SCENE-PRESERVING COUNTERFACTUAL DIFFUSION FOR LOW-LIGHT VIDEO RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-frame low-light restoration typically operates over predefined candidate temporal dependencies, yet relevant degradation factors and propagation paths may vary across targets as imaging conditions change. We formulate restoration as a scene-preserving counterfactual problem, asking how the target would appear if relevant degradation conditions shifted toward normal light while the factual scene remained fixed. We propose **Scene-Preserving Counterfactual Diffusion (SPCD)**, which infers a target-specific intervention structure. Using controlled responses and diffusion-score differences, SPCD selects a minimal set of degradation factors and propagation paths to guide reverse diffusion toward normal light. Experiments on real-world and controlled benchmarks across multiple dynamic regimes demonstrate improved quality. Code will be released after acceptance.

1 INTRODUCTION

Multi-frame low-light restoration aims to recover a normal-light target frame from neighboring observations. Existing methods have made substantial progress through temporal alignment, aggregation, recurrent propagation, and attention (Wang et al., 2021; Peng et al., 2022; Fu et al., 2023a). However, even input-adaptive weighting generally operates within a candidate temporal dependency structure prescribed by the input window or model architecture and shared across restoration targets.

This shared structure can be restrictive when imaging conditions vary over time. Illumination attenuation, exposure or gain, sensor noise, and camera response may have persistent or interacting effects across frames (Chen et al., 2025). Relevant degradation factors and temporal paths may therefore vary across targets (Zhao et al., 2024; Ghasemabadi et al., 2024). A limited temporal structure may miss earlier effects, while a dense one may include irrelevant conditions. The challenge is to identify the relevant factors and temporal paths for each target, beyond simply aggregating more frames.

A deeper limitation is *observational underdetermination*. Conventional restoration learns a mapping from a low-light window to its normal-light target (Wang et al., 2021; Fu et al., 2023a; Liu et al., 2023; Zhang et al., 2024; Wang et al., 2025). Even accurate prediction does not determine how the same factual scene would appear if a degradation factor changed: distinct latent mechanisms may yield the same input–target distribution yet respond differently to the same intervention. Adaptive weighting can improve prediction, but does not by itself identify which factors to change or how their effects should propagate.

We formulate dynamic low-light restoration as an instance-specific counterfactual problem under a finite-horizon causal model, shifting selected degradation conditions toward normal light while holding the factual scene fixed. We seek minimal interventions and target-sufficient propagation paths, rather than a unique physical graph. Our **Scene-Preserving Counterfactual Diffusion (SPCD)** anchors the scene through diffusion inversion and uses intervention responses to select relevant factors and verify sparse propagation paths for reverse diffusion. Our main contributions are threefold:

- We formulate dynamic low-light restoration as an instance-specific counterfactual problem and establish the observational underdetermination of its intervention responses.
- We characterize scene-preserving restoration through a minimal intervention set and a target-sufficient propagation structure.

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.

- We develop SPCD to screen factors via diffusion-score responses and verify inferred sparse propagation paths, guiding reverse diffusion toward normal light.

2 Causal Foundations of Dynamic Low-Light Restoration

As illustrated in Figure 1, we formulate dynamic low-light restoration as an inference problem under a finite-horizon causal model relating normal-light reference frames, degradation-related source conditions, and observed low-light frames. Video-instance indices are omitted for clarity.

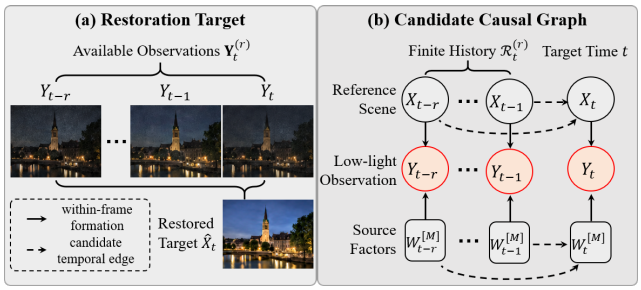

Figure 1: Dynamic low-light restoration as a causal graph.

2.1 Problem Setup

Consider a video of length T with M latent source conditions, and define $[T] = \{1, \dots, T\}$ and $[M] = \{1, \dots, M\}$. At time $t \in [T]$, X_t is the canonical normal-light frame of the factual scene, Y_t is its low-light observation, and $W_t = (W_t^1, \dots, W_t^M)$ collects source conditions such as effective illumination attenuation, camera gain and shot/read-noise statistics, etc. Here, "normal light" refers to the target domain defined by the reference data, not a universal physical illumination level. The sequence $\mathbf{X} = (X_1, \dots, X_T)$ forms the reference scene trajectory and may evolve because of object or camera motion.

For a finite modeling horizon r, let history $\mathcal{R}_t^{(r)} = \{i \in [T] : t - r \le i < t\}$ and input window $\mathcal{I}_t^{(r)} = \mathcal{R}_t^{(r)} \cup \{t\}$. The observed input window is $\mathbf{Y}_t^{(r)} = (Y_i)_{i \in \mathcal{I}_t^{(r)}}$. At inference, only $\mathbf{Y}_t^{(r)}$ is observed; $\{X_i, W_i : i \in \mathcal{I}_t^{(r)}\}$ remain latent, and no future frame is used. The restoration task is to recover the factual scene state at time t: the restored frame $\widehat{X}_t$ should follow the data-generating conditional $p^\star(\cdot \mid \mathbf{Y}_t^{(r)})$, rather than merely appear plausible in the target domain.

2.2 Latent Dynamics and Observation Model

The histories $\mathbf{X}_{\mathcal{R}_t^{(r)}} = (X_i)_{i \in \mathcal{R}_t^{(r)}}$ and $\mathbf{W}_{\mathcal{R}_t^{(r)}} = (W_i)_{i \in \mathcal{R}_t^{(r)}}$ collect the preceding scene states and source conditions. The term dynamics refers to finite-memory state evolution across video steps. Their dynamics and the resulting observation are modeled as

$$X_t = f_X(\mathbf{X}_{\mathcal{R}_t^{(r)}}, U_t^X), \qquad W_t = f_W(\mathbf{W}_{\mathcal{R}_t^{(r)}}, U_t^W), \quad \text{and} \quad Y_t = f_Y(X_t, W_t, U_t^Y). \tag{1}$$

The history arguments list possible dependencies; each mechanism may use only a subset of them. Here, f_X, f_W, and f_Y denote the scene-transition, source-condition-transition, and image-formation mechanisms respectively, and U_t^X, U_t^W, and U_t^Y are their associated exogenous variables.

Write $\mathbf{W} = (W_1, \dots, W_T)$ and $\mathbf{Y} = (Y_1, \dots, Y_T)$ for the source-condition and observation sequences. Together with the joint distribution of the exogenous variables, these mechanisms define a finite-horizon structural causal model $\mathcal{M}_r$. Assuming that the exogenous variables are independent across time and mechanisms, the data-generating distribution factorizes as

$$p^\star(\mathbf{X}, \mathbf{W}, \mathbf{Y}) = \prod_{t=1}^{T} p^\star(X_t \mid \mathbf{X}_{\mathcal{R}_t^{(r)}})\, p^\star(W_t \mid \mathbf{W}_{\mathcal{R}_t^{(r)}})\, p^\star(Y_t \mid X_t, W_t). \tag{2}$$

When $\mathcal{R}_t^{(r)} = \varnothing$, the first two factors reduce to initial-state distributions. Interventions act only on $\mathbf{W}$, while the realized scene trajectory $\mathbf{X}$ is shared between the factual and counterfactual worlds.

2.3 Finite-Horizon Candidate Causal Graph

These possible dependencies define a finite-horizon candidate causal graph $\mathcal{G}_r = (\mathcal{V}, \mathcal{E}_r)$, where $\mathcal{V} = \{X_t, Y_t : t \in [T]\} \cup \{W_t^m : t \in [T], m \in [M]\}$. For $t \in [T]$ and $i \in \mathcal{R}_t^{(r)}$, $\mathcal{E}_r$ contains the

scene-evolution edge $X_i \to X_t$ and, for $m, n \in [M]$, the source-propagation edges $W_i^n \to W_t^m$. The observation model adds $X_t \to Y_t$ and $W_t^m \to Y_t$. All temporal edges point forward and all within-frame edges point to Y_t, so $\mathcal{G}_r$ is acyclic. Exogenous variables are omitted.

Only a fixed subset $\mathcal{E}^{\mathrm{str}} \subseteq \mathcal{E}_r$ is realized by the structural mechanisms. For target time t, $\mathcal{E}^W_{r,t} \subseteq \mathcal{E}_r$ is the set of source-propagation edges within $\mathcal{I}_t^{(r)}$. The target-relevant edges satisfy $\mathcal{E}_t^{\mathrm{rel}} \subseteq \mathcal{E}^{\mathrm{str}} \cap \mathcal{E}^W_{r,t} \subseteq \mathcal{E}^W_{r,t}$. $\mathcal{I}_t^{(r)}$ determines $\mathcal{E}^W_{r,t}$, while the observed frames guide the selection of $\mathcal{E}_t^{\mathrm{rel}}$; scene-evolution and observation edges are retained.

3 From Observational to Counterfactual Restoration

Under $\mathcal{G}_r$, finite-horizon counterfactual restoration fixes $\mathbf{X}$, shifts selected conditions in $\mathbf{W}$ toward normal light, and enforces structural invariance and minimality. We establish observational underdetermination, then formalize restoration through abduction, action, and prediction.

3.1 Why Observational Restoration Is Insufficient

Conventional restoration estimates $p^\star(X_t \mid \mathbf{Y}_t^{(r)})$, marginalizing over latent source conditions $\mathbf{W}$. This captures observational associations but leaves degradation causes and their temporal propagation underdetermined. In Figure 2, either an earlier illumination change W_{t-1}^{illum} or a current camera change W_t^{cam} can explain the observed window and yield a plausible restoration, yet only the former predicts a response to an intervention on earlier illumination. This requires an instance-specific counterfactual query conditioned on the observed window and anchored to its factual scene.

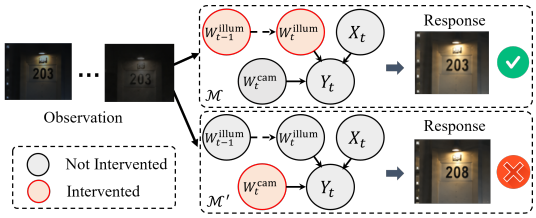

Figure 2: Observationally equivalent mechanisms may respond differently to the same intervention.

Let $\mathcal{C}_r$ be the class of SCMs compatible with $\mathcal{G}_r$ and $\mathcal{S}_t = \{W_i^m \in \mathcal{I}_t^{(r)}, m \in [M]\}$ the candidate intervention set. For $\mathcal{A} \subseteq \mathcal{S}_t$, define $\mathcal{M}[\mathcal{A}]$ by shifting the conditional mechanisms of $\mathcal{A}$ toward their normal-light counterparts while preserving temporal parent sets. Given $\mathbf{y}_t$, abduction yields the factual posterior $\mu_{t,\mathcal{M}} := p_{\mathcal{M}}(\mathbf{U}_t = \cdot \mid \mathbf{Y}_t = \mathbf{y}_t)$ over query-relevant exogenous variables $\mathbf{U}_t$. The response of $\mathcal{M}$ to $\mathcal{A}$ is

$$R_t^{\mathcal{M}}(\mathcal{A} \mid \mathbf{y}_t) := \mathrm{Law}_{\mathbf{U}_t}\left[Y_t^{\mathcal{M}[\mathcal{A}]}(\mathbf{U}_t)\right]. \tag{3}$$

The response depends on the factual exogenous posterior and intervened mechanisms, neither generally identifiable from the observational law alone.

Proposition (Observational underdetermination). Without interventional supervision or additional identifying assumptions, there may exist $\mathcal{M}, \mathcal{M}' \in \mathcal{C}_r$, $\mathcal{A} \subseteq \mathcal{S}_t$, and $\mathbf{y}_t^{(r)}$ in their common support such that

$$p_{\mathcal{M}}(X_t, \mathbf{Y}_t) = p_{\mathcal{M}'}(X_t, \mathbf{Y}_t), \qquad R_t^{\mathcal{M}}(\mathcal{A} \mid \mathbf{y}_t) \neq R_t^{\mathcal{M}'}(\mathcal{A} \mid \mathbf{y}_t). \tag{4}$$

A restoration rule based solely on the observed window cannot distinguish observationally equivalent mechanisms with different responses to the same intervention, even if it exactly fits $p^\star(X_t \mid \mathbf{Y}_t)$. See Appendix B for an explicit construction.

3.2 Scene Preservation and Structural Minimality

Without resolving the underdetermination in Eq. 4, the causal model imposes two requirements on target-specific restoration: (i) *Scene Preservation*: factual and counterfactual worlds share the same scene context; and (ii) *Structural Minimality*: $\mathcal{A}_t^\star \subseteq \mathcal{S}_t$ and $\mathcal{E}_t^{\mathrm{rel}} \subseteq \mathcal{E}^{\mathrm{str}} \cap \mathcal{E}^W_{r,t}$ are the smallest sets of source conditions and propagation paths that causally explain the target's observed degradation. We formalize both through abduction, action, and prediction.

Abduction. For *scene preservation*, abduction infers factual latent realizations consistent with the observed window, remaining fixed across counterfactual worlds. Conditioning on $\mathbf{Y}_t = \mathbf{y}_t$, define

$$\mu_t^\star(\mathbf{u}) = p_{\mathcal{M}_r^\star}(\mathbf{U}_t = \mathbf{u} \mid \mathbf{y}_t), \qquad \pi_t^\star(x) = p_{\mathcal{M}_r^\star}(X_t = x \mid \mathbf{y}_t), \tag{5}$$

where $\mathcal{M}_r^\star \in \mathcal{C}_r$ is the data-generating SCM inducing $p^\star$. Conditioned on the observed window, $\mu_t^\star$ and $\pi_t^\star$ describe the posteriors over latent realizations and the factual scene X_t, respectively, providing the factual basis for subsequent causal inference.

Action (minimal intervention). Assume the fully intervened response $R_t(\mathcal{S}_t)$ and the normal-light scene posterior $\pi_t^\star$ share the same image space. We enforce *structural minimality* by selecting the smallest intervention set $\mathcal{A}_t^\star$ whose response matches $\pi_t^\star$ within tolerance:

$$\mathcal{A}_t^\star \in \arg\min_{\mathcal{A} \subseteq \mathcal{S}_t} |\mathcal{A}| \quad \text{s.t.} \quad D_{\mathrm{JS}}\left(\pi_t^\star \parallel R_t(\mathcal{A})\right) \le \rho_{\mathrm{deg}}. \tag{6}$$

Here, $\rho_{\mathrm{deg}} \ge 0$ specifies the discrepancy tolerance between the counterfactual response $R_t(\mathcal{A})$ and the factual scene posterior $\pi_t^\star$.

Prediction. Given $\mathcal{A}_t^\star$, we select the sparsest source-condition subgraph whose target response matches the full-graph response within tolerance:

$$\mathcal{G}_t^\star \in \arg\min_{\mathcal{H}:\, \mathcal{E}_{\mathcal{H}} \subseteq \mathcal{E}_{r,t}^W} |\mathcal{E}_{\mathcal{H}}| \quad \text{s.t.} \quad D_{\mathrm{KL}}\left(R_t(\mathcal{A}_t^\star) \parallel R_t(\mathcal{A}_t^\star, \mathcal{H})\right) \le \rho_{\mathrm{prop}}, \tag{7}$$

where $\rho_{\mathrm{prop}} \ge 0$ controls propagation sufficiency and accommodates the relaxation from the unknown support $\mathcal{E}^{\mathrm{str}}$ to candidate edges $\mathcal{E}_{r,t}^W$. The selected edges $\mathcal{E}_t^{\mathrm{rel}} = \mathcal{E}_{\mathcal{G}_t^\star}$ form a minimal target-sufficient propagation structure and need not recover a unique physical graph.

In Figure 3, the first oracle and SPCD crossings below ρ_{prop} yield $\mathcal{G}_t^\star$ and $\widehat{\mathcal{G}}_t$, respectively. Temporal proximity and random sparsification do not match this sufficiency at equal edge budgets. The restricted response $R_t(\mathcal{A}_t^\star, \mathcal{G}_t^\star)$ preserves the counterfactual target $R_t(\mathcal{A}_t^\star)$ within this KL tolerance. See Appendix B for formal details and Section 4 for the diffusion approximation.

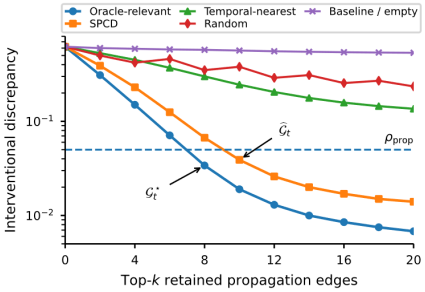

Figure 3: Selecting target-specific propagation structures under an edge budget.

4 Scene-Preserving Counterfactual Diffusion

SPCD anchors the scene via factual inversion and learns factor-specific changes from controlled responses. Diffusion-score responses guide structure selection, counterfactual generation (Figure 4).

4.1 From Counterfactual Responses to Diffusion Scores

We approximate the comparisons in Eqs. 6 and 7 using diffusion-score differences derived from factual and counterfactual denoiser responses.

Diffusion Conditioning. The encoder extracts scene evidence h_t and source codes $\mathcal{D}_t = \{d_i^m : i \in \mathcal{I}_t, m \in [M]\}$ from the observed window, with d_i^m representing W_i^m. For intervention set $\mathcal{A}$ and propagation graph $\mathcal{G}$, the diffusion condition $C_{t,\mathcal{G}}^{\mathcal{A}}$ is

$$(h_t, \mathcal{D}_t) = E_\phi\left(\mathbf{Y}_t\right), \qquad C_{t,\mathcal{H}}^{\mathcal{A}} = \left[h_t; F_{\mathrm{prop}}\left(\mathcal{T}_{\mathrm{nl}}\left(\mathcal{D}_t; \mathcal{A}\right); \mathcal{H}\right)\right], \tag{8}$$

where $\mathcal{T}_{\mathrm{nl}}$ applies learned transformations to selected source codes and F_{prop} propagates them through $\mathcal{H}$. Together, they encode soft interventions and temporal propagation while keeping scene context h_t fixed across diffusion-response comparisons.

Diffusion-Score Surrogate. Let $\widehat{\mathcal{G}}_t^{\mathrm{f}}$ be the factual graph predicted by the amortized predictor and $\mathcal{G}_{r,t}^W$ the full candidate source graph with edges $\mathcal{E}_{r,t}^W$. Inverting Y_t under $C_t^{\mathrm{f}} = C_{t,\widehat{\mathcal{G}}_t^{\mathrm{f}}}^{\varnothing}$ yields a fixed trajectory $\{z_\tau^{\mathrm{inv}}\}_{\tau=0}^N$, along which differences between factual and intervened diffusion responses serve as a counterfactual surrogate.

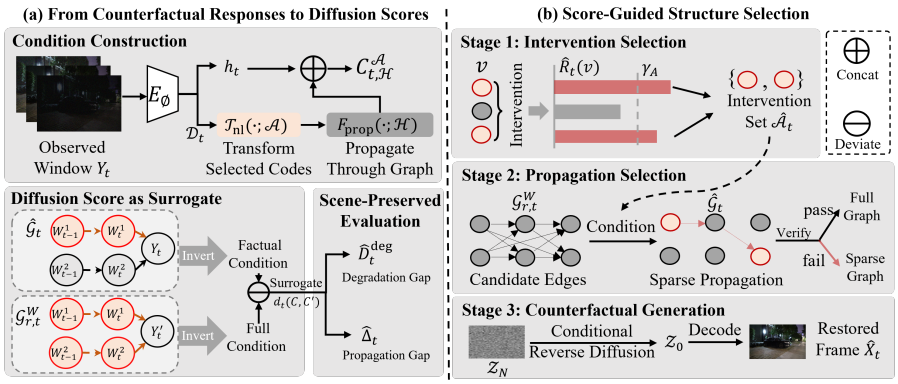

Figure 4: Overview framework of the proposed SPCD.

We use the full-graph reference $C_t^{\mathcal{A},\mathrm{ref}} := C^{\mathcal{A}}_{t,\mathcal{G}^W_{r,t}}$ to evaluate intervention responses and propagation sufficiency. For conditions C, C', define the diffusion-score discrepancy as

$$d_t(C, C') = \frac{1}{K}\sum_{k=1}^{K} w_{\tau_k} \left\| \epsilon_\theta\left(z^{\mathrm{inv}}_{\tau_k}, \tau_k; C\right) - \epsilon_\theta\left(z^{\mathrm{inv}}_{\tau_k}, \tau_k; C'\right)\right\|_2^2, \tag{9}$$

where $\{\tau_k\}_{k=1}^K$ are sampled diffusion steps, $w_{\tau_k} > 0$, and all responses are evaluated on z^{inv}_τ. For Gaussian reverse transitions with shared variance, one-step KL is proportional to squared noise-prediction differences, making d_t a computable surrogate for response discrepancies.

Scene-Preserved Evaluation. Motivated by diffusion-based reconstruction (Yi et al., 2023), we assume normalizing all modeled source conditions recovers the same factual scene's normal-light response. With accurate learning and factual abduction, the fully normalized condition C_t^{nl} induces a distribution approximating $\pi_t^\star$. We define

$$\widehat{D}_t^{\deg}(\mathcal{A}) = d_t\left(C_t^{\mathrm{nl}}, C_t^{\mathcal{A},\mathrm{ref}}\right), \qquad \widehat{\Delta}_t(\mathcal{H}; \mathcal{A}) = d_t\left(C_t^{\mathcal{A},\mathrm{ref}}, C^{\mathcal{A}}_{t,\mathcal{H}}\right). \tag{10}$$

The two scores approximate the degradation gap $D_{\mathrm{JS}}(\pi_t^\star \| R_t(\mathcal{A}))$ and propagation gap $D_{\mathrm{KL}}(R_t(\mathcal{A}) \| R_t(\mathcal{A}, \mathcal{H}))$, respectively. They provide tractable surrogates for instance-specific optimization of Eqs. 6 and 7.

4.2 Score-Guided Structure Selection

The surrogate guides target-specific structure selection and counterfactual image generation.

Stage 1: Intervention Selection. The objective seeks the smallest intervention set whose joint response approaches full normalization. We approximate this search by selecting conditions whose individual responses exceed a calibrated threshold, yielding $\widehat{\mathcal{A}}_t$:

$$\widehat{\mathcal{A}}_t = \{v : \widehat{R}_t(v) > \gamma_A\}, \tag{11}$$

where $\widehat{R}_t(v) := d_t(C_t^{\mathrm{ref}}, C_t^{\{v\},\mathrm{ref}})$ is the singleton intervention response, and γ_A is calibrated using factor-level macro F1 on the controlled validation set.

Stage 2: Propagation Selection. Given $\widehat{\mathcal{A}}_t$, SPCD seeks the sparsest propagation structure preserving the full-graph response within tolerance:

$$\widehat{\mathcal{G}}_t \in \arg\min \mathcal{H} :, \mathcal{E}_{\mathcal{H}} \subseteq \mathcal{E}^W_{r,t} \mathcal{L}_{\mathrm{sp}}(\mathcal{H}) \quad \text{s.t.} \quad \mathcal{L}_{\mathrm{resp}}(\mathcal{H}) \le \rho_{\mathrm{prop}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{resp}}(\mathcal{H})$ compares gated and full-graph predictions at the same noisy latent, and $\mathcal{L}_{\mathrm{sp}}(\mathcal{H})$ penalizes gate strength for sparsity. SPCD optimizes differentiable gates over $\mathcal{E}^W_{r,t}$ using both losses, then thresholds them at inference to obtain $\widehat{\mathcal{G}}_t$. If response tolerance is violated, only counterfactual propagation reverts to the full graph; the factual inversion trajectory under C_t^{f} and the scene and observation mechanisms remain fixed.

Table 1: Comparison on SDSD and zero-shot BVI-RLV, grouped by frame-wise, video-based, and diffusion-based methods. P/S/L: PSNR/SSIM/LPIPS; best/second-best: **bold**/underlined.

Type	Method(Configuration)	SDSD-In		SDSD-Out		BVI-RLV$_{ZS}$			Complexity	
		P↑	S↑	P↑	S↑	P↑	S↑	L↓	Para.	FLOPs
Frame	SNR-Net	27.30	0.84	23.23	0.82	20.55	0.72	0.29	4.01M	26.4G
	SNR-Net + CFIM	28.15	0.86	24.57	0.83	20.68	0.76	0.28	4.10M	27.4G
	Retinexformer	26.56	0.79	22.80	0.77	20.80	0.74	0.26	1.61M	15.6G
	SMG	27.82	0.86	25.17	0.81	22.01	0.78	0.24	7.8M	42.5G
	PairLIE	23.48	0.71	20.04	0.65	18.06	0.59	0.42	0.34M	44.7G
	ZERO-IG	26.19	0.74	23.37	0.76	20.53	0.66	0.30	1.80M	77.0G
	QuadPrior	25.53	0.71	22.42	0.73	19.88	0.72	0.35	1.30G	42.0T
Video	StableLLVE	25.32	0.70	22.47	0.65	19.65	0.68	0.33	4.32M	10.1G
	SDSDNet	26.81	0.75	23.08	0.71	20.25	0.71	0.30	4.30M	39.2G
	LAN	28.76	0.87	25.86	0.83	21.45	0.76	0.24	–	–
	RetinexMCNet	<u>29.16</u>	0.90	26.56	<u>0.86</u>	21.82	0.79	0.21	27.9M	25.2G
	STCD	28.93	0.88	26.32	0.82	22.05	0.80	0.20	1.81M	24.4G
	AMNet (RGB)	29.03	<u>0.92</u>	26.37	0.84	22.32	**0.84**	**0.16**	28.0M	86.0G
	SMID	24.84	0.72	23.30	0.67	18.22	0.62	0.40	85.6M	64.1G
	SMOID	24.63	0.70	22.25	0.68	18.21	0.62	0.41	12.2M	15.1G
	DP3DF	27.63	0.74	23.85	0.73	21.98	0.73	0.26	28.9M	77.5G
	LLVE-SEG	26.19	0.77	24.09	0.72	20.55	0.65	0.28	3.51M	3.28G
	BRVE	26.02	0.74	23.50	0.71	20.65	0.66	0.28	0.30M	1.49G
	Shift-Net	27.81	0.76	24.28	0.75	22.03	0.78	0.24	10.8M	146G
	VLLVE	28.84	0.87	**26.95**	0.84	21.92	0.76	0.22	3.16M	17.3G
Diffusion	Diff-Retinex	25.73	0.74	23.15	0.76	19.63	0.72	0.25	58.9M	3.74T
	ReDDiT	26.62	0.78	22.93	0.78	20.96	0.77	0.25	17.4M	547G
	MR. Illuminate	27.16	0.80	23.05	0.78	20.33	0.66	0.32	1.075G	58.0T
	EVDiffuser (RGB)	27.35	0.89	24.25	0.77	22.20	0.79	0.19	30.9M	7.90T
	SPCD (Joint-Code)	28.74	0.89	25.91	0.83	21.90	0.77	0.22	44.5M	3.80T
	SPCD (Full-Graph)	29.08	0.91	26.43	0.85	<u>22.48</u>	0.81	<u>0.18</u>	44.8M	3.90T
	SPCD	**29.47**	**0.93**	<u>26.86</u>	**0.88**	**22.96**	<u>0.83</u>	**0.16**	44.8M	6.10T

Stage 3: Counterfactual Generation. Starting from the inversion latent under C_t^{f}, SPCD performs reverse diffusion conditioned on the selected interventions and propagation structure:

$$z_N = z_N^{\mathrm{inv}}, \qquad z_{\tau-1} = \mathrm{Step}\left(z_\tau, \epsilon_\theta\left(z_\tau, \tau; C_t^{\mathrm{cf}}\right), \tau\right), \qquad \tau = N, \ldots, 1. \tag{13}$$

Decoding yields $\widehat{X}_t$. Shared h_t^{sc} and z_N^{inv} preserve the inferred factual scene, while $\widehat{\mathcal{A}}_t$ and $\widehat{\mathcal{G}}_t$ specify the intervention and its propagation. See Appendix B for derivation and implementation details.

5 Experiments

Benchmarks. We evaluate SPCD on SDSD (Wang et al., 2021), BVI-RLV (Lin et al., 2024), our Causal-V-SDSD constructed from SDSD, and three composite degradation scenarios.

Baselines. We compare three groups: (I) frame-wise methods: SNR-Net (Xu et al., 2022), Retinexformer (Cai et al., 2023), their CFIM variants (Xu et al., 2025), SMG (Xu et al., 2023b), PairLIE (Fu et al., 2023b), ZERO-IG (Shi et al., 2024), and QuadPrior (Wang et al., 2024); (II) video methods: StableLLVE (Zhang et al., 2021), SDSDNet (Wang et al., 2021), LAN (Fu et al., 2023a), RetinexMCNet (Wang et al., 2025), STCD (Xu et al., 2025), AMNet (Liang et al., 2026), SMID (Chen et al., 2019), SMOID (Jiang & Zheng, 2019), DP3DF (Xu et al., 2023a), LLVE-SEG (Liu et al., 2023), BRVE (Zhang et al., 2024), Shift-Net (Li et al., 2023), and VLLVE (Xu et al., 2026); (III) diffusion methods: Diff-Retinex (Yi et al., 2023), ReDDiT (Lan et al., 2025), MR. Illuminate (Cho et al., 2026), and EVDiffuser (Chen et al., 2025). We also construct two matched baselines: *Joint-Code Diffusion* and *Full-Graph Diffusion*.

Metrics. We report PSNR (P), SSIM (S) (Wang et al., 2004), LPIPS (L) (Zhang et al., 2018), Transition LPIPS (T-L), Factor-F1, edge-level F1, AUPRC (Davis & Goadrich, 2006), intervention

Table 2: Restoration under controlled dynamic degradation on Causal-V-SDSD.

Method	Smooth			Abrupt			Propagated		
	P↑	S↑	L↓	P↑	S↑	L↓	P↑	S↑	L↓
Retinexformer†	18.12	0.74	0.32	18.76	0.73	0.35	18.39	0.73	0.36
SDSDNet†	18.38	0.72	0.34	19.02	0.72	0.36	18.71	0.72	0.36
STCD†	19.24	0.79	0.28	19.91	0.79	0.30	19.61	0.79	0.31
AMNet (RGB)†	19.43	0.80	0.27	20.10	0.80	0.29	19.78	0.80	0.30
EVDiffuser (RGB)†	19.31	0.79	0.27	19.98	0.78	0.29	19.66	0.78	0.30
SPCD (Joint-Code)	16.97	0.76	0.27	18.23	0.77	0.29	18.21	0.77	0.29
SPCD (Full-Graph)	19.70	0.80	0.26	20.81	0.80	0.28	20.43	0.80	0.29
SPCD	**21.60**	**0.84**	**0.22**	**22.22**	**0.84**	**0.24**	**22.45**	**0.85**	**0.25**

† Comparisons are reproduced and evaluated under their original training protocols.

response error (IRE), GT-LPIPS, cross-intervention LPIPS (CI-L), FLOPs, runtime, NFE, and peak memory. See Appendix C for dataset, baseline, and evaluation details.

5.1 Main Comparisons

Comparison on SDSD and BVI-RLV. SPCD leads on SDSD-Indoor (+0.31 dB PSNR over RetinexMCNet, the strongest external baseline) and remains competitive on SDSD-Outdoor despite VLLVE's stronger video-specific temporal decomposition (Table 1). On zero-shot BVI-RLV, it surpasses EVDiffuser by 0.76 dB PSNR with better perceptual quality. With identical trainable parameters, SPCD improves all restoration metrics over Full-Graph, isolating the benefit of suppressing irrelevant factors and redundant propagation paths.

Comparison on Causal-V-SDSD. SPCD leads across all three dynamic regimes, exceeding the strongest external zero-shot baseline by 2.12–2.67 dB PSNR and matched Full-Graph by 1.41–2.02 dB, with higher SSIM and lower LPIPS (Table 2). Gains over Full-Graph peak under propagated degradation. Joint-Code's representation struggles with complex temporal degradation, while Full-Graph's improvement supports factorized conditioning. SPCD's further gains support target-specific structure selection over dense propagation for interacting degradations.

5.2 Ablation Studies

Conditioning structure ablation. Selective factor replacement improves restoration over replacing all factors (Table 3). SPCD achieves the lowest LPIPS across variants with 95.3% fewer edges than Selected + Full. Gains over empty-graph and cardinality-matched random controls support structure selection beyond sparsity alone. Lower full-graph response discrepancy does not consistently imply better restoration.

Table 3: Conditioning structure ablation on Causal-V-SDSD. $|\mathcal{A}_t|$, $|\mathcal{E}_t^{\mathrm{rel}}|$: mean selected factor/edge counts per window; $\overline{\Delta}_\epsilon$, $\Delta_\epsilon^{\max}$: mean/max response discrepancies.

Variant	$\lvert\mathcal{A}_t\rvert$ / $\lvert\mathcal{E}_t^{\mathrm{rel}}\rvert$	$\overline{\Delta}_\epsilon$	$\Delta_\epsilon^{\max}$	LPIPS↓
All + Full	14.06 / 81.56	0.00	0.00	0.285
Selected + Full	9.92 / 81.56	0.00	0.00	0.269
Selected + Static	9.92 / 11.06	0.31	0.56	0.270
Selected + Empty	9.92 / 0.00	0.49	1.40	0.269
Random Sparse	9.92 / 3.84	0.53	2.89	0.271
SPCD	9.92 / 3.84	0.74	2.27	**0.243**

Intervention-response validation. SPCD approximates generator responses across smooth, abrupt, and propagated changes, with low single-intervention but higher joint-intervention errors and an overall correlation of 0.4836 (Figure 5(a)). In panel (b), the diffusion score achieves $r_s^{\mathrm{CF}} = 0.8411$ at $K = 2$, versus -0.0416 for lagged correlation and 0.3710 for gradient attribution, supporting its use as a surrogate for explicit counterfactual responses.

Effectiveness of Scene Preservation. Table 4(a) compares degradation trajectories of the same factual scene. Removing scene conditioning (row 2) eliminates the explicit scene anchor, substantially degrading restoration quality and cross-trajectory consistency. SPCD maintains quality and low CI-L, showing scene-consistent counterfactual restoration through explicit scene conditioning.

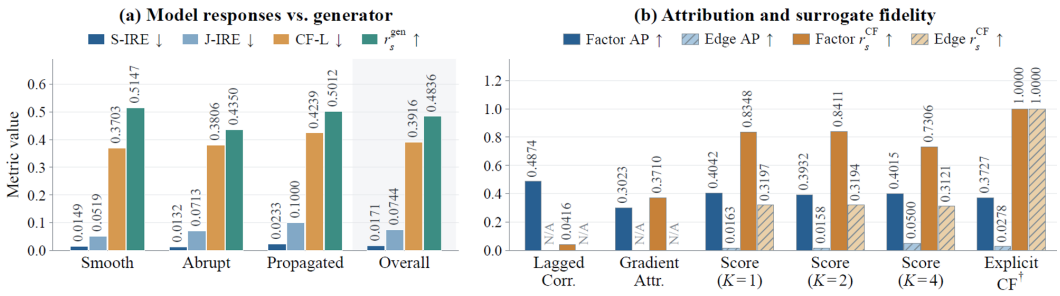

Figure 5: SPCD response validation. S/J-IRE: single/joint intervention-response error; F/E: factor/edge; r_s^{gen}, r_s^{CF}: Spearman correlations with generator/explicit model response magnitudes.

Table 4: Scene preservation and unseen-dynamics robustness of SPCD. GT-P/GT-L: clean-target PSNR/LPIPS; CI-L: cross-trajectory LPIPS (matched same-scene targets); IRE: matched joint-window soft-intervention response error.

(a) Scene Preservation

Configuration	GT-P↑	GT-L↓	CI-L↓
Selected + Full	19.214	0.269	0.140
w/o h_t^{sc}	7.876	0.766	0.239
w/o inversion init.	**22.141**	0.327	**0.139**
SPCD	20.001	**0.266**	0.140

(b) Unseen Dynamics

Configuration	Rate LPIPS/IRE↓	Composition LPIPS/IRE↓	Lag LPIPS/IRE↓
All + Full	0.284/**0.067**	0.247/**0.062**	0.296/0.075
Selected + Full	0.269/0.070	0.237/0.067	0.281/0.075
SPCD	**0.266**/0.072	**0.232**/0.073	**0.276/0.073**

Robustness to Unseen Dynamics. Table 4(b) tests unseen variation rates, factor compositions, and propagation lags within the same source-condition vocabulary. SPCD consistently achieves lower LPIPS than full-graph variants with comparable IRE, showing that its intervention and propagation selection generalizes beyond controlled training dynamics without new source types.

Table 6: Inference sensitivity to γ_A and τ_g. Edge F1: micro F1 using generator edge-support labels over valid target-window slots.

γ_A	τ_g	LPIPS↓	Edge F1↑
3×10^{-5}	0.30	0.2446	0.4408
6×10^{-5}	0.30	0.2352	0.4548
1.2×10^{-4}	0.30	0.2493	0.6306
6×10^{-5}	0.20	0.2350	0.4597
6×10^{-5}	0.40	0.2350	0.4515

Computational Efficiency. At matched NFE and nearly identical runtime, SPCD improves perceptual quality across all three degradation regimes, supporting target-specific conditioning (Table 5). The default $K = 2$ balances quality and efficiency: $K = 4$ adds 32.2% runtime with marginal, mixed quality changes, while $K = 1$ reduces cost with modest quality loss.

Parameter Sensitivity. At fixed $\tau_g = 0.30$, the default γ_A achieves the lowest LPIPS among tested factor thresholds, favoring selective intervention over indiscriminate expansion or restriction

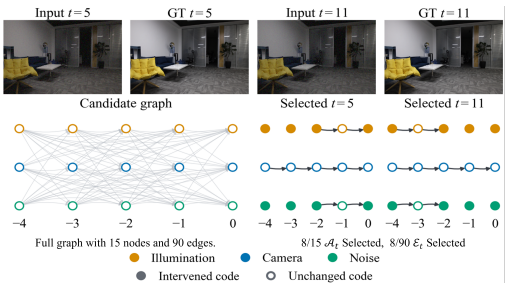

Figure 6: Target-specific structure selection.

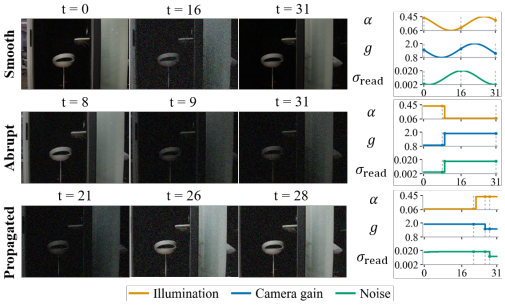

Figure 7: Controlled degradation dynamics.

Table 5: SPCD efficiency. K: response evaluation steps; NFE: function evaluations for inversion, selection, verification, and reverse sampling. Full-Graph(NFE) adjusts inversion/sampling schedules to match the default budget. FB: percentage of real targets affected by clip-level full-graph fallback.

Configuration	LPIPS↓			T-L↓	NFE↓ (/window)	Time↓ (s/target)	Memory↓ (GB)	FB (%)
	Smooth	Abrupt	Propagated					
Full-Graph	0.2686	0.2889	0.2986	0.2861	59	1.244	11.898	N/A
Full-Graph (NFE)	0.2814	0.3016	0.3118	0.3000	95	1.826	11.898	N/A
SPCD ($K = 1$)	0.2532	0.2712	0.2814	0.2685	77	1.540	11.902	0.00
SPCD ($K = 2$)	**0.2507**	0.2691	0.2774	**0.2648**	95	1.836	11.902	0.00
SPCD ($K = 4$)	0.2509	**0.2689**	**0.2773**	0.2650	131	2.428	11.902	0.00



Figure 8: Qualitative comparison on the Causal-V-SDSD benchmark.

(Table 6). LPIPS varies by only 0.0002 across $\tau_g \in [0.20, 0.40]$, supporting threshold stability. Restoration-oriented selection need not maximize edge-support agreement.

5.3 Visualization

Target-Specific Structure Selection. Figure 6 shows SPCD selecting eight intervention codes and eight propagation edges per target from the same 15-node, 90-edge candidate graph, with locations adapting across frames. Separating what to modify from how changes propagate enables selective intervention on degradation factors and relevant paths, yielding compact, adaptive conditioning for scene-preserving restoration.

Controlled Degradation Dynamics. Figure 7 shows three frames per row from the same source clip under smooth, abrupt, or propagated degradation. Delayed cross-factor responses in the propagated case motivate target-specific conditioning and temporal propagation.

Qualitative Comparison. Figure 8 compares five aligned frames under smooth degradation with frame-based SNR-Net+CFIM, video-based RetinexMCNet, and diffusion-based EVDiffuser. SPCD consistently restores balanced exposure, vehicle contours, and roadside details in both close-ups, matching the clean reference. This supports selective normalization and sparse temporal propagation, consistent with scene evidence and factual inversion anchoring targeted correction.

6 Conclusion

SPCD improves dynamic low-light video restoration through scene-preserving counterfactual diffusion and target-specific intervention structures, as validated experimentally.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation for language refinement, structural editing, and discussion of presentation and formulation. All technical claims, mathematical derivations, citations, experimental procedures, and final content are subject to author verification, and the authors take full responsibility for the submitted manuscript.

REFERENCES

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 IEEE/CVF International Conference on Computer Vision*, pp. 12470–12479, 2023.

Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 3184–3193, 2019.

Kanghao Chen, Zixin Zhang, Guoqiang Liang, Lutao Jiang, Zeyu Wang, and Yingcong Chen. Event-guided consistent video enhancement with modality-adaptive diffusion pipeline. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Kanghao Chen, Guoqiang Liang, Yunfan Lu, Hangyu Li, and Lin Wang. Evlight++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(2):1608–1625, 2026.

Joshua Cho, Sara Aghajanzadeh, Zhen Zhu, and David Forsyth. Mr. illuminate: Zero-shot low-light image enhancement with diffusion prior. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8460–8470, 2026.

Jesse Davis and Mark H. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of International Conference on Machine Learning*, pp. 233–240, 2006.

Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 12831–12840, 2023a.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22252–22261, 2023b.

Amirhosein Ghasemabadi, Muhammad Kamran Janjua, Mohammad Salameh, and Di Niu. Learning truncated causal history model for video restoration. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 7323–7332, 2019.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21277–21286, 2025.

Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9822–9832, 2023.

Hangfeng Liang, Yutao Hu, Yanhan Hu, Xiaohan Wu, Wenqi Shao, and Ying Fu. Anymod-llve: Low-light video enhancement with modality-agnostic inference. *CoRR*, abs/2606.11186, 2026.

Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 10581–10591, 2023.

Ruirui Lin, Nantheera Anantrasirichai, Guoxi Huang, Joanne Lin, Qi Sun, Alexandra Malyugina, and David R. Bull. BVI-RLV: A fully registered dataset and benchmarks for low-light video enhancement. *CoRR*, abs/2407.03535, 2024.

Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yanfeng Wang, and Qi Tian. Low-light video enhancement with synthetic event guidance. In *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 1692–1700, 2023.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

Bo Peng, Xuanyu Zhang, Jianjun Lei, Zhe Zhang, Nam Ling, and Qingming Huang. LVE-S2D: low-light video enhancement from static to dynamic. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8342–8352, 2022.

Rajat Rasal, Avinash Kori, Fabio De Sousa Ribeiro, Tian Xia, and Ben Glocker. Diffusion counterfactual generation with semantic abduction. In *Proceedings of International Conference on Machine Learning*, 2025.

Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. ZERO-IG: zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3015–3024, 2024.

Lei Tong, Zhihua Liu, Chaochao Lu, Dino Oglic, Tom Diethe, Philip Teare, Sotirios A. Tsaftaris, and Chen Jin. Causal-adapter: Taming text-to-image diffusion for faithful counterfactual generation. *CoRR*, abs/2509.24798, 2025.

Meiao Wang, Xuejing Kang, Yaxi Lu, and Jie Xu. Retinexmcnet: A memory controller dominated network for low-light video enhancement based on retinex. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 9716–9725, 2025.

Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 9680–9689, 2021.

Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26066, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Shenghao Wu, Wenbin Zhou, Minshuo Chen, and Shixiang Zhu. Counterfactual generative models for time-varying treatments. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, pp. 3402–3413, 2024.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17693–17703, 2022.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Deep parametric 3d filters for joint video denoising and illumination enhancement in video super resolution. In *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 3054–3062, 2023a.

Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9893–9903, 2023b.

Xiaogang Xu, Kun Zhou, Tao Hu, Jiafei Wu, Ruixing Wang, Hao Peng, and Bei Yu. Low-light video enhancement via spatial-temporal consistent decomposition. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 2134–2142, 2025.

Xiaogang Xu, Kun Zhou, Tao Hu, Jiafei Wu, Ruixing Wang, Hao Peng, and Bei Yu. Low-light video enhancement with an effective spatial-temporal decomposition paradigm. *CoRR*, abs/2602.08699, 2026.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with A generative diffusion model. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 12268–12277, 2023.

Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4967–4976, 2021.

Gengchen Zhang, Yulun Zhang, Xin Yuan, and Ying Fu. Binarized low-light raw video enhancement. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25753–25762, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Haiyu Zhao, Lei Tian, Xinyan Xiao, Peng Hu, Yuanbiao Gou, and Xi Peng. Avernet: All-in-one video restoration for time-varying unknown degradations. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

A Related Work

Low-Light Video Restoration. Low-light video restoration has evolved from frame-wise enhancement toward explicit temporal modeling through multi-frame alignment, recurrent propagation, and temporal consistency (Wang et al., 2021; 2025; Fu et al., 2023a). Event cameras further provide high-dynamic-range temporal cues for robust restoration under challenging illumination (Liang et al., 2023; Chen et al., 2026). More recently, diffusion priors and varying-illumination settings have been explored to improve restoration quality and robustness in dynamic environments (Chen et al., 2025). Despite these advances, temporal dependencies are still primarily modeled through observational correlation, alignment, or predefined connectivity.

Causal and Counterfactual Generative Modeling. Deep structural causal models combine learned generative mechanisms with explicit causal structure to support interventional and counterfactual inference (Pawlowski et al., 2020; Tong et al., 2025). Recent diffusion-based approaches incorporate causal representations or abduction mechanisms for counterfactual generation (Rasal et al., 2025), while researchers also use conditional generative models, including guided diffusion, to model outcome distributions under time-varying treatments (Wu et al., 2024). However, how intervention-relevant degradation dependencies vary across frames in multi-frame restoration remains largely unexplored.

B Additional Theory and Method Details

B.1 An Explicit Underdetermination Construction

The following construction shows why fitting the observational scene posterior alone cannot identify an intervention response. Consider two frames, a fixed scene $X_1 = X_2 = 1$, and latent source conditions (I_t, C_t). Let U, V be independent uniform random variables on $(0, 1)$ and let $0 < a < 1$. Both models use $I_1 = U$, $I_2 = C_1 = 0$, and $Y_t = X_t - C_t$, but their second-frame mechanisms differ:

$$M : \; C_2 = aI_1, \qquad M' : \; C_2 = aV. \tag{14}$$

Both belong to the same finite-horizon candidate class for $r \geq 1$, although their active dependencies differ. Their joint observational law is identical: $X_1 = X_2 = Y_1 = 1$ and $Y_2 \sim \mathrm{Unif}(1 - a, 1)$. In particular, both yield the same scene posterior.

Now apply the same soft intervention $I_1 \leftarrow (1-\eta)I_1$, where $0 < \eta < 1$ and zero is the reference source state. Within each model, retain the exogenous realization drawn from its own factual posterior. For almost every observed $y \in (1-a, 1)$,

$$Y_2^{\mathrm{cf}}(M) = (1-\eta)y + \eta, \qquad Y_2^{\mathrm{cf}}(M') = y. \tag{15}$$

Thus, observationally indistinguishable mechanisms have different responses to the same intervention, even when the scene is known. This abstract example establishes the underdetermination in the main text; it is not a camera-physics model. It motivates evaluating intervention responses separately from ordinary restoration accuracy.

B.2 Scene Preservation, Reachability, and Minimality

Fix the factual posterior $\mu_t^\star$ and an explicitly specified rule for restricting a mechanism to a candidate graph $\mathcal{H}$. The counterfactual response is the pushforward

$$R_t(\mathcal{A}, \mathcal{H}) = \mathrm{Law}_{U \sim \mu_t^\star}\left[Y_t^{M[\mathcal{A},\mathcal{H}]}(U)\right]. \tag{16}$$

Factual and counterfactual evaluations share the scene mechanisms and their exogenous realization; only designated source-condition mechanisms are modified. Consequently, the underlying scene realization is unchanged within this SCM formulation. Scene preservation does not require a static scene or identical pixels across time.

Normal-light reachability is the assumption

$$D_{\mathrm{JS}}\big(\pi_t^\star \,\big\|\, R_t(\mathcal{S}_t, \mathcal{G}_{r,t}^W)\big) \le \rho_{\mathrm{deg}}. \tag{17}$$

It makes the full intervention a feasible reference, rather than asserting that arbitrary learned code replacement necessarily recovers the clean-scene posterior. The full graph is likewise feasible for the propagation objective because its discrepancy from itself is zero. For finite candidate sets, these nonempty feasible sets admit cardinality-minimal solutions. Such solutions need not be unique or coincide with the physical generating graph. SPCD approximates these response-oriented objectives with learned representations, amortized gates, and finite score probes; shared scene evidence and inversion provide the corresponding practical anchors.

B.3 Representations and Matched Interventions

We use h_t^{sc} consistently for scene evidence; h_t in the main conditioning equation denotes the same quantity. The encoder produces scene feature maps and three factor-indexed code families: illumination, camera response, and noise. Camera response includes gain, response exponent, and white balance, rather than camera motion. The scene feature width is 64, each factor code has dimension 128, and the propagated graph representation has dimension 256. Clean-reference codes supervise learned factor-specific normal prototypes p_m.

For a selected factor slot, the implemented transformation is

$$\widetilde{d}_i^m = \big(1 - \eta \mathbf{1}[(i,m) \in \mathcal{A}]\big) d_i^m + \eta \mathbf{1}[(i,m) \in \mathcal{A}] p_m. \tag{18}$$

Restoration uses $\eta = 1$, while matched soft-response diagnostics use $\eta = 0.5$. Unselected codes and factual scene evidence remain unchanged. The graph module propagates the edited codes through two chronological message-passing sweeps with gated, degree-normalized aggregation and GRU updates. Code replacement does not delete a node or surgically cut its incoming messages. Six residual conditioning injections and factor tokens connect this representation to the frozen diffusion backbone.

Reproducible generator responses. Matched generator interventions share the clean source, factual parameter trajectories, spatial phase, and both Gaussian noise arrays. The response protocol explicitly defines a factual-anchored extension: for a dependent parameter θ_t, recompute

$$\theta_t^{\mathrm{cf}} = \mathrm{clip}\big[\theta_t^0 + F(p_t^{\mathrm{cf}}) - F(p_t^0)\big], \tag{19}$$

in chronological order, where p_t denotes the physical parent argument, not the learned prototype. The cross-factor functions are specified in Appendix D.2. Selected illumination and camera parameters move by η toward $(\alpha, a) = (1, 0)$ and $(g, \gamma, \mathbf{w}) = (1, 2.2, \mathbf{1})$, respectively. Noise intervention

scales read-noise amplitude by $1-\eta$ and replaces the shot parameter q by $q/(1-\eta)^2$. Downstream effects are reexecuted while the noise draws remain fixed. Suppressing a generator cross edge substitutes its parent's normal argument in that mechanism. This versioned extension is defined in addition to factual replay; it is not inferred uniquely from the observed trajectories.

The matched model edge probe replaces only the source argument of the queried message by its learned normal prototype, keeping gates and normalization denominators fixed. This isolates an edge response and is distinct from ordinary graph pruning. Agreement between model and generator responses is measured, not assumed from their matched labels. For the explicit-response diagnostic, factual inversion uses the learned thresholded graph. Both the empty-intervention reference and the soft-intervention samples then use the full graph of the same frozen SPCD checkpoint, sharing the inversion latent, seed, and sampling schedule.

B.4 Meaning and Implementation of the Score Surrogate

The diffusion score is a conditional log-density gradient. Under the forward process $z_\tau = \sqrt{\bar{\alpha}_\tau} z_0 + \sigma_\tau \epsilon$, an exact conditional noise predictor gives $s_C(z_\tau, \tau) = -\epsilon_C(z_\tau, \tau)/\sigma_\tau$. Consequently, for the noised response distributions p_τ^C and $p_\tau^{C'}$,

$$J(p_\tau^C \| p_\tau^{C'}) = \mathbb{E}_{p_\tau^C}\left[\sigma_\tau^{-2} \|\epsilon_C - \epsilon_{C'}\|_2^2\right]. \tag{20}$$

For a common forward SDE with diffusion coefficient $g(\tau)$, the relative de Bruijn identity, under the usual differentiability and integrability conditions, yields

$$\begin{aligned} D_{\mathrm{KL}}(p_0^C \| p_0^{C'}) = D_{\mathrm{KL}}(p_T^C \| p_T^{C'}) \\ + \frac{1}{2}\int_0^T g(\tau)^2 J(p_\tau^C \| p_\tau^{C'})\, d\tau. \end{aligned} \tag{21}$$

For stochastic Gaussian reverse transitions with common covariance $\nu_\tau^2 I$ and means $a_\tau z - b_\tau \epsilon_C$, the one-step KL is $b_\tau^2 \|\epsilon_C - \epsilon_{C'}\|_2^2/(2\nu_\tau^2)$. These identities motivate score-space comparisons. The finite probes on a fixed factual trajectory are a computable response surrogate, not an exact population KL estimate. The Gaussian transition statement does not assert a finite transition KL for deterministic DDIM sampling.

Prediction parameterization and normalization. The backbone predicts v, not ϵ directly. At identical latent and diffusion time,

$$\epsilon_C - \epsilon_{C'} = \sqrt{\bar{\alpha}_\tau}\,(v_C - v_{C'}). \tag{22}$$

Let MSE_D average the D latent coordinates of the scored target group and $q_k = 1/\max(1 - \bar{\alpha}_{\tau_k}, 10^{-4})$. Archived restoration thresholds operate on

$$d_v(C, C') = \frac{\sum_k q_k \,\mathrm{MSE}_D(v_C, v_{C'})}{\sum_k q_k}. \tag{23}$$

This is expressible in the main text's epsilon-space form with positive weights $w_k = Kq_k/(D\bar{\alpha}_{\tau_k}\sum_j q_j)$. For epsilon diagnostics, we instead report $d_\epsilon = \sum_k q_k \bar{\alpha}_{\tau_k} \mathrm{MSE}_D(v_C, v_{C'})/\sum_k q_k$. The two conventions are recorded separately; their numerical thresholds are not interchangeable.

The restoration runner probes forward-noised observed VAE latents, using shared noise and nominal times $\{0.35, 0.70\}$ before inversion. Same-checkpoint conditioning and explicit-response experiments instead probe the shared factual inversion trajectory. In both cases the two conditions are compared at identical latents and timesteps. The archived inversion-trajectory ablations retain the original selection thresholds as a fixed operating point, rather than claiming a newly calibrated epsilon-space tolerance.

B.5 Training and Differentiable Gates

Controlled samples provide normalized parameter targets and valid generator edge-support annotations. Real samples contribute paired restoration supervision without synthetic factor or edge labels.

Let $\mathcal{L}_{\text{fac}}$ be masked Smooth-L1 parameter regression, averaged equally across the three factor families, and let

$$\mathcal{L}_{\text{proto}} = \text{MSE}(d^{\text{clean}}, p), \qquad \mathcal{L}_{\text{sp}} = \frac{1}{|\mathcal{E}_{\text{valid}}|} \sum_{e \in \mathcal{E}_{\text{valid}}} g_e. \tag{24}$$

Structural supervision $\mathcal{L}_{\text{str}}$ averages valid-edge binary cross-entropy separately over same-factor and cross-factor strata, then weights the nonempty strata equally. Within each stratum, the positive-class weight is $\sqrt{N_-/N_+}$ when both classes occur, and one otherwise.

For paired scene maps S, S', scene supervision is

$$\begin{aligned} \mathcal{L}_{\text{sc}} = \; & \text{MSE}(\nu(S), \nu(S')) \\ & + \frac{1}{2C} \sum_{c=1}^{C} \left[(1 - \sigma_c(S))_+ + (1 - \sigma_c(S'))_+ \right], \end{aligned} \tag{25}$$

where ν normalizes each spatial–temporal feature vector across channels and $\sigma_c(S) = \sqrt{\text{Var}(S_c) + 10^{-4}}$. The variance term discourages an uninformative invariant representation.

Writing z_0 for the selected target's VAE latent, diffusion training uses

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\frac{\text{MSE}(\widehat{z}_0, z_0)}{\max(1 - \bar{\alpha}_\tau, 10^{-4})} \right], \tag{26}$$

$$\mathcal{L}_{x_0} = \text{MAE}(\widehat{z}_0, z_0), \tag{27}$$

$$\mathcal{L}_{\text{resp}} = \text{MSE}\Big(\widehat{z}_0^{\text{gated}}, \text{sg}[\widehat{z}_0^{\text{full}}] \Big). \tag{28}$$

The two predictions in $\mathcal{L}_{\text{resp}}$ share the noised latent, timestep, scene evidence, and target codes. This training consistency loss is distinct from the inference-time score discrepancy. The configured objectives are

$$\begin{aligned} \mathcal{L}_1 = \; & 0.10\mathcal{L}_{\text{sc}} + 0.50\mathcal{L}_{\text{fac}} + 0.20\mathcal{L}_{\text{str}} \\ & + 0.05\mathcal{L}_{\text{proto}} + 0.001\mathcal{L}_{\text{sp}}, \end{aligned} \tag{29}$$

$$\begin{aligned} \mathcal{L}_{2,3} = \; & \mathcal{L}_{\text{diff}} + 0.10\mathcal{L}_{\text{sc}} + 0.50\mathcal{L}_{\text{fac}} \\ & + 0.20(\mathcal{L}_{\text{resp}} + 0.5\mathcal{L}_{\text{str}}) + 0.05\mathcal{L}_{\text{proto}} \\ & + \lambda_{x_0}\mathcal{L}_{x_0} + 0.001\mathcal{L}_{\text{sp}}. \end{aligned} \tag{30}$$

Here $\lambda_{x_0} = 0.05$, increased to 0.20 in the final paired refinement. The pilot configuration evaluates full-graph consistency every eight optimizer steps. Diffusion training combines target-derived factor codes with degraded-input scene maps and a zero replacement mask; prototype alignment supplies the clean-domain reference used later for intervention.

Edge logits are learned through stretched logistic gates:

$$g_e = \text{clip}_{[0,1]} \left[1.2 \; \text{sigmoid}\left(\frac{\ell_e + \log u - \log(1-u)}{T_g} \right) - 0.1 \right]. \tag{31}$$

Training samples $u \sim \text{Unif}(0,1)$; evaluation uses $\text{clip}_{[0,1]}[1.2\,\text{sigmoid}(\ell_e) - 0.1]$ and thresholds it at τ_g. Thus, support supervision, response consistency, and sparsity have explicit, complementary roles. This amortized relaxation does not require retraining a separate model for every candidate graph.

B.6 Selection, Calibration, and Sampling

For $M = 3$ and $r = 4$, a complete target window contains 15 factor slots and 90 admissible directed edges; all edges point from earlier to later frames. Screening retains responses exceeding γ_A. Grouped probes edit one factor family and one source-index residue class modulo $r + 1$ at a time. This leaves at most one edited source in each local window and reduces query cost. It is a grouped approximation when the video denoiser couples different windows.

The counterfactual gates are recomputed after code replacement and checked against the full graph with the same intervention set. The restoration runner checks a clip-averaged discrepancy and falls

back for the evaluated batch if the tolerance fails. The target-level inversion diagnostic uses whole-clip fallback if any scored real target fails. Fallback changes propagation, not the selected factor set or the factual inversion. Its scope is retained in each experiment's protocol.

Calibration uses controlled validation data, maximizing factor positive/negative macro-F1 plus the mean of same-factor and cross-factor edge F1. Structural ties are resolved by validation LPIPS, followed by selected cardinalities and deterministic threshold ordering. The archived default is $\gamma_A = 6 \times 10^{-5}$, $\tau_g = 0.30$, and $\rho = 0.0025$ in the native response convention. The factual condition is inverted with a 30-timestep DDIM schedule (29 transitions); reverse sampling uses 30 deterministic steps. Scene maps, the initial inverted latent, and the chosen conditioning structure are fixed across matched counterfactual comparisons.

C Controlled Benchmark and Implementation Details

Controlled Dynamic Benchmark. Given a normal-light sequence $\mathbf{X} = \{X_1, \ldots, X_T\}$, we generate dynamic source conditions

$$\mathbf{W}_t = \{W_t^{\mathrm{ill}}, W_t^{\mathrm{cam}}, W_t^{\mathrm{noise}}\}, \tag{32}$$

where W_t^{ill}, W_t^{cam}, and W_t^{noise} denote illumination, camera exposure/gain state, and sensor-noise state, respectively.

The source conditions evolve through causal mechanisms across time. We consider both temporal persistence and cross-factor propagation, e.g.,

$$W_{t-1}^{\mathrm{ill}} \to W_t^{\mathrm{cam}}, \qquad W_{t-1}^{\mathrm{cam}} \to W_t^{\mathrm{noise}}, \qquad W_{t-1}^{m} \to W_t^{m}. \tag{33}$$

These dependencies follow the temporal acyclicity and candidate-parent structure defined in Section 2.

The corresponding degraded observation is generated according to

$$Y_t = \psi_{c(\mathbf{W}_t)}\left(b(\mathbf{W}_t) \odot X_t + a(\mathbf{W}_t)\right). \tag{34}$$

We consider three representative dynamic regimes: smooth variation, abrupt transition, and temporally propagated degradation. The strengths of individual mechanisms depend on the current imaging state, so their intervention relevance may vary across target frames. Consequently, the minimal intervention-relevant structure need not remain constant over time.

Ground-Truth Intervention Structure. The complete generating mechanism is known for the controlled benchmark. We therefore evaluate explicit interventions on source conditions within the same intervention domain used by SPCD.

For each target frame, the ground-truth intervention-relevant set $\mathcal{A}_t^\star$ is determined from the exact change induced by intervening on each candidate factor. The corresponding $\mathcal{G}_t^\star$ is defined as the minimal intervention-relevant structure that preserves these effects according to the criterion in Section 3.2. Thus, $\mathcal{G}_t^\star$ is not identified with the full generating graph: a physically present dependency may be excluded when its effect is interventionally negligible for the current target.

For counterfactual pairs, the factual scene sequence and the exogenous realization of the generating process are held fixed while only the designated source condition is intervened upon. This ensures that differences between factual and counterfactual observations are attributable to the intended intervention rather than unrelated stochastic variation.

Dataset Construction. Normal-light source sequences are obtained from the ground-truth videos of SDSD, with controlled variants inheriting their source-level train/validation/test assignments. The controlled training and validation sets contain 189 and 63 generated windows, respectively. The test set contains 60 generated clips derived from 20 source clips (10 indoor and 10 outdoor), with one Smooth, Abrupt, and Propagated trajectory per source. Each test clip contains 32 frames at 480×720 resolution, yielding 1,920 paired frames. The three trajectories of each source share identical clean targets. Dataset construction uses seed 2027.

Table 7: Active-factor parameter envelopes for controlled dynamic degradation. The original training and test sets share these envelopes; the separately evaluated unseen-dynamics settings modify temporal rates or schedules.

Source condition	Training range	Test range
Illumination W_t^{ill}	$\alpha \in [0.06, 0.45]$	$\alpha \in [0.06, 0.45]$
	$a \in [0, 0.25]$	$a \in [0, 0.25]$
Camera state W_t^{cam}	$g \in [0.8, 2.0]$	$g \in [0.8, 2.0]$
	$\gamma \in [1.8, 2.6]$	$\gamma \in [1.8, 2.6]$
	$w_R, w_B \in [0.90, 1.10]$	$w_R, w_B \in [0.90, 1.10]$
	$w_G \in [0.95, 1.05]$	$w_G \in [0.95, 1.05]$
Noise state W_t^{noise}	$q \in [50, 500]$	$q \in [50, 500]$
	$\sigma \in [0.002, 0.020]$	$\sigma \in [0.002, 0.020]$

Table 7 summarizes the active-factor parameter envelopes for the original controlled training and test sets. Their temporal trajectories differ across regimes: Smooth uses continuous variations, Abrupt introduces a discrete transition, and Propagated includes delayed cross-factor dependencies. Camera state refers to gain, response exponent, and white balance, rather than camera motion.

Here, α and a control illumination attenuation and spatial variation; g, γ, and $\mathbf{w}$ control camera response. Noise combines a signal-dependent Gaussian shot-noise approximation, with standard deviation $\sqrt{\max(S, 0)/q}$, and additive read noise with standard deviation σ. The balanced training set includes all seven nonempty factor subsets. Inactive illumination, camera, and noise factors use $(\alpha, a) = (1, 0)$, $(g, \gamma, \mathbf{w}) = (1, 2.2, \mathbf{1})$, and $(q, \sigma) = (10^8, 0)$, respectively. The latter is a near-noiseless setting.

Implementation Details. We use a temporal horizon of $r = 4$, factor-selection threshold $\gamma_A = 6 \times 10^{-5}$, training sparsity coefficient $\lambda_E = 10^{-3}$, and scene-invariance weight $\lambda_{\text{sc}} = 0.10$. The inference gate threshold is $\tau_g = 0.30$, and the propagation tolerance is $\rho = 0.0025$ under the archived score convention. The diffusion backbone is CogVideoX-2b. Its base weights and VAE are frozen, while rank-32 attention LoRA and the encoder, propagation, prototype, and conditioning modules are optimized.

Training uses 49-frame windows at 480×720 resolution on one NVIDIA RTX PRO 6000 GPU with 96-GB memory and BF16 precision. The per-device batch size is 2, with four gradient-accumulation steps, giving an effective batch size of 8. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-3}, and gradient clipping at 1. The learning rates for new modules and LoRA parameters are 10^{-4} and 5×10^{-5}, respectively; final PSNR-oriented refinement reduces them to 5×10^{-5} and 2.5×10^{-5}.

The reference configuration specifies stage budgets of 1,500 updates for encoder–graph pretraining, 4,000 for diffusion adaptation, and 500 for paired refinement. The separate PSNR-refinement configuration specifies a 1,000-update cosine-schedule horizon. Archived evaluations use the validation-selected `fine-200/checkpoint-200` checkpoint. These are configured budgets and a checkpoint-selection index, rather than a statement of the total executed training updates.

Inference uses 30-step DDIM reverse sampling with $\eta = 0$ and a 30-timestep factual inversion schedule. The default response screening uses $K = 2$ probe times, $\{0.35, 0.70\}$; these additional denoiser evaluations are included in the reported NFE. Short test clips are padded to the model window length, and padded targets are excluded from evaluation.

Unless explicitly varied in an ablation, internal comparisons use the same backbone, inversion procedure, sampling schedule, and input sequences. Same-checkpoint ablations additionally freeze all learned weights. Recomputed baseline comparisons use identical target frame IDs and metric implementations; method-specific resizing, padding, and temporal context are documented separately.

Table 8: Controlled-renderer parameter envelopes. Abrupt values are two levels, not continuous ranges. Propagated gain and noise follow the cross-factor equations below.

Parameter	Smooth	Abrupt initialization
Attenuation α	$[0.06, 0.45]$	$0.08, 0.42$
Spatial amplitude a	$[0, 0.25]$	smooth
Gain g	$[0.8, 2.0]$	$0.9, 1.9$
Exponent γ	$[1.8, 2.6]$	$1.9, 2.5$
White balance R/B	$[0.9, 1.1]$	smooth
White balance G	$[0.95, 1.05]$	smooth
Read scale σ	$[0.002, 0.020]$	$0.004, 0.018$
Shot parameter q	$[50, 500]$	derived

D Data, Configurations, and Evaluation Protocols

D.1 Datasets, Splits, and Evaluation Subsets

SDSD indoor and outdoor data are handled as separate domains, with source-sequence-level train/validation/test separation. Controlled variants inherit their clean source's split. The real-video evaluation pipeline uses registered low/normal-light pairs and records evaluated frame indices explicitly. BVI-RLV is an external zero-shot evaluation: no BVI-RLV adaptation or test-set checkpoint selection is part of the specified protocol. The specified SDSD test pipeline selects the first 30 frames per official test sequence. When predictions and references differ in resolution, the real-data evaluator explicitly bicubic-resizes the prediction to the native registered reference and records this choice.

Causal-V-SDSD uses normal-light SDSD frames as scene targets and generates three matched degradation trajectories per source. The frozen test set contains 10 indoor and 10 outdoor sources, each with 32 frames at 480×720, yielding 60 clips and 1,920 frame pairs. The construction seed is 2027. Eligible official-test sources and contiguous clip starts are selected deterministically by hashing. The three regimes share clean targets, but use separate degradation seeds. The selected indoor IDs are 13, 15, 23, 31, 33, 52, 58, 60, 68, and 70; the outdoor IDs are 1, 5, 36, 48, 49, 60, 62, 66, 75, and 76, using the release's `pair` identifiers.

The audited controlled training corpus contains 189 generated windows and 9,261 frame entries, with 63 windows per regime and 27 per nonempty factor subset. These are generated windows, not counts of independent source videos. Controlled validation contains 63 items. Training composition is constant within each window. The model processes 49-frame windows; shorter test clips are padded internally, with padded targets excluded from every reported metric.

Restoration, conditioning, scene-preservation, and unseen-schedule evaluations use 20 test sources. The reported explicit-response diagnostic uses one source, `indoor_pair13`, all three regimes, and targets $\{4, 16, 31\}$: nine matched packets in total. Repeated efficiency measurements use one common source with its three regimes, whereas quality and fallback coverage use all 20 sources.

D.2 Controlled Degradation Renderer

For clean RGB X_t, let L_t be its linearized sRGB value. With spatial coordinates $x, y \in [-1, 1]$ and $u_t = t/(F-1)$, define

$$F_t(x, y) = \tfrac{1}{2}\left[\sin(\pi x + 2\pi u_t + \psi) + \cos(\pi y - 2\pi u_t)\right], \tag{35}$$

$$A_t(x, y) = \alpha_t \max\{0.05, 1 + a_t F_t(x, y)\}, \tag{36}$$

$$S_t = L_t \odot A_t \odot g_t \mathbf{w}_t, \tag{37}$$

$$Y_t = \left[\mathrm{clip}_{[0,1]}\left(S_t + \sqrt{\max(S_t, 0)/q_t}\,\epsilon_t^{\mathrm{shot}} + \sigma_t \epsilon_t^{\mathrm{read}}\right)\right]^{1/\gamma_t}. \tag{38}$$

Both noise arrays are standard Gaussian. The signal-dependent term is a Gaussian shot-noise approximation. Rendered benchmark images are rounded to uint8 RGB; matched intervention-response metrics retain unquantized RGB values.

Smooth attenuation, gain, response exponent, and noise trajectories are sinusoids with a shared frequency $c \in [0.5, 2]$ cycles per window and factor-specific phase offsets. White balance uses c, while spatial amplitude uses $\max(0.5, c/2)$. Abrupt trajectories switch attenuation, gain, response exponent, and noise level once within the middle half of a clip; spatial amplitude and white balance retain smooth profiles. Propagated trajectories start from the abrupt initialization and add delayed illumination-to-camera and camera-to-noise mechanisms:

$$g_t = \mathrm{clip}_{[.8,2]}\left(.8 + 1.2[1 - \alpha_{t-\ell_{IC}}]\right), \tag{39}$$

$$\sigma_t = \mathrm{clip}_{[.002,.020]}\left(.002 + .018[g_{t-\ell_{CN}} - .8]/1.2\right), \tag{40}$$

$$q_t = 50 + 450\,\mathrm{clip}_{[0,1]}\left(1 - (\sigma_t - .002)/.018\right). \tag{41}$$

The two lags are sampled independently from $\{1, 2, 3, 4\}$ and remain constant within the original clip; unavailable parent positions retain their initialized values. Metadata records parameters, seeds, and generator support. Support annotations and measured intervention-response labels serve different evaluations: neither is automatically the cardinality-minimal oracle of the theoretical formulation. Inactive factors in balanced training and composition-schedule tests use $(\alpha, a) = (1, 0)$, $(g, \gamma, \mathbf{w}) = (1, 2.2, \mathbf{1})$, or $(q, \sigma) = (10^8, 0)$. The last is a near-noiseless convention, distinct from the soft noise-amplitude intervention.

D.3 Unseen-Dynamics Protocol

The training-support audit verifies all seven nonempty factor subsets, clip-constant compositions, and clip-constant cross-factor lags in the controlled training corpus. We therefore test three separately defined schedule shifts without retraining: (i) three cycles per window, outside the audited $[0.5, 2]$ rate range; (ii) four consecutive quarters using $IC \to CN \to IN \to ICN$; and (iii) both cross links switching from lag 1 to lag 4 halfway through the clip. Here I, C, N denote illumination, camera response, and noise. The latter two settings test unseen within-clip schedules, not unseen factor types or previously unobserved lag values. The three OOD cases share source-specific phases and noise draws, and use unchanged active parameter envelopes. Composition and lag schedules use one cycle per window. Rates are cycles per window, not physical Hz; novelty is established relative to the audited controlled corpus, not asserted for every real training video or the backbone's pretraining data.

Each case uses all 20 clean test sources. Its IRE protocol applies one soft joint intervention to all 15 candidate nodes at each target $t \in \{4, 16, 31\}$. Selected variants act on the intersection with their frozen selected set; empty intersections are retained. Structures remain fixed after fallback. IRE averages targets and then sources equally. This joint-window diagnostic is distinct from the single/joint query families of the explicit-response table.

D.4 Model and Comparison Configurations

The implementation uses CogVideoX-2b with frozen base and VAE weights, rank-32/alpha-32 attention LoRA, and trainable encoder, graph, prototype, and conditioning modules. The cached prompt is "high quality normal-light video." The single-GPU configuration uses BF16 on an NVIDIA RTX PRO 6000 with 96-GB memory, effective batch size 8, AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-3}, and gradient clipping at one. New-module/LoRA learning rates are $10^{-4}/(5 \times 10^{-5})$, reduced to $5 \times 10^{-5}/(2.5 \times 10^{-5})$ for final refinement. The controlled/real mixture changes from $1:1$ to $1:3$ during paired refinement. Actual stopping steps and checkpoint selection are distinct from the configured cosine-schedule horizon.

Joint-Code and Full-Graph are separately trained comparison models. Joint-Code replaces factorized codes and graph propagation with one 256-dimensional degradation code and its learned normal prototype. Full-Graph retains factorized conditioning but enables every valid edge and normalizes every valid factor slot at inference. Neither performs SPCD's response screening. In contrast, the conditioning and efficiency ablations modify the same frozen SPCD checkpoint: "All + Full" in these experiments is not the separately trained Full-Graph checkpoint.

External methods marked as unadapted use their original training protocols without Causal-V-SDSD fine-tuning. Recomputed predictions must correspond to the identical input and target frame IDs; any method-specific temporal context, padding, or spatial conversion belongs to its configuration

record. RGB-only variants must be identified by the actual executed path, rather than inferred from a multimodal model's total complexity.

D.5 Quality, Structure, and Response Metrics

Restoration and scene preservation. PSNR and SSIM are computed on aligned uint8 RGB with data range 255; LPIPS uses the VGG backbone with RGB normalized to $[-1, 1]$. The real-video evaluator reports an unweighted mean of per-clip means. For Causal-V-SDSD, metrics are computed per frame and averaged within clips, followed by equal source weighting; cross-regime summaries additionally weight regimes equally. GT-P and GT-L use the normal-light target. No implicit resize or crop is introduced by the controlled-benchmark evaluator. T-L averages framewise GT-LPIPS within the union of inclusive ± 2-frame windows around abrupt transitions, rather than comparing adjacent restored frames. CI-L compares outputs at the same physical frame under the three trajectories:

$$\text{CI–L} = \frac{1}{N_s} \sum_s \frac{1}{3} \sum_{\{g,g'\}} \frac{1}{F} \sum_t \text{LPIPS}(\widehat{X}_{s,g,t}, \widehat{X}_{s,g',t}). \tag{42}$$

The inner regime sum contains the three unordered pairs. Controlled benchmark confidence intervals use 10,000 paired source-level bootstrap resamples, seed 2027; the real-video evaluator resamples clips. Correlated frames are not treated as independent samples.

Factor and edge structure. Factor calibration reports the mean of positive- and negative-class F1. Structural gate calibration averages positive-class F1 for same-factor and cross-factor edges equally. The sensitivity table's Edge F1 instead pools TP, FP, and FN over valid post-fallback edge slots: $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$, with a zero denominator mapped to zero. These metrics use generator support labels. In conditioning ablations, $|A|$ counts prototype-replaced slots, not deleted graph nodes; $|E|$ counts enabled edges after fallback. Their means exclude padding and incomplete-window invalid slots. Random Sparse matches SPCD's effective per-target node and edge counts; Static retains lag-one same-factor edges, and Empty disables explicit graph messages while retaining the other conditioning pathways.

For each configuration, Δ_ϵ compares its prediction with the full-graph prediction under the *same* intervention set. We report the source/regime mean and the maximum over real targets using d_ϵ from Appendix B.4; displayed values are scaled by 10^5. Full-graph rows are zero by construction. Smaller discrepancy measures closer agreement with that reference, not necessarily better restoration or a more accurate physical graph.

Explicit intervention responses. Let y_t^a, y_t^0 be generator counterfactual/factual images and $\widehat{y}_t^a, \widehat{y}_t^0$ the matched model images in $[0, 1]$. For $D_{\text{RGB}} = 3HW$,

$$\text{IRE}(a, t) = \frac{\|(\widehat{y}_t^a - \widehat{y}_t^0) - (y_t^a - y_t^0)\|_2}{\sqrt{D_{\text{RGB}}}}. \tag{43}$$

CF-L is LPIPS between model and generator counterfactual outputs, not the clean target. At each of the three diagnostic targets we query all 15 single nodes and three joint sets: current illumination and camera; all current factors; and previous illumination with current noise. Single and joint IRE are averaged separately; CF-L weights these two query families equally before equal target, regime, and source aggregation.

Factor labels are positive when the generator response RMS exceeds $0.5/255$. Edge labels apply the same threshold to the response caused by suppressing one candidate edge under a fixed all-node soft intervention. All 90 admissible edges per packet, including absent or zero-response edges, remain in the evaluation. The single-source diagnostic therefore contains 135 single-node and 810 edge candidates. These are response labels, not a response-minimal structural oracle.

AP is non-interpolated average precision with tied scores grouped: $\text{AP} = \sum_j (P_j)(R_j - R_{j-1})$ over distinct descending score thresholds. We pool the targets and regimes within each source before computing AP, then average valid sources. Single-class source pools have undefined AP; positive/negative counts and valid-source coverage are reported. Spearman correlations use average ranks: r_s^{gen} compares model and generator response RMS, whereas r_s^{CF} compares an attribution score with the explicit model response RMS. The former averages valid regime correlations within

each source and then weights sources equally; the latter pools within each source before equal-source averaging. Constant-vector correlations are undefined and reported as N/A; Explicit CF is its own reference, not an independent perfect predictor. The zero-response baseline sets the model response difference to zero and uses the factual model output as every counterfactual prediction.

Attribution and score probes. Lagged Corr. averages the absolute channelwise correlation between factual factor-code RMS deviations from their prototypes and lag-aligned RGB intensity means, using only the factual prefix. It requires at least four pairs; undefined channels contribute zero. Gradient Attr. varies physical-node intervention strength u through both code interpolation and fractional adapter density:

$$G_{s,m} = \frac{\sum_k q_k |\eta\, \partial_{u_{s,m}} E_k(0)|}{\sum_k q_k}, \quad E_k(u) = \tfrac{1}{2}\,\mathrm{mean}\big(\epsilon_{k,g(t)}(u)^2\big). \tag{44}$$

Explicit CF ranks decoded model response RMS. Score probes use the epsilon squared-L_2 discrepancy, inverse-noise weights, and a $1/K$ average, with grids $\{.35\}$, $\{.35, .70\}$, and $\{.175, .35, .525, .70\}$ for $K = 1, 2, 4$. RGB frame zero maps to latent group zero; frame $t > 0$ maps to $g(t) = 1 + \lfloor (t-1)/4 \rfloor$. Each probe scores the entire target latent group, not an independently localized RGB-frame query.

Efficiency and sensitivity. NFE counts actual denoiser calls per model window, including inversion, selection, verification, and reverse sampling. Timing also includes encoding, VAE processing, conditioning, and decoding, but excludes model loading, disk I/O, and metric computation. We synchronize CUDA, perform one warmup per configuration, and take the median of three repeats per clip on the common profiling subset. Time is normalized by emitted real targets. Peak memory is allocated CUDA bytes divided by 10^9, not reserved memory. FB is the percentage of real targets affected by whole-clip fallback. The NFE-matched Full-Graph condition adjusts both inversion and reverse schedules to the same measured budget. Default $K_0 = 2$ and the $K = 2$ row are the same configuration; efficiency $K = 1$ uses time .70, unlike the explicit-response probe.

Threshold sensitivity varies γ_A or inference gate threshold τ_g one at a time, holding the checkpoint, ρ, selected sources, all three regimes, and sampling settings fixed. It does not vary the training sparsity coefficient λ_E. Parameter counts distinguish total from trainable parameters; FLOPs additionally require the stated spatial/temporal input, denoising budget, and counting convention. Transcribed complexity is not treated as a measurement of an unreported RGB-only inference path.