	15.79	0.824	0.361	15.27	0.661	0.257
	Exposure-slot (SICE L3)	15.94	0.743	0.280	15.93	0.825	0.360	15.14	0.695	0.272
	Exposure-slot (LCDP L2)	14.57	0.768	0.272	14.73	0.858	0.366	13.70	0.736	0.255
	Exposure-slot (LCDP L3)	16.00	0.785	0.271	15.45	0.865	0.361	15.55	0.743	0.259
	CoRest	25.40	0.840	0.173	21.30	0.902	0.351	23.87	0.818	0.181
50%	Exposure-slot (MSEC L2)	18.07	0.769	0.270	17.33	0.845	0.339	17.97	0.794	0.192
	Exposure-slot (MSEC L3)	17.51	0.765	0.272	16.93	0.843	0.342	18.34	0.796	0.178
	Exposure-slot (SICE L2)	16.48	0.741	0.265	15.28	0.803	0.352	16.01	0.728	0.189
	Exposure-slot (SICE L3)	16.34	0.747	0.290	15.51	0.814	0.351	15.87	0.727	0.199
	Exposure-slot (LCDP L2)	14.29	0.761	0.282	14.99	0.861	0.353	14.53	0.779	0.199
	Exposure-slot (LCDP L3)	15.79	0.778	0.282	15.62	0.866	0.348	16.34	0.786	0.209
	CoRest	24.80	0.830	0.185	21.02	0.898	0.349	24.57	0.857	0.139
75%	Exposure-slot (MSEC L2)	18.30	0.777	0.256	17.42	0.850	0.340	17.34	0.785	0.215
	Exposure-slot (MSEC L3)	17.76	0.773	0.258	16.87	0.846	0.346	17.38	0.782	0.206
	Exposure-slot (SICE L2)	16.37	0.739	0.253	14.89	0.792	0.362	15.90	0.720	0.212
	Exposure-slot (SICE L3)	16.24	0.750	0.274	15.30	0.807	0.358	15.73	0.728	0.225
	Exposure-slot (LCDP L2)	14.69	0.772	0.268	15.02	0.858	0.356	14.03	0.763	0.218
	Exposure-slot (LCDP L3)	16.10	0.788	0.270	15.60	0.861	0.352	16.15	0.772	0.226
	CoRest	23.99	0.832	0.188	20.71	0.896	0.353	23.09	0.835	0.160

Table 12. **Mixed-degradation restoration.** Targets receive motion blur, brightness alteration (under- or overexposure), or both in random order. Fractions denote degraded-view proportions. MoCE-IR and R2R use single-image inputs; *CoRest* uses four same-scene references and a separately trained mixed-degradation checkpoint, without inference-time VGGT. Colors indicate first and second place among three methods per dataset and fraction.

Fraction	Configuration	DL3DV			ScanNet++			ETH3D		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
25%	MoCE-IR	15.23	0.626	0.436	16.52	0.814	0.407	13.50	0.614	0.427
	R2R	14.96	0.617	0.441	16.43	0.816	0.410	12.90	0.602	0.435
	CoRest	20.71	0.635	0.304	18.79	0.830	0.439	21.67	0.710	0.264
50%	MoCE-IR	15.14	0.631	0.433	16.43	0.809	0.419	17.28	0.709	0.387
	R2R	14.92	0.622	0.436	16.27	0.812	0.420	17.09	0.689	0.390
	CoRest	20.71	0.640	0.307	18.64	0.822	0.445	21.49	0.724	0.267
75%	MoCE-IR	14.94	0.612	0.443	16.52	0.805	0.414	16.52	0.687	0.381
	R2R	14.79	0.605	0.448	16.35	0.807	0.417	16.20	0.680	0.388
	CoRest	19.87	0.616	0.335	18.37	0.818	0.453	21.12	0.706	0.293

Table 13. **Novel-view synthesis from mixed-degradation observations on DL3DV.** Mix combines motion blur and brightness alteration (under- or overexposure); a degraded view may receive one or both. Fractions denote degraded context views. DA3 reconstructs each scene from clean, degraded, or restored inputs; all six metrics score held-out rendered views. *CoRest* uses the mixed-degradation checkpoint without inference-time VGGT. DreamSim is $100(1 - \text{distance})$. Clean and degraded inputs are unranked references. Colors indicate first and second place among the three restoration methods within each fraction.

Fraction	Configuration	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↑	NIQE ↓	MUSIQ ↑
25%	Clean input	19.87	0.580	0.279	92.40	3.471	60.64
	Degraded input	19.00	0.561	0.305	89.86	3.415	57.55
	MoCE-IR	18.95	0.560	0.308	89.64	3.426	57.61
	R2R	18.86	0.558	0.309	89.44	3.441	57.27
	CoRest	19.84	0.579	0.290	91.66	3.426	59.54
50%	Clean input	19.87	0.580	0.279	92.40	3.471	60.64
	Degraded input	17.83	0.528	0.344	86.19	3.335	54.25
	MoCE-IR	17.88	0.531	0.333	87.69	3.384	55.54
	R2R	17.75	0.528	0.339	86.14	3.386	55.15
	CoRest	19.08	0.532	0.321	89.34	3.338	58.89
75%	Clean input	19.87	0.580	0.279	92.40	3.471	60.64
	Degraded input	16.44	0.486	0.384	83.87	3.300	49.24
	MoCE-IR	16.32	0.491	0.375	84.78	3.341	51.56
	R2R	16.31	0.490	0.374	84.83	3.357	51.45
	CoRest	18.33	0.494	0.364	86.44	3.288	56.71

puts and retain the same original clean context. DA3 uses 24 context views per scene to predict a Gaussian representation; eight additional views are held out for evaluation and never supplied to restoration or DA3.

DA3 processes inputs at resolution 504 without ground-truth camera conditioning. We align predicted context cameras to calibrated context cameras using a robust Umeyama similarity transform, then map held-out cameras into the predicted scene coordinates. The gsplat renderer uses the transformed poses and calibrated intrinsics to render at 480×270. Held-out RGB targets are resized with preserved aspect ratio and center-cropped to this resolution, with the same transform applied to their intrinsics. We evaluate the rendered images using PSNR, SSIM, LPIPS-VGG, DreamSim, NIQE, and MUSIQ, with DreamSim expressed as $100(1 - \text{distance})$. Scores are equal-weight means over the 400 held-out images for each condition. The three degraded-view fractions are reported separately.

Table 13 compares the same three restoration methods as the mixed-degradation restoration experiment, with clean and degraded context inputs as references. Since mix combines motion blur and brightness alterations, we use multi-degradation baselines and exclude SIR-Diff and GARD from this experiment. *CoRest* improves PSNR, SSIM, LPIPS, DreamSim, and MUSIQ over degraded inputs, MoCE-IR, and R2R at every fraction. At 75%, for example, PSNR increases from 16.44 to 18.33 dB and LPIPS decreases from 0.384 to 0.364 relative to degraded inputs. NIQE improves over both restoration baselines at 50% and 75%; relative to degraded inputs, it improves only at 75%. These results support improved held-out rendering on DL3DV across fidelity and perceptual metrics.

Fig. 8 shows eight selected DL3DV scenes at 50% degraded views. In these examples, restoring the context with *CoRest* preserves more distinct structures and more faithful appearance in the held-out renders than degraded inputs or either baseline.



Figure 4. **Depth consistency, set 1.** Two targets from each of DL3DV and ETH3D at 50% degraded views. Each two-row block shows the clean RGB view, the clean-input DA3 depth reference, and depth predictions from degraded or restored inputs. Each prediction retains its evaluation scene-level scale alignment, and all methods use a common color range within each block; displayed depths are divided by the clean-reference scene median, and the range spans the 1st–99th percentiles of the target reference depth. AbsRel is reported per target. The reference is a model prediction, not measured ground-truth depth. *CoRest* omits VGGT at inference; * denotes our SIR-Diff reproduction.

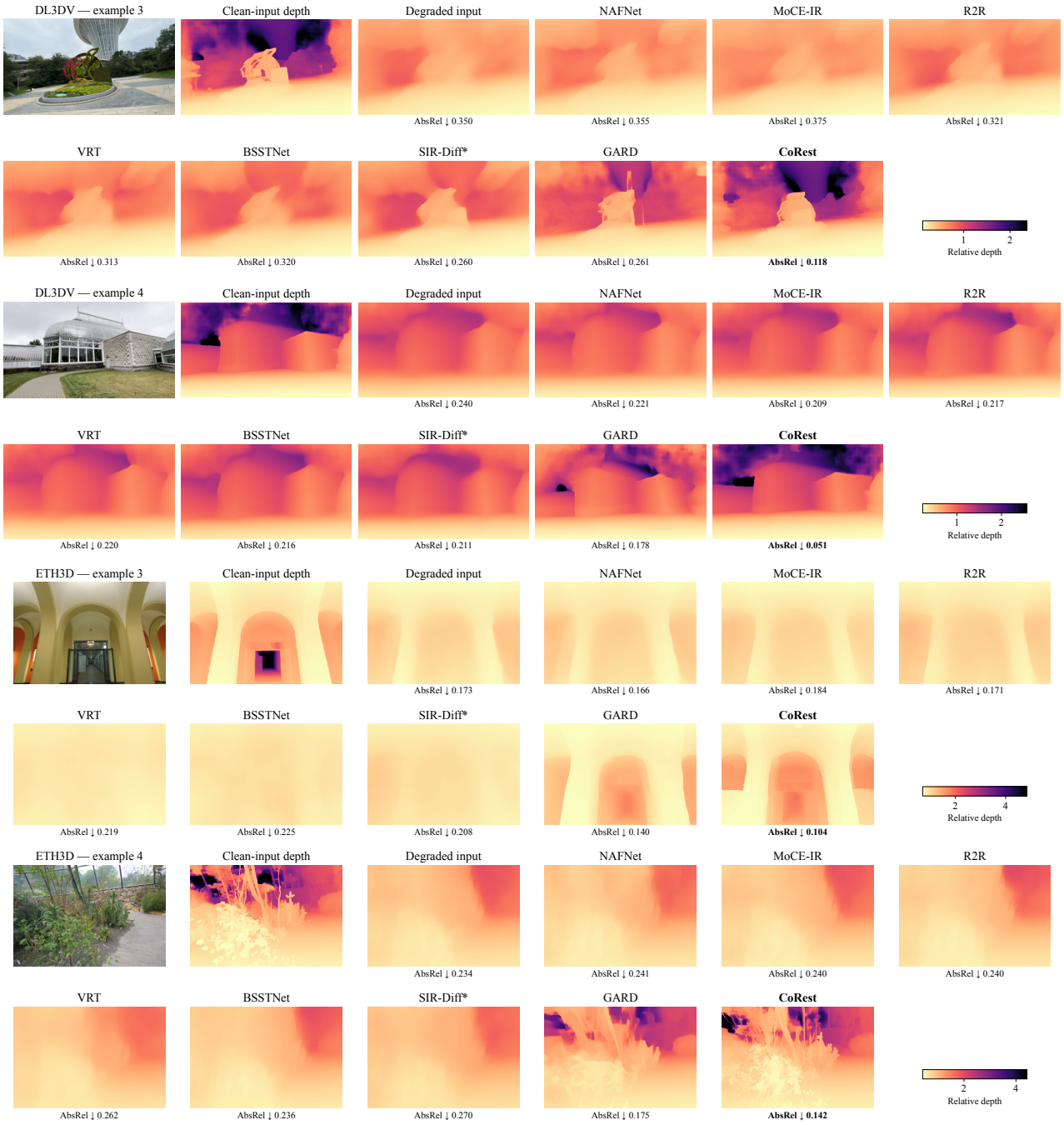

Figure 5. **Depth consistency, set 2.** Two targets from each of DL3DV and ETH3D at 50% degraded views. Each two-row block shows the clean RGB view, the clean-input DA3 depth reference, and depth predictions from degraded or restored inputs. Each prediction retains its evaluation scene-level scale alignment, and all methods use a common color range within each block; displayed depths are divided by the clean-reference scene median, and the range spans the 1st–99th percentiles of the target reference depth. AbsRel is reported per target. The reference is a model prediction, not measured ground-truth depth. *CoRest* omits VGGT at inference; * denotes our SIR-Diff reproduction.

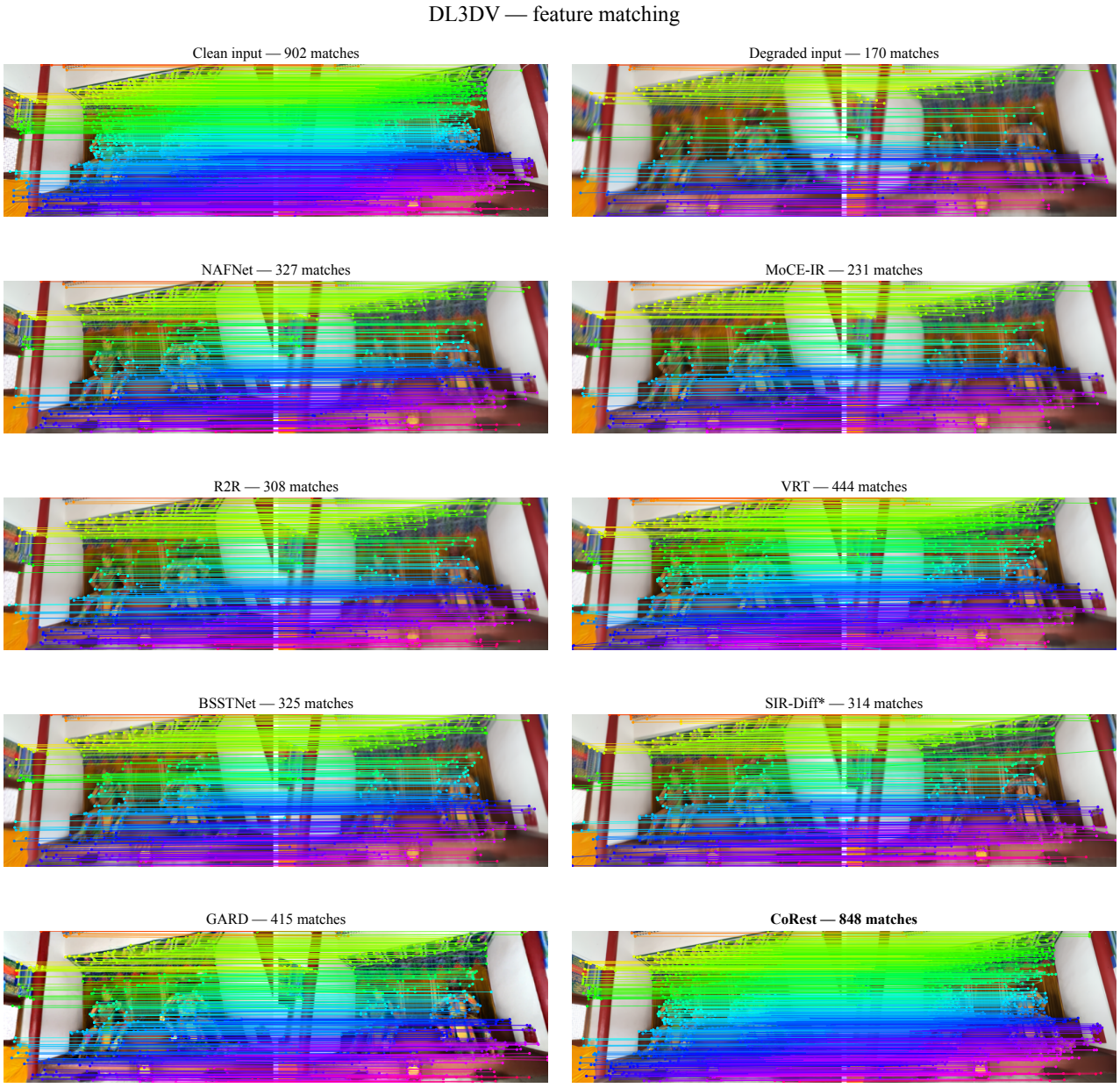

Figure 6. **Feature matching on DL3DV.** One selected image pair at 50% degraded views, held fixed across all methods. Panels are ordered from left to right and top to bottom and show all OmniGlue matches accepted at its native confidence threshold of 0.001, with counts above each panel. Colors distinguish correspondences by location, not correctness. Clean-input matches provide a reference. *CoRest* uses no VGGT at inference; * denotes our SIR-Diff reproduction.

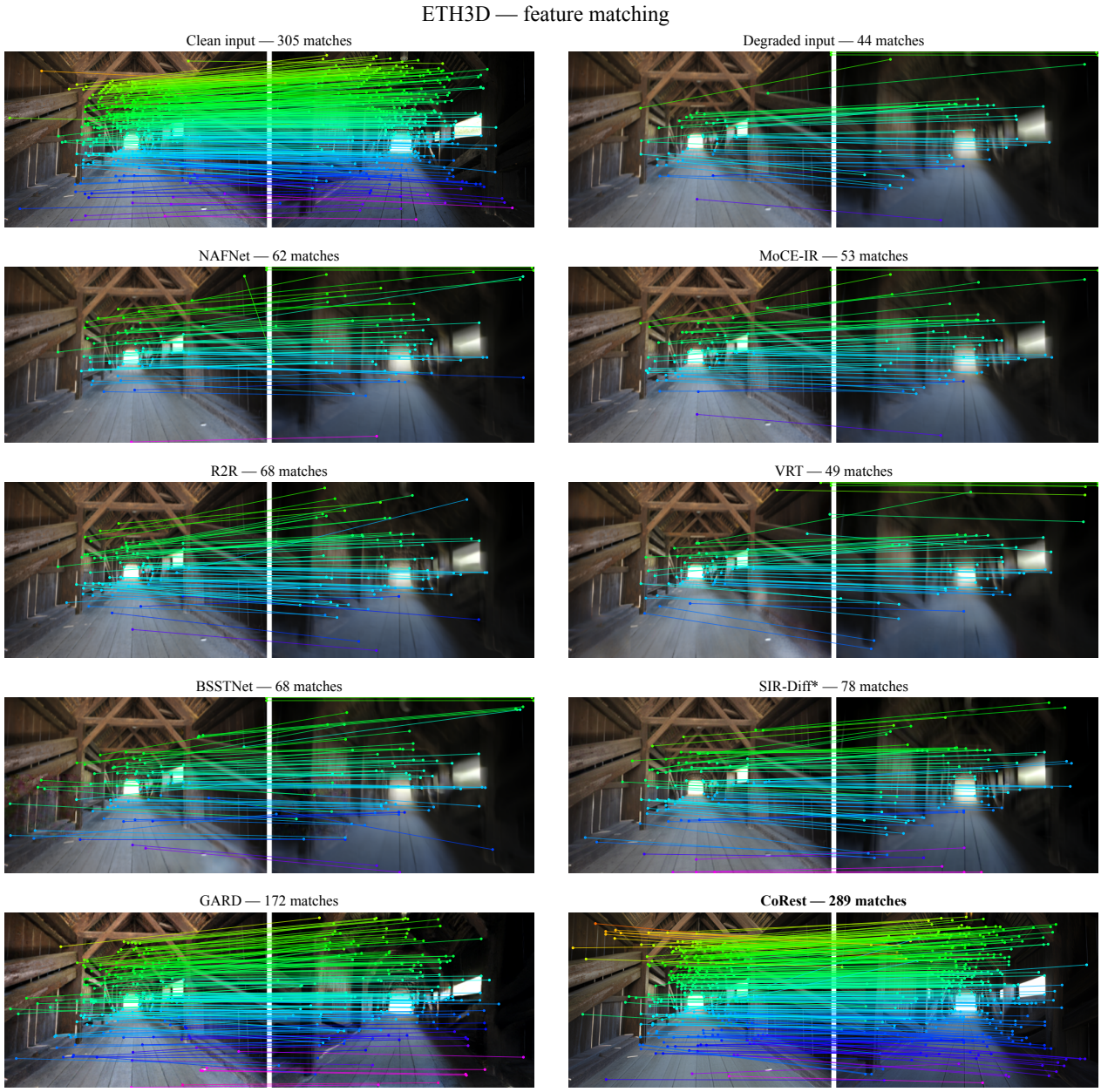

Figure 7. **Feature matching on ETH3D.** One selected image pair at 50% degraded views, held fixed across all methods. Panels are ordered from left to right and top to bottom and show all OmniGlue matches accepted at its native confidence threshold of 0.001, with counts above each panel. Colors distinguish correspondences by location, not correctness. Clean-input matches provide a reference. *CoRest* uses no VGGT at inference; * denotes our SIR-Diff reproduction.



Figure 8. **Novel-view synthesis from mixed-degradation observations.** Eight held-out views from different DL3DV scenes at 50% degraded context views. Columns compare DA3 renders from clean contexts, degraded contexts, and contexts restored by MoCE-IR, R2R, or *CoRest*, followed by the held-out GT photograph. Each row uses the same held-out camera and evaluation grid across methods. Numbers below renders are PSNR ↑ / LPIPS ↓ against the held-out GT. Clean-context renders are estimator outputs, not ground truth. *CoRest* uses the mix-trained checkpoint without inference-time VGGT; held-out views are never supplied to restoration or DA3.

G. Additional Visual Comparisons

Figs. 9 to 14 compare motion-blur restoration, and Figs. 15 to 18 compare brightness restoration. Each figure shows two targets per dataset at 50% degraded views. The motion-blur figures contain 12 distinct targets per dataset, and the brightness figures contain eight. These targets differ from those in the main paper's visual comparison. Motion-blur comparisons include all seven restoration baselines; brightness comparisons include all six Exposure-slot checkpoints. *CoRest* always uses inference without VGGT, with a checkpoint trained for the corresponding degradation.

We select illustrative cases from diverse scenes and visually review the resulting shortlist. These selected examples illustrate successful restoration, and quantitative tables include all evaluated targets.

Each figure follows the main paper's layout: datasets form column groups, methods form rows, and the top row shows GT contexts with marked crop regions. Crops cover approximately 35% of the GT image width at a 3:2 aspect ratio and are chosen from detailed regions of the GT image. Each original prediction is resized bicubically to the native GT grid before applying identical integer crop coordinates across all methods. All crops use the same display-resolution resampling when composing the figures. We apply no registration, color adjustment, sharpening, or generative editing to the comparison images.

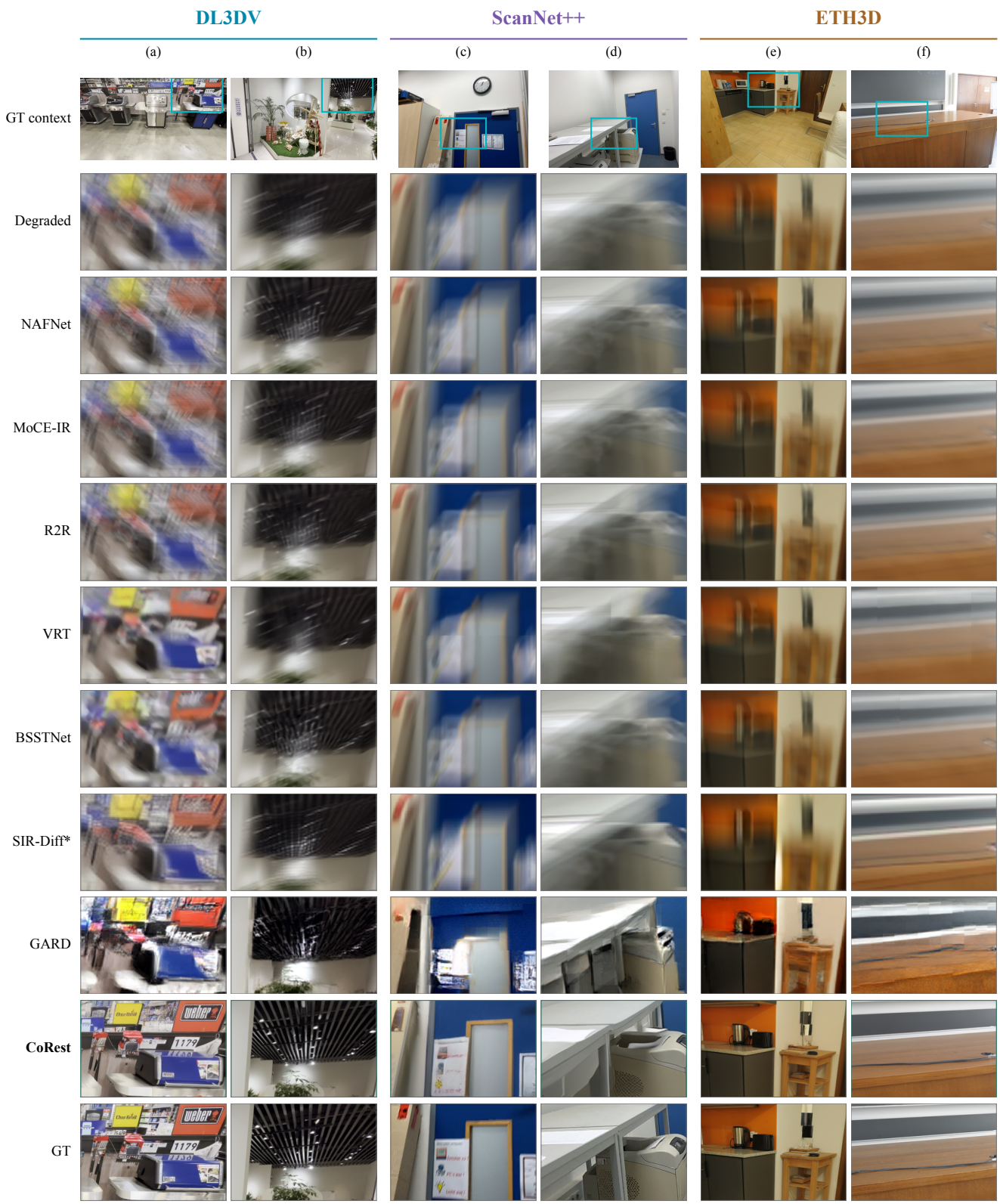

Figure 9. **Motion-blur visual comparison, set 1.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

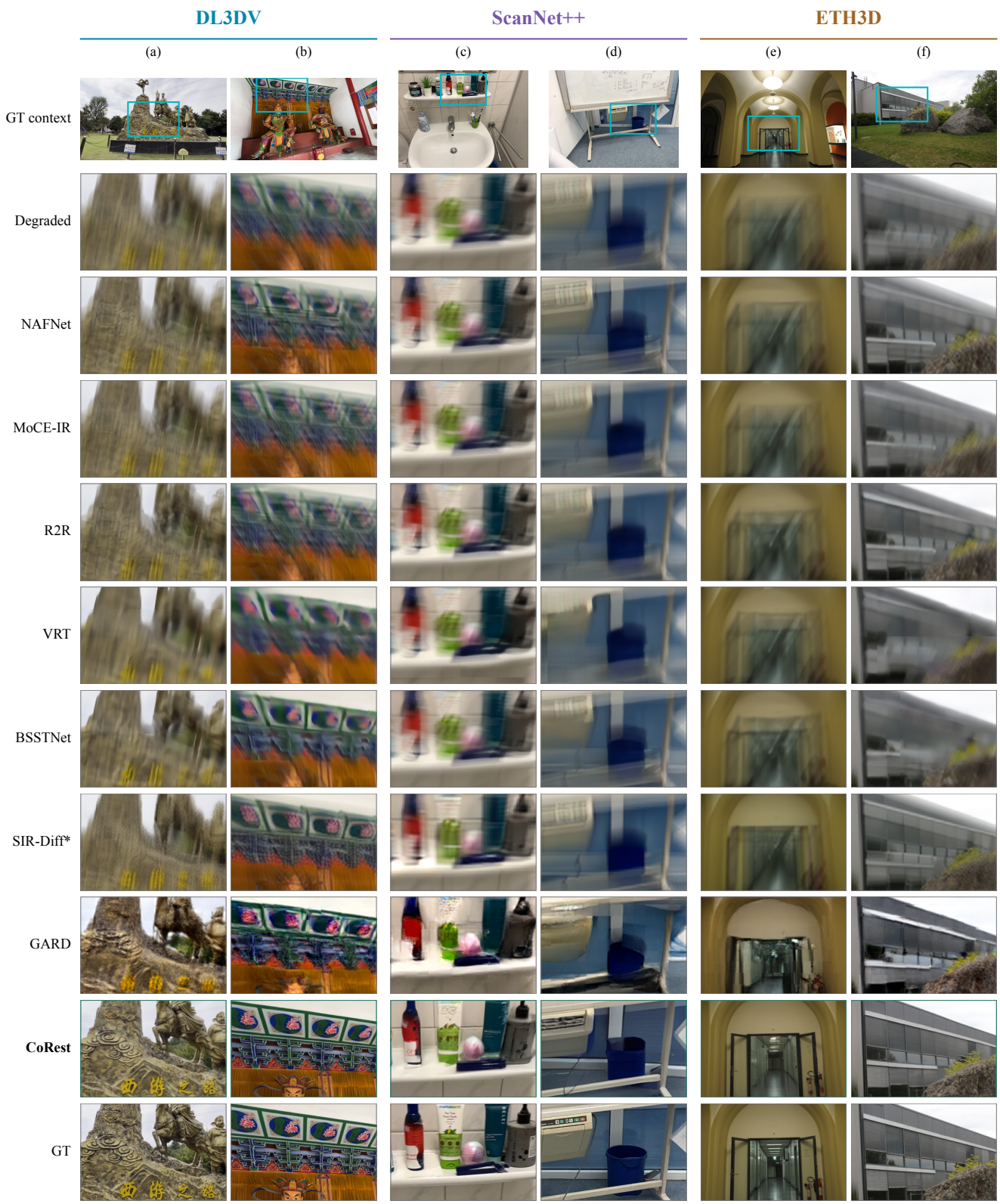

Figure 10. **Motion-blur visual comparison, set 2.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

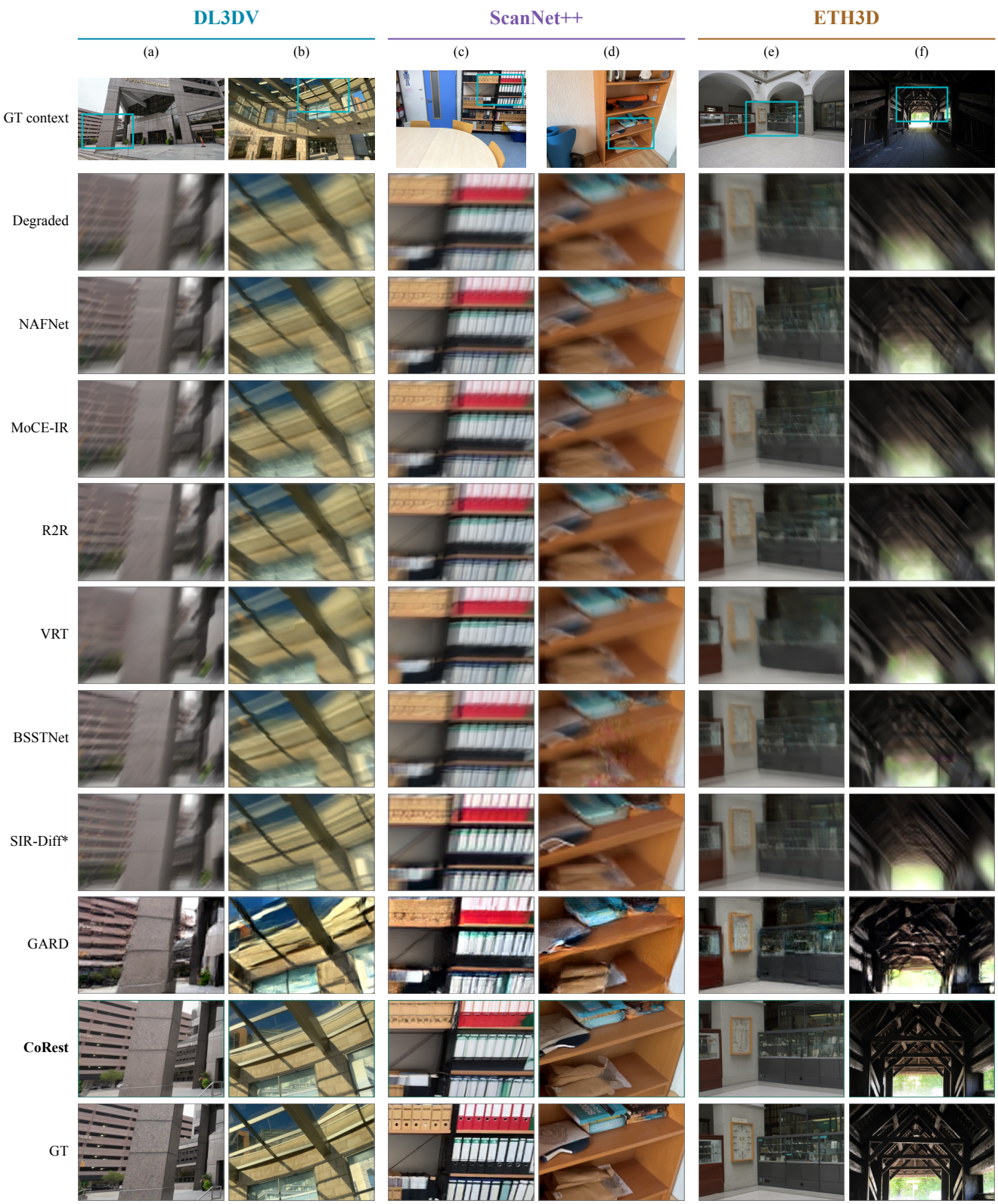

Figure 11. **Motion-blur visual comparison, set 3.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

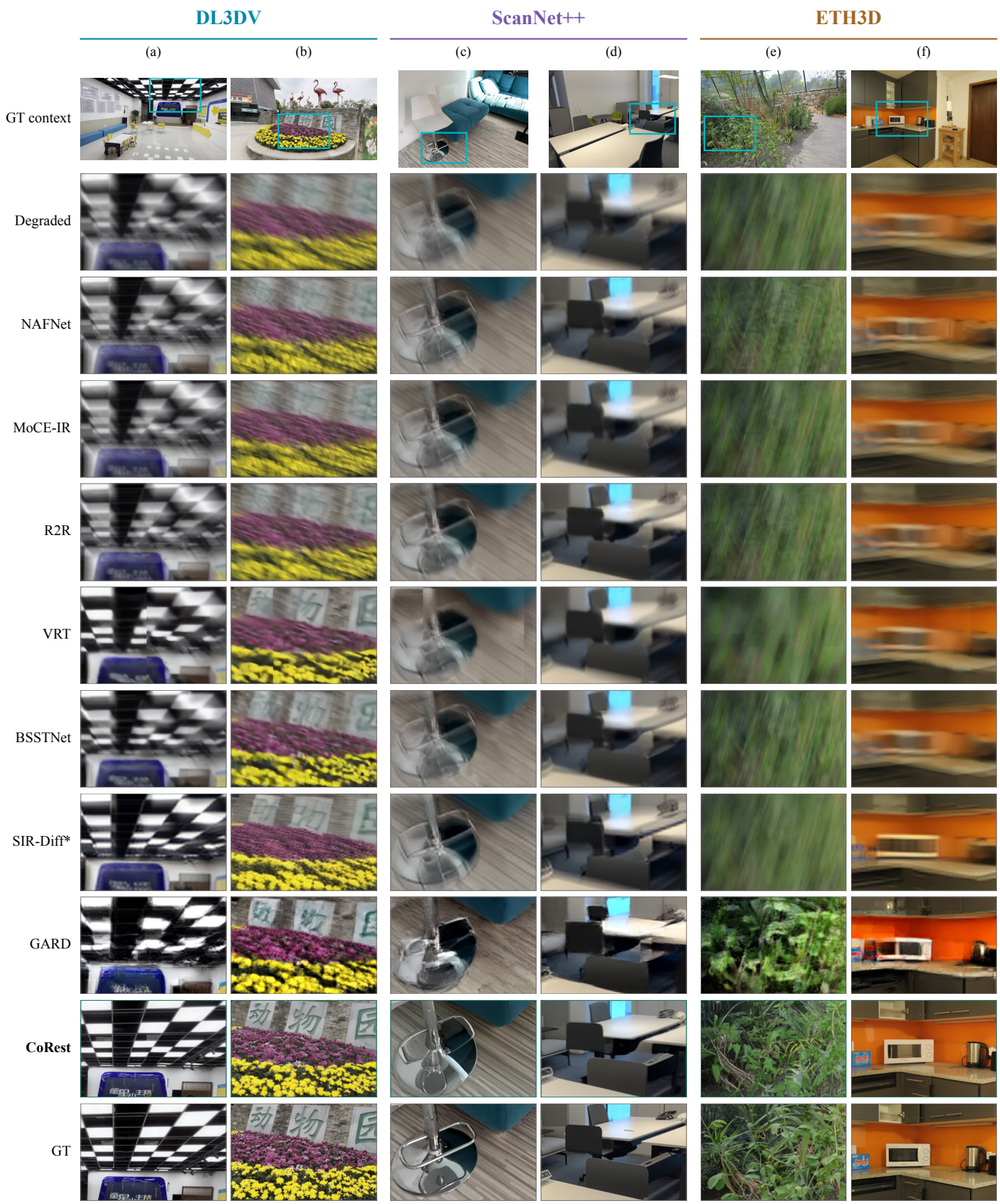

Figure 12. **Motion-blur visual comparison, set 4.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

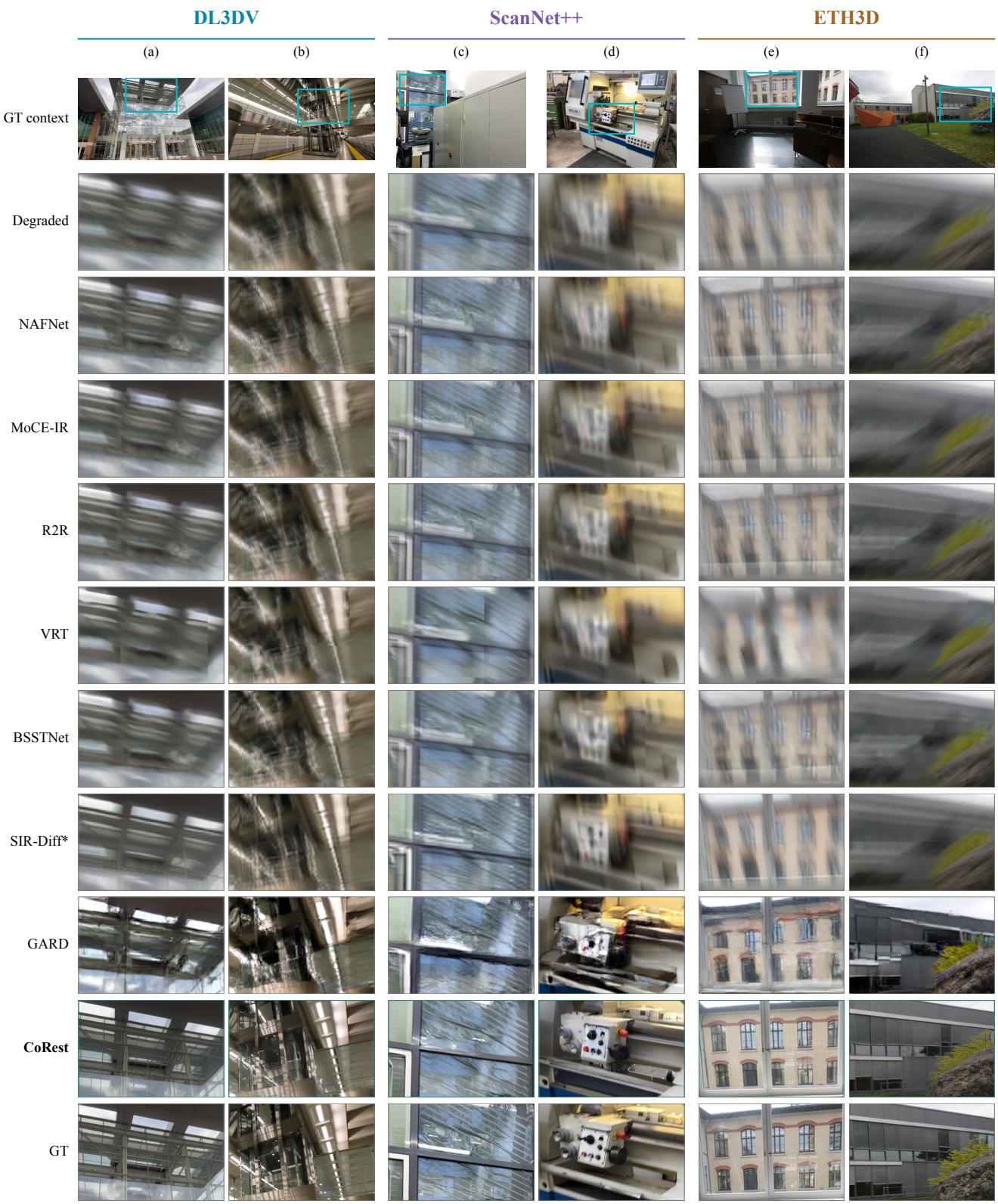

Figure 13. **Motion-blur visual comparison, set 5.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

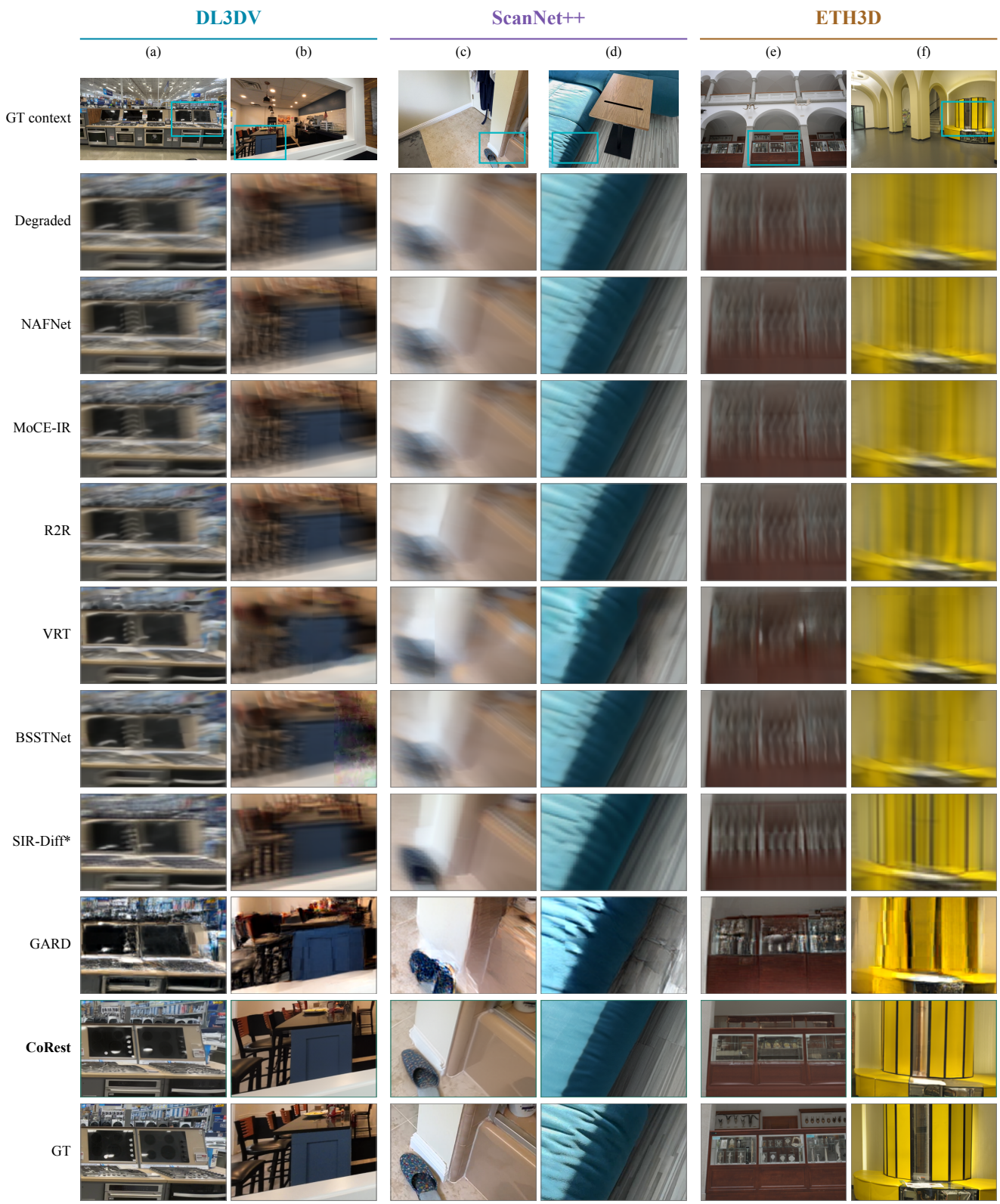

Figure 14. **Motion-blur visual comparison, set 6.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. Rows compare seven baselines with *CoRest*, which uses no VGGT at inference. * denotes our SIR-Diff reproduction.

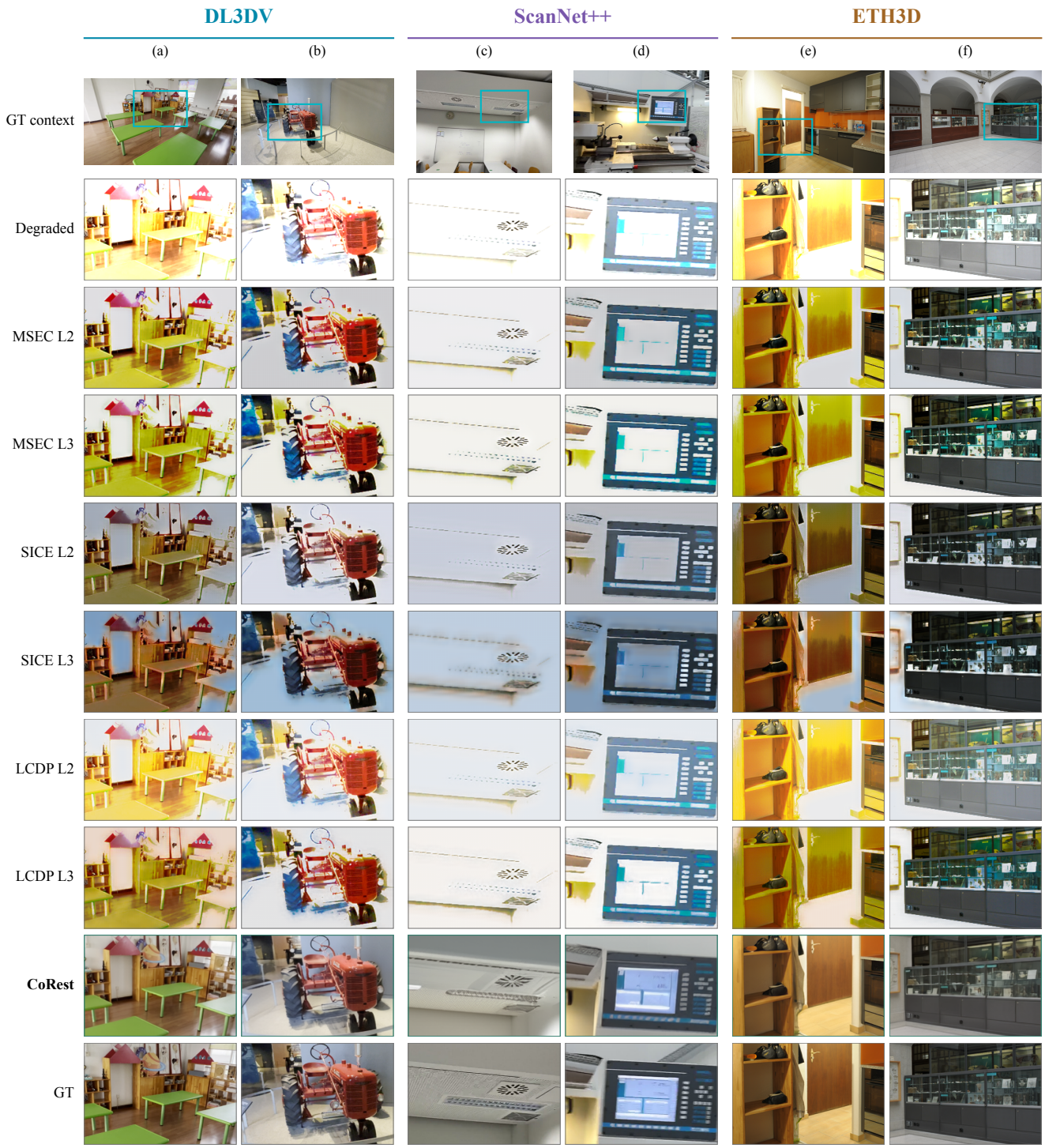

Figure 15. **Brightness visual comparison, set 1.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. MSEC, SICE, and LCDP rows show Exposure-slot checkpoints at levels 2 and 3. *CoRest* uses its brightness-trained checkpoint without inference-time VGGT.

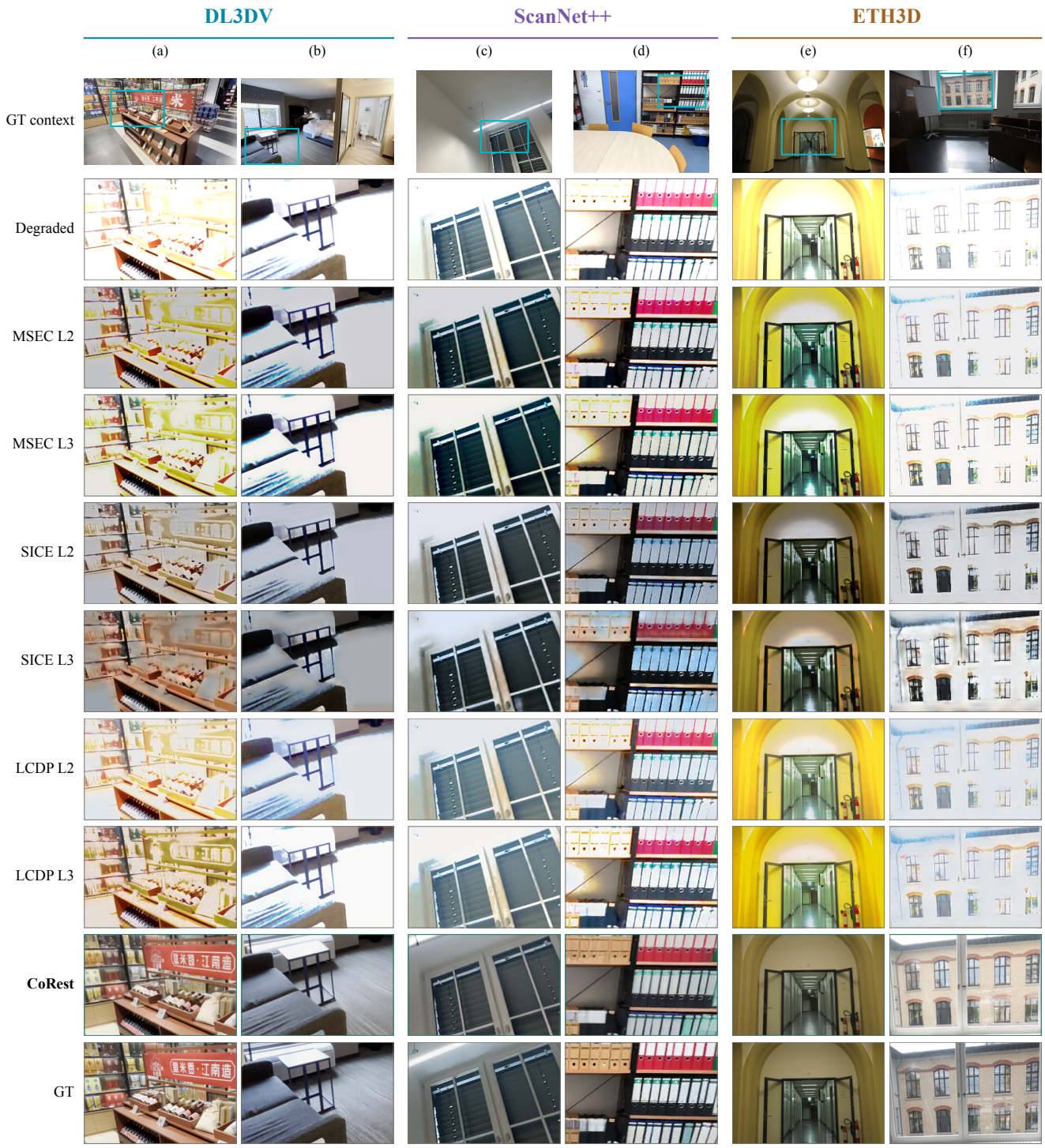

Figure 16. **Brightness visual comparison, set 2.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. MSEC, SICE, and LCDP rows show Exposure-slot checkpoints at levels 2 and 3. *CoRest* uses its brightness-trained checkpoint without inference-time VGGT.



Figure 17. **Brightness visual comparison, set 3.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. MSEC, SICE, and LCDP rows show Exposure-slot checkpoints at levels 2 and 3. *CoRest* uses its brightness-trained checkpoint without inference-time VGGT.

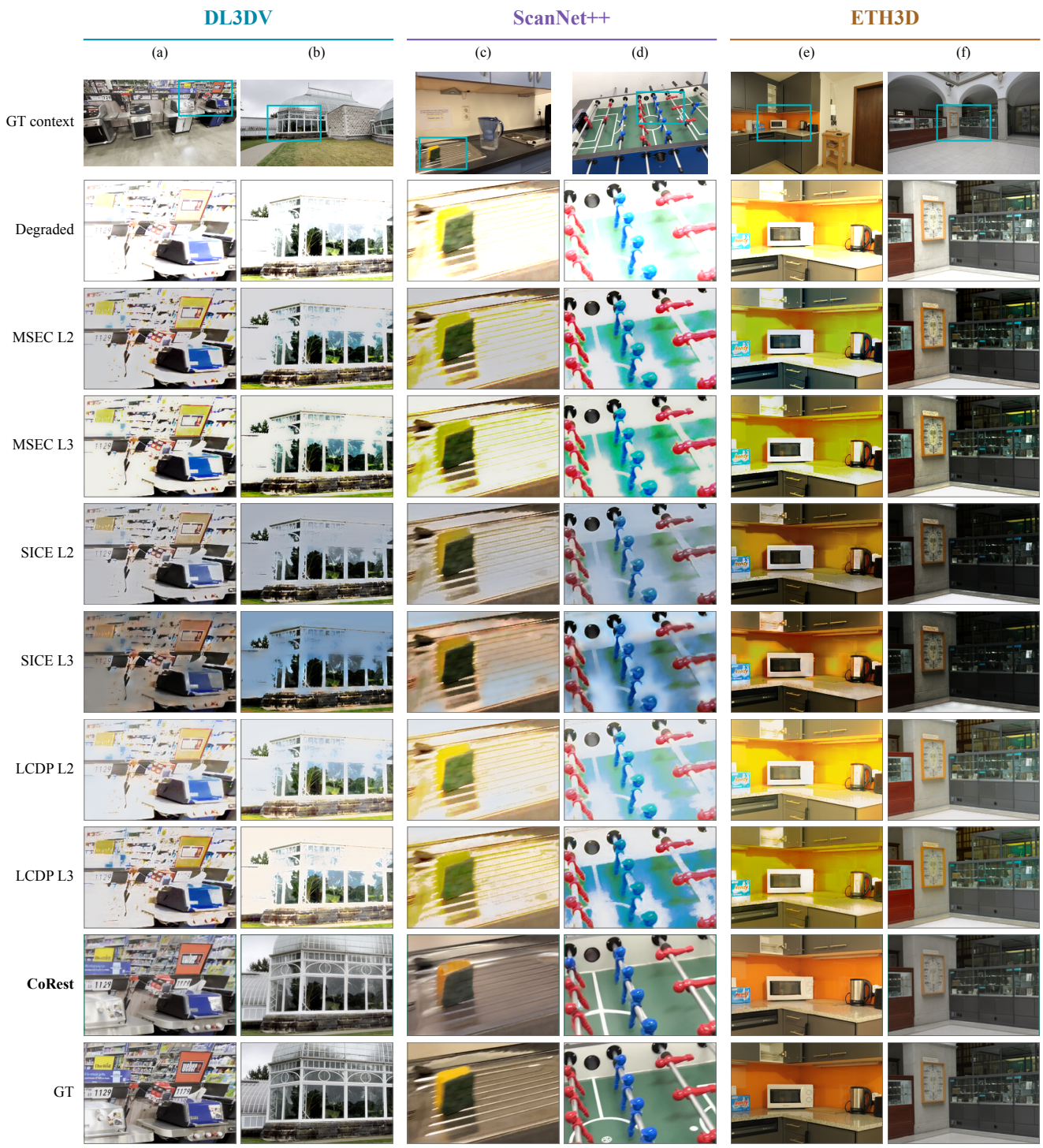

Figure 18. **Brightness visual comparison, set 4.** Each dataset has two distinct targets at 50% degraded views. GT contexts mark the shared crops. MSEC, SICE, and LCDP rows show Exposure-slot checkpoints at levels 2 and 3. *CoRest* uses its brightness-trained checkpoint without inference-time VGGT.