READING SCREENS FROM WALL REFLECTIONS: COALESCED EVIDENTIAL REPRESENTATION LEARNING FOR HIDDEN SCREEN CONTENT RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering electronic screen content from wall reflections exposes a practical optical side channel for hidden-screen observation. We develop a passive hidden-screen recovery framework that reconstructs screen content using only wall light spots captured by an ordinary camera and confronts three core challenges: diffuse reflection entangles screen structure with wall appearance and imaging noise; incomplete optical evidence causes reconstruction states to drift from the underlying screen layout; and locally plausible inversions produce globally inconsistent content. To address these challenges, we propose Coalesced Evidential Representation Learning (CERL), which coalesces complementary optical cues throughout reconstruction to recover screen structure from mixed wall reflections. CERL integrates Optical Evidence Representation Learning (OERL) to organize broad luminance, relative variation, boundaries, cross-scale changes, and local contrast; Inverse State Transport Learning (ISTL) to anchor progressive reconstruction to the observed evidence; and a Collective Consistency Module (CCM) to reconcile competing local corrections into a coherent screen estimate. Evaluated across five content categories, CERL surpasses a broad range of mainstream reconstruction methods across perceptual, structural, and pixel-level metrics. This consistent performance advantage validates that CERL effectively addresses the core challenges of this passive optical side-channel problem and enables recovery of hidden screen content from ordinary wall reflections.

1 INTRODUCTION

Screens may disclose their displayed content through light incident on nearby surfaces, even when the screen itself completely hidden from the observer. The screen-emitted light illuminates the wall and undergoes diffuse reflection, producing a low-contrast, blurred, and chromatically mixed light spot whose appearance is affected by the scene geometry and wall reflectance Bouman et al. (2017); Saunders et al. (2019). At the same time, the monitor chassis occludes the camera's direct view and appears as a dark region in the captured image. Consequently, the camera observes a composite image consisting of the monitor-back occlusion and the surrounding wall reflection. Recovering the frontal screen content from this indirect and severely degraded observation constitutes a passive optical side-channel problem with direct information-security implications.

The wall photograph mixes propagation, surface response, occlusion geometry, camera formation, and uncontrolled noise. Passive formulations can exploit corner geometry as a structured transport element Bouman et al. (2017), while modern restoration networks are typically designed for directly observed degradations Zamir et al. (2022); Wang et al. (2022b). Genuine screen traces may be weaker than wall texture or camera artifacts, intermediate decoder states can amplify an early interpretation error, and one local wall pattern can admit incompatible screen-space explanations. These coupled difficulties require evidence organization, anchored state evolution, and regional agreement.

To address the above challenges, we introduce Coalesced Evidential Representation Learning (CERL). CERL combines evidence across three connected reconstruction stages: Optical Evidence Representation Learning (OERL) organizes complementary photometric, boundary, and scale-dependent cues while bounding their injection into reconstruction features. Inverse State Transport Learning (ISTL)

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.

anchors each hierarchical transition to the observation and a reference state, limiting the propagation of unsupported intermediate updates. The Collective Consistency Module (CCM) aggregates alternative local corrections through soft agreement, encouraging spatially compatible regions to form a coherent estimate. Across the evaluated scenes and capture variations, this evidence-centered process recovers coarse screen brightness, region layout, and major contours, indicating that recognizable hidden-screen structure can be inferred from ordinary-camera wall photographs.

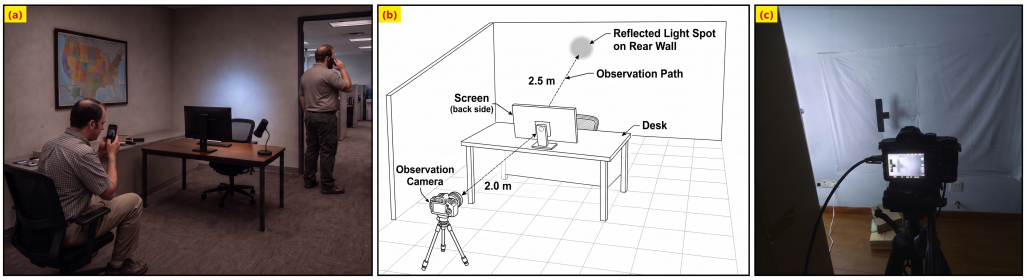

Figure 1: In the figure, (a), (b) and (c) correspond to the rendered scene, schematic diagram, and real-world scene respectively. An observer infers screen content via passive light projection. A light projection from the screen is cast onto a wall. By recording the wall's projection to reconstruct the original content non-invasively.

The main contributions are summarized as follows:

(1) We introduce OERL to turn complementary wall-reflection cues into a recovery-oriented evidence representation. Evidence-conditioned, bounded updates regulate how these cues shape the reconstruction features.

(2) We introduce ISTL to guide progressive reconstruction with an explicit reference state. Anchor-relative directional updates connect the current estimate to the observed evidence throughout hierarchical refinement.

(3) We introduce CCM to reconcile competing corrections through learned soft voting and agreement-guided updating, consolidating alternative local predictions into a coherent screen estimate.

2 Related Work

According to observation regime and output/readout space, existing studies can be broadly divided into passive indirect imaging and learned cross-domain recovery categories. The former asks what hidden information survives an indirect optical path, whereas the latter provides data-driven priors for mapping weak observations into target images.

Passive Indirect Imaging: Passive indirect imaging recovers hidden information from light emitted or reflected by the scene. Bouman et al. proposed a corner-camera method, which uses occlusion-induced transport structure to recover hidden scene variation Bouman et al. (2017). Saunders et al. proposed computational periscopy, which uses an ordinary camera and a diffuse wall to estimate hidden appearance from indirect illumination Saunders et al. (2019).

These methods establish that passive measurements can carry hidden-scene information, but they typically focus on different geometries, transport assumptions, or output spaces from dense screen-image recovery. The present task instead requires a learned mapping from wall appearance to aligned screen content while keeping every update tied to evidence in the current observation.

Learned Priors for Cross-Domain Recovery: Learned cross-domain recovery uses architectural priors and paired supervision to infer a target image from a degraded or transformed observation. Zamir et al. proposed Restormer, which uses efficient transformer operations in a hierarchical restoration model Zamir et al. (2022). Wang et al. proposed Uformer, which uses locally enhanced window processing to combine local and contextual information Wang et al. (2022b).

These methods establish that passive measurements can carry hidden-scene information. Building on this premise, Torralba & Freeman (2012) showed that accidental pinholes and pinspecks can

reveal scene content outside the direct field of view, while Klein et al. (2016) demonstrated hidden-object tracking from two-dimensional intensity images. Related studies inferred hidden light fields from cast shadows (Baradad et al., 2018) and recovered hidden scenes even when the occluder was unknown (Yedidia et al., 2019). In addition, Tancik et al. (2018) introduced a data-driven method for hidden-scene recovery using consumer-camera flash measurements, demonstrating the potential of learned priors while still relying on controlled active illumination. Despite this progress, existing approaches typically depend on specific geometries, target motion, calibrated transport models, or active illumination, and they do not directly recover spatially aligned screen content from passive wall observations. The present task instead requires a learned mapping from wall appearance to aligned screen content while keeping every update tied to evidence in the current observation.

3 METHOD

3.1 OVERALL ARCHITECTURE

CERL addresses three coupled ambiguities in hidden-screen recovery: which wall variations should inform reconstruction, how far an intermediate state should move, and which local corrections should be combined. Figure 2 connects these mechanisms: OERL organizes complementary evidence, ISTL conditions state transitions on that evidence and an anchor, and CCM reconciles competing RGB corrections.

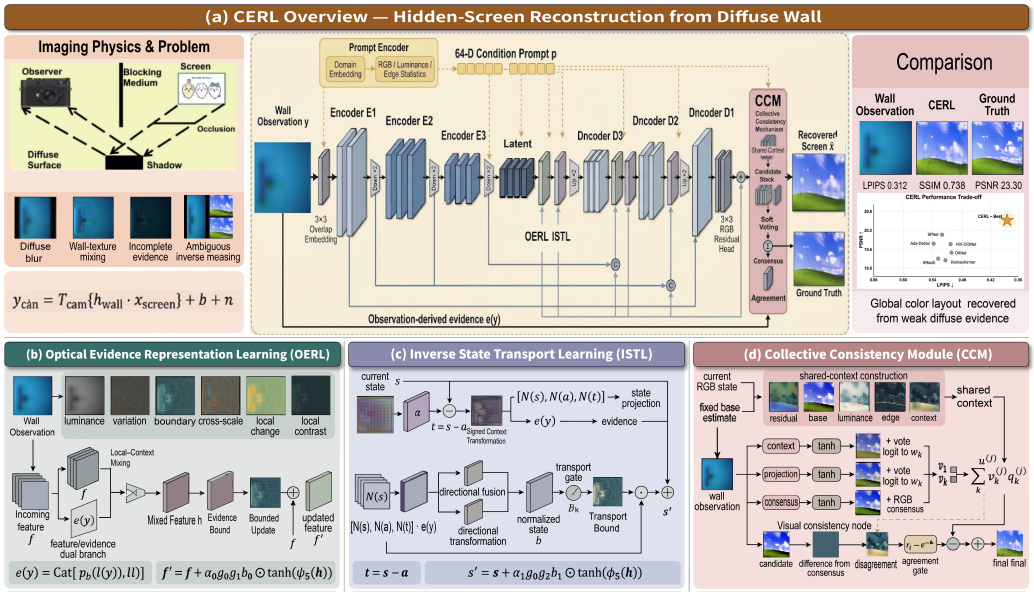

Figure 2: Overall CERL architecture. The shared wall observation conditions OERL evidence organization, ISTL anchor-relative state refinement, and CCM agreement-modulated RGB readout.

The observation-only mapping admits the finite-update decomposition

$$\hat{x} = \Phi_\theta(y) = x^{(0)} + \sum_{j=0}^{2} \alpha_2 g_0 g_3^{(j)} g_4^{(j)} r_j \odot u^{(j)}, \qquad (1)$$
$$x^{(0)} = y + o(f_{\text{out}}), \qquad y, \hat{x}, x^{(0)} \in \mathbb{R}^{B\times 3\times H\times W}.$$

Here f_{out} is the final feature of the OERL–ISTL hierarchy, o its learned RGB projection, and θ the learned parameters. Map products are pointwise, with scalar gates broadcast. Only y and learned parameters enter inference; the clean target x and valid mask m enter joint supervision.

3.2 Optical Evidence Representation Learning

Strong wall responses need not correspond to screen structure. OERL therefore separates complementary signal components before regulating their influence on features. Suppressing the batch index, let i index channels, $\mathbf{p}$ denote a position on the relevant grid, and Ω denote the full $H \times W$ image grid. For an RGB tensor z,

$$\begin{aligned} \ell(z) &= \frac{1}{3}\sum_{i=0}^{2} z_i, \qquad d_0(z) = \left[\sum_{\mathfrak{d}=0}^{1} (\mathcal{S}_{\mathfrak{d}} * \ell(z))^2 + \epsilon_0\right]^{1/2}, \\ \bar{d}_0(z)(\mathbf{p}) &= \frac{d_0(z)(\mathbf{p})}{\max\left\{(HW)^{-1}\sum_{\mathbf{p}'\in\Omega} d_0(z)(\mathbf{p}'), \epsilon_1\right\}}, \\ c(z) &= \left[\max\left\{p_c(\ell(z)^2) - p_c(\ell(z))^2, 0\right\} + \epsilon_0\right]^{1/2}. \end{aligned} \tag{2}$$

Fixed Sobel filters $\mathcal{S}_0, \mathcal{S}_1$ measure horizontal and vertical variation, indexed by $\mathfrak{d}$. Spatial normalization measures relative boundary strength, while c measures local dispersion. With bilinear resizing $\mathcal{I}_{P,Q}$, the ordered evidence is

$$\begin{aligned} e(y) = \mathcal{I}_{P,Q}\big[&p_b(\ell(y)),\ \ell(y) - p_b(\ell(y)),\ \tanh(\gamma_0 \bar{d}_0(y)), \\ &p_s(\ell(y)) - p_b(\ell(y)),\ \tanh(\gamma_1 c(y))\big]. \end{aligned} \tag{3}$$

Here $e = [e_0, e_1, e_2, e_3, e_4] \in \mathbb{R}^{B\times 5\times P\times Q}$ uses zero-based channel indices, and $p_b = p_9$, $p_c = p_5$, $p_s = p_3$ are same-size, zero-padded local means. Broad luminance retains the illumination envelope; relative variation and cross-scale differences expose departures from it; boundaries and contrast supply complementary local cues. These are observation-derived descriptors, not calibrated probabilities.

For $f \in \mathbb{R}^{B\times C\times P\times Q}$, projected normalized features and projected evidence undergo local/dilated mixing and response normalization to form h_{e} (Equation (B.11)). The evidence-dependent amplitude envelope is

$$b_0 = \beta_0 + \beta_1 \tanh(|e_1| + \beta_2|e_2| + |e_4|), \qquad \beta_0 \le b_0 \le \beta_0 + \beta_1. \tag{4}$$

Combining the observation gate g_0 and pooled-feature gate g_1, both sigmoid-valued, gives

$$\begin{aligned} f' &= f + \alpha_0 g_0 g_1 b_0 \odot \tanh(\psi_5(h_{\mathrm{e}})), \\ |f_i'(\mathbf{p}) - f_i(\mathbf{p})| &\le \alpha_0 g_0 g_1 b_0(\mathbf{p}) \le 0.036(\beta_0 + \beta_1). \end{aligned} \tag{5}$$

Here ψ_5 is the learned feature projection and $0 \le \alpha_0 \le 0.036$. Since $|\tanh| \le 1$, evidence bounds displacement while learned mixing determines direction; feature strength alone cannot exceed this envelope. The fixed nonnegative β coefficients, boundary/contrast scales γ_0, γ_1, and positive ϵ stabilizers are specified in the Appendix.

3.3 Inverse State Transport Learning

After evidence-conditioned refinement, ISTL addresses decoder drift using a current state s and aligned anchor a. Their signed displacement distinguishes departure from the reference from absolute feature strength:

$$\begin{aligned} t &= s - a, \qquad \varsigma = [\mathcal{N}(s), \mathcal{N}(a), \mathcal{N}(t)], \\ b_1 &= \beta_3 + \beta_4 \tanh\left(\mathrm{stop}\left(\frac{1}{C}\sum_{i=0}^{C-1} |t_i|\right) + \beta_5(|e_2| + |e_4|)\right). \end{aligned} \tag{6}$$

The states $s, a, t \in \mathbb{R}^{B\times C\times P\times Q}$ are observation-derived; $\mathcal{N}$ is per-pixel channel normalization. The stopped magnitude sets the allowance without backpropagating through that magnitude: displacement growth receives no gradient incentive through this branch, while signed t still conditions the update direction through ς.

Let $h_{\mathrm{t},1} = \phi_2(\phi_0(\varsigma) + \phi_1(e))$. Horizontal and vertical contextual responses then produce the transported state and its anchor-relative bound:

$$\begin{aligned} h_{\mathrm{t}} &= \mathcal{R}\left(\mathrm{SiLU}\left[h_{\mathrm{t},1} + \sum_{\mathfrak{d}=0}^{1} \phi_{\mathfrak{d}+3}(\mathrm{SiLU}(h_{\mathrm{t},1}))\right]\right), \\ s' &= s + \alpha_1 g_0 g_2 b_1 \odot \tanh(\phi_5(h_{\mathrm{t}})), \\ \|s'(\mathbf{p}) - a(\mathbf{p})\|_\infty &\le \|t(\mathbf{p})\|_\infty + \alpha_1 g_0 g_2(\mathbf{p}) b_1(\mathbf{p}). \end{aligned} \tag{7}$$

The learned ϕ maps and response normalization $\mathcal{R}$ are detailed in Equations (B.9) and (B.15); $g_2 \in (0,1)$ is spatial and $0 \le \alpha_1 \le 0.030$. By the triangle inequality, anchor-relative drift can increase only by the bounded increment; contraction is not assumed. These refined states support the RGB readout $x^{(0)}$ passed to CCM.

3.4 Collective Consistency Module

Even anchored states can admit incompatible corrections. At step j, the context $\upsilon^{(j)} \in \mathbb{R}^{B\times 24\times H\times W}$ stacks the current estimate $x^{(j)}$, fixed base $x^{(0)}$, observation y, and their residual, luminance, and edge cues (Equation (B.18)). Four learned candidates share the spatial envelope $b_2^{(j)}$, computed from local observation residuals and edges as in Equation (B.19). Writing their RGB components explicitly gives

$$\begin{aligned} q_{k,i}^{(j)} &= b_2^{(j)} \tanh\Big([\eta_k(\upsilon^{(j)})]_i\Big), \qquad v_k^{(j)} = \frac{\exp(w_k^{(j)})}{\sum_{k'=1}^{K}\exp(w_{k'}^{(j)})}, \\ u_i^{(j)} &= \sum_{k=1}^{K} v_k^{(j)} q_{k,i}^{(j)}, \qquad \bar{q}_i^{(j)} = \frac{1}{K}\sum_{k=1}^{K} q_{k,i}^{(j)}, \qquad K = 4, \\ \delta_j &= \frac{1}{3K}\sum_{k=1}^{K}\sum_{i=0}^{2}(q_{k,i}^{(j)} - u_i^{(j)})^2 \\ &= \frac{1}{3K}\sum_{k=1}^{K}\sum_{i=0}^{2}(q_{k,i}^{(j)} - \bar{q}_i^{(j)})^2 + \frac{1}{3}\sum_{i=0}^{2}(\bar{q}_i^{(j)} - u_i^{(j)})^2, \quad r_j = e^{-\delta_j}. \end{aligned} \tag{8}$$

Here k indexes candidates, η_k is a learned RGB projection, and $w_k^{(j)}$ are learned one-channel vote logits. The arithmetic mean $\bar{q}^{(j)}$ is introduced only for the identity. Since $\sum_k (q_{k,i}^{(j)} - \bar{q}_i^{(j)}) = 0$, cross terms cancel, separating candidate spread from consensus displacement. Even low-vote candidates contribute to disagreement; r_j measures internal compatibility, not target correctness.

Since the votes are nonnegative and sum to one, $|u_i^{(j)}| \le b_2^{(j)}$. The finite evolution consequently satisfies

$$\begin{aligned} x^{(j+1)} &= x^{(j)} + \alpha_2 g_0 g_3^{(j)} g_4^{(j)} r_j \odot u^{(j)}, \qquad j \in \{0,1,2\}, \\ \|\hat{x}(\mathbf{p}) - x^{(0)}(\mathbf{p})\|_\infty &\le \sum_{j=0}^{2} \alpha_2 g_0 |g_3^{(j)}(\mathbf{p}) g_4^{(j)}| r_j(\mathbf{p}) b_2^{(j)}(\mathbf{p}), \qquad \hat{x} = x^{(3)}. \end{aligned} \tag{9}$$

The shared step obeys $0 \le \alpha_2 \le 0.037$; context, spatial gate $g_3^{(j)}$, image-level gate $g_4^{(j)}$, candidates, and agreement are recomputed at each step. Voting forms the correction, agreement softly attenuates disputed updates, and the envelope bounds cumulative refinement rather than guaranteeing convergence.

Joint training. All three modules are optimized together. For a valid mask $m \in \{0,1\}^{B\times 1\times H\times W}$ and loss map $\varrho \in \mathbb{R}_{\ge 0}^{B\times C\times H\times W}$, the channel-repeated mask m_C yields

$$\langle \varrho \rangle_m = \frac{\sum_{n=1}^{B}\sum_{i=0}^{C-1}\sum_{\mathbf{p}\in\Omega} m_{n,\mathbf{p}}\varrho_{n,i,\mathbf{p}}}{C\sum_{n=1}^{B}\sum_{\mathbf{p}\in\Omega} m_{n,\mathbf{p}} + \epsilon_2}, \qquad [m_C]_{n,i,\mathbf{p}} = m_{n,\mathbf{p}}. \tag{10}$$

Here n indexes images and the singleton mask channel is suppressed. The same valid elements enter numerator and denominator. The seven-term objective is

$$\begin{aligned} L &= L_0 + \sum_{\mathfrak{l}=1}^{6} \lambda_{\mathfrak{l}} L_{\mathfrak{l}}, \\ (L_0, L_1, L_2) &= \left(\left\langle \sqrt{(\hat{x}-x)^2 + 10^{-6}} \right\rangle_m, \langle (\hat{x}-x)^2 \rangle_m, \langle |\ell'(\hat{x}) - \ell'(x)| \rangle_m \right). \end{aligned} \tag{11}$$

The fixed luminance projection is $\ell' = l_1$ in Equation (B.3), distinct from the evidence mean $\ell = l_0$. Sobel loss L_3 and perceptual loss L_5 are given in Equation (B.24); the fixed nonnegative weights are specified in Equation (B.25). For $\iota = \ell'(\hat{x})$ and $\iota' = \ell'(x)$, the local structural term is

$$\begin{aligned} S(\iota, \iota') &= \frac{(2\mu\mu' + \xi_0)(2\kappa + \xi_1)}{(\mu^2 + \mu'^2 + \xi_0)(\nu + \nu' + \xi_1) + \epsilon_3}, \\ L_4 &= \left\langle \frac{1}{2} \left[1 - \mathrm{clip}(S(\iota, \iota'), -1, 1) \right] \right\rangle_m. \end{aligned} \tag{12}$$

Here μ, μ' are local means, ν, ν' variances, and κ covariance under p_7; positive ξ_0, ξ_1 stabilize this SSIM-based ratio Wang et al. (2004). Exact moments and constants are given in Equation (B.24). Finally, the observation-edge leakage term is

$$\begin{aligned} d_1(z) &= \left[\sum_{\mathfrak{d}=0}^{1} (\nabla_{\mathfrak{d}} \ell'(z))^2 + \epsilon_4 \right]^{1/2}, \\ \bar{d}_1(z)(\mathbf{p}) &= \frac{d_1(z)(\mathbf{p})}{(HW)^{-1} \sum_{\mathbf{p}' \in \Omega} d_1(z)(\mathbf{p}') + \epsilon_5}, \\ L_6 &= \left\langle d_1(\hat{x}) \, \mathrm{clip}\big(\bar{d}_1(y) - \tau_0 \bar{d}_1(x), 0, \tau_1\big) \right\rangle_m. \end{aligned} \tag{13}$$

Forward differences ∇_0, ∇_1 differ from the Sobel evidence filters; τ_0, τ_1 set target-edge subtraction and clipping. The fixed observation–target edge factor penalizes predicted gradients more where observation edges exceed target-supported responses. Structural supervision encourages target organization while L_6 counterweights copying wall transitions. Neither target nor mask enters inference in Equation (1).

4 Experiments and Results

4.1 Experimental Setup

Five paired datasets: Screen, OShome, Figure, Game, and Photo were employed to emulate desktop and system interfaces, operating-system lock-screen and wallpaper scenes, data-visualization charts, video-game frames, and natural-photo scenarios, randomized into training, validation, and test subsets in an 8:1:1 ratio. All experiments were implemented in PyTorch on an NVIDIA RTX 3080 GPU cluster, using Adam optimizer with a fixed learning rate of 1×10^{-4} and a batch size of 16; other hyperparameters were set to their default values unless stated otherwise.

4.2 Comparison with State-of-the-Art

Evaluated across five benchmarks, namely Screen, Figure, Game, OShome, and Photo, CERL is juxtaposed with general-purpose restoration frameworks, models for related degradations, and classical baselines under identical data partitions, evaluation protocols, and preprocessing settings. Table 1 reports results on LPIPS, SSIM, and PSNR: specifically, CERL reduces LPIPS on Screen by 6.1% relative to OKNet and increases PSNR on Photo by 1.2% relative to SFNet, while achieving the best reported point estimate across all 15 dataset–metric combinations. Qualitative illustrations (Fig. 3) reveal more consistent restoration of edges, textures, and fine structures, with fewer residual artifacts across the supplied Screen samples.

The visual distinction lies in the structure recovered from the reflection. CERL forms recognizable window regions and principal boundaries, whereas several comparison outputs retain the broad

Table 1: Comparisons on five benchmarks. LPIPS is lower-is-better (↓); SSIM and PSNR are higher-is-better (↑). Best/second-best are marked in red/blue. *Source* indicates venue/year.

Methods	Source	Screen			Figure			Game			OShome			Photo		
		LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
General Purpose End-to-End Image Restoration Frameworks																
OKNet Cui et al. (2024b)	AAAI,24	**0.4445**	**0.7316**	20.243	0.5964	0.7211	19.245	0.5756	0.6625	17.892	0.6781	0.5553	14.664	0.8316	0.3944	12.577
ConvIR Cui et al. (2024a)	TPAMI,24	0.4495	0.7283	**20.271**	0.6293	0.7151	18.540	0.6013	0.6515	17.642	0.6867	0.5925	14.911	0.8447	0.3904	11.692
Uformer Wang et al. (2022a)	CVPR,22	0.4639	0.7068	19.204	0.6260	0.6187	17.586	0.6142	0.6470	17.547	0.6743	0.5652	15.149	0.8338	0.3802	12.096
AST Zhou et al. (2024)	CVPR,24	0.4668	0.7050	19.391	0.6324	0.6386	17.891	0.6158	0.6453	17.456	0.6920	0.5366	15.072	0.8210	0.3622	12.330
DGUNet Mou et al. (2022)	CVPR,22	0.4818	0.6960	18.962	0.6229	0.6952	18.038	0.5950	0.6596	17.888	0.6514	**0.6236**	**16.521**	0.8258	0.4001	12.669
MIRNetv2 Zamir et al. (2023)	TPAMI,23	0.4923	0.6917	18.237	0.6536	0.7133	18.340	0.6120	0.6559	17.679	0.6575	0.6042	15.211	0.8436	0.4086	11.980
OneRestore Guo et al. (2024)	ECCV,24	0.7142	0.4642	10.409	0.7044	0.5863	12.529	0.7474	0.5062	10.313	0.7212	0.5221	12.237	0.8725	0.3756	11.114
DeepSN-Net Deng et al. (2025)	TPAMI,25	0.7787	0.2954	10.489	0.7005	0.4993	14.401	0.7390	0.4440	10.868	0.7711	0.4228	12.836	0.8817	0.3173	12.224
Models for Related Degradations																
IRNeXt Cui et al. (2023a)	ICML,23	0.4542	0.7234	19.956	0.6326	0.6679	17.479	0.5735	**0.6681**	**18.136**	0.6785	0.6082	16.134	0.8499	0.3816	12.070
Retinexformer Cai et al. (2023)	ICCV,23	0.4554	0.6732	19.567	0.6186	0.6905	17.895	0.5924	0.6556	17.925	0.6506	0.5586	14.600	**0.7911**	0.3983	13.539
SFNet Cui et al. (2023b)	ICLR,23	0.4618	0.7164	19.626	**0.5940**	**0.7236**	19.340	**0.5699**	0.6603	17.991	**0.6374**	0.6132	16.376	0.8088	0.4122	**13.641**
Ada-Deblur Wu et al. (2023)	TIP,23	0.4637	0.7067	19.425	0.6078	0.7206	19.257	0.5878	0.6577	17.753	0.6608	0.6086	16.273	0.8223	**0.4145**	13.490
HVI-CIDNet Yan et al. (2025)	CVPR,25	0.4724	0.7178	19.692	0.5965	0.7210	**19.415**	0.5979	0.6555	17.774	0.6714	0.5758	15.606	0.8119	0.4055	13.249
DehazeFormer Song et al. (2023)	TIP,23	0.4794	0.6931	19.366	0.6157	0.7086	18.917	0.6074	0.6494	17.774	0.6687	0.5375	14.873	0.8230	0.3589	12.485
DarkIR Feijoo et al. (2025)	CVPR,25	0.8673	0.1749	9.221	0.9290	0.0384	4.185	0.9361	0.0409	4.465	0.8040	0.0949	8.787	0.9073	0.0522	6.369
INFWIDE Zhang et al. (2023)	TIP,23	0.5965	0.6187	18.745	0.6174	0.7195	19.054	0.6330	0.6408	16.807	0.7829	0.4527	11.753	0.8658	0.4116	12.022
SFHformer Jiang et al. (2024)	ECCV,24	0.8066	0.3940	8.817	0.7134	0.5970	13.205	0.7598	0.4620	10.010	0.7404	0.5095	12.461	0.8867	0.3768	11.611
Classical / Generic Baselines																
Pixel2Pixel Isola et al. (2017)	CVPR,17	0.6637	0.4816	10.324	0.6379	0.6145	15.476	0.6917	0.4502	11.140	0.7303	0.5178	12.405	0.7986	0.3207	11.012
DnCNN Zhang et al. (2017)	TIP,17	0.7245	0.3466	8.883	0.6639	0.5738	14.204	0.7258	0.4275	8.738	0.7221	0.4787	11.818	0.8386	0.3171	10.290
UNet Ronneberger et al. (2015)	MICCAI,15	0.8061	0.2401	9.184	0.6675	0.5658	11.815	0.7581	0.4715	10.285	0.7618	0.4874	12.180	0.8533	0.3243	10.817
CERL (Full)	**Ours**	**0.4172**	**0.7384**	**20.584**	**0.5657**	**0.7259**	**19.478**	**0.5414**	**0.6682**	**18.151**	**0.6235**	**0.6237**	**16.547**	**0.7382**	**0.4228**	**13.809**

appearance of the input light pattern. The recovery therefore reaches the screen image domain: the predicted regions follow the hidden display layout. We next examine how evidence organization, state anchoring, and candidate fusion contribute to this transition.

4.3 Ablation Experiment

Module ablation (Table 2) is conducted on Screen, Figure, Game, and OShome using LPIPS, SSIM, and PSNR, which respectively characterize perceptual similarity, structural preservation, and pixel-level fidelity. For each stage, we either remove the corresponding module or replace it with a general-purpose restoration network while keeping the remaining pipeline unchanged. These comparisons examine whether each recovery stage benefits from its task-specific design when the surrounding pipeline is retained. The full model leads every category–metric comparison. Its LPIPS is 7.4%–20.1% lower than that of the strongest ablated alternative in each category, while SSIM and PSNR improve consistently.

Table 2: Ablation on Screen, Figure, Game, and OShome dataset. Best and second-best results are marked in red and blue.

Method	Metrics											
	Screen			Figure			Game			OShome		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
OERL Variants												
w/o OERL	0.6263	0.7012	18.519	0.8491	0.6893	17.524	0.8127	**0.6345**	16.329	0.7388	0.5922	14.886
OERL → Uformer	**0.5221**	**0.7159**	**19.317**	**0.7077**	**0.7038**	**18.108**	**0.6773**	0.5479	16.873	0.7801	**0.6047**	15.382
OERL → OKNet	0.7504	0.6863	17.915	0.7349	0.6747	16.939	0.8327	0.6297	15.784	0.8391	0.5797	14.389
ISTL Variants												
w/o ISTL	0.6681	0.6974	18.312	0.8036	0.6856	17.328	0.7382	0.6311	16.147	0.7892	0.5892	14.723
ISTL → ConvIR	0.5429	0.7122	19.033	0.7361	0.7001	18.016	0.7044	0.5445	16.783	0.8113	0.6016	15.299
ISTL → DnCNN	0.8142	0.6826	17.283	0.7829	0.6711	16.354	0.6982	0.6177	**16.966**	0.6782	0.5765	13.891
CCM Variants												
w/o CCM	0.7099	0.6937	18.107	0.8372	0.6819	17.133	0.8321	0.5739	16.692	0.7384	0.5895	14.554
CCM → U-Net	0.5827	0.7085	18.932	0.7927	0.6965	17.912	0.7568	0.5837	14.876	0.8192	0.5084	15.216
CCM → SFNet	0.7648	0.6789	16.871	0.7263	0.6674	15.963	0.7632	0.6143	15.239	**0.6732**	0.4928	**15.928**
CERL (Full)	**0.4172**	**0.7384**	**20.584**	**0.5657**	**0.7259**	**19.478**	**0.5414**	**0.6682**	**18.151**	**0.6235**	**0.6237**	**16.547**

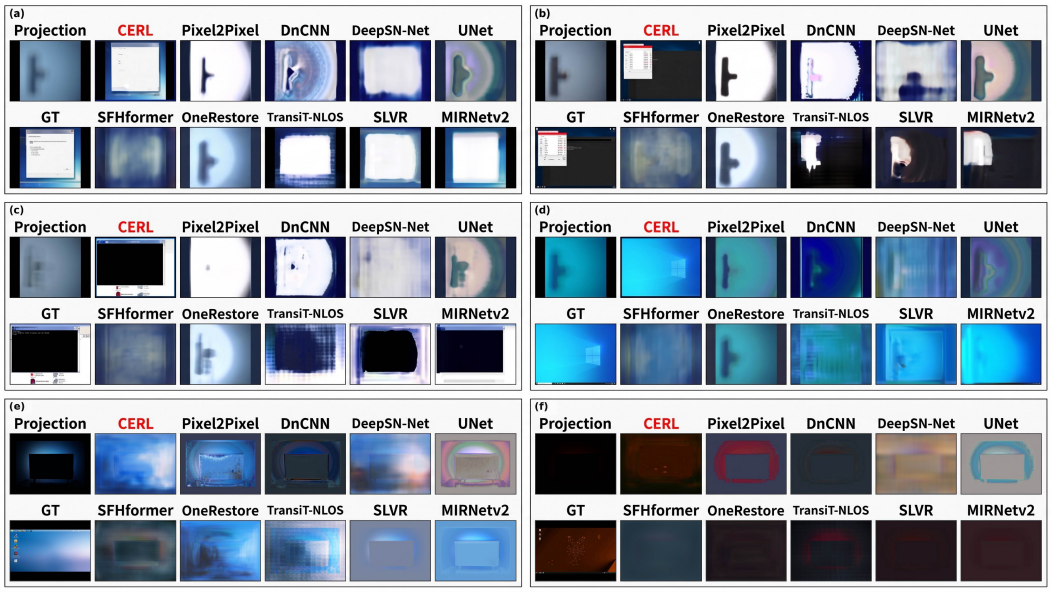

Figure 3: Visual comparison of CERL and baseline methods on four examples. Our model yields visually more faithful restorations across various scenes

4.4 Distance Robustness Evaluation

To evaluate robustness to variations in camera–screen distance, a controlled acquisition protocol was established, as illustrated in Fig. 4. The screen faced the reflecting surface, and the camera was positioned behind the screen. The camera was then translated along a predefined path, with the camera–screen distance progressively increased from 0.5 m to 4.0 m at 0.5 m intervals; 1.0 m was selected as the reference distance. Under identical display and acquisition conditions, images were captured at each distance.

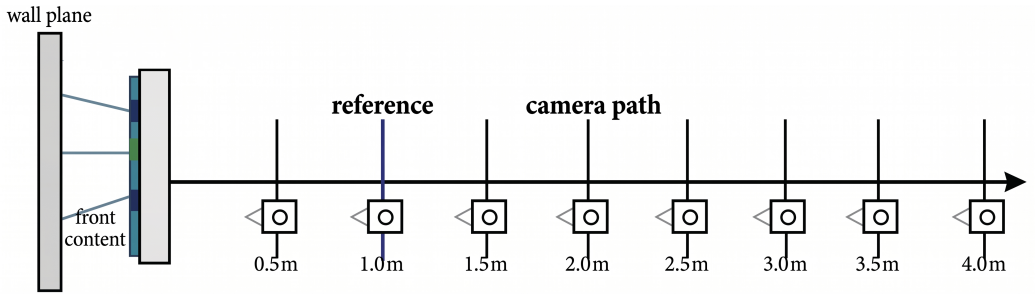

Figure 4: The camera is translated along the rear-side trajectory over distances of 0.5–4.0 m, with 1.0 m designated as the reference position.

As summarized in Fig. 5, pronounced performance degradation emerged for several baselines as the camera–screen distance increased. For instance, OKNet exhibited a sharp increase in LPIPS and simultaneous declines in SSIM and PSNR beyond approximately 2.0 m, whereas CERL (ours) showed only marginal variation across the entire distance range. Retinexformer also experienced noticeable fluctuations and performance degradation at longer distances. In contrast, CERL consistently achieved the lowest LPIPS and the highest SSIM and PSNR values, indicating substantially greater stability under distance variation.

Visual evidence in Fig. 6 indicates that, at near and intermediate distances, CERL produced reconstructions with coherent scene structures, clear cloud and grass boundaries, and faithful color transitions. At the far distance of 4.0 m, competing methods generated severe structural distortions, dark artifacts, color deviations, or blurred contours, particularly for HVI-CIDNet and Retinexformer.

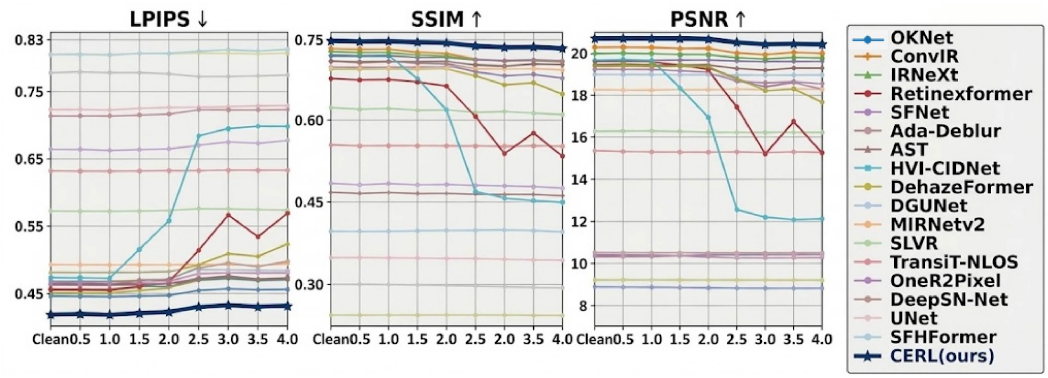

Figure 5: Quantitative comparison of LPIPS, SSIM, and PSNR under increasing camera–screen distances. CERL exhibits consistently superior and more stable performance.

By comparison, CERL maintained consistent edge geometry, texture continuity, and overall visual fidelity across near, mid, and far conditions.

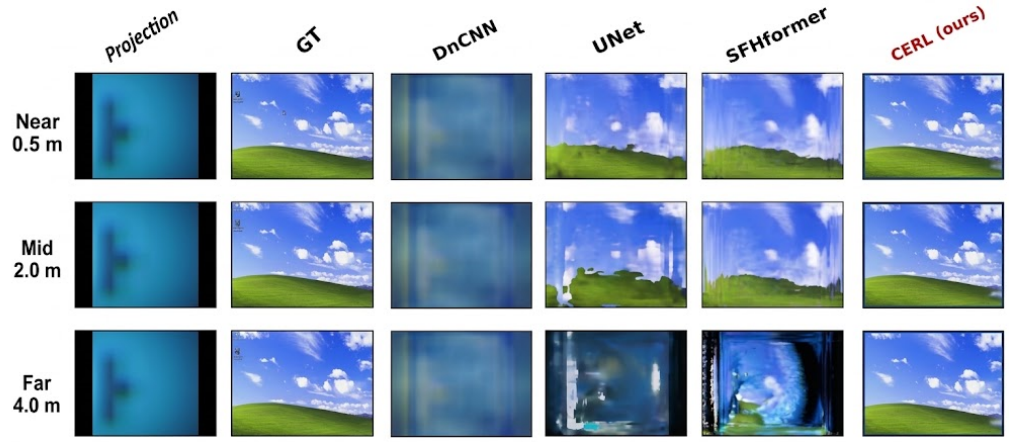

Figure 6: Visualization of representative recovery results at near (0.5 m), mid (2.0 m), and far (4.0 m) camera–screen distances. CERL better preserves scene structure and visual fidelity, particularly at long distances.

5 Conclusion

CERL formulates passive hidden-screen recovery as an observation-constrained inverse problem in which reconstruction is driven by an explicit chain of observation-derived evidence. OERL decomposes the wall measurement into complementary photometric and structural cues and controls their incorporation through a bounded feature interface. ISTL transports this evidence across hierarchical states by anchoring each transition to the current observation and a reference state, while CCM evaluates alternative corrections and aggregates them according to learned spatial agreement. The resulting chain links evidence extraction, state transport, and correction consolidation before the final image readout. Across the paired benchmarks and capture variations considered, this formulation recovers coarse screen luminance, region layout, and major contours from diffuse wall photographs, providing empirical support for the feasibility of ordinary-camera hidden-screen reconstruction under the tested conditions.

REFERENCES

Manel Baradad, Vickie Ye, Adam B. Yedidia, Frédo Durand, William T. Freeman, Gregory W. Wornell, and Antonio Torralba. Inferring light fields from shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6267–6275, 2018. doi: 10.1109/CVPR.2018.00656.

Katherine L. Bouman, Vickie Ye, Adam B. Yedidia, Fredo Durand, Gregory W. Wornell, Antonio Torralba, and William T. Freeman. Turning corners into cameras: Principles and methods. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. doi: 10.1109/ICCV.2017.249.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12504–12513, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cai_Retinexformer_One-stage_Retinex-based_Transformer_for_Low-light_Image_Enhancement_ICCV_2023_paper.html.

Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. IRNeXt: Rethinking convolutional network design for image restoration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6545–6564. PMLR, 2023a. URL https://proceedings.mlr.press/v202/cui23d.html.

Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=tyZ1ChGZIKO.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Revitalizing convolutional network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9423–9438, 2024a. doi: 10.1109/TPAMI.2024.3419007.

Yuning Cui, Wenqi Ren, and Alois Knoll. Omni-Kernel Network for image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1426–1434, 2024b. doi: 10.1609/aaai.v38i2.27907.

Xin Deng, Chenxiao Zhang, Lai Jiang, Jingyuan Xia, and Mai Xu. DeepSN-Net: Deep semi-smooth newton driven network for blind image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2632–2646, 2025. doi: 10.1109/TPAMI.2024.3525089.

Daniel Feijoo, Juan C. Benito, Alvaro Garcia, and Marcos V. Conde. DarkIR: Robust low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10879–10889, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Feijoo_DarkIR_Robust_Low-Light_Image_Restoration_CVPR_2025_paper.html.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. OneRestore: A universal restoration framework for composite degradation. In *Computer Vision – ECCV 2024*, pp. 255–272. Springer, 2024. doi: 10.1007/978-3-031-72655-2_15.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_CVPR_2017_paper.html.

Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *Computer Vision – ECCV 2024*, pp. 381–402. Springer, 2024. doi: 10.1007/978-3-031-72995-9_22.

Jonathan Klein, Christoph Peters, Jaime Martín, Martin Laurenzis, and Matthias B. Hullin. Tracking objects outside the line of sight using 2d intensity images. *Scientific Reports*, 6(1):32491, 2016. doi: 10.1038/srep32491.

Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17399–17410, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Mou_Deep_Generalized_Unfolding_Networks_for_Image_Restoration_CVPR_2022_paper.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Charles Saunders, John Murray-Bruce, and Vivek K. Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740), 2019. doi: 10.1038/s41586-018-0868-6.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. doi: 10.1109/TIP.2023.3256763.

Matthew Tancik, Guy Satat, and Ramesh Raskar. Flash photography for data-driven hidden scene recovery. *arXiv preprint arXiv:1810.11710*, 2018. doi: 10.48550/arXiv.1810.11710.

Antonio Torralba and William T. Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 374–381, 2012. doi: 10.1109/CVPR.2012.6247698.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general U-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Uformer_A_General_U-Shaped_Transformer_for_Image_Restoration_CVPR_2022_paper.html.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004. doi: 10.1109/TIP.2003.819861.

Qiucheng Wu, Yifan Jiang, Junru Wu, Victor Kulikov, Vidit Goel, Nikita Orlov, Humphrey Shi, Zhangyang Wang, and Shiyu Chang. Broad spectrum image deblurring via an adaptive super-network. *IEEE Transactions on Image Processing*, 32:5270–5282, 2023. doi: 10.1109/TIP.2023.3312912.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. HVI: A new color space for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5678–5687, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yan_HVI_A_New_Color_Space_for_Low-light_Image_Enhancement_CVPR_2025_paper.html.

Adam B. Yedidia, Manel Baradad, Christos Thrampoulidis, William T. Freeman, and Gregory W. Wornell. Using unknown occluders to recover hidden scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12223–12231, 2019. doi: 10.1109/CVPR.2019.01251.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1934–1948, 2023. doi: 10.1109/TPAMI.2022.3167175.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017. doi: 10.1109/TIP.2017.2662206.

Zhihong Zhang, Yuxiao Cheng, Jinli Suo, Liheng Bian, and Qionghai Dai. INFWIDE: Image and feature space wiener deconvolution network for non-blind image deblurring in low-light conditions. *IEEE Transactions on Image Processing*, 32:1390–1402, 2023. doi: 10.1109/TIP.2023.3244417.

Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2952–2963, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_Adapt_or_Perish_Adaptive_Sparse_Transformer_with_Attentive_Feature_Refinement_CVPR_2024_paper.html.

SUPPLEMENTARY MATERIAL FOR READING SCREENS FROM WALL REFLECTIONS: COALESCED EVIDENTIAL REPRESENTATION LEARNING FOR HIDDEN SCREEN CONTENT RECOVERY

The appendix presents a more concrete description of the scene for real data collection, the used datasets, and derivations. Furthermore, additional experiments and results are reported and analyzed with considering the robustness to different brightness and noise and the learning performance discussion under different wall material and partial occlusion of captured light spots.

A DESCRIPTION OF THE EXPERIMENTAL SETUP

A.1 DESCRIPTION OF THE SCENE

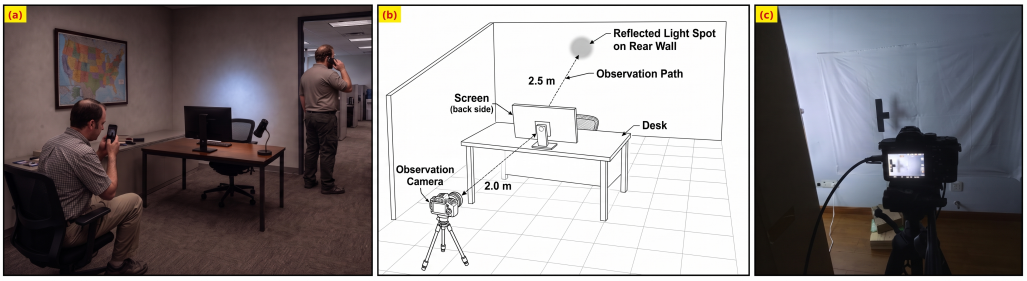

Figure S1: In the figure, (a), (b) and (c) correspond to the rendered scene, schematic diagram, and real-world scene respectively. An observer infers screen content via passive light projection. A light projection from the screen ("Wanted information") is cast onto a wall. By recording the wall's projection without viewing the screen, hacking, or capturing signals, the observer attempts to reconstruct the original content non-invasively.

Figure S1(a) presents a 3D rendering of the CERL scene configured as an interview-room side-channel attack scenario. During the interview, the interviewer is temporarily distracted while answering a phone call near the doorway. The interviewee, acting as an unauthorized observer, covertly raises a smartphone toward the back of the monitor to capture both the the screen light reflected by the wall and the reflected screen light on the wall. Because the monitor content is not directly accessible from the interviewee's normal viewpoint, the smartphone records only the indirect optical leakage produced by the screen–wall interaction. This arrangement models a realistic opportunistic attack in which sensitive display information is passively acquired without hacking, signal tapping, or other active interference.

Figure S1(b) presents a two-dimensional schematic of the optical path and imaging logic. It illustrates the indirect transmission of screen-related information from the rear side of the monitor to the wall and subsequently to the observation camera. The schematic indicates an approximately 2.0 m separation between the camera and the monitor, while the observation path toward the reflected light spot on the rear wall is approximately 2.5 m. The dashed lines denote the propagation paths. Diffuse reflection, attenuation, and spatial spreading along this screen–wall–camera channel substantially degrade and compress the original visual information.

Figure S1(c) depicts the actual experimental environment. A standard computer monitor displays critical information to simulate a practical side-channel attack scenario. The screen is placed approximately 0.9 m away from the wall behind it.

A.2 DESCRIPTION OF THE DATASETS

A.2.1 THE SCREEN DATASET

This data set is collected from the real indoor reflection scene: a 60 Inch self luminous LED screen is placed facing the white diffuse reflection wall, with a distance of about 0.5 meters. The light emitted

by the screen directly irradiates the wall to form a bright spot. The observation end uses the Sony A7SII camera to shoot from the direction facing the back of the screen. The picture includes the outline of the back of the screen and the reflected light spot on the wall as the recorded light spot image, and synchronously records the real content displayed on the screen at that time (including web pages, tables, charts, UI, password characters, etc.) as the matching label. Each sample in the dataset is composed of "composite image including the spot on the back of the screen and the wall" and "real screen image". The task goal is to reverse restore the specific content displayed on the screen with only a single composite image.



Figure S2: Real scene pictures

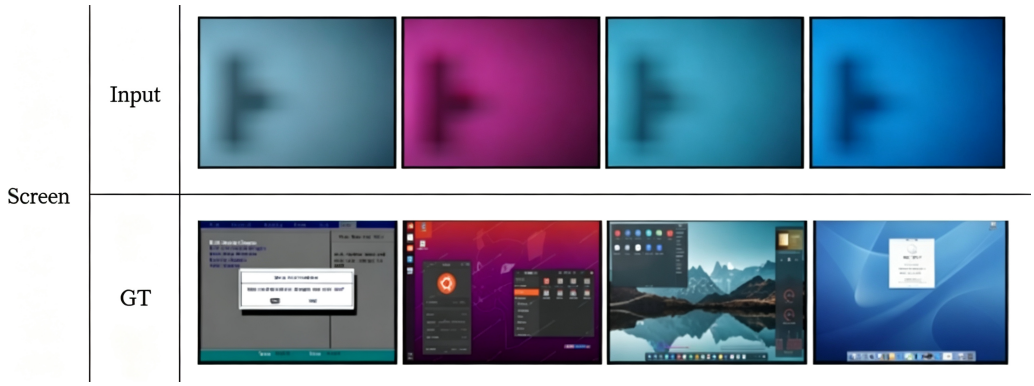

Figure S3: The samples of the Screen dataset

A.2.2 THE FIGURE DATASET, GAME DATASET, OSHOME DATASET AND PHOTO DATASET

These datasets are collected from the real indoor reflection scene: a 60 Inch self luminous LED screen is placed facing the white diffuse reflection wall, with a distance of about 0.5 meters. The light emitted by the screen directly irradiates the wall to form a bright spot. The observation end uses the Sony A7SII camera to shoot from the direction facing the back of the screen. The picture includes the outline of the back of the screen and the reflected light spot on the wall as the recorded light spot image, and synchronously records the real content displayed on the screen at that time (including web pages, tables, charts, UI, password characters, etc.) as the matching label. Each sample in the dataset is composed of "composite image including the spot on the back of the screen and the wall" and "real screen image". The task goal is to reverse restore the specific content displayed on the screen with only a single composite image.



Figure S4: Real scene pictures

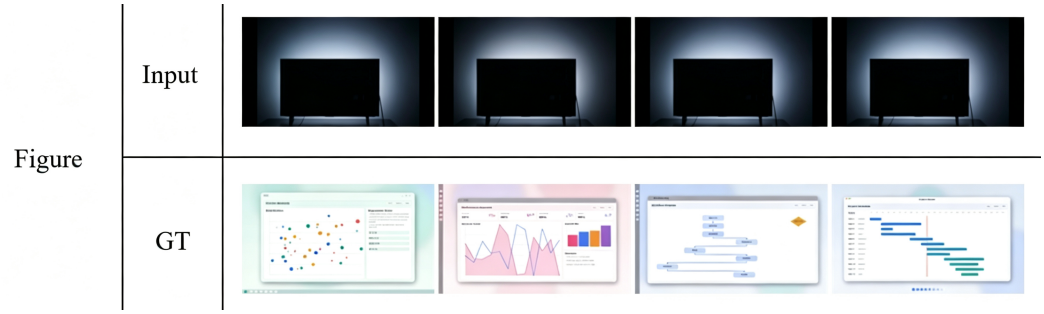

Figure S5: The samples of the Figure dataset

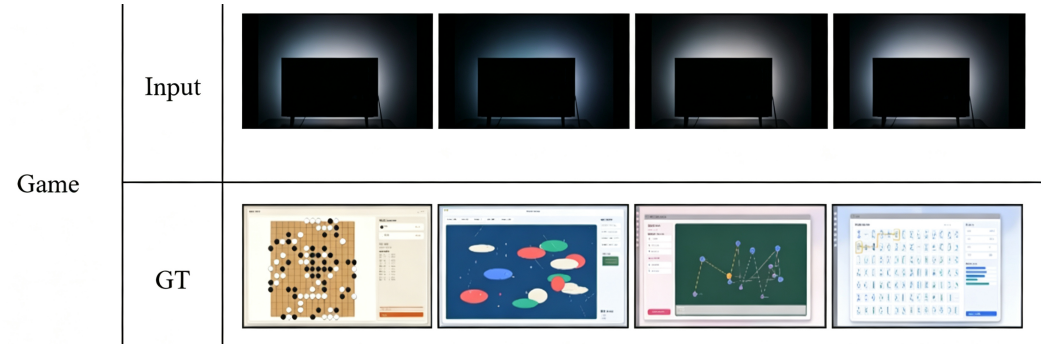

Figure S6: The samples of the Game dataset

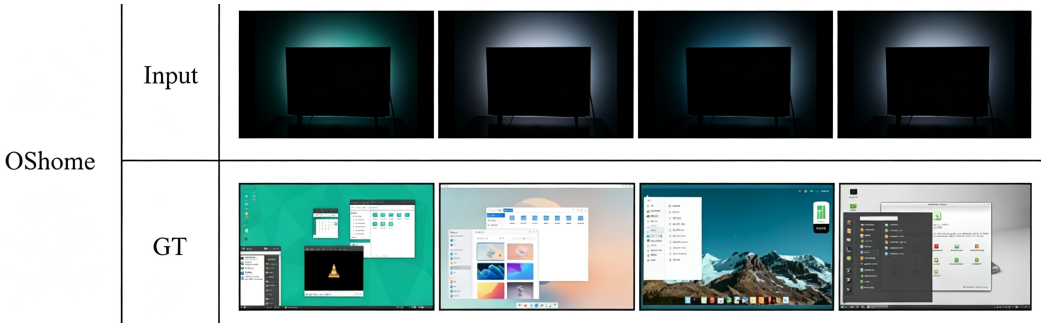

Figure S7: The samples of the OShome dataset

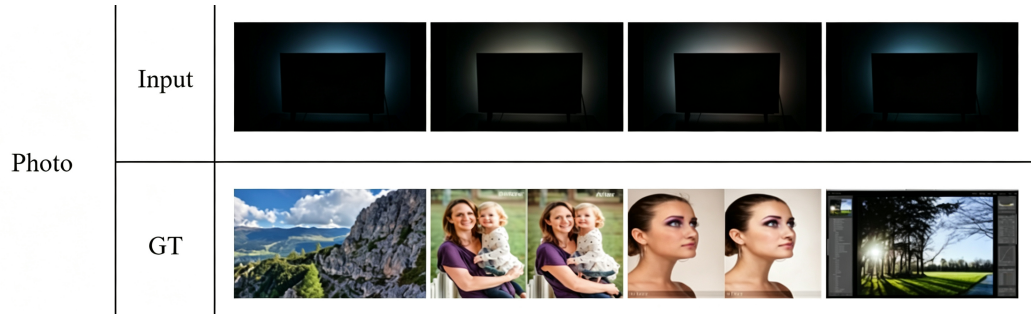

Figure S8: The samples of the Photo dataset

B MECHANISTIC DERIVATIONS

B.1 OBSERVATION AND SUPERVISION CONTRACT

The observation-only prediction mapping is

$$\hat{x} = \Phi_\theta(y), \qquad y, \hat{x} \in \mathbb{R}^{B\times 3\times H\times W}. \tag{B.1}$$

Equation (B.1) defines y as a batch of RGB wall observations, $\hat{x}$ as the aligned RGB reconstruction, and Φ_θ as a neural mapping with learned parameters θ. The observation and learned parameters are available at inference, while no target enters this mapping. The paper-level implementation correspondence is one RGB input followed by one final consensus-refined RGB output.

The paired training tuple and valid-region reduction are

$$(y, x, m), \qquad m \in \{0,1\}^{B\times 1\times H\times W}, \qquad m_C = \mathrm{repeat}_c(m), \qquad \langle \varrho \rangle_m = \frac{\sum m_C \odot \varrho}{\sum m_C + 10^{-8}}. \tag{B.2}$$

Equation (B.2) adds the RGB clean target x, one-channel valid mask m, and arbitrary loss map $\varrho \in \mathbb{R}^{B\times C\times H\times W}$, with $m_C \in \{0,1\}^{B\times C\times H\times W}$ denoting the channel-repeated mask and $y, x \in \mathbb{R}^{B\times 3\times H\times W}$. Both the numerator and denominator use the same repeated mask; for a one-channel loss, the repetition is the identity. The target, mask, loss map, and reduction are supervision- and evaluation-only and absent at inference, whereas y remains inference-visible. The implementation correspondence is the exact masked normalized sum used by the explicitly extracted pixelwise components L_0 through L_4 and L_6.

B.2 OERL DERIVATION

Two deterministic luminance operators serve different roles:

$$l_0(z) = \frac{1}{3}\sum_{i=0}^{2} z_i, \qquad l_1(z) = 0.299 z_0 + 0.587 z_1 + 0.114 z_2. \tag{B.3}$$

Equation (B.3) maps an RGB tensor $z \in \mathbb{R}^{B\times 3\times H\times W}$ to one-channel maps $l_0(z), l_1(z)$. The first is the exact channel mean used for inference-visible evidence and CCM context, while the second is the fixed luminance projection used by training-only losses when z is a prediction or target. The implementation correspondence is deterministic channel reduction.

For an odd positive neighborhood width k, the local mean is

$$p_k(z) = \frac{1}{k^2}\sum_{(i,j)\in\Omega_k} z(i,j). \tag{B.4}$$

Equation (B.4) defines p_k as a stride-one same-size local mean over the index set Ω_k, with i, j denoting spatial offsets and z denoting a one-channel or channelwise tensor. The operator is fixed, preserves batch and channel domains, and is inference-visible whenever its input is observation-derived. Its implementation correspondence is same-size average filtering with zero padding; The local mean aggregates the observation over a fixed spatial neighborhood.

The normalized Sobel response is

$$d_0(z) = \sqrt{(\mathcal{S}_0 * l_0(z))^2 + (\mathcal{S}_1 * l_0(z))^2 + 10^{-6}}, \qquad \bar{d}_0(z) = \frac{d_0(z)}{\max(\mathrm{mean}_p\, d_0(z), 10^{-4})}. \tag{B.5}$$

Equation (B.5) uses fixed horizontal and vertical Sobel filters $\mathcal{S}_0, \mathcal{S}_1$, convolution $*$, spatial index p, raw magnitude $d_0(z)$, and normalized response $\bar{d}_0(z)$, all with one-channel domain $\mathbb{R}^{B\times 1\times H\times W}_{\geq 0}$.

The constants 10^{-6} and 10^{-4} are the respective settings of ϵ_0 and ϵ_1 in Equation (2). With $z = y$, all terms are deterministic and inference-visible. The implementation correspondence is the exact stabilized spatial-mean normalization; $\bar{d}_0$ is a signal response.

Local contrast is

$$c(z) = \sqrt{\max(p_5(l_0(z)^2) - p_5(l_0(z))^2, 0) + 10^{-6}}. \tag{B.6}$$

Equation (B.6) defines the nonnegative one-channel map $c(z)$ from the local second moment and squared local mean of $l_0(z)$; p_5 is the contrast-window operator p_c from Equation (2), and 10^{-6} is the same ϵ_0 setting used for raw magnitudes. For $z = y$, the complete expression is deterministic and inference-visible. Its implementation correspondence is the exact stabilized local standard deviation used by the evidence extractor; the local standard deviation supplies complementary contrast information.

The five-channel evidence tensor is

$$e(y) = \left[p_9(l_0(y)),\ l_0(y) - p_9(l_0(y)),\ \tanh(0.20\bar{d}_0(y)),\ p_3(l_0(y)) - p_9(l_0(y)),\ \tanh(3c(y))\right]. \tag{B.7}$$

Equation (B.7) gives $e(y) \in \mathbb{R}^{B\times 5\times P\times Q}$ after bilinear resizing to a requested feature grid, where P and Q denote grid height and width. Its indexed channels contain broad luminance, relative variation, normalized boundary response, cross-scale difference, and local contrast; every channel is deterministically derived from the inference-visible observation. The implementation correspondence is the exact ordered concatenation and resizing used at each OERL and ISTL scale.

Per-pixel channel normalization is

$$\mathcal{N}(z) = \frac{z - \mu}{\sqrt{\nu + 10^{-6}}} \odot \vartheta_0 + \varpi_0, \qquad \mu = \mathrm{mean}_c\, z, \qquad \nu = \mathrm{mean}_c(z - \mu)^2. \tag{B.8}$$

Equation (B.8) preserves $z \in \mathbb{R}^{B\times C\times P\times Q}$, with one-channel per-pixel mean μ, variance ν, fixed stabilizer 10^{-6}, and learned channelwise affine parameters ϑ_0, ϖ_0. The operator $\mathcal{N}$ denotes this complete per-pixel channel normalization. Feature inputs and learned parameters are available at inference. The implementation correspondence is exact normalization across channels before feature projection; normalization aligns a neural representation .

Channelwise spatial-response normalization is

$$\rho_c = \frac{\|z_c\|_2}{\max(\mathrm{mean}_i \|z_i\|_2, 10^{-6})}, \qquad \mathcal{R}(z) = z + \vartheta_1 \odot (z \odot \rho) + \varpi_1. \tag{B.9}$$

Equation (B.9) defines a broadcast response ratio $\rho \in \mathbb{R}^{B\times C\times 1\times 1}$ for channel index c and a normalized output $\mathcal{R}(z)$ with the domain of z; i indexes channels and ϑ_1, ϖ_1 are learned affine parameters. All quantities are learned intermediates or parameters available at inference. The implementation correspondence is exact spatial norm comparison followed by an affine residual.

The external observation descriptor supplies a scene gate through

$$\zeta = \chi_{\theta_0}(y) \in \mathbb{R}^{B\times 64}, \qquad g_0 = \sigma(\omega_0^\top \zeta + \omega_1). \tag{B.10}$$

Equation (B.10) defines the descriptor ζ, learned encoder χ_{θ_0} with parameter subset θ_0, learned projection vector ω_0, learned scalar bias ω_1, logistic nonlinearity σ, and one broadcast scalar gate g_0 per image. The observation and learned quantities are inference-visible. The implementation correspondence is the supplied descriptor interface and scalar projection.

The OERL mixed representation is

$$h_0 = \psi_0(\mathcal{N}(f)) + \psi_1(e), \qquad h_1 = \psi_2(h_0) + \psi_3(\mathrm{SiLU}(h_0)), \qquad h = \mathcal{R}(\mathrm{SiLU}(h_1)). \tag{B.11}$$

Equation (B.11) takes feature $f \in \mathbb{R}^{B\times C\times P\times Q}$, evidence e, and normalized representation $\mathcal{N}(f)$, producing $h_0, h_1, h \in \mathbb{R}^{B\times 48\times P\times Q}$. The learned maps ψ_0, ψ_1 project feature and evidence, ψ_2, ψ_3 mix local and dilated contextual responses into h_1, and $\mathcal{R}$ from Equation (B.9) normalizes $\mathrm{SiLU}(h_1)$. All inputs and parameters are inference-visible. The implementation correspondence is projected summation, local and contextual mixing, activation, and response normalization in their recorded order; this is a neural representation.

OERL confidence and its analytic evidence bound are

$$g_1 = \sigma(\psi_4(\mathrm{pool}(h))), \qquad b_0 = 0.010 + 0.105\tanh(|e_1| + 0.35|e_2| + |e_4|)\,. \tag{B.12}$$

Equation (B.12) defines learned per-image confidence g_1 from pooled hidden state h and the spatial bound $b_0 \in \mathbb{R}^{B\times 1\times P\times Q}$ from evidence channels e_1, e_2, e_4; ψ_4 is learned and pool is deterministic spatial averaging. Both terms are inference-visible. The implementation correspondence is the neural confidence branch paired with the exact analytic amplitude bound.

The OERL residual transition is

$$f' = f + \alpha_0 g_0 g_1 b_0 \odot \tanh(\psi_5(h)), \qquad 0 \le \alpha_0 \le 0.036. \tag{B.13}$$

Equation (B.13) returns f' in the domain of f, with learned candidate map ψ_5, learned clipped step α_0, fixed cap 0.036, scene gate g_0, confidence g_1, and analytic bound b_0; $\odot$ denotes elementwise multiplication. All operands are available from the observation-conditioned forward path at inference. The implementation correspondence is the exact bounded residual structure used at the deepest and first decoder states.

B.3 ISTL Derivation

Anchor-relative transport and normalized packing are

$$t = s - a, \qquad \varsigma = [\mathcal{N}(s), \mathcal{N}(a), \mathcal{N}(t)]. \tag{B.14}$$

Equation (B.14) uses current state s, aligned anchor a, and signed difference t, all in $\mathbb{R}^{B\times C\times P\times Q}$, and concatenates their normalized forms into packed state $\varsigma \in \mathbb{R}^{B\times 3C\times P\times Q}$. These learned intermediates arise from the observation-driven path and are inference-visible. The implementation correspondence is exact anchor alignment, subtraction, normalization, and channel packing; interpreting t as inverse transport is explanatory.

Directional transport mixing is

$$\begin{aligned} &h_0 = \phi_0(\varsigma) + \phi_1(e), \qquad h_1 = \phi_2(h_0), \\ &h_2 = h_1 + \phi_3(\mathrm{SiLU}(h_1)) + \phi_4(\mathrm{SiLU}(h_1)), \qquad h = \mathcal{R}(\mathrm{SiLU}(h_2)). \end{aligned} \tag{B.15}$$

Equation (B.15) projects packed state ς and evidence e into $h_0, h_1, h_2, h \in \mathbb{R}^{B\times 48\times P\times Q}$. The learned maps ϕ_0, ϕ_1 project state and evidence, ϕ_2 produces the transport response h_1, and ϕ_3, ϕ_4 apply horizontal and vertical mixing to $\mathrm{SiLU}(h_1)$ before their residual sum forms h_2; $\mathcal{R}$ then normalizes $\mathrm{SiLU}(h_2)$. All inputs and parameters are inference-visible. The implementation correspondence is the exact local, row, column, activation, and response-normalization order.

The ISTL allowance and update are

$$\begin{aligned} b_1 &= 0.010 + 0.075\tanh(\mathrm{stop}(\mathrm{mean}_c\,|t|) + 0.25(|e_2| + |e_4|))\,, \\ s' &= s + \alpha_1 g_0 g_2 b_1 \odot \tanh(\phi_5(h)), \qquad 0 \le \alpha_1 \le 0.030. \end{aligned} \tag{B.16}$$

Equation (B.16) defines spatial allowance $b_1 \in \mathbb{R}^{B\times 1\times P\times Q}$, updated state $s' \in \mathbb{R}^{B\times C\times P\times Q}$, learned candidate map ϕ_5, learned scalar step α_1 with fixed cap 0.030, scalar scene gate g_0, and spatial transport gate $g_2 \in (0,1)^{B\times 1\times P\times Q}$. The operator stop fixes the gradient contribution of the mean transport magnitude without changing its inference value, while e_2, e_4 remain observation-derived.

Every operand is produced by the observation-driven forward path or learned parameters and is inference-visible. The implementation correspondence is the exact analytic bound and neural residual update at the documented state scales.

The initial RGB readout before collective refinement is

$$x^{(0)} = y + o(f). \tag{B.17}$$

Equation (B.17) defines the initial RGB state $x^{(0)} \in \mathbb{R}^{B\times 3\times H\times W}$ as observation y plus the learned projection $o(f)$ of final decoder feature f. The feature, observation, and learned operator are inference-visible. The implementation correspondence is the exact residual RGB readout passed to CCM.

B.4 CCM Derivation and Training Objective

At collective step j, the RGB context bundle is

$$\begin{aligned} \upsilon^{(j)} = \big[&x^{(j)}, x^{(0)}, y, y - x^{(j)}, x^{(j)} - x^{(0)}, \\ &p_5(y - x^{(j)}), l_0(x^{(j)}), l_0(y), l_0(y - x^{(j)}), \\ &\tanh(0.20\bar{d}_0(x^{(j)})), \tanh(0.20\bar{d}_0(y)), \tanh(0.20\bar{d}_0(y - x^{(j)}))\big]. \end{aligned} \tag{B.18}$$

Equation (B.18) concatenates current RGB state $x^{(j)}$, fixed base state $x^{(0)}$, observation y, their RGB differences and local mean, three channel means, and three normalized edge responses into $\upsilon^{(j)} \in \mathbb{R}^{B\times 24\times H\times W}$. Every term is deterministic and inference-visible, using $p_5, l_0, \bar{d}_0$ from Equations (B.4), (B.3), and (B.5). The implementation correspondence is the exact feature bundle supplied to the learned CCM mixer; the bundle organizes neural context .

The candidate bound and RGB candidates are

$$\begin{aligned} b_2^{(j)} &= 0.010 + 0.090 \tanh\Big(p_5(\mathrm{mean}_c\,|y - x^{(j)}|) + 0.05\,\mathrm{clip}(p_5(\bar{d}_0(y)), 0, 3)\Big), \\ q_k^{(j)} &= b_2^{(j)} \odot \tanh(\eta_k(\upsilon^{(j)})). \end{aligned} \tag{B.19}$$

Equation (B.19) defines the one-channel spatial bound $b_2^{(j)}$ from observation residual magnitude and locally averaged observation edges, and the RGB candidate $q_k^{(j)} \in \mathbb{R}^{B\times 3\times H\times W}$ from learned map η_k for candidate index k. The bound is deterministic, the candidate map is learned, and both use only inference-visible quantities. The implementation correspondence is the exact analytic candidate ceiling followed by parallel neural candidate projection.

Normalized votes and their consensus are

$$v_k^{(j)} = \frac{\exp(w_k^{(j)})}{\sum_{k'=1}^{K} \exp(w_{k'}^{(j)})}, \qquad u^{(j)} = \sum_k v_k^{(j)} q_k^{(j)}. \tag{B.20}$$

Equation (B.20) uses learned one-channel logits $w_k^{(j)}$, normalized nonnegative votes $v_k^{(j)}$, RGB candidates $q_k^{(j)}$, and RGB consensus $u^{(j)}$; the dummy index k' and the output index k both range over the finite candidate set at each pixel. Logits are neural intermediates and aggregation is deterministic, with all terms inference-visible. The implementation correspondence is exact soft voting over parallel candidates.

Candidate disagreement and agreement are

$$\delta_j = \frac{1}{3K} \sum_{k=1}^{K} \sum_{i=0}^{2} ([q_k^{(j)}]_i - [u^{(j)}]_i)^2, \qquad r_j = \exp(-\delta_j), \qquad K = 4. \tag{B.21}$$

Equation (B.21) defines the one-channel mean squared disagreement δ_j across candidate index k and RGB index i, agreement $r_j \in (0,1]^{B\times 1\times H\times W}$, and the recorded candidate count K. Both maps are deterministic functions of inference-visible candidates and consensus. The implementation correspondence is the exact unweighted mean over candidate and RGB axes followed by an exponential.

Collective evolution and the final readout are

$$x^{(j+1)} = x^{(j)} + \alpha_2 g_0 g_3^{(j)} g_4^{(j)} r_j \odot u^{(j)}, \qquad j \in \{0,1,2\}, \qquad \hat{x} = x^{(3)}, \qquad 0 \le \alpha_2 \le 0.037. \tag{B.22}$$

Equation (B.22) updates full-resolution RGB state $x^{(j)}$ with learned step α_2 capped at 0.037, shared scene gate g_0, step-specific spatial gate $g_3^{(j)}$, step-specific image-level gate $g_4^{(j)}$, agreement r_j, and consensus $u^{(j)}$, then reads the terminal state as $\hat{x}$. The two step-specific gates are recomputed from the current hidden representation at every iteration. All operands are inference-visible and the gates and step are learned. The implementation correspondence is the exact recorded three-step update and final readout.

The first supervised components are

$$L_0 = \left\langle \sqrt{(\hat{x}-x)^2 + 10^{-6}} \right\rangle_m, \qquad L_1 = \langle (\hat{x}-x)^2 \rangle_m, \qquad L_2 = \langle |l_1(\hat{x}) - l_1(x)| \rangle_m. \tag{B.23}$$

Equation (B.23) defines masked Charbonnier loss L_0, squared error L_1, and luminance error L_2 as nonnegative batch scalars, using prediction $\hat{x}$, training-only target x, mask m, and fixed luminance operator l_1 from Equation (B.3). The implementation correspondence is exact valid-region reduction by Equation (B.2).

The Sobel, structural, and perceptual components are

$$\begin{aligned}
\mu = p_7(\iota), \quad \mu' = p_7(\iota'), \quad \nu = p_7(\iota^2) - \mu^2, \quad \nu' = p_7(\iota'^2) - \mu'^2, \quad \kappa = p_7(\iota\iota') - \mu\mu', \\
S_7(\iota,\iota') = \frac{(2\mu\mu' + 0.01^2)(2\kappa + 0.03^2)}{(\mu^2 + \mu'^2 + 0.01^2)(\nu + \nu' + 0.03^2) + 10^{-8}}, \\
L_3 = \langle |\mathcal{S}_0 * l_1(\hat{x}) - \mathcal{S}_0 * l_1(x)| + |\mathcal{S}_1 * l_1(\hat{x}) - \mathcal{S}_1 * l_1(x)| \rangle_m, \\
L_4 = \left\langle \frac{1}{2}\left(1 - \mathrm{clip}(S_7(l_1(\hat{x}), l_1(x)), -1, 1)\right) \right\rangle_m, \qquad L_5 = \varphi(\hat{x}, x, m).
\end{aligned} \tag{B.24}$$

Equation (B.24) defines local structural map $S_7 \in \mathbb{R}^{B\times 1\times H\times W}$ for $\iota = l_1(\hat{x})$ and target-derived $\iota' = l_1(x)$, with local means μ, μ', variances ν, ν', and covariance κ in the same spatial domain. The operator p_7 is the stride-one 7 by 7 mean from Equation (B.4) with zero padding, matching the displayed moment definitions. The equation also defines nonnegative scalar Sobel loss L_3, clipped structural loss L_4, and external perceptual loss L_5. The fixed filters $\mathcal{S}_0, \mathcal{S}_1$, target x, and mask m are used only in training. The learned perceptual interface φ receives a valid-mask crop resized to a 224 by 224 canvas. The correspondence is exact for L_3, L_4 and for the known preprocessing boundary of L_5.

Finally, forward-difference leakage and the weighted total are

$$\begin{aligned}
d_1(z) = \sqrt{(\nabla_0 l_1(z))^2 + (\nabla_1 l_1(z))^2 + 10^{-8}}, \qquad \bar{d}_1(z) = \frac{d_1(z)}{\mathrm{mean}_p\, d_1(z) + 10^{-4}}, \\
L_6 = \left\langle d_1(\hat{x})\, \mathrm{clip}\big(\bar{d}_1(y) - 0.65\bar{d}_1(x), 0, 4\big) \right\rangle_m, \qquad L = L_0 + \sum_{i=1}^{6} \lambda_i L_i.
\end{aligned} \tag{B.25}$$

Equation (B.25) defines fixed horizontal and vertical forward differences ∇_0, ∇_1, one-channel gradient magnitude $d_1(z)$ and normalized magnitude $\bar{d}_1(z)$ in $\mathbb{R}^{B\times 1\times H\times W}$, scalar leakage L_6, scalar

total objective L, and fixed nonnegative weights λ_i. The constants 10^{-8} and 10^{-4} are the respective settings of ϵ_4 and ϵ_5 in Equation (13). The observation y, prediction $\hat{x}$, training-only target x, and mask m supply the operands, while observation and target edge weights are held fixed during differentiation. The loss maps, target, mask, and reduction are training-only. The objective settings are $(\lambda_1, \ldots, \lambda_6) = (1.75, 0.38, 0.012, 1.18, 0.015, 0.020)$, with unit weight on L_0. The implementation correspondence is the exact leakage construction and weighted seven-term objective.

C ADDITIONAL EXPERIMENTS AND ANALYSES

C.1 ABLATION ON OERL

To assess the contributions of optical evidence and the bounded gated update to hidden-screen reconstruction, comparative experiments were conducted across four datasets, namely Screen, Figure, Game, and OSlock . For optical evidence, the complete evidence representation was removed or replaced with low frequency information, high pass information, and Sobel edge information. For the update mechanism, the proposed bounded gated update was removed or replaced with ResBlock, SE Block, and CBAM. The objective was to examine the effectiveness of optical guidance and constrained feature updates under varying projection conditions.

Table 3: **Ablation study of optical evidence and bounded gated updates across four projection scenarios. CERL (Full) consistently achieves superior restoration quality across all evaluation metrics, demonstrating the effectiveness of complementary optical cues and bounded iterative updates under heterogeneous projection degradations.**

Method	Metrics											
	Screen			Figure			Game			OShome		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
OERL w/o Optical Evidence	0.6473	0.6429	17.842	0.7462	0.6103	16.517	0.7248	0.5516	15.104	0.8372	0.4847	13.426
Evidence → Low-frequency Only	0.4926	0.5927	18.536	0.6684	0.6432	17.263	0.6417	0.5823	15.742	0.7461	0.5173	14.137
Evidence → High-pass Only	0.5641	0.6307	17.481	0.7735	0.5964	16.142	0.7532	0.5323	14.728	0.8714	0.4682	13.041
Evidence → Sobel Edge Only	0.4715	0.6882	18.874	0.6347	0.6561	17.645	0.6103	0.5967	16.126	**0.7084**	0.5346	**14.523**
w/o Bounded Gated Update	0.6824	0.5183	17.127	0.8016	0.5827	15.783	0.7826	0.5234	14.361	0.8013	0.4516	12.674
Bounded Gated Update → ResBlock	0.5237	0.6519	18.163	0.7125	0.6246	16.847	0.6874	0.5642	15.435	0.9012	0.4981	13.674
Bounded Gated Update → SE Block	0.4842	0.5928	18.653	0.6531	0.6487	17.426	0.6241	0.5886	15.917	0.7946	0.5264	13.742
Bounded Gated Update → CBAM	**0.4586**	**0.6973**	**19.084**	**0.6074**	**0.6673**	**17.913**	**0.5835**	**0.6082**	**16.434**	0.7361	**0.5527**	14.316
CERL (Full)	**0.4172**	**0.7384**	**20.584**	**0.5657**	**0.7259**	**19.478**	**0.5414**	**0.6682**	**18.151**	**0.6235**	**0.6237**	**16.547**

Table 3 reports metric specific outcomes. CERL demonstrates consistently favorable performance across all evaluation criteria. Compared with the variant without optical evidence, the complete model reduces LPIPS by approximately 24.2% to 35.5%, while improving SSIM by 14.9% to 28.7% and PSNR by 15.4% to 23.2%. Among the individual evidence variants, Sobel edge information produces the most consistent restoration quality. For the update mechanism, replacing the proposed design with conventional residual or attention modules results in clear performance degradation. Although CBAM provides the strongest alternative, the complete model remains superior across all scenarios and evaluation metrics. These patterns indicate that optical evidence and the bounded gated update contribute consistently to restoration quality under heterogeneous projection conditions.

This behavior may arise from the complementary roles of optical evidence and constrained iterative updates. Optical evidence provides stable structural and frequency cues for distinguishing degradation patterns from image content. Meanwhile, the bounded gated update adaptively regulates the magnitude of feature refinement, suppressing excessive corrections and promoting coherent information propagation. Consequently, their joint integration enables CERL to preserve global appearance and local details while maintaining stable restoration across diverse projection scenarios.

C.2 ABLATION ON ISTL

In the first group, the anchor is removed or assigned to the current state, the encoder state, or the decoder base state while all other settings remain fixed. In the second group, directional gated transport is removed or replaced with a ConvBlock, a ResBlock, or a linear blending operation. The quantitative results are summarized in Table 4.

Removing the anchor consistently degrades the restoration results across all scenarios, producing higher LPIPS and lower SSIM and PSNR. Using the current state or the encoder state recovers part of the lost performance, whereas anchoring to the decoder base state provides the strongest results among the anchor variants. This observation suggests that the decoder base state offers a more stable reference for subsequent feature refinement and is better aligned with the reconstruction objective.

The importance of directional gated transport is further demonstrated by the second group of experiments. Disabling this mechanism leads to a clear decline across all evaluation metrics. Replacing it with conventional feature transformation modules partially restores the performance, with the ResBlock variant generally performing better than the ConvBlock and linear blending alternatives. Nevertheless, the complete CERL model remains superior across all scenarios. These results indicate that directional gated transport performs a functional role in regulating information propagation and preserving structural consistency.

This behavior may result from the complementary roles of state anchoring and directional information transport. The decoder base state provides a reliable initialization for iterative refinement, while directional gated transport selectively propagates useful corrections through the network. Their joint design constrains unstable updates, maintains coherent feature evolution, and improves the robustness of ISTL.

Table 4: **Ablation study of anchor states and directional gated transport across four projection scenarios. CERL (Full) consistently achieves the best restoration performance, highlighting the importance of decoder base state anchoring and directional gated transport for stable feature propagation.**

Method	Metrics											
	Screen			Figure			Game			OShome		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
ISTL w/o Anchor	0.6478	0.5426	16.128	0.8834	0.4917	14.205	0.8691	0.4302	12.847	1.0263	0.3415	11.082
Anchor → Current State	0.5861	0.5843	17.042	0.8072	0.5364	15.163	0.7825	0.4736	13.719	0.9375	0.3824	11.936
Anchor → Encoder State	0.5207	0.6271	18.175	0.7216	0.5823	16.241	0.6947	0.5185	14.832	0.8402	0.4276	13.054
Anchor → Decoder State	**0.4735**	**0.6708**	**19.263**	**0.6384**	**0.6357**	**17.486**	**0.6128**	**0.5641**	**16.075**	**0.7316**	**0.4853**	**14.271**
w/o Directional Gated Transport	0.6842	0.5104	15.416	0.9357	0.4532	13.478	0.9164	0.3947	12.136	1.0871	0.3028	10.315
Directional Transport → ConvBlock	0.5483	0.6015	17.634	0.7649	0.5571	15.627	0.7352	0.4963	14.208	0.8846	0.4017	12.463
Directional Transport → ResBlock	0.4916	0.6527	18.807	0.6753	0.6048	16.792	0.6419	0.5426	15.461	0.7754	0.4532	13.628
Directional Transport → Linear Blend	0.6135	0.5632	16.571	0.8428	0.5146	14.684	0.8207	0.4518	13.253	0.9827	0.3609	11.474
CERL (Full)	**0.4172**	**0.7384**	**20.584**	**0.5657**	**0.7259**	**19.478**	**0.5414**	**0.6682**	**18.151**	**0.6235**	**0.6237**	**16.547**

C.3 Ablation on CCM

In the first group, the consensus mechanism is removed or replaced with uniform averaging, max vote selection, or single mode prediction, while other settings remain fixed. In the second group, the agreement guided collective update is removed or replaced with a ResBlock, an SE Block, or CBAM. These experiments evaluate the roles of consensus estimation and coordinated feature updating under different projection degradations.

Table 5 shows that CERL (Full) consistently achieves the best performance across all scenarios and metrics. Removing consensus causes clear degradation, while alternative strategies recover only part of the performance. Single mode prediction performs best among the alternatives but remains inferior to the complete model, confirming that consensus modeling integrates complementary predictions and reduces restoration ambiguity.

Removing the agreement guided collective update also degrades performance across all metrics. Although residual and attention based replacements provide partial improvements, the SE Block performs best among them and still falls below the complete model. These results demonstrate that consensus modeling and collective updating jointly suppress inconsistent corrections, coordinate feature propagation, and preserve structural coherence.

C.4 Results On Robustness to Brightness

To evaluate luminance robustness, we perturb the input brightness symmetrically around the normal exposure level, ranging from 24% below to 24% above the nominal condition, and compute LPIPS,

Table 5: **Ablation study of consensus modeling and agreement-guided collective updates across four projection scenarios. CERL (Full) consistently achieves the best restoration quality across all evaluation metrics, demonstrating the effectiveness of consensus modeling and collective feature updates.**

Method	Metrics											
	Screen			Figure			Game			OShome		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
CCM w/o Consensus	0.5421	0.6122	16.821	0.6843	0.5417	14.263	0.6725	0.4831	13.174	0.7714	0.4026	11.352
Consensus → Uniform Average	0.4964	0.6475	17.742	0.6318	0.5824	15.135	0.6207	0.5246	14.028	0.7183	0.4471	12.267
Consensus → Max-vote Selection	0.5173	0.6281	17.105	0.6572	0.5603	14.641	0.6481	0.5012	13.516	0.7426	0.4218	11.743
Consensus → Single-mode Prediction	0.4758	0.6634	18.237	0.6104	0.6015	15.627	0.5983	0.5473	14.583	0.6947	0.4705	12.831
w/o Agreement-Guided Collective Update	0.5687	0.5846	16.148	0.7126	0.5138	13.572	0.7014	0.4527	12.465	0.7952	0.3743	10.628
Collective Update → ResBlock	0.4625	0.6753	18.642	0.5947	0.6182	16.034	0.5826	0.5621	15.017	0.6781	0.4924	13.305
Collective Update → SE Block	**0.4483**	**0.6871**	**19.056**	**0.5813**	**0.6307**	**16.418**	**0.5694**	**0.5786**	**15.424**	**0.6602**	**0.5106**	**13.742**
Collective Update → CBAM	0.4547	0.6815	18.823	0.5876	0.6244	16.205	0.5748	0.5703	15.208	0.6675	0.5013	13.516
CERL (Full)	**0.4172**	**0.7384**	**20.584**	**0.5657**	**0.7259**	**19.478**	**0.5414**	**0.6682**	**18.151**	**0.6235**	**0.6237**	**16.547**

SSIM, and PSNR without any brightness-specific adaptation or re-tuning. LPIPS exhibits an approximately U-shaped trend, whereas SSIM and PSNR follow inverted U-shaped trends, with the strongest reconstruction quality obtained near the normal brightness level. This behavior is consistent with the underlying imaging process. Under dim illumination, the reflected screen signal becomes weak relative to sensor noise, environmental illumination, and wall-induced interference, reducing the reliability of color and evidence. Under excessive illumination, dynamic-range compression, highlight clipping, and nonlinear camera responses suppress local contrast and distort the relative intensities of screen regions.

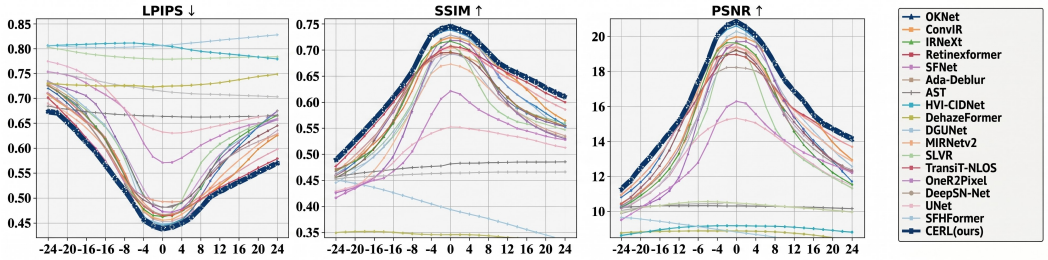

Figure S9: Qualitative comparison under decreasing projection brightness.

CERL retains the strongest reconstruction quality across the brightness sweep, with the best scores near the nominal input level. The darker examples in Figure S10 show the practical meaning of this advantage: the broad window region and its main boundary remain recognizable as the input weakens. OERL makes intensity cues available alongside normalized boundary responses and local contrast, allowing reconstruction to use complementary aspects of the light pattern. The bounded updates then regulate how these cues alter the estimate. This combination helps preserve screen organization as the observation brightness changes.

Based on this representation, ISTL progressively refines the screen state through bounded, gated updates anchored to both the current observation and a reference state. The anchoring mechanism prevents feature transport from deviating from the available optical evidence, while bounded updates avoid excessively amplifying weak signals in dark inputs or propagating large errors from saturated regions in bright inputs. CCM further compares local restoration hypotheses and reinforces corrections that are jointly supported by neighboring regions, thereby suppressing isolated brightness-induced artifacts and inconsistent regional predictions. Together, evidence consolidation, constrained inverse transport, and collective verification explain why CERL preserves structural and perceptual quality and degrades more smoothly as the input brightness deviates from the normal exposure level.

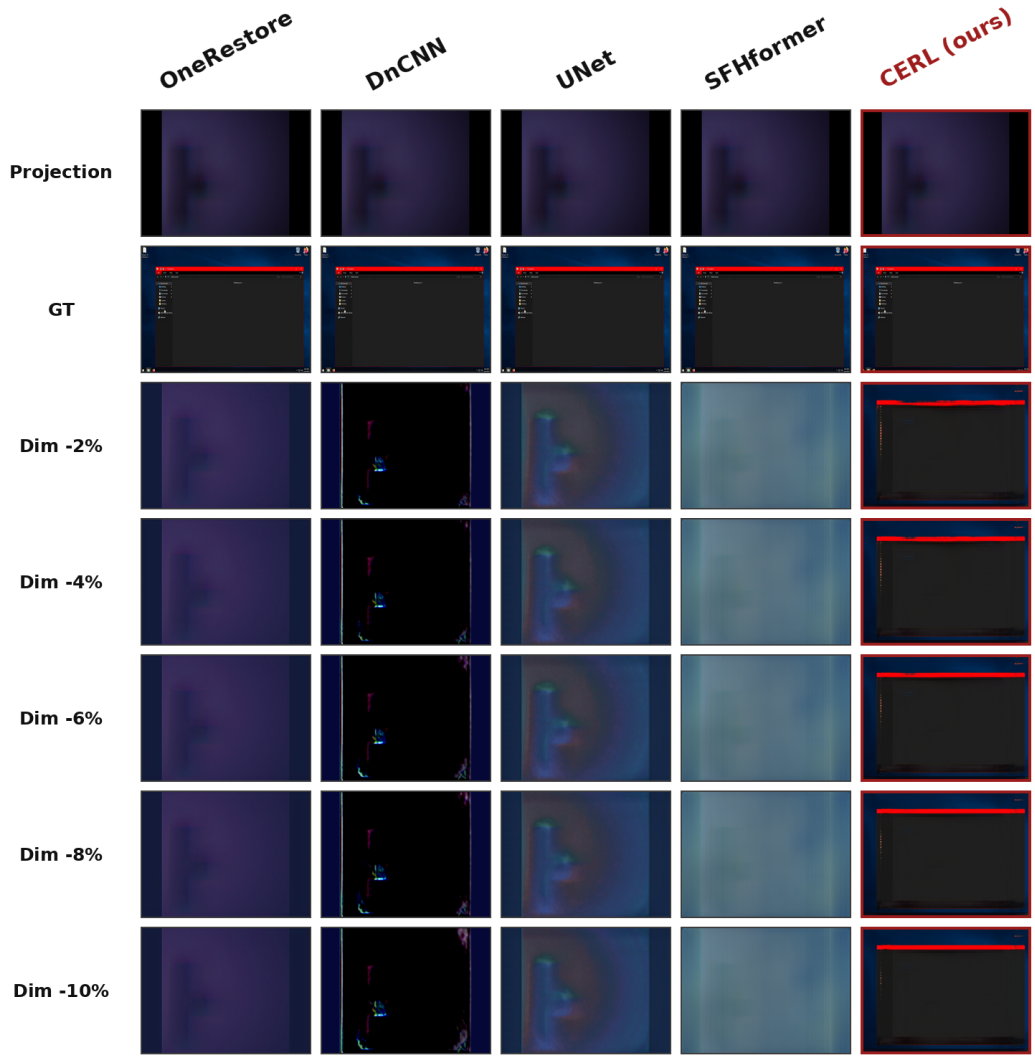

Figure S10: Qualitative comparison under decreasing projection brightness.

C.5 Results On Robustness to Noise

Wall reflection, occluded light transport, and finite-aperture imaging jointly constitute a strongly low-pass optical channel, substantially attenuating the high-frequency structures of the screen before they reach the camera. Additive noise further corrupts these weak reflection cues, causing conventional inverse mappings to amplify perturbations along directions associated with near-zero singular values. We add Gaussian noise to the clean image following the signal-to-noise ratio (SNR) definition, where the noise variance is computed according to the target SNR ranging from 20 dB to 40 dB, and the noisy image is clipped to a valid pixel range after noise injection.CERL maintains stable perceptual and structural reconstruction performance across all noise levels.

This robustness can be attributed to the fact that OERL first exploits the stability of reflected responses and local structural consistency to disentangle physically meaningful optical evidence from stochastic noise. ISTL then continuously corrects the state-transition trajectory using observational evidence during progressive inverse reconstruction, thereby suppressing the accumulation of errors induced by ill-conditioned inversion. Finally, CCM imposes collective consistency across local recovery states to eliminate spurious structures unsupported by shared evidence. CERL therefore avoids unconstrained high-frequency amplification and instead selectively restores attenuated structural components only

where sufficient physical evidence is available, achieving a favorable balance between detail recovery and noise suppression.

The quantitative noise-robustness results are summarized in Table 6.

Table 6: **Noise robustness comparison on the Screen test set under additive Gaussian noise. LPIPS is lower is better, while SSIM and PSNR are higher is better.**

Rank	Method	Clean			SNR 20 dB			SNR 25 dB			SNR 30 dB			SNR 35 dB			SNR 40 dB		
		LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
1	**CERL (ours)**	**0.417**	**0.738**	**20.58**	**0.471**	**0.679**	**19.38**	**0.445**	**0.696**	**20.21**	**0.440**	**0.724**	**20.44**	**0.440**	**0.735**	**20.61**	**0.440**	**0.739**	**20.65**
2	OKNet	**0.445**	**0.729**	**20.38**	0.480	0.598	18.75	**0.449**	0.688	**20.13**	**0.444**	**0.720**	**20.36**	**0.444**	**0.731**	**20.53**	**0.444**	**0.735**	**20.57**
3	ConvIR	0.449	0.728	20.27	0.516	0.558	19.19	0.458	0.667	20.09	0.448	0.707	20.23	0.448	0.721	20.26	0.449	0.725	20.27
4	Ada-Deblur	0.464	0.707	19.43	**0.475**	0.673	18.92	0.468	**0.692**	19.30	0.465	0.700	19.40	0.464	0.704	19.42	0.464	0.705	19.43
5	SFNet	0.462	0.716	19.63	0.493	0.574	**19.30**	0.465	0.660	19.50	0.460	0.696	19.58	0.461	0.709	19.60	0.461	0.713	19.61
6	Uformer	0.464	0.707	19.20	0.527	0.586	18.85	0.482	0.662	19.10	0.468	0.692	19.18	0.465	0.701	19.20	0.464	0.704	19.20
7	AST	0.467	0.705	19.39	0.522	0.573	18.97	0.481	0.654	19.29	0.470	0.686	19.36	0.467	0.698	19.38	0.467	0.702	19.39
8	DehazeFormer	0.479	0.693	19.37	0.496	0.665	18.83	0.483	0.684	19.24	0.480	0.690	19.32	0.479	0.692	19.34	0.479	0.692	19.35
9	IRNeXt	0.454	0.723	19.96	0.583	0.509	16.46	0.483	0.647	18.71	0.456	0.701	19.80	0.453	0.716	19.94	0.454	0.720	19.96
10	DGUNet	0.482	0.696	18.96	0.512	0.582	18.77	0.488	0.652	18.90	0.482	0.680	18.94	0.481	0.690	18.96	0.481	0.693	18.96
11	Retinexformer	0.455	0.673	19.57	0.570	0.504	15.51	0.489	0.616	18.42	0.470	0.645	18.99	0.460	0.659	19.31	0.456	0.666	19.45
12	MIRNetv2	0.492	0.692	18.24	0.493	**0.675**	18.04	0.492	0.683	18.11	0.492	0.687	18.18	0.492	0.689	18.22	0.492	0.690	18.23
13	SLVR	0.571	0.620	16.29	0.671	0.455	16.14	0.606	0.549	16.25	0.578	0.593	16.28	0.572	0.610	16.29	0.571	0.616	16.29
14	TransiT-NLOS	0.633	0.552	15.34	0.635	0.541	15.22	0.632	0.547	15.30	0.632	0.550	15.32	0.632	0.551	15.33	0.632	0.551	15.33
15	pix2pix	0.664	0.482	10.32	0.683	0.436	9.68	0.683	0.447	9.71	0.685	0.453	9.88	0.674	0.466	10.08	0.668	0.474	10.20
16	HVI-CIDNet	0.472	0.718	19.69	0.759	0.348	12.16	0.719	0.403	12.26	0.704	0.429	12.31	0.699	0.442	12.36	0.693	0.453	12.44
17	DnCNN	0.724	0.347	8.88	0.769	0.263	8.81	0.739	0.311	8.88	0.726	0.334	8.89	0.725	0.342	8.89	0.725	0.345	8.89
18	OneRestore	0.714	0.464	10.41	0.801	0.344	10.44	0.751	0.411	10.43	0.724	0.444	10.42	0.717	0.457	10.42	0.715	0.461	10.41
19	DeepSN-Net	0.779	0.295	10.49	0.867	0.219	10.34	0.818	0.264	10.41	0.794	0.284	10.45	0.785	0.291	10.47	0.781	0.294	10.48
20	UNet	0.806	0.240	9.18	0.825	0.213	9.29	0.810	0.228	9.22	0.805	0.235	9.19	0.805	0.237	9.18	0.806	0.239	9.18
21	SFHformer	0.807	0.394	8.82	0.844	0.344	8.80	0.823	0.376	8.81	0.813	0.388	8.81	0.809	0.392	8.81	0.807	0.393	8.82

Figure S11 presents qualitative comparisons under different noise levels. The wall observation is not a degraded copy of the screen image but the result of occlusion geometry and diffuse reflective transport. This transport strongly attenuates high-frequency information and mixes the screen boundary, textual content, and interface layout into a smooth, low-frequency shadow pattern. Additive noise further weakens the already limited correspondence between local wall intensities and screen structures, making direct inversion highly ill-posed, particularly at low SNRs.

The comparison reveals two characteristic failure modes caused by the mismatch between the wall-observation domain and the target screen-image domain. The first is noise amplification. HVI-CIDNet treats noise-corrupted intensity fluctuations as recoverable details, transforming weak measurement noise into pronounced textures and spatially irregular artifacts. Although SFHformer suppresses some high-frequency noise through spatial aggregation, the same operation disperses the already weak boundary cues needed to distinguish the screen-support region from its surroundings. Consequently, its outputs remain spatially blurred and lack a stable screen structure.

The second failure mode is projection-pattern retention. OneRestore and Pixel2Pixel largely preserve the smooth, low-frequency shadow profile of the input rather than reconstructing the latent screen content. Diffuse reflection removes or mixes much of the high-frequency screen information, so a generic restoration prior cannot recover the screen state simply by enhancing local contrast or suppressing noise. Without explicit constraints on the inverse transport from wall reflections to screen content, these methods remain confined to the observation domain. Thus, conventional restoration approaches tend either to amplify unreliable fluctuations or to over-smooth the limited structural evidence, preventing a stable inverse mapping.

In contrast, CERL formulates reconstruction as an evidence-constrained inverse process. Even at 20 dB SNR, it recovers an approximately complete screen-support region and a recognizable rectangular boundary. This result indicates that CERL does not infer the screen from isolated intensity fluctuations; instead, it exploits spatial cues that receive consistent support across the wall observation. As the SNR increases, the available optical evidence becomes more reliable, leading to a more uniform reconstructed background, clearer screen boundaries, and more stable internal responses.

This robustness originates first from OERL, which consolidates complementary cues that repeatedly support the same latent structure while suppressing features dominated by measurement noise or wall interference. ISTL subsequently transports this evidence from the wall-observation state toward the screen state through a sequence of bounded intermediate updates. By progressively refining edge,

chromatic, and local structural representations, ISTL avoids the abrupt noise amplification that can occur in single-step inversion and keeps the recovery trajectory tied to the measured evidence.

CCM provides an additional consistency constraint by examining whether local recovery hypotheses collectively support the same screen configuration. Regional predictions that agree on the screen boundary and spatial layout are reinforced, whereas isolated artifacts and mutually conflicting corrections are suppressed. Through evidence consolidation, progressive inverse transport, and collective consistency, CERL avoids both noise-induced texture hallucination and low-frequency projection retention, enabling a structurally coherent reconstruction from weak and noise-contaminated wall reflections.

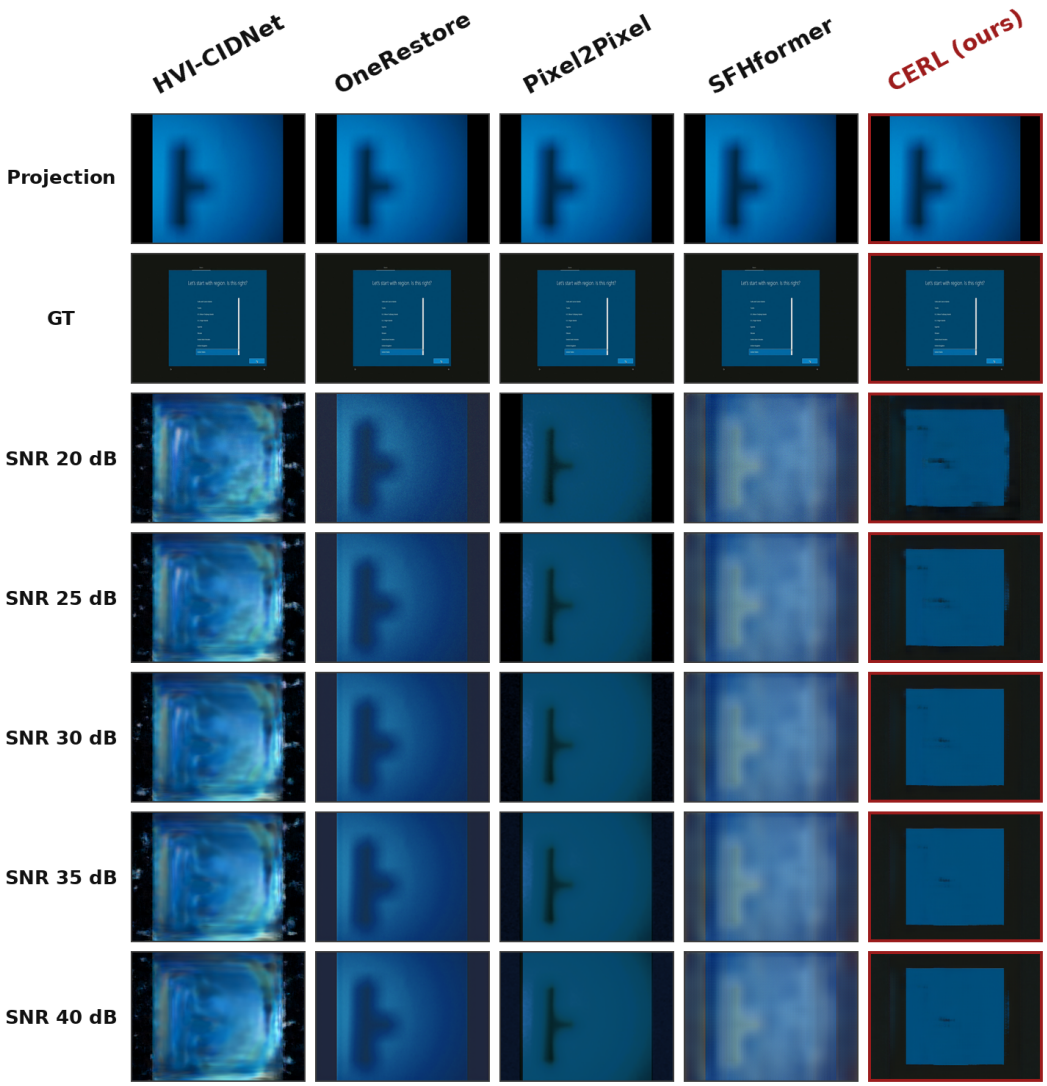

Figure S11: Qualitative comparison under additive Gaussian noise at different SNR levels. CERL consistently recovers the screen region and its geometric structure while suppressing noise-induced artifacts, whereas competing methods exhibit severe distortion, excessive smoothing, or residual projection patterns.

C.6 Reconstruction Performance Across Different Materials

To evaluate robustness to marble-like wall appearance, we introduce controlled surface variations characterized by multi-scale intensity fluctuations, irregular bright and dark veins, mild chromatic

shifts, and broad polished-surface responses. Unlike spatially independent noise, these patterns modify both local contrast and global irradiance distribution. In particular, marble veins may produce strong edge responses that resemble screen boundaries or internal structures, while nonuniform reflectance and surface highlights can distort the low-frequency illumination profile used to localize the screen region. Marble interference therefore challenges a reconstruction method to distinguish screen-derived evidence from structured variations introduced by the reflecting surface.

As the marble-pattern strength increases, surface-induced veins and reflectance variations become more strongly entangled with the already weak screen evidence. Conventional methods may interpret these wall patterns as target structures or suppress them together with valid screen cues, resulting in vein leakage, local dark artifacts, blurred boundaries, and spatial deformation.Structured surface patterns introduce competing boundaries into the wall observation. CERL preserves the screen region and its principal layout as these patterns become stronger, showing the value of combining boundary responses with complementary luminance and contrast cues.

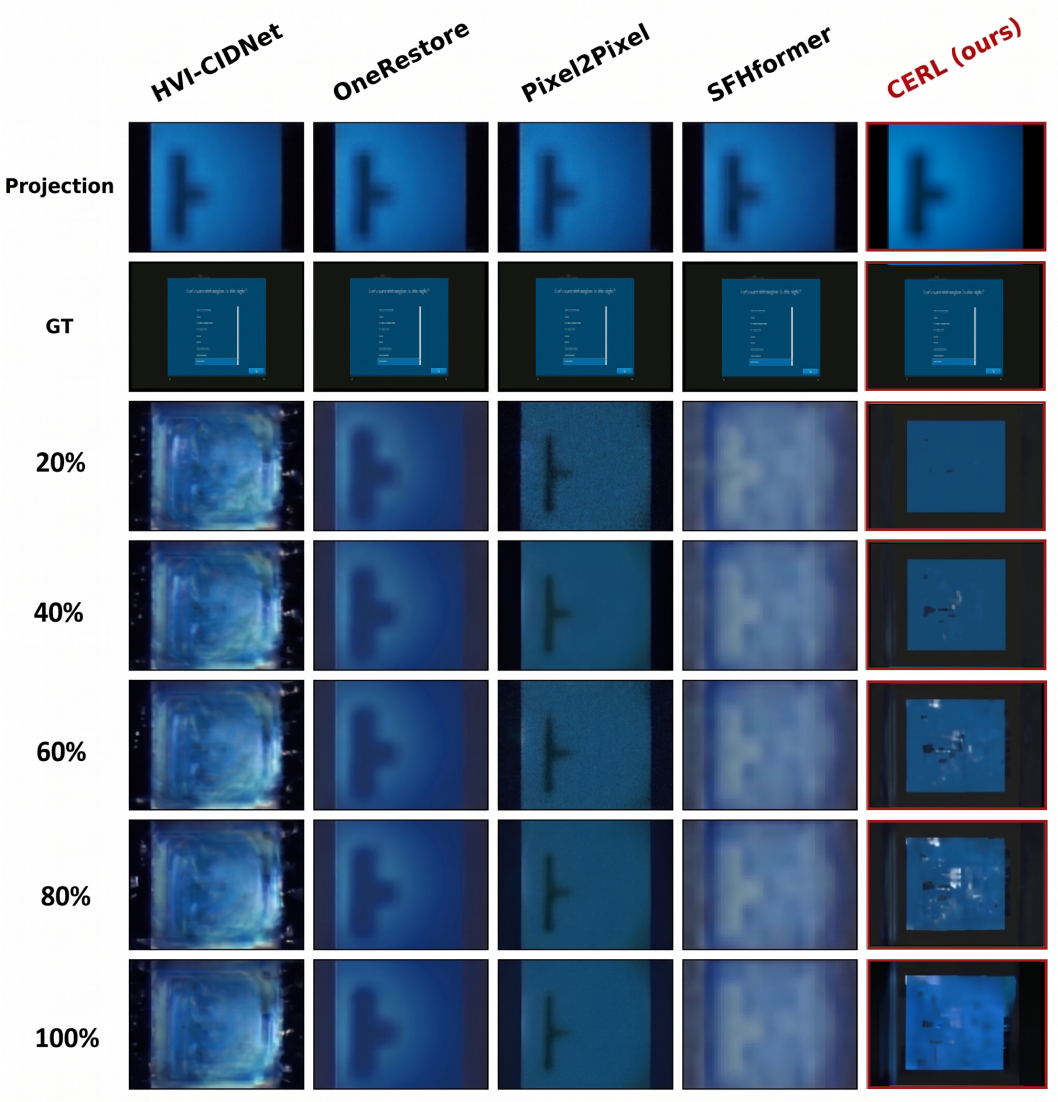

Figure S12: Marble wall appearance.

To evaluate robustness to fiberized wall appearance, we introduce controlled surface variations characterized by periodic warp-and-weft ridges, cross-thread patterns, mild chromatic shifts, and broad surface sheen. Unlike random noise, these directional textures generate repeated edge and contrast responses that can overlap with the weak structural cues carried by wall reflections. The

woven patterns may therefore be misidentified as screen boundaries or internal content, while the nonuniform reflectance and surface sheen can alter the luminance distribution used to locate the screen region.

As fiberization becomes stronger, the periodic ridges increasingly compete with genuine screen evidence and can be propagated into the reconstruction as false edges, repeated textures, or block-like artifacts. Methods that aggressively suppress these patterns may simultaneously remove weak but valid screen cues, leading to blurred boundaries and incomplete content. CERL addresses this interference through its coordinated evidence-recovery chain: OERL consolidates complementary screen-related cues while reducing the influence of isolated fiber responses; ISTL constrains progressive feature updates using the current observation and reference state, preventing repetitive textures from redirecting the recovery trajectory; and CCM suppresses regional corrections that are incompatible with the overall screen structure. Consequently, CERL degrades more gradually and retains clearer, more spatially coherent screen structures as fiber-pattern interference increases.

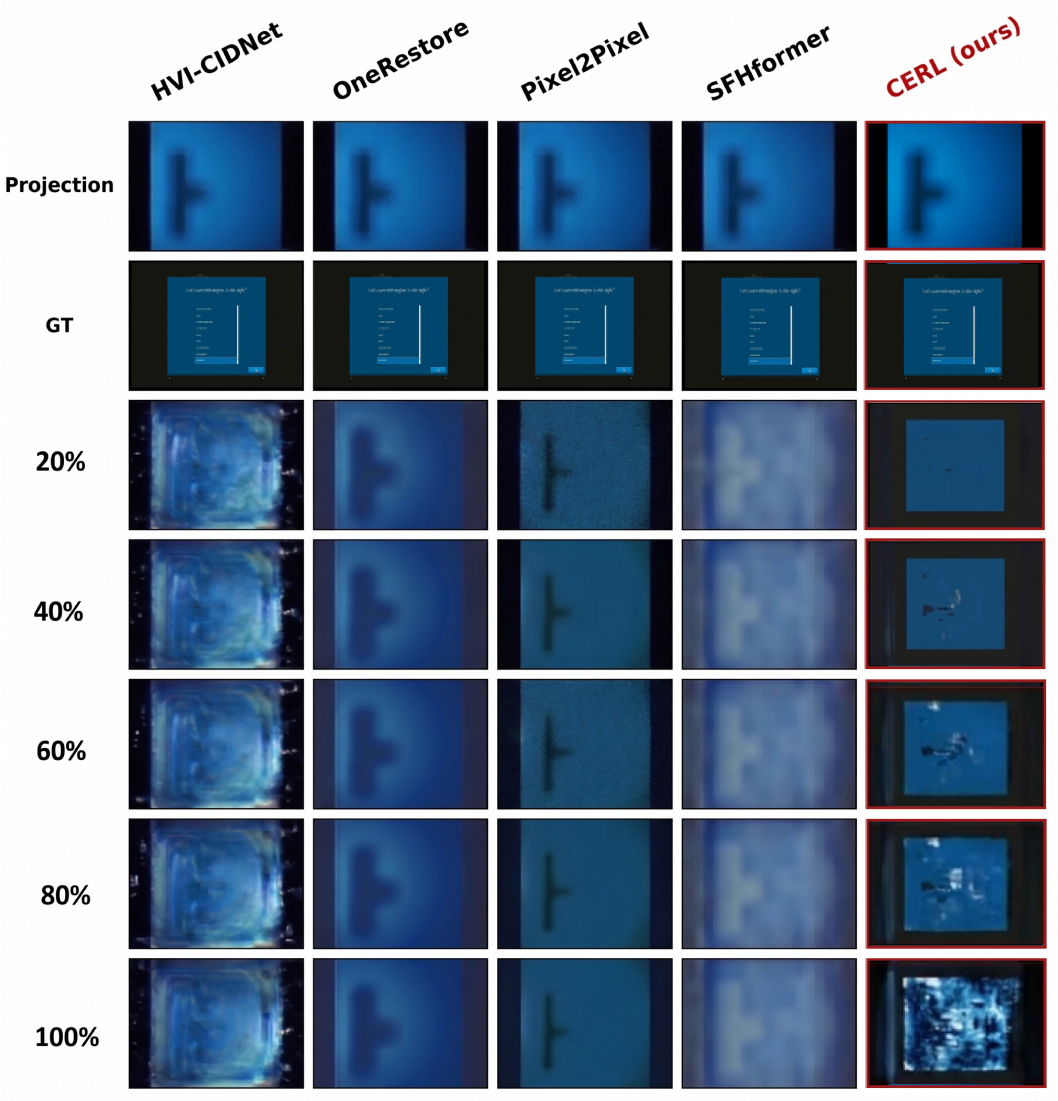

Figure S13: Fibrous wall appearance.

C.7 Local Occlusion Analysis

Localized occlusion reveals how different wall regions contribute to hidden-screen reconstruction. As shown in the 3×3 grid in Figure S14, equal-sized masks placed at different locations produce clearly location-dependent recovery errors. Occlusions over the central and right-central regions lead to extensive corruption of the screen support, boundary, and interface layout. Several peripheral occlusions preserve a larger portion of the target structure. Since the mask size remains fixed across all cases, these differences reflect the distinct evidential value of each wall region and confirm that CERL conditions its prediction on the spatial content of the observation.

This sensitivity pattern arises from diffuse optical transport, through which screen emissions overlap and form spatially mixed, low-frequency responses on the wall. Wall regions containing pronounced illumination transitions and shadow gradients carry stronger constraints on the screen extent, foreground–background separation, and internal organization. Removing these regions weakens several coupled constraints simultaneously, leading to boundary displacement, missing content, and erroneous intensity propagation. Smoother peripheral regions contain more spatially redundant information that can be partially recovered from neighboring observations. The nine-grid results therefore demonstrate that CERL integrates wall evidence according to its spatial relevance and uses complementary measurements to maintain a coherent reconstruction.

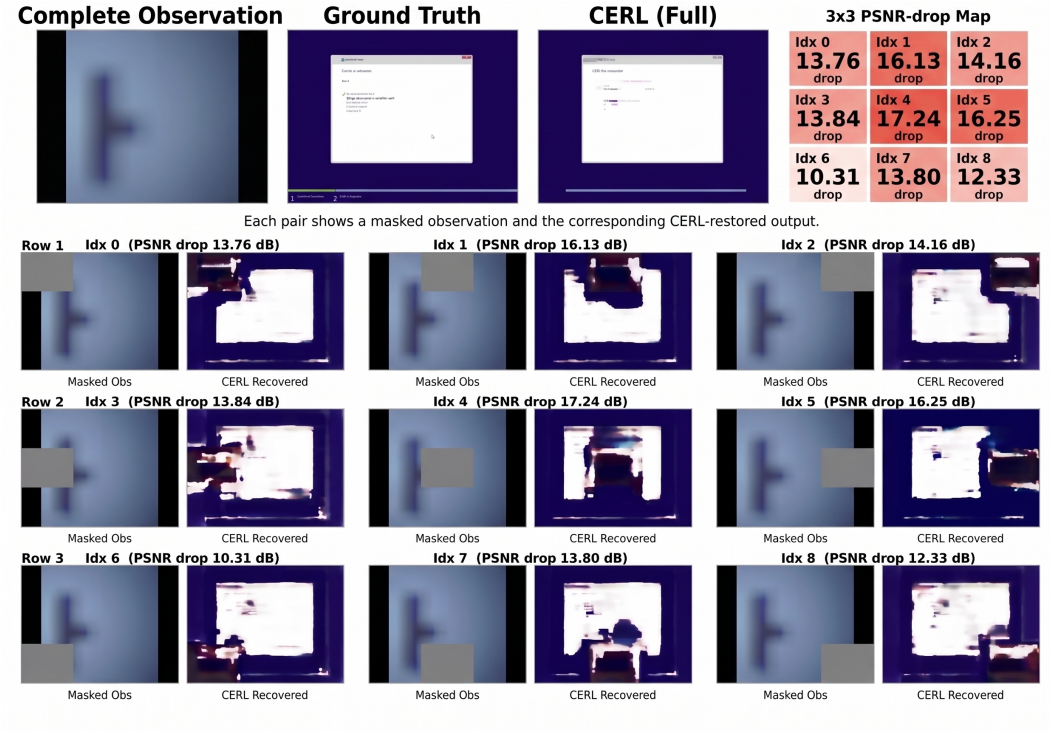

Figure S14: We mask each of the 3×3 spatial patches on the observation and measure PSNR drop. Larger PSNR drop indicates the patch carries critical information for screen content inversion.

C.8 Image Cropping Experiment

This experiment evaluates the robustness of CERL under progressively increasing right-side cropping. Starting from the complete observation, the cropping strength is gradually increased along the horizontal direction while the target image and evaluation protocol remain unchanged. This operation reduces the spatial coverage of the wall measurement and removes part of the diffuse optical evidence available for reconstruction. As the retained region becomes narrower, the observation contains fewer illumination gradients, boundary responses, and local contrast variations, thereby increasing the uncertainty of the inverse recovery process.

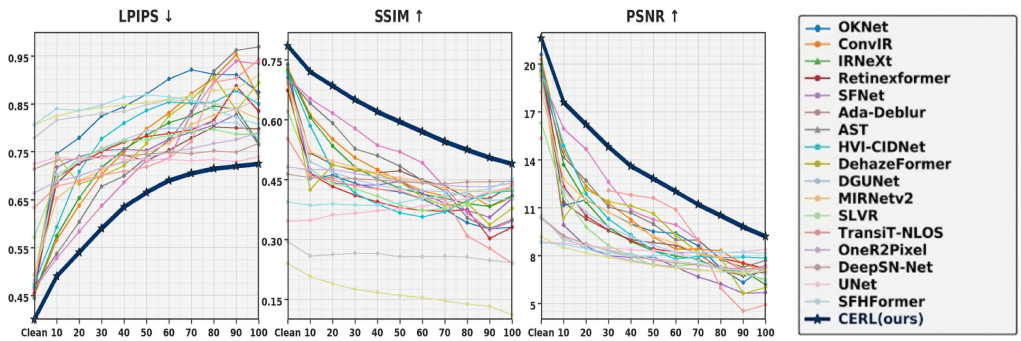

Figure S15: CERL exhibits stronger robustness under limited observational information.

The effect of cropping extends beyond the direct loss of pixels. Under diffuse optical transport, each observed wall region contains spatially mixed contributions from multiple screen regions. Removing the right side therefore weakens a distributed set of constraints associated with the screen extent, intensity distribution, and internal layout. Stronger cropping also changes the balance between the retained low-frequency illumination field and the remaining structural transitions. These changes make it increasingly difficult to distinguish screen-related evidence from wall appearance and measurement noise, leading to accumulated errors in boundary localization, color recovery, and structural reconstruction.

The metric curves in Figure S15 show an overall increase in LPIPS and corresponding decreases in SSIM and PSNR as cropping becomes more severe. Many competing methods exhibit rapid deterioration or pronounced fluctuations because their predictions are sensitive to the loss of particular spatial cues. CERL maintains the most favorable overall trajectory across the tested range, with smoother changes in all three metrics. Its stable performance indicates that the reconstruction draws on complementary evidence distributed throughout the visible observation. When part of the measurement is removed, cues retained in other regions continue to constrain the screen support and its principal structures.

C.9 Comparison of Visual Representations under a Frozen U-Net.

We compare how different visual representations support hidden-screen reconstruction through a shared frozen U-Net.To isolate the intrinsic contribution of the representation, we feed different visual representations into the same frozen U-Net while keeping the decoder, reconstruction target, and evaluation protocol fixed. The compared features cover geometric information from Depth Anything, object-level structure from SAM, semantic alignment from SigLIP, and diffusion correspondence from DIFT. Under this controlled setting, any performance difference primarily reflects how effectively each representation organizes information relevant to recovering hidden screen content.

Edges are extracted from reconstructions and ground truth with a unified edge detector to calculate Edge F1, Edge leakage, and Target-edge correlation. Edge F1 reports the F1 score between predicted and ground-truth edge maps. Target-edge correlation evaluates the correlation between reconstructed target edge profiles and ground-truth edges.

CERL exhibits a recovery-oriented, boundary-sensitive, and perceptually informative representation. As reported in Table 7, it achieves the highest Edge F1 and the lowest LPIPS, indicating that its representation retains weak target-related transitions while preserving information aligned with the perceptual appearance of the screen. This characteristic directly matches the requirements of wall-reflection inversion: diffuse transport attenuates screen boundaries, mixes local structures, and leaves only weak luminance and contrast variations in the observation. CERL organizes these remaining optical cues into a form that a fixed decoder can readily translate into recognizable screen structures.

CERL remains competitive in PSNR and SSIM, while its higher edge leakage reveals a clear sensitivity trade-off. The representation responds strongly to attenuated structural evidence, which improves target-boundary recovery but can also preserve nuisance transitions introduced by wall texture or measurement interference. Overall, the comparison characterizes CERL as a task-specific

Table 7: **Representation comparison using the same frozen U-Net.**

Representation	Source	Encoder input	Metrics				
			PSNR↑	SSIM↑	LPIPS↓	Edge F1↑	Edge leakage↓
Depth Anything	CVPR 2024	518×518	9.1532	0.23796	0.80714	0.18145	0.07318
SAM	ICCV 2023	1024×1024	**9.1856**	**0.23979**	**0.80602**	**0.18323**	**0.06989**
SigLIP	ICCV 2023	224×224	9.1759	0.23805	0.81124	0.18166	0.07102
DIFT	NeurIPS 2023	512×384	9.1805	0.23890	0.80668	0.18298	0.07242
CERL	Ours	512×384	**9.1979**	**0.23995**	**0.80138**	**0.18717**	**0.05254**

Table 8: **Learning Objective Comparison with a Frozen U-Net.**

Learning objective	Metrics				
	LPIPS↓	SSIM↑	PSNR↑	Edge leakage↓	Target-edge corr.↑
L1	0.781426	0.436496	10.301	0.202341	0.004563
MSE	0.796660	0.429946	10.028	0.394197	0.001167
Huber	0.782248	0.423970	10.193	0.316470	0.000319
Charbonnier	**0.780001**	0.427233	10.182	0.312491	0.001714
VGG Perceptual	0.796347	0.439590	10.054	0.277613	-0.000172
LPIPS	0.782651	0.354509	10.056	**0.179855**	0.004309
SSIM	0.797622	0.444033	10.090	0.223067	0.005817
MS-SSIM	0.797987	**0.470518**	**10.561**	0.317112	0.005385
Gradient	0.781542	0.434880	10.280	0.221148	0.001611
Laplacian	0.785673	0.433651	10.217	0.276953	-0.000536
Total Variation	0.781890	0.433074	10.301	0.196112	**0.006764**
Frequency	0.783294	0.423002	10.373	0.283324	0.004063
Wavelet	0.785307	0.441745	10.356	0.248706	0.004682
CERL (Ours)	**0.738076**	**0.473849**	**10.600**	**0.165930**	**0.018571**

optical-evidence representation that emphasizes recoverable structure and perceptual relevance, with particularly strong capability for exposing weak screen boundaries to downstream reconstruction.

C.10 Comparison of Visual Learnings under a Frozen U-Net.

We use the same U-shaped structure and compare it with different learning methods.The comparison covers pixel-domain objectives, perceptual objectives, structural-similarity objectives, gradient regularization, and frequency-domain constraints. This controlled setting directly measures how each learning objective organizes the weak and spatially mixed information carried by wall reflections.

Conventional objectives exhibit clear specialization. Pixel-wise losses preserve average intensity fidelity but provide limited guidance for separating screen-related structure from wall interference. Perceptual losses improve visual similarity while showing weaker structural consistency, whereas SSIM and MS-SSIM favor global organization at the cost of local perceptual fidelity. Gradient-, Laplacian-, and frequency-based objectives strengthen high-frequency responses, yet these responses contain both attenuated screen boundaries and nuisance edges introduced by the wall. Their supervision therefore lacks a reliable mechanism for determining which structures should be recovered and which should be suppressed.

CERL achieves the best results across LPIPS, SSIM, PSNR, edge leakage, and target-edge correlation, demonstrating an evidence-selective and structurally aligned learning characteristic. The substantial increase in target-edge correlation, together with the lowest edge leakage, is particularly important: CERL learns to reinforce boundary responses that agree with the target screen while suppressing unrelated transitions originating from wall texture and measurement interference. Its simultaneous

advantage in perceptual, structural, and pixel-level metrics further shows that the learning process coordinates complementary constraints instead of optimizing a single image property in isolation. As summarized in Table 8, CERL provides a task-oriented learning strategy that converts weak wall observations into target-consistent evidence, balances global appearance with local structure, and directs the fixed reconstruction network toward stable hidden-screen recovery.

C.11 Robustness experiment of camera position movement

This experiment introduces continuous lateral camera position offsets ranging from -12% to +12% to judge geometric imaging mismatch caused by camera alignment errors in real-world scenarios, and systematically quantifies the robustness of different reconstruction algorithms against acquisition perturbations. From the overall trend of the curves, all algorithms achieve their peak performance at the zero-offset ideal position; as the offset magnitude increases, the LPIPS metric rises monotonically, while SSIM and PSNR decrease synchronously, showing a consistent pattern of performance degradation. The physical essence of this phenomenon is that lateral camera offset disrupts the spatial correspondence of non-line-of-sight light paths and invalidates photon trajectory constraints, which in turn leads to inaccurate feature matching, loss of high-frequency details, and artifact proliferation during the reconstruction process.

Horizontal comparison across baseline algorithms reveals clear differentiation in robustness among different technical routes. Traditional convolutional architectures such as UNet and DnCNN present relatively low overall performance but gentle performance degradation, while some Transformer-based methods deliver outstanding performance under ideal conditions but suffer from rapid performance drop as the offset increases, reflecting that such models rely excessively on feature learning under ideal geometric alignment and lack sufficient generalization ability against physical acquisition perturbations.

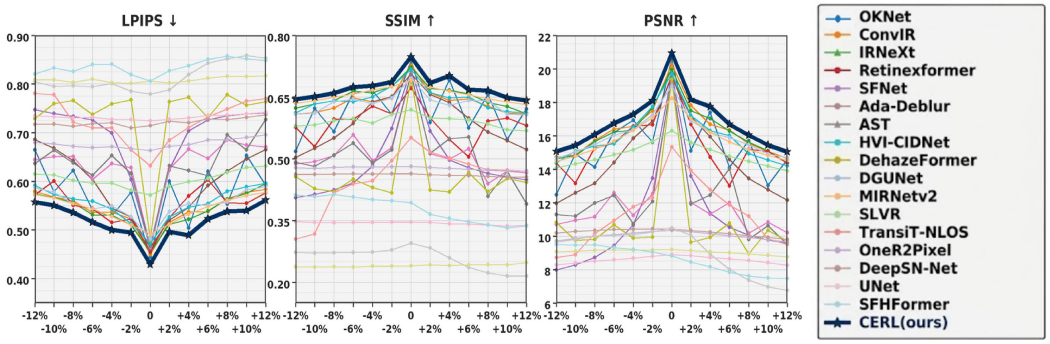

Figure S16: Robustness evaluation of non-line-of-sight reconstruction algorithms under lateral camera position offsets, where the horizontal axis denotes the relative offset ratio from the reference position.

The proposed CERL method maintains optimal performance across the entire offset range, demonstrating significantly stronger tolerance to pose perturbations than baseline methods. Under the zero-offset reference condition, CERL substantially outperforms all competing algorithms in PSNR and SSIM, while achieving the lowest LPIPS perceptual distortion globally. When the offset expands to the extreme condition of ±12%, the performance attenuation of CERL is noticeably milder than that of other methods, and it consistently holds the upper bound of overall performance. This advantage stems from the explicit modeling of underlying physical priors by the optical evidence representation learning module in the CERL framework, combined with the global correction effect of the consistency constraint module , which enables it to effectively extract reliable optical features under geometric mismatch and suppress reconstruction distortion caused by pose deviations.

C.12 Robustness experiment of camera jitter

The lateral displacement is synthesized by shifting the image within a range of $\pm 12\%$ of the image width, where the percentage is relative to the full image dimension.With the progressive intensification of camera shake, the reconstruction performance of all types of image restoration methods declines to varying extents. This trend verifies that the aggravation of motion blur continuously erodes the

valid information dimensions of images, and imposes incrementally growing challenges on the feature decoupling capability and prior generalization capacity of restoration algorithms. Distinct divergences exist in the performance decay rate and robustness boundary across different methods. Some algorithms undergo rapid performance degradation in the high-intensity shake interval, which reveals notable limitations of their restoration modeling mechanisms in adapting to extreme blur interference, making them incompetent to handle severe motion degradation.

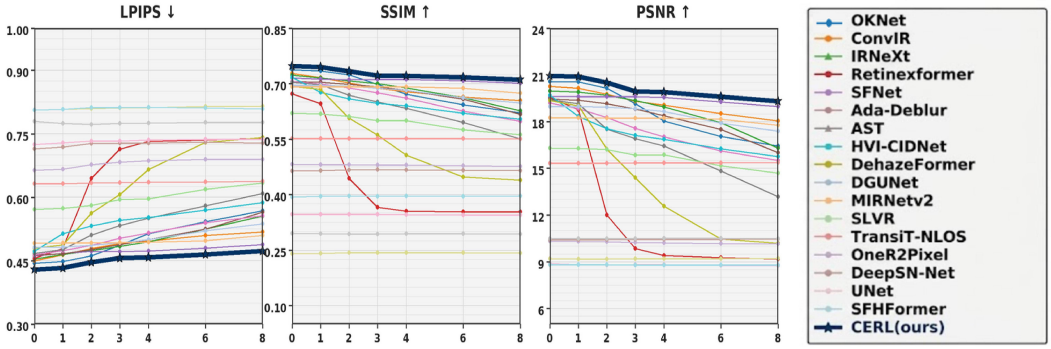

Figure S17: Robustness comparison curves of various image restoration methods under increasing camera shake intensity. The horizontal axis represents the gradient of shake radius.

The proposed CERL method sustains performance leadership throughout the whole range of shake intensity, and its magnitude of performance decay along with rising blur severity is remarkably lower than that of peer comparison algorithms, manifesting outstanding shake robustness and scene adaptability. Such performance advantage stems from the optical evidence representation learning framework and iterative optimization strategy devised in this method. It can stably mine and reuse valid restoration cues during the continuous deepening of blur, alleviate the cumulative effect of information loss induced by motion degradation, and deliver a more stable technical solution for high-quality image restoration under strong interference conditions.