HARD, YET REDUCIBLE: CONTROLLED FORWARD TRANSFER FOR SYNTHETIC DEGRADATION CURATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting synthetic degradations for dense prediction requires an estimate of their training utility, the generalization gain they bring under a finite training budget. Clean and degraded twins share content and labels, suggesting a score based on how much short training reduces the excess error caused by degradation. However, this gap can also shrink when clean performance deteriorates. Measuring the improvement on degraded images alone avoids that confound, but it still credits progress that the same amount of clean training would have produced. We propose the **controlled Reducible Degradation Gap** (cRDG) for regions defined by degradation type and severity. From a common checkpoint, cRDG runs two budget-matched probes that differ only in one augmentation slot, which holds either a synthetic degradation or a clean augmentation. The score is the gain on held-out degraded images relative to the clean-control probe. Clean harm is a separate feasibility constraint. cRDG reveals a correctable severity band in which training on the degradation yields high controlled gain under the available budget, and the band moves with the predictor, the starting checkpoint, and the training budget. **Curation of Reducible Bands** (CURB) uses cRDG to select synthetic data without changing the predictor. On semantic segmentation and salient object detection, CURB improves representative predictors under matched synthetic-data budgets and training schedules, extends to existing data-generation pipelines, and preserves clean performance. Code and supporting materials will be publicly released.

1 INTRODUCTION

Generators can now produce unlimited variants of labelled data. The binding question is no longer how to synthesize but what to spend a finite training budget on (He et al., 2024; 2026a). Under a fixed schedule, every degraded sample occupies a slot that could hold a clean augmentation instead. We take this alternative as the reference, and the value of a degradation is the extra generalization from degradation.

Data selection often conflates *visual quality*, *instantaneous difficulty*, and *training utility*. The first two describe the current sample or the model's response to it, whereas training utility is the gain obtained by actually training on that sample. Utility is therefore the quantity relevant to selection, and it depends on the model state, the optimizer, and the training budget. We study this quantity for synthetic degradation in dense prediction, where degraded variants of clean labelled images keep scene content, geometry, and labels fixed (Rombach et al., 2022; Zhang et al., 2023; Rajagopalan et al., 2025).

Higher visual fidelity does not guarantee higher downstream utility, and recent studies attribute this mismatch to limited support and diversity in generated data (Adamkiewicz et al., 2026; Zhang et al., 2026). Within a plausible candidate pool, realism acts as a support constraint rather than a ranking criterion. Difficulty is also a poor indicator of utility, because a strong degradation can remove the evidence needed to recover the mask, leaving samples that are hard yet carry little learnable signal under the available budget. Selecting by loss or uncertainty (Mindermann et al., 2022; Askari-Hemmat et al., 2025) can then waste the budget or cause negative transfer.

We define the utility of a degradation as its held-out gain beyond equally long clean training under the same probe budget, subject to a clean-retention constraint. Forward transfer here means the gain on held-out source scenes that carry the same degradation after a short probe, and transfer to real adverse domains is evaluated separately (Sec. 4.4). Curation uses only source images and their synthetic twins, never target-domain observations. We group candidates into *regions* g defined by

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.

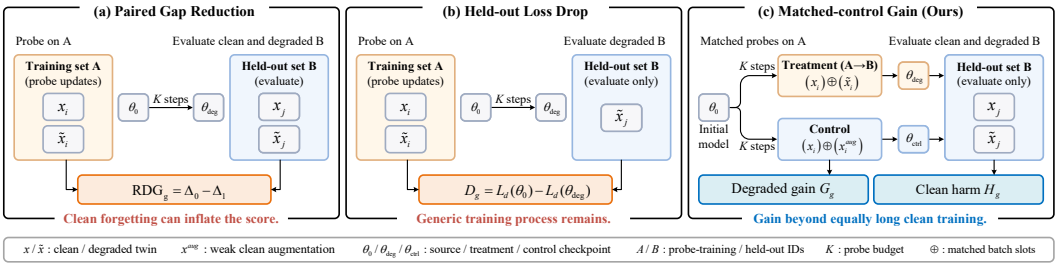


Figure 1: **Why Matched Clean Control Matters.** (a) Paired gap reduction can be inflated by clean forgetting. (b) Held-out degraded loss drop avoids this inflation but still credits generic training progress. (c) Our matched-control formulation underlying cRDG measures degraded gain G_g beyond equally long clean training, while checking clean harm H_g separately.

degradation type and severity bin, so the policy chooses kinds of degradation rather than individual scenes. Full retraining for every region would be costly, so short probes estimate held-out gains per region. Synthetic degradation also offers a natural control, since each valid degraded image has a clean twin with the same scene and label. Existing valuation scores, such as influence and gradient alignment (Pruthi et al., 2020; Zhu et al., 2024), holdout-loss prioritization (Mindermann et al., 2022), early-training statistics (Paul et al., 2021; Toneva et al., 2019), and learnability-guided generation (Chan-Santiago & Shah, 2026; Askari-Hemmat et al., 2025), do not combine this twin structure with a clean-control stream and held-out transfer at the region level.

This pairing suggests comparing degraded and clean loss for the same content. Let $L_d(\theta_t)$ and $L_c(\theta_t)$ be the mean losses on the degraded twins of a region and on their clean counterparts at checkpoint θ_t. The gap $\Delta_t = L_d(\theta_t) - L_c(\theta_t)$ is the excess error caused by the degradation. One could score a region by how much a short probe shrinks this gap:

$$\Delta_0 - \Delta_1 = \underbrace{[L_d(\theta_0) - L_d(\theta_1)]}_{\text{degraded loss falls}} - \underbrace{[L_c(\theta_0) - L_c(\theta_1)]}_{\text{clean loss falls}}. \tag{1}$$

The second bracket exposes the problem. A probe that forgets clean scenes also shrinks the gap, so clean forgetting inflates the score (Proposition 1). Dropping the clean term and scoring only the fall in degraded loss removes this inflation. It still credits progress that equally long clean training would have produced anyway. The quantity we need is the gain *beyond* that clean-control trajectory (Fig. 1).

We estimate this gain with matched treatment and clean-control probes. Both start from the same checkpoint and differ only in whether the designated augmentation slot contains a degradation or a clean augmentation. They are evaluated on held-out twins, and the two halves are swapped to obtain the **controlled Reducible Degradation Gap** (cRDG). Scores stay signed so negative transfer remains visible, and a separate clean-harm constraint excludes regions that damage clean performance.

cRDG shows that utility is not monotonic in severity. Each degradation type has a *correctable band* in which training on the degradation yields high controlled gain under the budget, and the severity centroid of this band shifts with probe budget K, starting checkpoint θ_0, and predictor setting, and differs across degradation types. Per-region retraining at three final-training budgets shows the same shift in real-target gains, and proportionally scaled probes track it (Sec. 4.4). A fixed middle-severity rule (Fixed-Mid) serves as the sanity check for this claim throughout the experiments.

Around cRDG we build **Curation of Reducible Bands** (CURB), a curation framework that adds degradation-aware selection to existing dense predictors without changing their architectures. Scores are recomputed for the operating checkpoint and budget. The band only summarizes the profile. Selection ranks feasible positive-score regions directly and fills a fixed synthetic budget in rank order, with validity gating before ranking and k-center sampling for the last partial region.

Our contributions are as follows.

- We formulate *budget-conditioned correctability*. Training utility need not track severity, realism, or difficulty, and the useful severity region migrates with predictor, checkpoint, and probe budget.
- We propose cRDG, a held-out, budget-matched estimate of the gain from training on a degradation region relative to an equally long clean-control probe, with clean harm handled separately.

- We introduce CuRB, a plug-and-play framework that adds degradation-aware curation to existing dense predictors without changing their architectures.
- Under a matched protocol, CuRB outperforms region-level adaptations of existing selection methods. Across SemSeg and SOD, its gains extend to the evaluated data-generation settings while preserving clean performance.

2 Related Work

Synthetic Degradation and Degradation-Aware Training. Diffusion-based methods synthesize degradation patterns and weather conditions (Rajagopalan et al., 2025; Nguyen et al., 2026). Restoration-driven methods (Guan et al., 2026) inject enhancement knowledge into downstream segmentation. Synthesis frameworks such as DGInStyle (Jia et al., 2024) and CA-LoRA (Park et al., 2026) generate labelled training images for segmentation. These approaches decide *how* training data are generated or used. We ask *which* synthetic degradations deserve a finite training budget. Higher fidelity alone does not guarantee training value. Adamkiewicz et al. (2026) link the mismatch to limited coverage and diversity, and Zhang et al. (2026) show that scene composition and instance fidelity matter for segmentation. We instead hold content and labels fixed and vary only degradation type and severity, which allows a controlled estimate of degradation-specific utility. CuRB is a curation layer between an existing generator and an existing predictor and changes neither. We evaluate it with ISSA (Li et al., 2024b), procedural weather corruption (Hendrycks & Dietterich, 2019), and Gen4Seg (Yin et al., 2026) under their original settings. UDA methods (Hoyer et al., 2022; Tranheden et al., 2021; Sakaridis et al., 2019) instead use unlabelled target images during adaptation. Our setting uses only source labels and synthetic twins during curation and training.

Data Valuation and Learnability-Guided Selection. Influence- and gradient-based methods (Pruthi et al., 2020; Zhu et al., 2024) estimate how much a sample contributes to learning. RHO-Loss (Mindermann et al., 2022) prioritizes examples by their effect on holdout loss. Early-training statistics such as EL2N, GraNd, and forgetting events (Paul et al., 2021; Toneva et al., 2019) find informative examples from short trajectories. Learnability-guided generation (Chan-Santiago & Shah, 2026; Askari-Hemmat et al., 2025) steers synthesis with learner feedback. Our setting differs in that utility is estimated at the region level using content-matched twins, a clean-control probe, and held-out transfer under a fixed budget (Sec. 3.3). Diffusion Curriculum (Liang et al., 2025) is closest in spirit and adapts the guidance level of generated data using validation feedback. It does not compare against an equally long clean-control probe, which is the counterfactual our score is built on.

3 Controlled Reducibility for Degradation Curation

CuRB selects synthetic degradations for a given predictor and training budget without changing the predictor. Fig. 2 shows the pipeline. Matched probes estimate a region-level utility, cRDG, and selection ranks regions by cRDG under a clean-harm constraint and a fixed synthetic-data budget.

3.1 Candidate Regions and Validity

We are given a labelled clean source set $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{N}$. A generator $\mathcal{G}$ produces degraded twins $\tilde{x}_{i,d,s} = \mathcal{G}(x_i; d, s, z)$ with degradation type d, severity level s, and stochastic detail z. We use four appearance degradations that preserve object geometry after validity filtering. Low light and backlight cover illumination changes, and fog and rain cover weather effects. Each family is divided into eight ordered severity levels, giving 32 regions. This discretization resolves severity-dependent effects while keeping the probe set small. A *degradation region* $g = (d, s)$ contains candidates from one family and one severity bin, where "region" denotes a group in the candidate pool rather than a spatial image region. CuRB ranks regions, not individual scenes. The region definition is shared across architectures. Its estimated utility is recomputed for each model and budget.

Validity gate. Severe degradation does not by itself invalidate the mask. We therefore separate label validity from reducibility. A binary gate rejects candidates only when generation changes scene structure. It checks frozen DINO feature correspondence near ground-truth boundaries and passes a uniformly darker image with intact geometry. Details and the gate audit are in App. B.

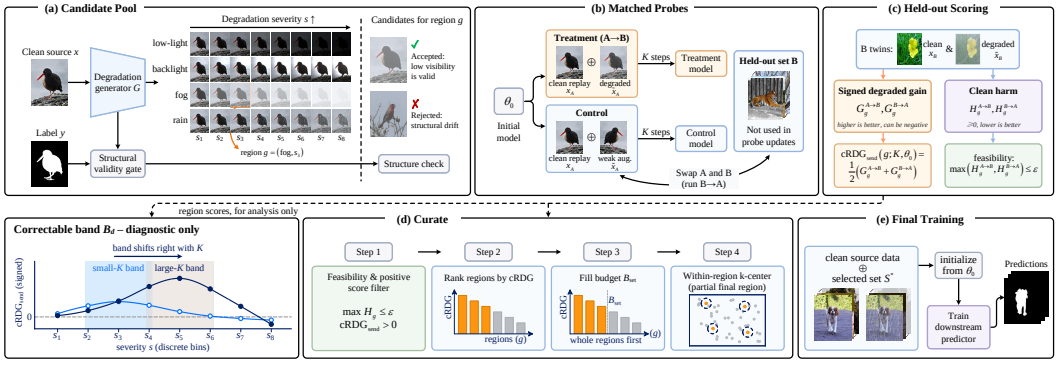


Figure 2: **Overview of CuRB.** (a) Degradation-severity regions are generated and validity-filtered. (b) Two matched probes start from θ_0 and differ only in the augmentation slot, where $\oplus$ denotes the union of the clean-replay stream and the slot stream. (c) Held-out evaluation yields degraded gain G_g and clean harm H_g. Averaging G_g over the two split directions gives cRDG. H_g determines feasibility, while the correctable band is diagnostic only. (d) Feasible positive-score regions are ranked and fill the budget B_{sel} in rank order, with k-center sampling for the last partial region. (e) Final training restarts from θ_0 on clean data, with $\mathcal{S}^\star$ in the synthetic slots under the fixed schedule.

The twin gap. Write $L_d^S(\theta; g)$ for the mean loss on the degraded twins of an index set S and $L_c^S(\theta)$ for the mean loss on their clean images. At the source checkpoint θ_0, the gap $\Delta_0(g) = L_d^S(\theta_0; g) - L_c^S(\theta_0)$ isolates the error caused by the degradation, since twins share scene, geometry, and label. It measures what is not yet mastered, rather than what can be mastered within a given budget.

3.2 Matched-Control Probes

Consider a short probe $\theta_0 \to \theta_1$ on region g, scored by the shrinkage $\Delta_0 - \Delta_1$. By Eq. (1) this equals the drop in degraded loss minus the drop in clean loss, so it rewards clean forgetting:

Proposition 1 (Gap reduction can reward clean forgetting) $\Delta_0 - \Delta_1 > 0$ *can hold even when the degraded loss worsens,* $L_d(\theta_1) > L_d(\theta_0)$*, whenever* $L_c(\theta_1) - L_c(\theta_0) > L_d(\theta_1) - L_d(\theta_0) > 0$.

Scoring the degraded-loss drop alone removes this inflation but still credits progress that equally long clean training would produce. We therefore compare each degradation probe with an equally long clean-control probe. We report the uncontrolled score as the baseline RDG, defined together with cRDG in Sec. 3.3.

To measure transfer rather than fit to the probe samples, we split source image IDs once into halves A and B, train on one half, and evaluate on the other (cross-fitting). The held-out half is excluded from the current probe update, although it may have contributed to training θ_0. The split is reused for every region. Let $U(\theta, \mathcal{S}; K)$ denote K optimizer steps from θ on a data stream $\mathcal{S}$, and let $\oplus$ denote the union of the clean-replay stream and the augmentation-slot stream in each probe batch. From the common start θ_0 we run two probes on half A,

$$\theta_{A,g}^{\text{deg}} = U\big(\theta_0,\ A_{\text{clean}} \oplus A_{\text{deg},g};\ K\big), \qquad \theta_A^{\text{ctrl}} = U\big(\theta_0,\ A_{\text{clean}} \oplus A_{\text{aug}};\ K\big). \tag{2}$$

The treatment probe fills the slot with twins from region g. The clean-control probe fills the same slot with a weak photometric augmentation of the same base view (App. C). Both probes share source IDs, steps, batch composition, clean-replay ratio, optimizer, schedule, data order, and initialization. The control does not depend on g, so it is run once per direction and shared across all regions. If no valid twin exists after n_z regeneration draws, the control view is kept in the treatment slot, so the match is never broken. Held-out twins that fail the gate are dropped from both evaluations.

Operating rule for K. We set K to $1/80$ of the final training schedule and scale it with the schedule, so the probe-to-training budget ratio stays fixed across operating points.

3.3 MATCHED-CONTROL REDUCIBILITY

Both probes are evaluated on the held-out half B. We define the *degraded gain* and the *clean harm* as

$$G_g^{A\to B} = L_d^B(\theta_A^{\text{ctrl}}; g) - L_d^B(\theta_{A,g}^{\text{deg}}; g), \qquad H_g^{A\to B} = \Big[L_c^B(\theta_{A,g}^{\text{deg}}) - L_c^B(\theta_A^{\text{ctrl}})\Big]_+. \tag{3}$$

G is positive when degradation training outperforms the matched clean control on held-out degraded images. H is positive when the treatment increases the held-out clean loss. We then swap the two halves and average the two held-out gains,

$$\text{cRDG}(g; K, \theta_0) = \tfrac{1}{2}\big(G_g^{A\to B} + G_g^{B\to A}\big). \tag{4}$$

A region is *feasible* iff $\max\big(H_g^{A\to B}, H_g^{B\to A}\big) \le \epsilon$. Infeasible regions are excluded from selection, and signed scores remain available for analysis. We use a hard constraint rather than a penalty $G_g - \lambda H_g$ because it enforces clean retention without another trade-off parameter. The constraint bounds probe-time harm and does not guarantee final clean accuracy, which Sec. 4.2 checks empirically.

Baselines from the same probes. The same held-out losses define two baselines. The paired gap reduction is $\text{RDG}_g^{A\to B} = [L_d^B(\theta_0; g) - L_d^B(\theta_{A,g}^{\text{deg}}; g)] - [L_c^B(\theta_0) - L_c^B(\theta_{A,g}^{\text{deg}})]$, and the held-out loss drop is $D_g^{A\to B} = L_d^B(\theta_0; g) - L_d^B(\theta_{A,g}^{\text{deg}}; g)$. Like G_g, both are averaged over the directions $A \to B$ and $B \to A$. D_g differs from cRDG only in what is subtracted, the initial loss instead of the clean-control loss, so it is the direct ablation of the control branch, whereas RDG additionally credits clean forgetting. cRDG is a short-horizon, source-side estimate of controlled degradation utility. Its relation to full-training gains on real target domains is evaluated empirically in Sec. 4.4.

3.4 CURATION AND FINAL TRAINING

Selection. CURB first removes infeasible and non-positive regions. The remaining regions are ranked by cRDG and filled in descending order until B_{sel} is reached. Each region is exhausted before moving to the next. If only part of the final region is needed, we use feature-space k-center sampling. Any unfilled slots revert to clean data, so the total number of training samples is unchanged.

Final objective. The curation procedure can be used with different final-training objectives, and all matched-curation methods share the one below. Starting again from θ_0, we train on clean data plus $\mathcal{S}^\star$ with

$$\mathcal{L} = \mathcal{L}_{\text{seg}}(f(x_i), y_i) + \lambda_s\, \mathcal{L}_{\text{seg}}(f(\tilde{x}_i), y_i) + \lambda_c \sum_p w_i(p)\, \text{KL}\big(f(x_i)_p \,\|\, f(\tilde{x}_i)_p\big). \tag{5}$$

Here $w_i(p) = M_i(p)\,(1 + \lambda_b B_{y_i}(p))$, $B_{y_i}(p)$ marks a ground-truth boundary band, and $M_i(p)$ is the cosine similarity between the clean DINOv2 feature at p and the degraded feature at its matched location q_p, mapped linearly to $[0, 1]$. The KL term compares class-probability maps. Matched-curation experiments use Eq. (5) unchanged, and plug-in experiments keep each host's own objective.

Cost. With M regions and R split directions, curation costs MR treatment probes plus R shared controls. Changing the predictor or the schedule requires recomputing cRDG but never changes the architecture. Sec. 4.2 reports compute matched between CURB and the baselines.

3.5 THE DIAGNOSTIC SEVERITY BAND

cRDG depends on the probe budget and the checkpoint, not on severity alone. For a type d with bins $s_1 < \cdots < s_{N_s}$, let $\mathcal{F}_d$ contain the feasible bins with positive cRDG. The *correctable band* is the set of bins whose score is at least half of the family maximum,

$$\mathcal{B}_d(K, \theta_0) = \big\{s \in \mathcal{F}_d : \text{cRDG}\big((d, s); K, \theta_0\big) \ge \kappa \max_{s' \in \mathcal{F}_d} \text{cRDG}\big((d, s'); K, \theta_0\big)\big\}, \tag{6}$$

with $\kappa = 1/2$. For smoother comparison across operating conditions, we also report the cRDG-weighted severity centroid $\bar{s}_d(K, \theta_0) = \sum_{s\in\mathcal{F}_d} s\, \text{cRDG}((d, s); K, \theta_0) \big/ \sum_{s\in\mathcal{F}_d} \text{cRDG}((d, s); K, \theta_0)$, which is undefined when $\mathcal{F}_d$ is empty. Severity is represented by normalized bin centers $s_k = (2k-1)/16$, $k = 1, \ldots, 8$. Both are diagnostics, and selection ranks regions directly by cRDG. The **band-migration** hypothesis is that $\bar{s}_d$ shifts with the probe budget K, the checkpoint, and the predictor, and differs across types. Sec. 4.4 tests whether the real band shifts with the training budget and whether scaled-K cRDG tracks it.

Method	$M \downarrow$	$F_\beta^{\max} \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
WFANet (Chen et al., 2025)	.016	.908	.953	.925
+ Uncurated twins (B_{sel})	.015	.912	.956	.926
+ CuRB (B_{sel})	**.013**	**.923**	**.963**	**.933**
CSSAM (Cong et al., 2026)	.011	.937	.972	.940
+ Uncurated twins (B_{sel})	.011	.939	.975	.942
+ CuRB (B_{sel})	**.009**	**.952**	**.987**	**.947**
NUN (He et al., 2026b)	.014	.920	.965	.933
+ Uncurated twins (B_{sel})	.013	.924	.968	.935
+ CuRB (B_{sel})	**.011**	**.934**	**.978**	**.938**

(a) WXSOD-real (Chen et al., 2025).

Method	$M \downarrow$	$F_\beta^{\max} \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
SI-EDN (Li et al., 2024a)	.062	.817	.841	.810
+ Uncurated twins (B_{sel})	.058	.835	.858	.817
+ CuRB (B_{sel})	**.052**	**.853**	**.873**	**.831**
CSSAM (Cong et al., 2026)	.035	.887	.916	.886
+ Uncurated twins (B_{sel})	.034	.891	.919	.888
+ CuRB (B_{sel})	**.030**	**.912**	**.934**	**.893**
NUN (He et al., 2026b)	.038	.857	.896	.875
+ Uncurated twins (B_{sel})	.036	.867	.902	.877
+ CuRB (B_{sel})	**.032**	**.880**	**.917**	**.886**

(b) CSOD10K (Cong et al., 2026).

Table 1: **Evaluation on SOD Benchmarks** following the official protocol (Sec. 4.1). At the same augmentation budget, comparing uncurated sampling with CuRB isolates the gain from curation.

Data-generation host	Published host setting	ACDC mIoU ↑		Dark Zurich mIoU ↑		Cityscapes
		Original	+CuRB	Original	+CuRB	val Δ ↑
single-generator hosts						
ISSA (Li et al., 2024b)	SegFormer	52.45	**53.92**	27.39	**28.76**	+0.24
H-Weather (Hendrycks & Dietterich, 2019)	SegFormer	49.21	**50.63**	23.44	**24.65**	+0.36
Gen4Seg (Yin et al., 2026)	SegFormer	51.04	**52.39**	25.63	**27.28**	+0.45
composite DG host						
ISSA (Li et al., 2024b)	RobustNet (DeepLabV3+)	47.55	**48.73**	23.09	**24.35**	+0.18

Table 2: **CuRB as a plug-in to existing data-generation settings.** Published references for ISSA, H-Weather, and RobustNet+ISSA are taken from Li et al. (2024b). Gen4Seg follows Yin et al. (2026) (its Dark Zurich value is obtained with the released pipeline). Our "+CuRB" runs apply CuRB under the corresponding host setting and synthetic-data budget. The composite host stacks RobustNet (Choi et al., 2021) with ISSA on DeepLabV3+. Host-specific region construction: App. D.

4 Experiments

4.1 Setup

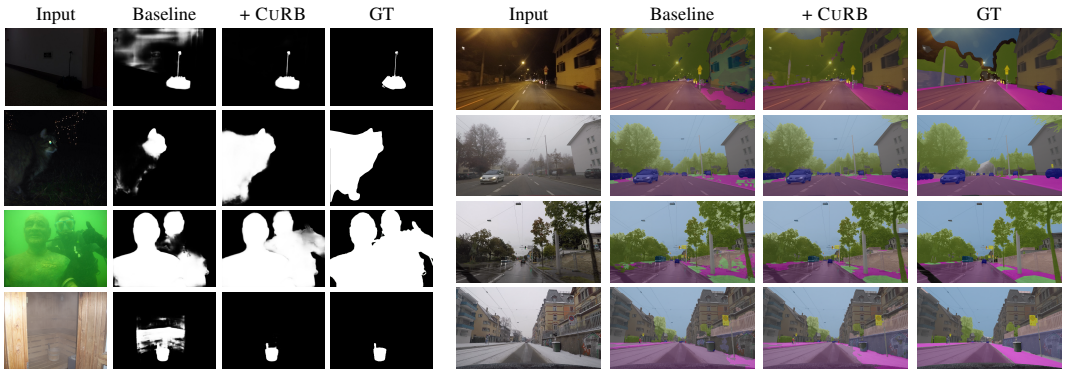
Data and protocols. We use two core evaluation protocols. For SOD, WXSOD (Chen et al., 2025) and CSOD10K (Cong et al., 2026) follow their official splits and metrics, with each base model using its published training recipe. Augmented variants use the same budget of degraded twins from clean DUTS-TR (Wang et al., 2017) and the same final-training schedule (App. E). For SemSeg, we follow the Cityscapes→ACDC / Dark Zurich protocol of ISSA (Li et al., 2024b), also used by Gen4Seg (Yin et al., 2026). Cityscapes (Cordts et al., 2016) is the only source, and SegFormer follows the released ISSA configuration (App. A). The matched synthetic pool contains low light, backlight, fog, and rain at eight severity levels per family. Low light is aligned with the night condition, while fog and rain have direct counterparts in ACDC (Sakaridis et al., 2021). Backlight is an additional source-side illumination family, and snow is held out entirely for unseen-family evaluation. We report the in-family mean over fog, night, and rain, and evaluation also includes Dark Zurich (Sakaridis et al., 2019) and Cityscapes val. Plug-in experiments apply CuRB under each host's published setting and synthetic-data budget. Operational curation uses one fixed probe seed and one fixed $A|B$ partition. Two additional partitions are used only for stability analysis (App. C).

Comparison groups. *Benchmark baselines*: WFANet, CSSAM, and NUN on WXSOD, and SI-EDN, CSSAM, and NUN on CSOD10K. *Data-generation hosts*: ISSA (Li et al., 2024b) with SegFormer and RobustNet (Choi et al., 2021) / DeepLabV3+ (Chen et al., 2018), Hendrycks-Weather (Hendrycks & Dietterich, 2019), and Gen4Seg (Yin et al., 2026). *Selection baselines*: realism (NR-IQA / CLIPScore), highest loss / uncertainty, RHO-Loss-style selection (Mindermann et al., 2022), gradient contribution (Zhu et al., 2024), and a DisCL-style severity preference (Liang et al., 2025) (App. G). *Twin-based scores*: Δ_0, uncontrolled RDG, held-out loss drop D_g, and cRDG. **Fixed-Mid**, the middle severity bins of every family, is a sanity control. Source-only and All-gen. are no-ranking anchors. Full-pool exposure appears only in the compute analysis.

Matched protocol. All selectors use the same gated pool, B_{sel}, clean/degraded composition, initialization θ_0, objective, and training schedule. Sample-level baseline scores are averaged within each region before the common k-center step (App. G), so the comparison isolates region ranking. Compute-matched baselines spend CuRB's curation budget on extended final training (Sec. 4.2).

Method	ACDC in-family			in-fam.	Snow	Dark	Cityscapes
	Fog	Night	Rain	mean	(held out)	Zurich	val
Source-only	60.61	28.42	50.35	46.46	48.79	24.08	67.84
matched curation protocol: same pool, gate, B_{sel}, objective, and schedule							
All-gen. (uniform, B_{sel})	63.82	31.51	52.78	49.37	50.15	26.55	68.21
Realism (NR-IQA / CLIPScore)	63.34	30.97	53.02	49.11	50.31	26.12	68.44
Highest loss / uncertainty	64.05	31.88	52.41	49.45	49.97	26.71	67.93
RHO-Loss style (Mindermann et al., 2022)	64.19	32.24	52.96	49.80	50.28	27.06	68.31
Grad. contrib. (Zhu et al., 2024)	64.73	31.85	53.42	50.00	50.55	26.84	68.58
DisCL-style selection (Liang et al., 2025)	64.48	32.36	53.27	50.04	50.79	27.02	68.40
RDG (uncontrolled)	65.11	32.29	53.76	50.39	50.47	27.45	67.41
Fixed-Mid	65.02	32.91	53.58	50.50	50.68	27.33	68.87
CuRB (cRDG)	**66.73**	**34.26**	**54.93**	**51.97**	**51.38**	**28.75**	**70.13**

Table 3: **Controlled curation comparison on Cityscapes→ACDC / Dark Zurich.** All methods share the candidate pool, validity gate, B_{sel}, final-training objective, and schedule, differing in their sample-selection policies. Fog, night, and rain are treated as the in-family conditions. Snow is held out as an unseen family, with no snow samples entering the candidate pool, curation, or final training.



(a) SOD (b) SemSeg

Figure 3: **Qualitative comparison.** (a) Degraded SOD with NUN. (b) adverse-condition SemSeg with ISSA SegFormer. Baseline and +CuRB predictions are shown for both tasks.

4.2 Main Results

Benchmark-level performance. Tab. 1 attaches CuRB to three published base models on SOD benchmarks, including the SAM-based CSSAM, under corresponding protocols. The uncurated row adds the same budget of synthetic twins without selection and brings only small gains, with almost none on the strongest base model. Selecting the same budget with cRDG improves all four metrics on every base model and both benchmarks. The margin over uncurated twins remains positive throughout, including for CSSAM, where uncurated data provide little benefit.

Across data-generation frameworks. Tab. 2 evaluates CuRB across several published data-generation settings at their original synthetic-data budgets. The "+CuRB" variants obtain $+1.2$ to $+1.5$ higher ACDC mIoU and $+1.2$ to $+1.7$ higher Dark Zurich mIoU than the corresponding originals, while maintaining or slightly improving Cityscapes performance. The gains span style synthesis (ISSA), procedural weather corruption (H-Weather), and diffusion-based editing (Gen4Seg), and persist for the composite RobustNet+ISSA host. The same insertion also improves SOD baselines in Tab. 1. Qualitative examples for both tasks and hosts are shown in Fig. 3.

Controlled data-selection comparison. Tab. 3 compares selection criteria under one shared recipe, so the differences reflect what is selected. CuRB is best on every column. It reaches 51.97 in-family mIoU versus 50.50 for the strongest control and 50.04 for the strongest literature baseline, 28.75 on Dark Zurich, and 51.38 on held-out snow, where improvement reflects cross-family transfer. Cityscapes performance is also preserved, reaching 70.13 versus 67.84 for source-only. The uncontrolled twin score is the only selector below the source-only model. Sec. 4.4 examines the ranking quality behind these differences. Published augmentation and adaptation results are not mixed into this table because their training exposure differs from our held-out-family protocol. Some use synthetic snow or target images. Tab. 2 instead compares CuRB within each host's native setting.

Variant	in-fam. mIoU↑	City val↑	ρ_{glob} vs. U_{real} ↑
CuRB (full)	**51.97**	**70.13**	**0.79**
paired gap reduction (RDG)	50.39	67.41	0.61
held-out loss drop D_g	51.02	68.72	0.70
in-sample (no cross-fitting)	49.21	69.58	0.44
scored on training twins (no held-out)	50.51	69.85	0.58
w/o clean-harm constraint (H free)	50.68	67.18	0.79

Table 4: **Core ablations** (Cityscapes→ACDC). RDG and D_g replace cRDG by the paired gap reduction and the held-out loss drop. Other rows remove cross-fitting, held-out transfer, and the clean-harm constraint. ρ_{glob} is over all 32 regions.

Methods	$M \downarrow$	$F_\beta^{max} \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
Source-only	.021	.892	.960	.920
All-gen. (uniform, B_{sel})	.022	.889	.958	.918
Grad. contrib. (Zhu et al., 2024)	.022	.890	.959	.919
RDG (uncontr.)	.023	.885	.955	.914
Fixed-Mid	.021	.891	.960	.919
cRDG, H unconstrained	.025	.879	.950	.908
CuRB	.021	.893	.961	.921

Table 5: **Clean retention on SOD** (DUTS-TE, NUN/PVTv2-B2). Source-only trains on clean DUTS-TR. Augmented variants add the same synthetic budget selected by each method, so any drop is the clean cost of that selection.

Metric	Adapted valuation baselines			Twin-based scores			
	Highest loss / uncertainty	RHO-Loss style (Mindermann et al., 2022)	Grad. contrib. (Zhu et al., 2024)	Δ_0 (gap)	RDG (uncontrolled)	D_g (held-out drop)	CuRB (cRDG)
ρ_{glob} with U_{real} ↑	0.25	0.49	0.52	0.39	0.61	0.70	**0.79**
ρ_{macro} with U_{real} ↑	0.17	0.35	0.37	0.27	0.43	0.56	**0.74**
ρ_{DZ} with U_{DZ} ↑	0.21	0.43	0.45	0.33	0.53	0.62	**0.72**
in-fam. mIoU ↑	49.45	49.80	50.00	49.28	50.39	51.02	**51.97**

Table 6: **Region-level score comparison** under matched pool, gate, diversity, budget, objective, and schedule (Cityscapes→ACDC, ISSA SegFormer). ρ_{glob} is Spearman correlation with U_{real} over all 32 regions, ρ_{macro} averages within-family correlations, and ρ_{DZ} evaluates the same ranking against per-region retraining gains on Dark Zurich. Δ_0 is the unmastered gap and D_g the held-out loss drop.

Matched compute. With $K = 1/80$ of the final schedule, the 66 probes cost $0.59\times$ one final-training run, or $1.59\times$ with final training. Candidate generation, held-out scoring, and selection are excluded. At matched probe-plus-training compute, All-gen. and Fixed-Mid reach 49.86 and 51.02 in-family mIoU, versus 51.97 for CuRB. Full-pool training at $3.1\times$ reaches only 50.02 (App. F).

4.3 Ablation Study

The ablations isolate the main components of cRDG (Tab. 4). Replacing cRDG with paired gap reduction lowers ρ_{glob} from 0.79 to 0.61, while removing only the clean-control subtraction (D_g) lowers it to 0.70, showing the value of both the matched control and the controlled difference. In-sample estimation and training-twin scoring further weaken the ranking, supporting held-out cross-fitted evaluation. Removing the H constraint leaves ρ_{glob} unchanged at 0.79 but drops Cityscapes accuracy from 70.13 to 67.18, indicating that reducibility and clean feasibility capture different aspects of the selection problem. Tab. 5 shows the same effect on SOD: CuRB preserves clean DUTS-TE performance, whereas the H-free variant falls on every metric, including S_α (.921 → .908).

4.4 Analysis: Training Utility and the Correctable Band

Does cRDG predict real utility? We first obtain an empirical retraining utility. Holding everything but the region fixed, we retrain once per region g and record the *signed*, unnormalized gain on ACDC in-family, $U_{\text{real}}(g) = \text{mIoU}_{\text{real}}(M_g) - \text{mIoU}_{\text{real}}(M_{\text{clean}})$. Fig. 4a overlays it with Δ_0, RDG, and cRDG for one family. We test whether RDG remains high when real utility becomes negative and whether cRDG preserves the ordering of U_{real}, including regions with negative utility. The decision unit and the correlation unit are both the *region* (per-sample real gain would need one retraining per sample). A global rank correlation can be inflated by mean differences *between* families, so Tab. 6 reports ρ_{glob} over all 32 regions and $\rho_{\text{macro}} = \frac{1}{|\mathcal{D}|}\sum_d \rho_d$, where ρ_d is the Spearman correlation over the 8 regions of family d. Both use the operational ranking. Because cRDG is computed on source twins only, the table also correlates the same ranking with U_{DZ}, the per-region retraining gain on Dark Zurich. Correlation with a second real target supports transfer beyond the benchmark used to define U_{real}. The held-out loss drop D_g is the no-subtraction counterpart, and cRDG improves ρ_{glob} from 0.70 to 0.79 and in-family mIoU from 51.02 to 51.97 over it, and ρ_{DZ} from 0.62 to 0.72.

How does the band move? Fig. 4c examines variation with probe budget K, source checkpoint θ_0, and predictor setting. Doubling K shifts the centroid by about 0.10, well beyond the 0.03–0.04

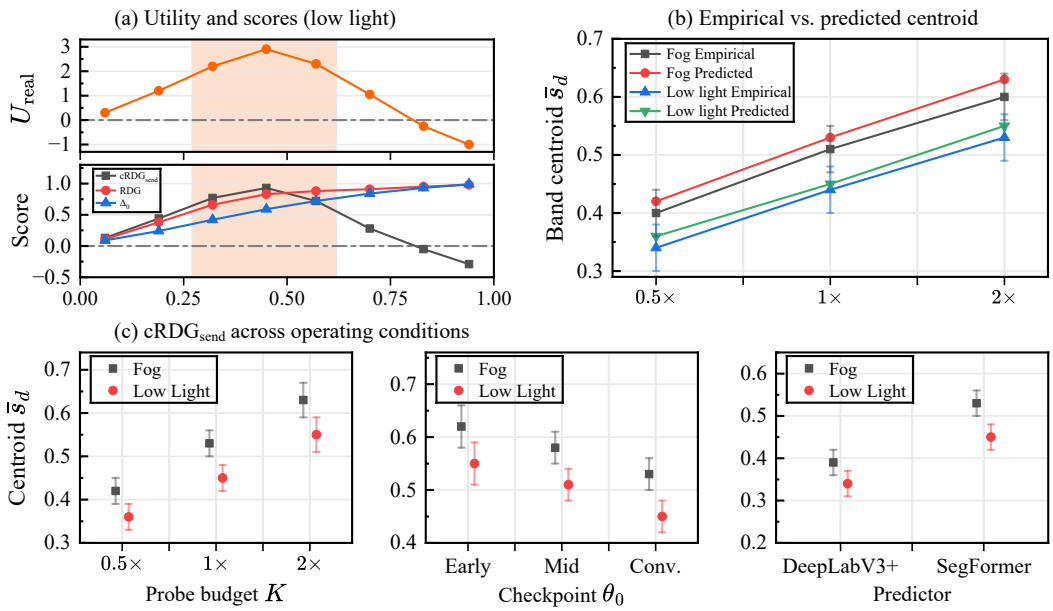


Figure 4: **Training utility and the correctable band.** (a) Per-bin retraining gains (top) and scores (bottom) for low light. Shading marks the predicted band. (b) Empirical severity centroids at three final budgets with scaled-K cRDG predictions overlaid (paired-bootstrap 95% CI). (c) cRDG centroids across probe budgets, checkpoints, and predictors. Error bars are s.d. over three partitions.

variation across partitions. The centroid also changes across families and predictor settings, and mid-training checkpoints are about 0.05 higher than converged ones. Panel (b) asks whether the useful severity region changes with the final-training budget rather than remaining fixed. At each budget, the empirical severity centroid is computed over the probe-feasible regions, replacing cRDG by positive U_{real}, using per-bin retraining at $0.5\times$, $1\times$, and $2\times$ the final budget. Within those regions, its centroid shifts by $+0.20$ for fog and $+0.19$ for low light from $0.5\times$ to $2\times$, while the scaled-K cRDG prediction remains within 0.03 of the empirical centroid at every budget.

5 Discussion and Limitations

The twin-based comparisons isolate two different confounds. Paired gap reduction is affected by clean forgetting, whereas D_g still includes progress obtainable from clean training. Subtracting the matched clean-control trajectory gives cRDG, which best tracks the signed retraining utility in our experiments. The correctable band depends on the learner and the budget rather than on severity alone, and it shifts with the final budget as the scaled-K score predicts. The procedure transfers across the tested predictors unchanged, whereas the scores are recomputed for each model.

Limitations. cRDG is an empirical surrogate for the full-training gain relative to clean-control training, tested at three budgets on one backbone. Residual drift near the validity threshold may introduce mild label noise despite the structural gate. Unseen families such as snow benefit only through cross-family transfer. The probe horizon follows a fixed budget-scaling rule rather than adapting to each region. The SOD twin pool comes from external clean images. Also, all newly trained models are single runs, and we do not report seed variance.

6 Conclusion

We study how to curate synthetic degradations under a finite training budget. cRDG compares degradation training with an equally long clean-control probe on held-out twins and treats clean harm separately. Using this score, CuRB selects a fixed-budget training set and improves SOD and semantic segmentation across the tested predictors and generation settings. We also find that the useful severity range changes with model state and training budget instead of remaining fixed.

AI USE STATEMENT

Generative AI was used in two roles. First, the synthetic degradation pipeline uses the diffusion model described in App. B as part of the experimental method. Second, LLM-based tools assisted with critique of manuscript drafts, discussion of experimental design and result interpretation, manuscript editing and LaTeX formatting, and the design and refinement of schematic figures. LLM tools did not execute experiments or generate or alter quantitative results. All AI-assisted suggestions, text, and figures were reviewed by the authors, who take responsibility for the content.

ETHICS STATEMENT

This work uses only publicly available academic datasets, *e.g.*, Cityscapes, ACDC, Dark Zurich, and CSOD10K, under their respective licenses. No new human-subject data or annotations were collected. The synthetic degradations are derived from clean source images and are intended for robustness research. We do not release a generator tuned to imitate any real individual, place, or sensor.

REPRODUCIBILITY STATEMENT

The estimator and the curation procedure are specified in Secs. 3.3 and 3.4 and Alg. 1. Implementation details, the probe protocol, the generator and validity gate, benchmark protocols, and cost accounting are given in Apps. A to F. Code and curation manifests will be released.

REFERENCES

Krzysztof Adamkiewicz, Brian B Moser, Stanislav Frolov, Tobias Christian Nauen, Federico Raue, and Andreas Dengel. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. In *CVPR*, pp. 36660–36669, 2026.

Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. *ICML*, 2025.

Jeffrey A Chan-Santiago and Mubarak Shah. Learnability-guided diffusion for dataset distillation. In *CVPR*, pp. 41657–41666, 2026.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pp. 833–851. Springer, 2018.

Quan Chen, Xiong Yang, Bolun Zheng, Rongfeng Lu, Xiaokai Yang, Qianyu Zhang, Yu Liu, and Xiaofei Zhou. Wxsod: A benchmark for robust salient object detection in adverse weather conditions. *arXiv preprint arXiv:2508.12250*, 2025.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, pp. 11575–11585. IEEE, 2021.

Runmin Cong, Zhiyang Chen, Hao Fang, Sam Kwong, and Wei Zhang. Breaking barriers, localizing saliency: A large-scale benchmark and baseline for condition-constrained salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2026.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Kai Guan, Rongyuan Wu, Shuai Li, Wentao Zhu, Wenjun Zeng, and Lei Zhang. Restoration adaptation for semantic segmentation on low quality images. *Int. J. Comput. Vis.*, 134(5):229, 2026.

Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *ICLR*, 2024.

Chunming He, Rihan Zhang, Fengyang Xiao, Dingming Zhang, Zhiwen Cao, and Sina Farsiu. Refining context-entangled content segmentation via curriculum selection and anti-curriculum promotion. *ICML*, 2026a.

Chunming He, Rihan Zhang, Dingming Zhang, Fengyang Xiao, Deng-Ping Fan, and Sina Farsiu. Nested unfolding network for real-world concealed object segmentation. *ECCV*, 2026b.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, pp. 9914–9925. IEEE, 2022.

Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *ECCV*, pp. 91–109. Springer, 2024.

Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Runmin Cong, Xiaochun Cao, and Qingming Huang. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection. *ICML*, 2024a.

Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-& extra-source exemplar-based style synthesis for improved domain generalization. *Int. J. Comput. Vis.*, 132(2):446–465, 2024b.

Yijun Liang, Shweta Bhardwaj, and Tianyi Zhou. Diffusion curriculum: Synthetic-to-real data curriculum via image-guided diffusion. In *ICCV*, pp. 1697–1707. IEEE, 2025.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *ICML*, pp. 15630–15649, 2022.

Thang-Anh-Quan Nguyen, Moussab Bennehar, Luis Guillermo Roldao Jimenez, Nathan Piasco, Dzmitry Tsishkou, Laurent Caraffa, Jean-Philippe Tarel, and Roland Brémond. Cyclone: Diffusion model for cycle-consistent weather editing from unpaired driving data. *arXiv preprint arXiv:2607.13927*, 2026.

Minho Park, Sunghyun Park, Jungsoo Lee, Hyojin Park, Kyuwoong Hwang, Fatih Porikli, Jaegul Choo, and Sungha Choi. Concept-aware lora for domain-aligned segmentation dataset generation. In *CVPR*, pp. 39858–39868, 2026.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *NeurIPS*, pp. 20596–20607, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *NeurIPS*, pp. 19920–19930, 2020.

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N Paranjape, and Vishal M Patel. Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration. In *CVPR*, pp. 28144–28154, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685, 2022.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pp. 7374–7383, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pp. 10745–10755. IEEE, 2021.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *ICLR*, 2019.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, pp. 1378–1388. IEEE, 2021.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pp. 3796–3805. IEEE, 2017.

Zijin Yin, Bing Li, Kongming Liang, Hao Sun, Zhongjiang He, Zhanyu Ma, and Jun Guo. Benchmarking semantic segmentation models via appearance and geometry attribute editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jinjin Zhang, Xiefan Guo, Yizhou Jin, Nan Zhou, and Di Huang. What makes synthetic data effective in image segmentation. *ICML*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3813–3824. IEEE, 2023.

Muzhi Zhu, Chengxiang Fan, Hao Chen, Yang Liu, Weian Mao, Xiaogang Xu, and Chunhua Shen. Generative active learning for long-tailed instance segmentation. *ICML*, 2024.

A Implementation Details

Backbones and optimization. SemSeg follows the released ISSA configuration (Li et al., 2024b) for SegFormer. The DeepLabV3+ experiments follow the official RobustNet recipe (Choi et al., 2021). SOD clean-retention runs use NUN on a PVTv2-B2 backbone with its released training configuration. CuRB leaves the host optimizer, input resolution, augmentation, and final-training schedule unchanged in every experiment. All runs use a single NVIDIA RTX PRO 6000 GPU with 96 GB of memory.

Curation budget and probe settings. For the Cityscapes experiments, $B_{\text{sel}} = 5{,}950$ synthetic samples are filled in rank order over 32 regions (four families, eight severity bins), with a 1:1 clean/degraded slot ratio. Each region starts from $2{,}975 \times 2$ candidates before validity filtering. Probes optimize the first two terms of Eq. (5) at the same slot ratio and $0.25\times$ the base learning rate, for K equal to $1/80$ of the corresponding final-training schedule. Plug-in experiments keep each host's synthetic budget, and cRDG is recomputed per host. Under this budget, rank-greedy filling is concentrated. The top-ranked region contributes over 90% of $\mathcal{S}^\star$ and the remainder comes from the next-ranked region. For WXSOD and CSOD10K the synthetic budget is filled entirely from the top-ranked feasible region, so k-center controls diversity within that region.

Thresholds. λ_s=1.0, λ_c=0.5, λ_b=1.0 in Eq. (5). The clean-retention tolerance ϵ is set to one standard error of the clean-harm estimate from degradation-free matched probes on a held-out source split, where the standard error is taken over the per-image loss differences on the held-out images. The rule is fixed across experiments, and the value is estimated once for each task and predictor on its own source data. We use κ=1/2 for the diagnostic band and n_z=4 regeneration draws in the validity fallback. Nothing is tuned on a target set.

Evaluation protocol. SemSeg follows the ISSA evaluation protocol on Cityscapes val, the four ACDC conditions, and Dark Zurich. SOD benchmarks use their official metrics. Source-only and M_{clean} are obtained from the same θ_0 with the same final-training schedule, with the synthetic slot replaced by clean data. Unless otherwise noted, newly trained models are reported from single runs, and published reference values are taken from the cited sources. Operational curation uses one fixed probe seed and one fixed $A|B$ partition. Two additional $A|B$ partitions serve only the partition-stability analysis and the error bars of Fig. 4c and never enter curation (App. C).

B Generator and Validity Gate

Structure-preserving generation. We use Stable Diffusion 2.1 in image-to-image mode with depth- and Canny-conditioned ControlNet guidance together with the source image, but not the ground-truth mask, which avoids mask-aligned synthesis artifacts. We instantiate low light, backlight, fog, and rain. Each family contains eight ordered severity bins obtained by uniformly sampling its family-specific degradation-strength parameter, with two stochastic draws per source image. The strength variable of each family and its eight values are provided in the released generation configuration.

Validity gate. The gate rejects candidates whose structure has drifted. Given $(x, \tilde{x}, y)$ it measures frozen DINOv2 ViT-B/14 feature correspondence and boundary displacement in pixels within a band around ground-truth boundaries. For SOD it additionally checks that the salient object is neither deleted nor invented. It is a local structural test rather than a global appearance distance, so a uniformly darker or foggier image with intact geometry passes. Thresholds are set to the 99th percentile of the corresponding statistics on clean source pairs under geometric re-augmentation, after compensating for the known geometric transform, and reused across datasets. In a manual audit of 500 candidates per family (Tab. S1), acceptance is 98–100% at mild levels and remains 90–96% at severe levels, so the gate does not simply remove the stronger degradations from the pool.

C Probe Protocol and Estimator Details

Probe matching. Operational curation uses one fixed random split of source image IDs into equal halves A and B, shared across all regions. For each region g and split direction, let $S \in \{A, B\}$ denote the probe-update half. The treatment branch trains on $S_{\text{clean}} \oplus S_{\text{deg},g}$ and the control branch on $S_{\text{clean}} \oplus S_{\text{aug}}$, both from θ_0. The two branches share K, batch composition, clean-replay ratio,

Algorithm 1 CURB: controlled curation by reducibility

Require: clean source set $\mathcal{D}_s$, generator $\mathcal{G}$, checkpoint θ_0, probe budget $K = \frac{1}{80}$ of the final schedule, tolerance ϵ, sample budget B_{sel}

split source IDs once into $A|B$ and generate gated degradation regions $g = (d, s)$
$\theta_A^{\text{ctrl}} \leftarrow U(\theta_0, A_{\text{clean}} \oplus A_{\text{aug}}; K)$ and $\theta_B^{\text{ctrl}} \leftarrow U(\theta_0, B_{\text{clean}} \oplus B_{\text{aug}}; K)$ ▷ shared controls
for all regions $g = (d, s)$ **do**
    $\theta_{A,g}^{\text{deg}} \leftarrow U(\theta_0, A_{\text{clean}} \oplus A_{\text{deg},g}; K)$ and likewise $\theta_{B,g}^{\text{deg}}$
    compute $G_g^{A\to B}, H_g^{A\to B}$ on B and $G_g^{B\to A}, H_g^{B\to A}$ on A
    $\text{cRDG}(g; K, \theta_0) \leftarrow \frac{1}{2}(G_g^{A\to B} + G_g^{B\to A})$
    $\text{feasible}(g) \leftarrow [\max(H_g^{A\to B}, H_g^{B\to A}) \le \epsilon]$
end for
rank feasible regions with positive $\text{cRDG}(g; K, \theta_0)$ and fill B_{sel} in descending order, exhausting each region before moving to the next
use k-center when only part of a region is needed
use clean data for any unfilled budget and obtain $\mathcal{S}^\star$
train from θ_0 on clean data $\cup\, \mathcal{S}^\star$ with Eq. (5)

Degradation	gate prec.	gate recall	accept. s_1–s_4	accept. s_5–s_8
Low light	0.94	0.91	0.99	0.95
Backlight	0.92	0.89	0.99	0.92
Fog	0.95	0.92	1.00	0.96
Rain	0.91	0.88	0.98	0.90

Tab. S1: **Validity-gate audit** on 500 sampled candidates per family. Structural drift is the positive class for gate precision and recall. Acceptance remains high at severe levels (0.90–0.96), indicating that the gate does not simply remove stronger degradations.

optimizer and schedule, geometric views, and data order. The region-independent control is run once per direction and shared across regions. One fixed probe seed is used operationally. Two additional $A|B$ partitions serve only to assess stability and provide the error bars of Fig. 4c, and probe horizons of $0.5\times$ and $2\times$ the default serve only the sensitivity analysis. Alg. 1 gives the full procedure. The *training-twin* variant scores a region by the loss drop on the treatment probe's own degraded training twins rather than by held-out transfer. The *in-sample* variant keeps the controlled gain G_g but evaluates both branches on the probe-update half instead of the held-out half.

Matched views and validity fallback. Treatment and control views share the crop and flip of each source image. The control applies weak photometric augmentation (colour jitter, gamma in $[0.8, 1.25]$, and Gaussian blur with $\sigma \le 0.6$), while the treatment uses a generated twin from region g. Twins that fail the gate are resampled up to $n_z{=}4$ times. If no valid training twin is found, the control view is retained in that slot, and invalid evaluation twins are dropped from both branches.

Statistical reporting. cRDG is reported signed rather than rectified, preserving negative transfer. Aggregate statistics use paired bootstrap over regions. The bootstrap of Fig. 4b resamples ACDC evaluation images with replacement within each condition, using matched indices for every retrained model and the clean model. Checkpoints are kept fixed, and U_{real}, band membership, and the centroid are recomputed per replicate. Error bars in Fig. 4c are standard deviations over the three partitions.

Sensitivity to ϵ. Tab. S2 varies the clean-retention tolerance around its default of one standard error. Degraded-domain performance is stable over 0.5–2 standard errors, while a looser tolerance admits more regions at the cost of a small drop in clean accuracy on Cityscapes val.

D PLUG-AND-PLAY FRAMEWORK COMPARISONS

Protocol. Tab. 2 follows each host's published pipeline and synthetic-data budget. Published references for ISSA, H-Weather, and RobustNet+ISSA are taken from Li et al. (2024b). Gen4Seg

ϵ	0	0.5 SE	1 SE (default)	2 SE
feasible-region fraction	0.41	0.63	0.78	0.91
in-family mIoU ↑	51.43	51.86	**51.97**	51.79
Cityscapes mIoU ↑	70.21	70.24	70.13	69.82

Tab. S2: **Sensitivity to the clean-retention tolerance ϵ.** Degraded-domain performance is stable around the default, while loosening the constraint permits a small clean-performance drop.

follows Yin et al. (2026), with its Dark Zurich value obtained using the released pipeline. For our "+CuRB" runs, CuRB is applied under the corresponding host setting and synthetic-data budget, with cRDG recomputed for each host.

Region construction. Regions are defined by family and severity bin, with eight bins per family unless the host provides fewer discrete levels. ISSA (Li et al., 2024b) uses style-exemplar groups and style-mixing strength. Hendrycks-Weather (Hendrycks & Dietterich, 2019) uses corruption type and its predefined levels 1–5. Gen4Seg (Yin et al., 2026) uses weather type and a family-specific image-space statistic relative to the source (contrast attenuation for fog, luminance ratio for night, high-frequency energy gain for rain, bright-region coverage for snow), with prompt strength not used for region assignment. RobustNet + ISSA uses the ISSA regions with cRDG recomputed on the RobustNet model.

E SOD Benchmark Protocols

WXSOD (Chen et al., 2025) and CSOD10K (Cong et al., 2026) use their official splits and metrics, with each base model following its published training recipe. The augmented variants add twins synthesized from clean DUTS-TR with the four families and eight severity levels of App. B, including its validity gate. The uncurated and CuRB variants use the same synthetic budget and the base model's final-training schedule. The former samples uniformly, and the latter selects with cRDG. We use $B_{\text{sel}} = 4{,}000$ on WXSOD and 3,750 on CSOD10K.

F Compute and Cost Accounting

For the reported SegFormer configuration, the 66 probe updates cost approximately $0.59\times$ one final-training run under the $K = T/80$ scaling rule. Candidate-pool construction, held-out scoring, and selection are not included, and final training is accounted separately in Sec. 4.2.

Quantity	Value
Regions M (four families $\times$ eight severity bins)	32
Split directions R	2
Probe horizon K	$1/80$ of the final schedule
Treatment probes (MR)	64
Shared control probes (R)	2
Probe-step cost / final-step cost (measured)	≈ 0.71
Probe-update cost / one final run	$0.59\times$
Control overhead vs. uncontrolled probes ($2/64$)	3.1%

Tab. S3: **Curation cost.** Operational selection uses one $A|B$ partition. The region-independent control adds only two probes, and the normalized probe-update cost is $(66/80) \times 0.71 \approx 0.59$ of one final-training run, under the $K = T/80$ scaling rule.

G Baseline Adaptation

All selection baselines rank the same 32 gated regions and use the same B_{sel} and within-region k-center step. Candidate-level scores are averaged within each region before ranking, and the final training recipe is identical for every method. RHO-Loss-style selection uses the reducible loss of

a candidate relative to a small model trained on held-out source data. Gradient contribution uses the alignment of a candidate's gradient with the mean source-validation gradient. The DisCL-style baseline is converted to a static region score. Its curriculum rule is run on the source-validation response of the current model to obtain a severity preference per family, which is then used to rank regions once before final training rather than to alter the training schedule.