ESCAPING THE TEXTURE TRAP: INTERVENTION-GUIDED ADAPTIVE RGB–GEOMETRY FUSION FOR INDOOR 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust indoor RGB–geometry segmentation can fail when misleading appearance conflicts with valid geometric evidence. We introduce an intervention-guided framework that couples geometry-preserving texture interventions with input-adaptive fusion. Structured Texture Swapping (STS) creates paired clean and intervened observations while preserving labels and point correspondence. A Geometry–Texture Interaction Gated Unit (GT-IGU) predicts point- and channel-wise routing weights from ordered cross-modal evidence. Candidate-Oracle Routing Distillation (CORD) further supervises GT-IGU using label-aware, candidate-relative routing targets. Structure-Aware Texture-Swapping Invariance (SA-TSI) regularizes intervention-induced feature changes over compatible neighborhoods. All teacher and graph operations are training-only. On S3DIS and ScanNetV2, the system improves affected-region performance under synthetic texture conflicts across both backbones. On the reported appearance, geometry, and joint shifts, it improves full-scene mIoU over the evaluated RGB–geometry baselines. Ablations identify paired intervention–adaptive fusion as the primary driver, with CORD and SA-TSI providing further gains at near-baseline inference cost.

1 INTRODUCTION

Three-dimensional semantic understanding supports applications including autonomous driving (Geiger et al., 2012; Mei et al., 2022; Caesar et al., 2020), robotics (Yeshwanth et al., 2023; Grauman et al., 2024; Cadena et al., 2016), and augmented reality. We study robust indoor RGB–geometry 3D semantic segmentation, where complementary appearance and geometric cues are fused through image-to-point aggregation, bidirectional interaction (Jaritz et al., 2019; Hu et al., 2021), or projection-based, unified-space, and Transformer architectures (Vora et al., 2020; Liu et al., 2023; Bai et al., 2022). Despite strong clean-data performance, these models may over-rely on appearance when texture becomes unreliable because of illumination changes, misalignment, sensor artifacts, or misleading surface patterns. We target two coupled failures: dataset-level dependence on texture shortcuts and instance-level misallocation of fusion weights under appearance shifts.

Texture bias arises when semantic categories co-occur with characteristic appearances, encouraging models to exploit texture–label shortcuts (Geirhos et al., 2019). Such shortcuts become harmful when appearance contradicts otherwise valid geometry. We introduce Structured Texture Swapping (STS), which constructs controlled clean/intervened pairs by changing RGB while preserving geometry, point correspondence, and semantic supervision. These pairs expose localized cross-modal conflicts and provide controlled supervision for learning how fusion should adapt when appearance becomes unreliable.

Using these paired observations, we develop a Geometry–Texture Interaction Gated Unit (GT-IGU) that predicts point- and channel-wise fusion weights from ordered cross-modal evidence. GT-IGU is jointly optimized on clean and intervened observations and requires neither intervention masks nor corruption identities at inference. Because segmentation objectives provide only indirect supervision for relative modality allocation, Candidate-Oracle Routing Distillation (CORD) evaluates alterna-

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.

tive fusion hypotheses through their clean and intervened semantic risks and distills a point-wise geometry-preference target into the channel-mean output of GT-IGU.

Texture replacement may also introduce abrupt feature residuals near intervention boundaries. We therefore propose Structure-Aware Texture-Swapping Invariance (SA-TSI), which anchors affected fused representations to their clean references and regularizes intervention-induced changes over spatially, geometrically, and semantically compatible neighborhoods (Belkin et al., 2006; Shuman et al., 2013; Defferrard et al., 2016). Together, STS and GT-IGU form the core intervention-guided adaptive fusion framework, while CORD and SA-TSI provide task-grounded preference supervision and structure-aware regularization, respectively. All auxiliary teacher and graph operations are removed at inference.

Our contributions are threefold:

- We formulate texture-robust RGB–geometry segmentation as jointly exposing appearance shortcuts and learning input-dependent modality allocation, and introduce STS to construct paired geometry-preserving texture interventions.
- We develop GT-IGU for point- and channel-wise adaptive fusion, CORD for task-grounded supervision of its mean modality preference, and SA-TSI for structure-aware regularization of intervention-induced feature changes.
- Experiments across two datasets and two backbones show robust gains under structured appearance conflicts, external material projection, and RGB corruptions, together with cross-corruption gains under geometry-side and joint degradation.

2 Related Work

RGB–geometry segmentation and robust fusion. Robust fusion addresses unreliable or dominant modalities with uncertainty estimates (Ma et al., 2023; Gao et al., 2024), modality dropout (Neverova et al., 2016; Ma et al., 2022), and conflict-aware interaction (Yang et al., 2021b; Liao et al., 2025; Deregnaucourt et al., 2025; Cai et al., 2025; Zheng et al., 2025). Expert routing and teacher–student transfer offer related tools (Shazeer et al., 2017; Park et al., 2025; Hinton et al., 2015; Tarvainen & Valpola, 2017; Tan et al., 2026). CORD instead derives point-wise, label-aware geometry preferences from fixed mixtures evaluated on paired texture interventions.

Data augmentation and intervention-inspired learning. Counterfactual and intervention objectives have been explored in 2D vision (Wang et al., 2020; Niu et al., 2021; Sauer & Geiger, 2021; Pitis et al., 2020). In 3D, ST3D/ST3D++ align domains (Yang et al., 2021a; 2023), whereas PolarMix, LaserMix, and QuadMix diversify point or region mixtures (Xiao et al., 2022; Kong et al., 2023; Zhang et al., 2025). STS instead creates appearance conflicts while preserving geometry, correspondence, and labels.

Geometry-aware representation regularization. PointNeXt and PTv3 provide strong geometry encoders (Qian et al., 2022; Wu et al., 2024). Generic Laplacian or consistency regularization may blur semantic boundaries across incompatible neighborhoods (Belkin et al., 2006; Shuman et al., 2013; Tarvainen & Valpola, 2017). STS creates local appearance discontinuities, motivating SA-TSI's spatial, normal, and semantic compatibility weights.

3 Methodology

As illustrated in Fig. 1, our framework couples structured texture interventions with adaptive RGB–geometry fusion. Structured Texture Swapping (STS) creates paired clean and texture-intervened observations while preserving geometry, labels, and point correspondences. The Geometry–Texture Interaction Gated Unit (GT-IGU) uses ordered cross-modal evidence to predict point- and channel-wise routing weights. Candidate-Oracle Routing Distillation (CORD) evaluates a fixed bank of scalar geometry–texture mixtures with a label-aware EMA teacher under a paired semantic-risk objective and uses the resulting candidate-relative soft oracle to supervise GT-IGU's channel-mean geometry-routing coefficient. Structure-Aware Texture-Swapping Invariance (SA-TSI) regularizes intervention-induced changes in the fused representation over compatible neighborhoods. At inference, the model retains only the modality encoders, GT-IGU, and the post-fusion prediction network.

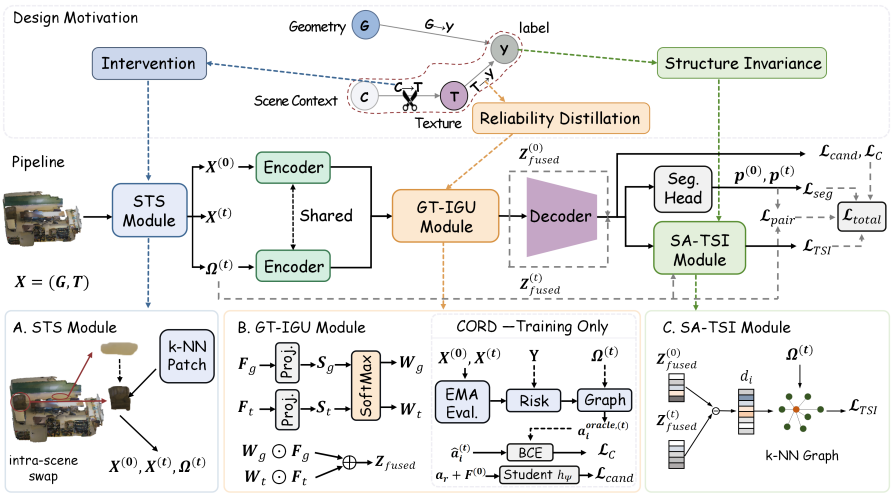

Figure 1: Overview of the proposed framework. STS generates paired texture interventions, GT-IGU performs adaptive fusion, CORD supervises its mean geometry preference, and SA-TSI regularizes fused-feature changes.

3.1 Problem Formulation

We consider point-wise semantic segmentation of a point cloud $\mathcal{P} = (G, T, Y)$, where $G \in \mathbb{R}^{N \times d_g}$ and $T \in \mathbb{R}^{N \times d_t}$ denote geometry and RGB observations, and $Y = \{y_i\}_{i=1}^{N}$ contains labels from C classes. From a clean observation $\mathcal{X}^{(0)} = (G, T)$, STS constructs a paired intervention $\mathcal{X}^{(t)} = (G, \widetilde{T})$ that preserves geometry, labels, and point correspondences. We denote its affected point set by $\Omega^{(t)}$. For $v \in \{0, t\}$, let $f_i^{g,(v)}, f_i^{t,(v)} \in \mathbb{R}^D$ denote the aligned modality features. The clean and intervened branches share the modality encoders, GT-IGU, and post-fusion prediction network. They are jointly optimized using semantic supervision and paired consistency, while CORD provides additional supervision only for the point-wise channel mean of the geometry weight.

3.2 Structured Texture Swapping

STS replaces a local texture patch while preserving the underlying geometry and target labels. In each scene, we sample distinct target and source categories and construct category-constrained neighborhoods Ω_{tgt} and Ω_{src}. After independently centering and scaling their coordinates, each target point receives the RGB value of its nearest source point:

$$\pi(i) = \arg\min_{j \in \Omega_{\text{src}}} \left\| \bar{g}_i^{\text{tgt}} - \bar{g}_j^{\text{src}} \right\|_2, \qquad \widetilde{t}_i = \begin{cases} t_{\pi(i)}, & i \in \Omega^{(t)}, \\ t_i, & \text{otherwise}, \end{cases} \quad \Omega^{(t)} = \Omega_{\text{tgt}}. \tag{1}$$

Let $m_i = \mathbb{I}[i \in \Omega^{(t)}]$ denote the intervention mask. This places source-category appearance on target-category geometry while preserving target labels and point correspondence. The student predicts a class distribution $p_i^{(v)}$ for each $v \in \{0, t\}$. Both clean and intervened observations receive direct semantic supervision through the average point-wise cross-entropy loss, denoted by $\mathcal{L}_{\text{seg}}$. We additionally encourage the intervened prediction to remain consistent with its clean reference:

$$\mathcal{L}_{\text{pair}} = \frac{1}{Z_{\text{pair}}} \sum_{i=1}^{N} w_i D_{\text{KL}}\left(\text{sg}[p_i^{(0)}] \,\middle\|\, p_i^{(t)} \right), w_i = m_i + \gamma(1 - m_i) \tag{2}$$

where $Z_{\text{pair}} = \sum_i w_i$, $p_i^{(v)}$ is the student class distribution, sg denotes stop-gradient, and $0 \le \gamma < 1$ assigns a smaller weight to unchanged context points. The stop-gradient affects only $\mathcal{L}_{\text{pair}}$ and makes the clean prediction its reference; it does not remove the direct semantic supervision on either branch.

3.3 GT-IGU: Conflict-Aware Geometry–Texture Fusion

At inference, GT-IGU predicts fusion weights from the modality features alone, without labels, intervention masks, or corruption identities. Given layer-normalized features $\bar{F}^g$ and $\bar{F}^t$, define

$$\begin{aligned} q(X,Y) &= \mathrm{Concat}\big(X, Y, X-Y, (X-Y)^2, X \odot Y\big), \\ E^g &= q(\bar{F}^g, \bar{F}^t), \qquad E^t = q(\bar{F}^t, \bar{F}^g). \end{aligned} \tag{3}$$

The ordered inputs retain modality identity, while their difference and product encode cross-modal disagreement and agreement. A shared estimator r_ϕ then produces channel-wise weights and the fused representation:

$$\begin{aligned} [W^g, W^t] &= \mathrm{Softmax}_{\mathrm{mod}}\big([r_\phi(E^g), r_\phi(E^t)]\big), \\ Z^{\mathrm{fused}} &= W^g \odot F^g + W^t \odot F^t. \end{aligned} \tag{4}$$

Sharing r_ϕ induces exchange-consistent scoring. This equation borrows the normalized competitive form of inverse-variance fusion; r_ϕ is trained discriminatively and is not a variance estimate. We therefore interpret W^g and W^t as routing coefficients.

GT-IGU is applied to the clean and intervened branches with shared parameters. The module itself receives neither intervention masks nor corruption identities; such information is used only to construct and weight training-time objectives. Consequently, the paired semantic and consistency objectives train the gate to respond to cross-modal disagreement without requiring intervention metadata at inference. For $v \in \{0, t\}$, we define its point-wise channel-mean geometry preference as $\widehat{a}_i^{(v)} = D^{-1} \sum_{d=1}^{D} W_{i,d}^{g,(v)}$. CORD applies additional supervision only to $\widehat{a}_i^{(t)}$ within the affected region, while channel-wise deviations remain task-driven.

3.4 Candidate-Oracle Routing Distillation

Because segmentation loss only indirectly supervises modality allocation, CORD uses a label-aware EMA teacher to evaluate scalar geometry–texture candidates on paired observations. The resulting point-wise task risks induce a candidate-relative soft oracle for GT-IGU's channel-mean geometry-routing coefficient. Here, candidate oracle denotes the soft target induced by label-aware task risks over the fixed bank $\mathcal{A}$; we refer to this task-grounded, candidate-relative signal as reliability supervision. Let $(\bar{\theta}, \bar{\psi})$ denote the EMA parameters of the modality encoders and the complete post-fusion prediction network $h_{\bar{\psi}}$, respectively; GT-IGU itself is not copied into the teacher. The teacher is updated after each student step by exponential moving average.

We use $\mathcal{A} = \{a_r\}_{r=1}^{M_a} = \{0, 0.25, 0.5, 0.75, 1\}$. To reduce the distribution mismatch between GT-IGU outputs and the scalar candidates, each batch exposes the student post-fusion network to one uniformly sampled clean candidate. The encoder features are detached, so the resulting $\mathcal{L}_{\mathrm{cand}}$ updates only the post-fusion decoder and prediction head.

For each $v \in \{0, t\}$, the EMA teacher evaluates every candidate through the complete post-fusion network:

$$\ell_{i,r}^{(v)} = \mathrm{CE}\Big(h_{\bar{\psi}}\Big(a_r F_{\bar{\theta}}^{g,(v)} + (1-a_r) F_{\bar{\theta}}^{t,(v)}\Big)_i, y_i\Big). \tag{5}$$

We combine intervened risk, clean risk, and positive intervention-induced degradation to obtain a candidate-relative soft oracle:

$$\begin{aligned} c_{i,r}^{(t)} &= \ell_{i,r}^{(t)} + \lambda_{\mathrm{clean}} \ell_{i,r}^{(0)} + \eta [\ell_{i,r}^{(t)} - \ell_{i,r}^{(0)}]_+, \\ \pi_{i,:}^{(t)} &= \mathrm{Softmax}_r(-c_{i,:}^{(t)}/\tau), \qquad \widetilde{a}_i^{(t)} = \sum_r \pi_{i,r}^{(t)} a_r. \end{aligned} \tag{6}$$

Here, $[x]_+ = \max(0, x)$, $\lambda_{\mathrm{clean}}, \eta \geq 0$, and $\tau > 0$. The raw target $\widetilde{a}_i^{(t)}$ is the expected geometry coefficient under the candidate distribution. We assign it a confidence weight $\chi_i^{(t)} \in [0, 1]$ that decreases with candidate entropy and absolute candidate cost.

Within $\Omega^{(t)}$, we refine the raw preference target using a confidence-anchored graph whose affinities encode spatial proximity and surface compatibility. This operation suppresses isolated candidate decisions while preserving geometric boundaries. The resulting target is denoted by $\mathbf{a}^{\mathrm{oracle},(t)}$;

the complete optimization objective, affinity definition, and numerical solver are provided in Appendix B. The resulting oracle is detached and supervises GT-IGU through:

$$\mathcal{L}_{\mathrm{C}} = \frac{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)}\, \mathrm{BCE}\Big(\widehat{a}_i^{(t)}, \mathrm{sg}[a_i^{\mathrm{oracle},(t)}]\Big)}{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)} + \epsilon}. \tag{7}$$

Although each candidate uses a scene-level scalar mixture, its point-wise task losses provide localized preference targets for supervising GT-IGU's heterogeneous routing field.

3.5 Structure-Aware Texture-Swapping Invariance

Semantic and paired-prediction objectives do not directly constrain how fused representations change within and around texture-intervened regions. Let $z_{i,\mathrm{fuse}}^{(0)}$ and $z_{i,\mathrm{fuse}}^{(t)}$ denote the clean and intervened GT-IGU fused features, respectively. For $v \in \{0, t\}$, we define $\bar{z}_{i,\mathrm{fuse}}^{(v)} = z_{i,\mathrm{fuse}}^{(v)}/(\|z_{i,\mathrm{fuse}}^{(v)}\|_2 + \epsilon)$ and the intervention residual $d_i = \bar{z}_{i,\mathrm{fuse}}^{(t)} - \mathrm{sg}[\bar{z}_{i,\mathrm{fuse}}^{(0)}]$. The graph node set $\mathcal{V}_{\mathrm{S}}^{(t)} = \Omega^{(t)} \cup \mathcal{N}_1(\Omega^{(t)})$ contains the swapped region and its one-hop geometric neighborhood. We construct $\mathcal{G}_{\mathrm{S}}^{(t)} = (\mathcal{V}_{\mathrm{S}}^{(t)}, \mathcal{E}_{\mathrm{S}}^{(t)})$ by retaining edges for which at least one endpoint lies in $\Omega^{(t)}$. The fixed weights $A_{\mathrm{S},ij}^{(t)}$ encode semantic agreement, spatial proximity, and surface-normal compatibility. With $Z_{\mathrm{p}} = \sum_i m_i + \epsilon$ and $Z_{\mathrm{e}} = \sum_{(i,j)\in\mathcal{E}_{\mathrm{S}}^{(t)}} A_{\mathrm{S},ij}^{(t)} + \epsilon$, SA-TSI anchors affected features to their clean references and smooths intervention residuals over compatible surfaces:

$$\mathcal{L}_{\mathrm{TSI}} = \frac{1}{Z_{\mathrm{p}}} \sum_i m_i \|d_i\|_2^2 + \frac{\beta_{\mathrm{m}}}{Z_{\mathrm{e}}} \sum_{(i,j)\in\mathcal{E}_{\mathrm{S}}^{(t)}} A_{\mathrm{S},ij}^{(t)} \|d_i - d_j\|_2^2. \tag{8}$$

Unlike the detached scalar refinement in CORD, SA-TSI acts on fused representations and backpropagates through the student network.

Overall objective. The complete student objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{pair}}\mathcal{L}_{\mathrm{pair}} + \lambda_{\mathrm{cand}}\mathcal{L}_{\mathrm{cand}} + \lambda_{\mathrm{C}}(e)\mathcal{L}_{\mathrm{C}} + \lambda_{\mathrm{TSI}}\mathcal{L}_{\mathrm{TSI}}. \tag{9}$$

Here, the subscript C denotes the CORD distillation term, e is the current training epoch, and $\lambda_{\mathrm{C}}(e)$ is activated after warm-up and increased linearly. The segmentation and paired-consistency terms constitute the basic paired intervention-training objective. SA-TSI regularizes the resulting fused representations, whereas candidate exposure and $\mathcal{L}_{\mathrm{C}}$ implement the additional CORD supervision pathway. Implementation-level definitions and graph-solver details omitted from the main text are collected in Appendices B and C.

4 Experiments

Our experiments examine five aspects: robustness to structured appearance conflicts, sources of improvement, CORD's incremental gains and localized preference responses, transfer to unseen appearance shifts, and robustness under geometry-side and joint degradation.

4.1 Experimental Setup

We evaluate PointNeXt and Point Transformer V3 (PTv3) on S3DIS (Armeni et al., 2016) Area 5 and the ScanNetV2 (Dai et al., 2017) validation split. On clean scenes, we report overall accuracy (OA), mean class accuracy (mAcc), and mean intersection-over-union (mIoU). Under semantic texture conflict, Target Retention (TR) is the percentage of affected points retaining their target category, Texture Deception Rate (TDR) is the percentage predicted as the texture-source category, and $\mathrm{mIoU}_{\mathrm{conf}}$ is computed over affected points and evaluated target categories. For sensor corruptions, we report mIoU at fixed severity levels. All controlled comparisons use identical data splits, intervention masks, training schedules, and checkpoint-selection criteria. Implementation details and per-class results are included in the Appendix.

Table 1: Main comparison under global semantic texture conflicts. Clean is full-scene mIoU, whereas $\mathrm{mIoU}_{\mathrm{conf}}$ is computed only over affected points and evaluated target categories.

		S3DIS				ScanNetV2			
Backbone	**Method**	**Clean↑**	**TR↑**	**TDR↓**	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑	**Clean↑**	**TR↑**	**TDR↓**	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑
PointNeXt	Baseline	70.4	52.70	**1.42**	49.89	**70.8**	33.56	2.01	14.86
	Ours	**74.0**	**86.59**	2.05	**84.53**	69.3	**89.75**	**1.55**	**88.64**
PTv3	Baseline	69.9	54.79	3.23	47.75	77.2	69.08	6.20	52.71
	Ours	**72.0**	**67.84**	**1.67**	**59.61**	**77.6**	**84.31**	**4.83**	**66.88**

Table 2: Controlled comparison on the ScanNetV2 global semantic-conflict benchmark with PointNeXt. Clean is full-scene mIoU; $\mathrm{mIoU}_{\mathrm{conf}}$ is restricted to affected target-category points. † denotes a point-cloud adaptation of MS-FER under the matched STS-pair recipe.

Method	**Clean↑**	**TR↑**	**TDR↓**	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑
Baseline	70.79	33.56	2.01	14.86
Geometry-only	68.04	82.37	**1.08**	72.46
Modality dropout	70.31	76.84	1.74	65.37
RSMix-style	70.57	68.93	3.74	38.95
MS-FER† + STS	**70.96**	77.40	1.97	70.53
Class-agnostic donor	68.93	82.35	1.84	81.54
Ours	69.29	**89.75**	1.55	**88.64**

4.2 Robustness to Semantic Texture Conflicts

Protocol. STS replaces local intra-scene texture patches during training. We evaluate transfer beyond this augmentation using a stronger global protocol that replaces the appearance of an entire object or semantic region with texture from a distinct source category, while preserving geometry, point correspondence, and target labels. All models are frozen during evaluation, and the category labels used to construct the stress test are never provided to the evaluated model. Additional same-class, randomized-texture, and paired clean–conflict controls are reported in Appendix D.4.

Controlled alternatives. Table 2 compares geometry-only input, modality dropout, RSMix-style mixing, an MS-FER-based matched alternative(Zheng et al., 2025), and a class-agnostic-donor ablation. MS-FER retains fixed fusion and its published prediction- and four-scale feature-level regularizers, but receives the same STS pairs and L_{pair} as ours; it uses no GT-IGU, candidate exposure, CORD, or SA-TSI. The class-agnostic variant keeps the complete learner and changes only donor selection to a real patch sampled without consulting its semantic label. All alternatives use matched optimization steps and clean-validation checkpoint selection.

Results. Tables 1 and 2 show cross-backbone gains and matched controls. MS-FER + STS has the highest Clean mIoU (70.96) but reaches 77.40 TR and 70.53 $\mathrm{mIoU}_{\mathrm{conf}}$; Ours gains 12.35 TR and 18.11 $\mathrm{mIoU}_{\mathrm{conf}}$ at a cost of 1.67 Clean points. The class-agnostic full learner reaches 82.35 TR and 81.54 $\mathrm{mIoU}_{\mathrm{conf}}$; enforcing cross-class donors adds 7.40 and 7.10 points with only +0.36 Clean. This supports an operational benefit beyond generic real-texture exposure without causally identifying donor semantics. Ours also exceeds geometry-only by 7.38 TR and 16.18 $\mathrm{mIoU}_{\mathrm{conf}}$. The baseline's low TDR (2.01) alongside high OER (64.43) reveals two complementary effects: broad appearance-shift errors dominate, while TDR captures the smaller source-directed deception component. Table 5 separately decomposes the complete learner.

Beyond the aggregate results, Fig. 2 provides scene-level comparisons under the same global appearance-conflict protocol. The clean observations are shown only as visual references; all predictions are generated from the texture-intervened observations alone. The baseline exhibits substantial errors after the appearance intervention, whereas the complete model better preserves the underlying semantic layout and object identities.

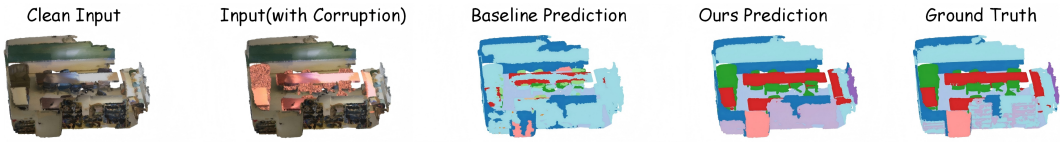

Figure 2: Qualitative results under global appearance conflicts. Columns show the clean reference, intervened input, baseline prediction, our prediction, and ground truth. Predictions use only the intervened input; geometry and labels remain unchanged.

Table 3: Clean accuracy and robustness to sensor corruptions.

		S3DIS						ScanNetV2					
Backbone	**Method**	**OA**	**mAcc**	**mIoU**	**N-3**	**J-3**	**D-3**	**OA**	**mAcc**	**mIoU**	**N-3**	**J-3**	**D-3**
PointNeXt	Baseline	90.7	76.8	70.4	42.3	48.0	44.2	87.3	80.8	70.8	51.0	46.1	44.7
	CorrAug	90.8	77.1	70.6	58.4	56.6	51.0	**87.5**	81.0	**71.0**	58.1	54.3	51.5
	Ours	**92.2**	**79.6**	**74.0**	**68.5**	**71.0**	**68.1**	86.6	**81.3**	69.3	**62.5**	**68.2**	**68.8**
PTv3	Baseline	90.4	76.0	69.9	43.8	39.4	46.0	92.1	84.8	77.2	65.0	62.6	67.8
	CorrAug	90.5	76.2	70.1	58.9	48.2	49.5	92.2	84.9	77.4	67.1	65.9	68.3
	Ours	**91.5**	**77.0**	**72.0**	**68.5**	**63.1**	**58.5**	**92.7**	**85.2**	**77.6**	**75.2**	**70.1**	**72.2**

4.3 Robustness to Sensor Corruptions

The preceding experiments focus on semantically misleading appearance. We next evaluate the same frozen complete-model checkpoints under Gaussian coordinate noise, bounded coordinate jitter, and random point dropout. Because the proposed framework is trained and selected without exposure to these geometry-corruption families, this protocol measures cross-corruption generalization beyond its texture-side training interventions.

Results. Table 3 compares the full framework with the unaugmented baseline and matched corruption augmentation. Relative to the unaugmented baseline, our method improves the average mIoU across the three corruptions by 24.4 and 20.3 points with PointNeXt and PTv3 on S3DIS, respectively, and by 19.2 and 7.4 points on ScanNetV2. The corresponding improvements over CorrAug are 13.9 and 11.2 points on S3DIS, and 11.8 and 5.4 points on ScanNetV2. On ScanNetV2 with PointNeXt, for example, CorrAug obtains 58.1, 54.3, and 51.5 mIoU under noise, jitter, and dropout, whereas our method achieves 62.5, 68.2, and 68.8. Although CorrAug achieves higher clean mIoU in this setting (71.0 vs. 69.3), its advantage does not persist under the three severity-3 corruptions. With PTv3 on ScanNetV2, our method improves both clean mIoU (77.6 vs. 77.4) and average corrupted mIoU (72.5 vs. 67.1) relative to CorrAug. Although CorrAug is trained with matched geometry corruptions, the frozen complete model outperforms both comparators on every reported severity-3 geometry corruption across both datasets and backbones. These results support whole-model cross-corruption robustness under this protocol; isolating CORD or another individual component would require matched ablations under the same geometry corruptions.

4.4 Generalization Beyond the Training Intervention

Beyond the global semantic-conflict and sensor-corruption protocols above, we evaluate frozen models under three additional unseen settings: tri-planar projection of held-out external materials, RGB-space corruptions, and joint appearance–geometry degradation. External texture projection differs from STS in texture source, mapping rule, and perturbation granularity. RGB-C averages fog, brightness, and hue-shift sweeps, while Joint combines external texture with severity-3 geometry corruptions.

Results. Table 4 provides the aggregate comparison. The complete model obtains 55.4 under complete external-texture projection and 63.88 over unseen RGB corruptions. Under joint appearance

Table 4: Generalization to intervention operators unseen during training on ScanNetV2. Ext.-Tex. uses $\alpha = 1.0$; RGB-C averages the complete severity sweeps; Joint combines external texture with the three severity-3 geometry corruptions. Results are mIoU (%).

Method	**Ext.-Tex.↑**	**RGB-C↑**	**Joint↑**
Baseline	39.6	56.84	28.12
CorrAug	43.5	57.42	34.62
Ours	**55.4**	**63.88**	**47.17**

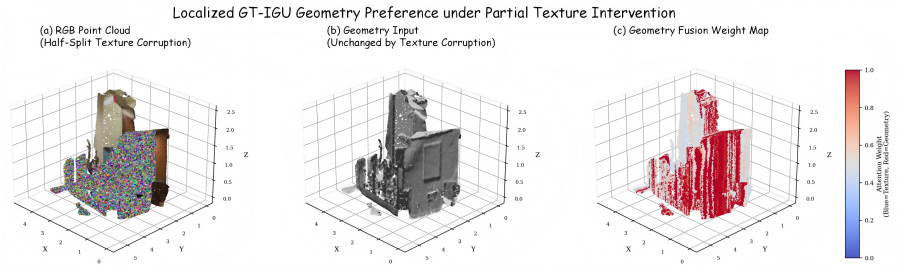

Figure 3: GT-IGU geometry-preference response to localized texture interventions. Only RGB is modified, and all maps share a fixed $[0, 1]$ scale. Higher values indicate stronger geometry routing; reported values are regional means of $\widehat{a}_i$.

and geometry degradation, the complete model obtains 47.17 mIoU, compared with 34.62 for CorrAug and 28.12 for the baseline. Together with Table 3, these results show improvements over the evaluated RGB–geometry baselines under the tested unseen operators. External material projection and RGB-C modify RGB only. If point sampling and preprocessing are RGB-independent, a geometry-only model would be expected to retain its clean full-scene performance. However, this has not been measured on the shifted samples, so these results do not establish superiority over discarding RGB.

4.5 Ablation and Routing Diagnostics

All ablations use PointNeXt on ScanNetV2 with the same training schedule. Table 5 provides a progressive system-level construction of the complete framework. The STS stage includes $\mathcal{L}_{\text{pair}}$, the adaptive-fusion stage jointly introduces GT-IGU, candidate exposure, and $\mathcal{L}_{\text{C}}$, and the final stage adds SA-TSI. Because GT-IGU and CORD are introduced together in this table, their contributions are subsequently decomposed in Table 6.

Component analysis. STS with paired consistency increases clean mIoU from 70.79 to 71.97 but provides only a limited improvement in $\text{mIoU}_{\text{conf}}$, from 14.86 to 17.82, indicating that intervention-based augmentation alone is insufficient. Adding the GT-IGU + CORD adaptive-fusion stage produces the largest system-level robustness improvement, raising TR from 25.10 to 85.67 and $\text{mIoU}_{\text{conf}}$ from 17.82 to 86.34, although clean mIoU decreases to 67.73. SA-TSI subsequently recovers 1.56 points of clean mIoU and further improves TR and $\text{mIoU}_{\text{conf}}$ by 4.08 and 2.30 points, respectively. Table 5 measures the joint GT-IGU+CORD stage, while Table 6 isolates the incremental contribution of the CORD pathway under the same GT-IGU and training recipe.

Stepwise CORD ablation. The component ablation above introduces GT-IGU and CORD together. To assess the incremental contribution of the CORD training pathway, we keep the backbone, STS pairs, GT-IGU, $\mathcal{L}_{\text{pair}}$, SA-TSI, training schedule, and checkpoint-selection criterion fixed, and compare progressively stronger routing-supervision variants. The CORD-free baseline omits candidate exposure, whereas the remaining variants enable it as part of the CORD training pathway. Consequently, the comparison with the first row measures the complete CORD pathway, while the binary-target, EMA-soft-oracle, and graph-refinement variants isolate the design of the preference target under the same candidate-exposure setting.

Table 5: Progressive system-level ablation on ScanNetV2. STS includes $\mathcal{L}_{\text{pair}}$; the GT-IGU + CORD stage includes candidate exposure and $\mathcal{L}_{\text{C}}$; and the final stage adds SA-TSI. Clean columns report full-scene performance, sensor-corruption columns report mIoU at severity 3, and $\text{mIoU}_{\text{conf}}$ is computed over affected semantic-conflict regions. All values are percentages.

Components			Clean Data			Sensor-Corruption mIoU			Semantic Conflict		
STS	**GT-IGU**	**SA-TSI**	**OA↑**	**mAcc↑**	**mIoU↑**	**Noise-3↑**	**Jitter-3↑**	**Drop.-3↑**	**TR↑**	**TDR↓**	$\text{mIoU}_{\text{conf}}$ ↑
			87.26	80.76	70.79	50.99	46.07	44.70	33.56	2.01	14.86
✓			**87.80**	81.20	**71.97**	47.09	42.68	39.77	25.10	3.19	17.82
✓	✓		85.70	79.10	67.73	59.73	59.27	64.24	85.67	1.64	86.34
✓	✓	✓	86.66	**81.29**	69.29	**62.50**	**68.22**	**68.78**	**89.75**	**1.55**	**88.64**

Table 6: Progressive ablation of the CORD training pathway on ScanNetV2. The CORD-free baseline omits candidate exposure, whereas all remaining variants include it.

Reliability design	**Clean↑**	**TR↑**	$\text{mIoU}_{\text{conf}}$ ↑	**Ext.-Tex.↑**
No CORD pathway	**69.56**	83.84	82.16	49.73
Binary intervention target	69.47	87.18	85.76	52.37
EMA soft oracle + confidence	69.22	89.31	88.09	54.79
Full CORD (+ graph refinement)	69.29	**89.75**	**88.64**	**55.40**

Routing-supervision analysis. The intervention-trained fusion model without CORD already reaches 83.84 TR and 82.16 $\text{mIoU}_{\text{conf}}$, showing that paired intervention and adaptive fusion account for the main robustness gain. The complete CORD pathway further improves TR by 5.91 points, $\text{mIoU}_{\text{conf}}$ by 6.48 points, and external-texture mIoU by 5.67 points. Among variants sharing the same candidate-exposure mechanism, the EMA soft target with confidence weighting raises $\text{mIoU}_{\text{conf}}$ from 85.76 to 88.09, while graph refinement yields the highest point estimate of 88.64. Together, these results position CORD as a task-grounded refinement of an already strong intervention-trained fusion model. **Preference response.** With geometry fixed, the mean student geometry-routing coefficient increases from 0.52 to 0.84 inside affected regions as conflict severity rises from 0 to 5, while remaining nearly unchanged outside (0.50–0.52) and tracking the teacher target (0.51–0.88). Fig. 3 localizes this increase to the intervened RGB region, consistent with a localized routing response rather than a scene-wide gate change. Complete teacher–student routing diagnostics are provided in Appendix D.1.

4.6 Efficiency and Qualitative Analysis

The EMA modality features are reused across fusion candidates. Evaluating both clean and intervened candidate risks nevertheless requires $2M_a = 10$ complete EMA post-fusion decoder/head evaluations per paired sample during training. Candidate evaluation, oracle refinement, and SA-TSI graph regularization are used only during training; inference retains only the modality encoders, GT-IGU, and post-fusion prediction network. For PointNeXt and PTv3, the complete framework requires $1.33\times$ and $2.76\times$ the baseline training time and increases peak training memory by 28.4% and 47.7%, respectively. Inference latency increases by at most 1.5%. Further fusion, efficiency, and qualitative analyses appear in Appendices D.3 and G.

5 Conclusion

We presented an intervention-guided framework for adaptive RGB–geometry fusion under texture conflict. Across two datasets and backbones, it improves robustness to structured conflicts, external-material projection, and RGB corruptions, and also improves performance under the tested geometry and joint degradations with at most 1.5% inference overhead. Ablations show that paired intervention–adaptive fusion provides most of the gain, with additional gains from CORD and SA-TSI. Future work will address training cost, clean accuracy, and natural shifts.

AI USE STATEMENT

We used ChatGPT solely to improve the grammar, clarity, and readability of the manuscript. The authors reviewed and edited all AI-assisted changes and verified that the technical content, results, citations, and conclusions were preserved. The authors take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Section 3 specifies the proposed method and training objective, with further implementation details in Appendices B and C. Section 4 describes the datasets, backbones, evaluation protocols, and metrics. Appendix H provides the data splits, software and hardware environment, training hyperparameters, baseline settings, corruption protocols, and checkpoint-selection criterion. Additional ablations and per-class results are reported in Appendices D and I.

REFERENCES

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1534–1543, 2016.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hangbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1080–1089, 2022.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7(85):2399–2434, 2006.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

Jiaxin Cai, Jingze Su, Qi Li, Wenjie Yang, Shu Wang, Tiesong Zhao, Shengfeng He, and Wenxi Liu. Keep the balance: A parameter-efficient symmetrical framework for RGB+X semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10587–10598, June 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2432–2443, 2017.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Adv. Neural Inform. Process. Syst.*, volume 29, 2016.

Lucas Deregnaucourt, Hind Laghmara, Alexis Lechervy, and Samia Ainouz. A conflict-guided evidential multimodal fusion for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1373–1382, February 2025.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26876–26885, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3354–3361, 2012.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Int. Conf. Learn. Represent.*, 2019.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, and et al. Malik. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19383–19400, 2024.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14373–14382, June 2021.

Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, October 2019.

Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21706–21716, 2023.

Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1576–1586, June 2025.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pp. 2774–2781, 2023.

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23429–23450, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *Eur. Conf. Comput. Vis.*, pp. 53–72, 2022.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12700–12710, 2021.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6729, 2025.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *Adv. Neural Inform. Process. Syst.*, volume 35, pp. 23192–23204, 2022.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

David I Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Jiaqi Tan, Xu Zheng, and Yang Liu. Towards robust multi-modal semantic segmentation with teacher-student framework and hybrid prototype distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27665–27675, June 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inform. Process. Syst.*, volume 30, 2017.

Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4840–4851, 2024.

Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. In *Adv. Neural Inform. Process. Syst.*, volume 35, pp. 11035–11048, 2022.

Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10368–10378, 2021a.

Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):6354–6371, 2023.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3340–3349, June 2021b.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Int. Conf. Comput. Vis.*, pp. 12–22, 2023.

Zhe Zhang, Gaochang Wu, Jing Zhang, Xiatian Zhu, Dacheng Tao, and Tianyou Chai. Unified domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(8):6731–6748, 2025.

Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Reducing unimodal bias in multi-modal semantic segmentation with multi-scale functional entropy regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21166–21176, October 2025.

SUPPLEMENTARY ROADMAP

The supplement follows the same evidence chain as the main paper. Appendices A–C complete the method definitions for GT-IGU, STS, CORD, and SA-TSI. Appendix D decomposes the component gains and analyzes benchmark difficulty through matched appearance-shift controls. Appendices E and F then test transfer across datasets and intervention families, respectively. Appendix G provides qualitative diagnostics, while Appendices H and I consolidate the fixed experimental protocol, baseline-versus-ours per-class breakdowns, and every evaluated semantic-conflict target category.

A STRUCTURAL MOTIVATION FOR GT-IGU

This section derives the inverse-variance normalization that motivates the structure of GT-IGU; the learned gate itself remains discriminative.

A.1 INVERSE-VARIANCE INTERPRETATION

Let $F^g, F^t \in \mathbb{R}^D$ be two aligned, unbiased observations of a latent representation Z. For motivation only, suppose that their noises are conditionally independent and Gaussian:

$$\begin{aligned} p(F^g \mid Z) &= \mathcal{N}(F^g; Z, \Sigma_g), \\ p(F^t \mid Z) &= \mathcal{N}(F^t; Z, \Sigma_t). \end{aligned} \tag{10}$$

The maximum-likelihood estimate is

$$Z^* = (\Sigma_g^{-1} + \Sigma_t^{-1})^{-1}(\Sigma_g^{-1} F^g + \Sigma_t^{-1} F^t). \tag{11}$$

Under diagonal covariances, this decomposes channel-wise. For isotropic variances σ_g^2 and σ_t^2, the normalized weights become

$$w_g = \frac{\sigma_g^{-2}}{\sigma_g^{-2} + \sigma_t^{-2}}, \qquad w_t = \frac{\sigma_t^{-2}}{\sigma_g^{-2} + \sigma_t^{-2}} = 1 - w_g. \tag{12}$$

Equivalently, a modality-wise softmax over hypothetical scores $r_k = -\log \sigma_k^2$ recovers these normalized inverse-variance weights:

$$\mathrm{Softmax}_{\mathrm{mod}}([r_g, r_t])_k = \frac{\sigma_k^{-2}}{\sigma_g^{-2} + \sigma_t^{-2}}. \tag{13}$$

GT-IGU retains the normalized competitive form motivated by Eq.13 and learns its scores discriminatively from ordered cross-modal evidence. CORD supervises the point-wise channel-mean geometry-routing coefficient. Accordingly, the diagnostics below analyze coefficient responses on the learned feature scale. Because these coefficients multiply learned, unnormalized modality features, their values are routing variables rather than scale-invariant feature attributions. The diagnostics below therefore analyze coefficient responses, not calibrated uncertainty or causal modality contribution.

B ADDITIONAL DETAILS OF STS AND CORD

This section provides the implementation-level definitions omitted from the main paper. It first specifies the local correspondence used by Structured Texture Swapping (STS), and then details candidate exposure, EMA updates, and the graph refinement used by Candidate-Oracle Routing Distillation (CORD). The final CORD training path has three fixed stages: clean candidate exposure trains the post-fusion predictor on the finite candidate family; the EMA teacher evaluates that family on paired clean and intervened observations; and a detached, quality-weighted graph refinement constructs the target for the student gate. Appendix D.1 ablates these stages in the same order.

B.1 STS Patch Correspondence

For a target patch Ω_{tgt} and a source patch Ω_{src}, we center and scale their coordinates independently to obtain $\bar{g}_i^{\mathrm{tgt}}$ and $\bar{g}_j^{\mathrm{src}}$. Each target point receives RGB values from its nearest source point in the normalized local coordinate system:

$$\begin{aligned} \pi(i) &= \arg\min_{j\in\Omega_{\mathrm{src}}} \left\|\bar{g}_i^{\mathrm{tgt}} - \bar{g}_j^{\mathrm{src}}\right\|_2, \\ T_i^{\mathrm{src}} &= t_{\pi(i)}, \qquad i \in \Omega_{\mathrm{tgt}}. \end{aligned} \tag{14}$$

Only valid points are retained when a neighborhood reaches a scene boundary. Nearest-neighbor reuse is allowed when the two patches contain different numbers of points. Consequently, STS changes only the RGB values at the target points and preserves their coordinates, labels, point indices, and clean–intervened correspondence.

B.2 Clean Candidate Exposure

CORD evaluates fusion hypotheses that need not follow the distribution produced by GT-IGU. To reduce this decoder-distribution mismatch, we expose the student post-fusion network to one randomly sampled clean candidate per batch. For $r \sim \mathrm{Unif}(\{1,\dots,M_a\})$, we define

$$\begin{aligned} Z_{\mathrm{cand},r}^{(0)} &= a_r\,\mathrm{sg}[F_\theta^{g,(0)}] + (1-a_r)\,\mathrm{sg}[F_\theta^{t,(0)}], \\ \mathcal{L}_{\mathrm{cand}} &= \mathbb{E}_r\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{CE}\Big(h_\psi(Z_{\mathrm{cand},r}^{(0)})_i, y_i\Big)\right]. \end{aligned} \tag{15}$$

The stop-gradient operator blocks updates to both modality encoders, so $\mathcal{L}_{\mathrm{cand}}$ updates only h_ψ. Candidate exposure uses the clean observation and therefore does not train the decoder to recognize a particular synthetic intervention.

B.3 EMA Teacher and Candidate Evaluation

The teacher contains EMA copies of the modality encoders and the complete post-fusion prediction network, but not GT-IGU. After each student update, its parameters are updated as:

$$\bar{\theta} \leftarrow \beta_{\mathrm{ema}}\bar{\theta} + (1-\beta_{\mathrm{ema}})\theta, \qquad \bar{\psi} \leftarrow \beta_{\mathrm{ema}}\bar{\psi} + (1-\beta_{\mathrm{ema}})\psi. \tag{16}$$

For every clean and intervened observation, all candidates are passed through the complete EMA post-fusion network. Teacher predictions, candidate risks, soft candidate probabilities, quality weights, and refined targets are detached from gradient propagation. Thus, only the student preference $\widehat{a}_i^{(t)}$ receives gradients from $\mathcal{L}_{\mathrm{C}}$, where the subscript C abbreviates the CORD reliability-distillation term.

B.4 Quality-Weighted CORD Graph Refinement

Given the candidate distribution and costs defined in the main paper, we use the quality weight

$$\begin{aligned} \chi_i^{(t)} &= \left(1 - \frac{H(\pi_i^{(t)})}{\log M_a}\right)\exp\left(-\frac{\min_r c_{i,r}^{(t)}}{\tau_c}\right), \\ H(\pi) &= -\sum_r \pi_r \log \pi_r. \end{aligned} \tag{17}$$

where the entropy term measures preference concentration and the exponential term suppresses comparisons for which every candidate has high absolute cost. Thus, $\chi_i^{(t)}$ is a candidate-comparison quality weight that combines preference concentration with absolute candidate cost.

The raw target $\widetilde{a}_i^{(t)}$ is refined on the CORD graph $\mathcal{G}_{\mathrm{C}}^{(t)} = (\Omega^{(t)}, \mathcal{E}_{\mathrm{C}}^{(t)})$ induced by the swapped region. Since the swapped patch is category-constrained, edge affinities need only encode spatial and surface compatibility. For neighboring points, we use:

$$A_{\mathrm{C},ij}^{(t)} = \exp\left(-\frac{\|g_i - g_j\|_2^2}{2\sigma_{g,c}^2} - \frac{1-|n_i^\top n_j|}{2\sigma_{n,c}^2}\right), (i,j) \in \mathcal{E}_{\mathrm{C}}^{(t)}, \tag{18}$$

where g_i and n_i denote the coordinate and locally estimated normal of point i. The refined target is:

$$\mathbf{a}^{\star,(t)} = \arg \min_{\mathbf{u}\in[0,1]^{|\Omega^{(t)}|}} \Bigg[\sum_{i\in\Omega^{(t)}} (\chi_i^{(t)} + \epsilon_a)(u_i - \widetilde{a}_i^{(t)})^2 + \mu \sum_{(i,j)\in\mathcal{E}_{\mathrm{C}}^{(t)}} A_{\mathrm{C},ij}^{(t)} (u_i - u_j)^2 \Bigg]. \tag{19}$$

The positive anchor ϵ_a makes the quadratic objective strictly convex and ensures a unique minimizer even when candidate quality is low. We approximate that minimizer with a fixed-step solver. Let $q_i = \chi_i^{(t)} + \epsilon_a$ and initialize $u_i^{(0)} = \widetilde{a}_i^{(t)}$. A projected Jacobi iteration is:

$$u_i^{(k+1)} = \Pi_{[0,1]} \left(\frac{q_i \widetilde{a}_i^{(t)} + \mu \sum_{j\in\mathcal{N}_{\mathrm{C}}(i)} A_{\mathrm{C},ij}^{(t)} u_j^{(k)}}{q_i + \mu \sum_{j\in\mathcal{N}_{\mathrm{C}}(i)} A_{\mathrm{C},ij}^{(t)}} \right). \tag{20}$$

With $K_{\mathrm{J}} = 5$, we define the implemented detached target as $a_i^{\mathrm{oracle},(t)} := u_i^{(K_{\mathrm{J}})}$. Thus, $\mathbf{a}^{\mathrm{oracle},(t)}$ is a fixed-step approximation to $\mathbf{a}^{\star,(t)}$, not a claim that five iterations recover the exact argmin. The graph solver is used only to construct training targets and is removed at inference.

C Additional Details of SA-TSI

Structure-Aware Texture-Swapping Invariance (SA-TSI) regularizes the change caused by STS rather than directly smoothing the fused representation. This distinction prevents the regularizer from forcing all semantic features within a connected component toward a constant value.

C.1 Residual and Graph Construction

Let $z_{i,\mathrm{fuse}}^{(0)}$ and $z_{i,\mathrm{fuse}}^{(t)}$ denote the clean and texture-intervened GT-IGU fused features. The normalized intervention residual is:

$$d_i = \frac{z_{i,\mathrm{fuse}}^{(t)}}{\|z_{i,\mathrm{fuse}}^{(t)}\|_2 + \epsilon} - \mathrm{sg}\left[\frac{z_{i,\mathrm{fuse}}^{(0)}}{\|z_{i,\mathrm{fuse}}^{(0)}\|_2 + \epsilon} \right]. \tag{21}$$

The clean branch is detached so that the reference cannot drift toward the intervened representation.

We construct the SA-TSI graph $\mathcal{G}_{\mathrm{S}}^{(t)} = (\mathcal{V}_{\mathrm{S}}^{(t)}, \mathcal{E}_{\mathrm{S}}^{(t)})$ over the swapped region and its one-hop geometric neighborhood, counting each undirected edge once. An edge is retained only if at least one endpoint belongs to $\Omega^{(t)}$. Its affinity is:

$$A_{\mathrm{S},ij}^{(t)} = \mathbb{I}[y_i = y_j] \exp\left(-\frac{\|g_i - g_j\|_2^2}{2\sigma_g^2} - \frac{1 - |n_i^\top n_j|}{2\sigma_n^2} \right), \quad (i,j) \in \mathcal{E}_{\mathrm{S}}^{(t)}. \tag{22}$$

The semantic indicator excludes annotated category boundaries, while the coordinate and normal terms attenuate edges across distant points and geometric discontinuities. Labels, coordinates, normals, and STS masks are treated as fixed quantities when differentiating the loss.

C.2 Complete Objective and Residual-Space Gradient

With $m_i = \mathbb{I}[i \in \Omega^{(t)}]$, $Z_{\mathrm{p}} = \sum_i m_i + \epsilon$, and $Z_{\mathrm{e}} = \sum_{(i,j)\in\mathcal{E}_{\mathrm{S}}^{(t)}} A_{\mathrm{S},ij}^{(t)} + \epsilon$, the complete loss is:

$$\mathcal{L}_{\mathrm{TSI}} = \frac{1}{Z_{\mathrm{p}}} \sum_i m_i \|d_i\|_2^2 + \frac{\beta_{\mathrm{m}}}{Z_{\mathrm{e}}} \sum_{(i,j)\in\mathcal{E}_{\mathrm{S}}^{(t)}} A_{\mathrm{S},ij}^{(t)} \|d_i - d_j\|_2^2. \tag{23}$$

Because the graph quantities are independent of the learned features, differentiation with respect to the residual gives

$$\frac{\partial \mathcal{L}_{\mathrm{TSI}}}{\partial d_i} = \frac{2m_i}{Z_{\mathrm{p}}} d_i + \frac{2\beta_{\mathrm{m}}}{Z_{\mathrm{e}}} \sum_{j\in\mathcal{N}_{\mathrm{S}}(i)} A_{\mathrm{S},ij}^{(t)} (d_i - d_j). \tag{24}$$

The first term pulls an affected feature toward its paired clean reference. The second moves its intervention residual toward the affinity-weighted residuals of compatible neighbors. Gradients then propagate through the normalization of $z^{(t)}_{i,\mathrm{fuse}}$; no gradient passes through $z^{(0)}_{i,\mathrm{fuse}}$.

C.3 Localization, Boundaries, and Degenerate Cases

SA-TSI is localized in three ways. First, the anchor is applied only where $m_i = 1$. Second, graph edges require at least one intervention-affected endpoint. Third, $\mathbb{I}[y_i = y_j]$ removes cross-category edges, while the spatial and normal factors attenuate geometrically incompatible neighbors. Consequently, the objective neither enforces global feature constancy nor explicitly smooths across annotated semantic boundaries.

The affinity satisfies $0 \le A^{(t)}_{\mathrm{S},ij} \le 1$. Normalization by Z_{p} and Z_{e} reduces sensitivity to the affected-region size and number of active edges. If a sample has no valid graph edge, the graph numerator is zero and the second term is defined as zero through the ϵ-stabilized denominator.

Boundary evaluation band. Let $\mathcal{E}_{K_{\mathrm{S}}}$ be the same $K_{\mathrm{S}} = 16$ geometric KNN graph underlying SA-TSI before affinity and semantic weighting. The inside boundary seeds are

$$\partial\Omega^{(t)} = \left\{ i \in \Omega^{(t)} : \exists j \notin \Omega^{(t)},\ (i,j) \in \mathcal{E}_{K_{\mathrm{S}}} \right\}.$$

The fixed boundary band used for Boundary mIoU contains affected points whose shortest-path distance to $\partial\Omega^{(t)}$ in this graph is at most two hops. IoU counts are restricted to this band and macro-averaged over the evaluated target categories.

C.4 Relation to Paired Prediction Consistency

$\mathcal{L}_{\mathrm{pair}}$ and SA-TSI operate at different levels. The former aligns the semantic distributions of corresponding clean and intervened points, with a smaller context weight for unchanged points. SA-TSI instead anchors normalized fused features within the swapped region and constrains how their intervention residual varies over compatible geometric neighborhoods. Their combination therefore provides output-level paired consistency and feature-level structure-aware invariance.

Shared Evaluation Rules

All experiments use the S3DIS Area 5 or ScanNetV2 validation protocol, clean-validation checkpoint selection, and pre-generated evaluation samples shared across compared methods. Evaluation-only shifts are applied to frozen checkpoints and are never used for model selection or early stopping. Appendix H gives the complete common protocol; the result sections below specify only experiment-specific changes.

D Detailed Ablation Studies

This section complements the system-level ablation in the main paper by following four questions in order. First, how much does CORD add when STS, GT-IGU, and SA-TSI are held fixed? Second, do the two terms of SA-TSI function as a coupled regularizer? Third, does ordered adaptive fusion outperform parameter-matched alternatives? Fourth, does the semantic-conflict benchmark contain a category-directed component beyond generic appearance shift?

D.1 Reliability-Supervision Design of CORD

Throughout this supplement, candidate oracle denotes a label-aware routing target induced over the fixed candidate bank $\mathcal{A}$, while reliability supervision refers to this candidate-relative, task-grounded signal. We compare candidate exposure, a binary intervention target, hard and soft candidate oracles, confidence weighting, and graph refinement while keeping the remaining components fixed. We then examine the candidate-cost terms and the response of the learned preference to intervention severity. Preference MAE is computed between the detached oracle and the predicted mean

Variant	Clean↑	TR↑	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑	Ext.-Tex.↑	Train cost↓
No candidate exposure/reliability supervision	69.56	83.84	82.16	49.73	**1.00**×
Candidate exposure only	**69.71**	84.62	82.91	50.48	1.03×
Binary intervention target	69.47	87.18	85.76	52.37	1.04×
Hard candidate oracle	69.16	88.36	87.05	53.41	1.13×
Soft candidate oracle	69.38	88.87	87.61	54.06	1.14×
Soft oracle + confidence weighting	69.22	89.31	88.09	54.79	1.14×
Full CORD (+ graph refinement)	69.29	**89.75**	**88.64**	**55.40**	1.16×

Table 7: Ablation of the CORD reliability-distillation design on ScanNetV2. Ext.-Tex. denotes mIoU under external texture projection at $\alpha = 1.0$. End-to-end training cost is measured under identical hardware, batch-size, point-count, precision, and data-loading settings and is normalized to the reliability-free first row. This normalization differs from the comparison with the original backbone baseline reported in Table 12.

geometry preference over affected points:

$$\mathrm{MAE}_{\mathrm{pref}} = \frac{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)} |\widehat{a}_i^{(t)} - a_i^{\mathrm{oracle},(t)}|}{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)} + \epsilon}. \tag{25}$$

For the candidate-family regret diagnostic, we map the student prediction to its nearest candidate, $r_i^{\mathrm{stu}} = \arg\min_r |a_r - \widehat{a}_i^{(t)}|$, and use

$$\mathrm{Regret}_{\mathrm{cand}} = \frac{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)} \left(c_{i,r_i^{\mathrm{stu}}}^{(t)} - \min_r c_{i,r}^{(t)} \right)}{\sum_{i\in\Omega^{(t)}} \chi_i^{(t)} + \epsilon}. \tag{26}$$

All variants share the same modality encoders, STS pairs, GT-IGU, paired-consistency objective, SA-TSI, optimization schedule, and checkpoint-selection rule. Candidate exposure and the reliability target are added progressively. The first row represents the paired intervention–fusion framework without candidate exposure or reliability supervision, while the remaining rows decompose the additional CORD training pathway. Candidate regret is defined only relative to the finite candidate family and does not imply regret with respect to a continuous heterogeneous optimum.

Ablation-target definitions. The binary intervention target assigns geometry coefficient 1 to every affected point and applies no preference loss outside $\Omega^{(t)}$. The hard oracle instead selects

$$r_i^\star = \min \arg\min_r c_{i,r}^{(t)}, \qquad a_i^{\mathrm{hard},(t)} = a_{r_i^\star},$$

where the outer minimum gives a deterministic lower-coefficient tie break. The soft oracle is the candidate-cost expectation defined in the main paper. The soft-oracle row uses neither confidence weighting nor graph refinement; the next row adds confidence weighting, and Full CORD retains that weighting and additionally applies graph refinement.

The reliability-free paired intervention–fusion framework already achieves 83.84 TR and 82.16 $\mathrm{mIoU}_{\mathrm{conf}}$, compared with 33.56 and 14.86 for the original RGB–geometry baseline in the main paper. Thus, the dominant robustness gain arises before CORD. Candidate exposure alone provides improvements, increasing TR, $\mathrm{mIoU}_{\mathrm{conf}}$, and external-texture mIoU to 84.62, 82.91, and 50.48, respectively, with a relative training cost of 1.03×. Adding the binary intervention target further improves the three robustness metrics to 87.18, 85.76, and 52.37 at 1.04× cost.

The point estimates increase progressively from the hard to the soft candidate oracle. The increase from 1.04× to 1.13× mainly arises from evaluating the candidate bank with the EMA teacher, whereas soft target construction and confidence weighting introduce little additional measured overhead. The full CORD pathway, including graph refinement, achieves 89.75 TR, 88.64 $\mathrm{mIoU}_{\mathrm{conf}}$, and 55.40 external-texture mIoU at 1.16× the training cost of the reliability-free model. Relative to the first row, these correspond to improvements of 5.91, 6.48, and 5.67 points, respectively, with a 0.27-point reduction in clean mIoU. After graph refinement, the point estimates are higher by 0.44 TR, 0.55 $\mathrm{mIoU}_{\mathrm{conf}}$, and 0.61 external-texture mIoU than for the confidence-weighted soft oracle, while normalized training cost changes from 1.14× to 1.16×. These sub-point differences are descriptive under the fixed paired seed rather than evidence of statistical separation.

Candidate cost	Clean↑	TR↑	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑	Ext.-Tex.↑	Regret↓
Full cost	69.29	89.75	**88.64**	**55.40**	**0.084**
Without clean-risk term	68.92	**89.93**	88.37	54.82	0.096
Without intervention-regret term	**69.43**	88.96	87.69	53.88	0.121

Table 8: Ablation of the candidate-cost terms on ScanNetV2.

Severity	Oracle	Affected	Unaffected	χ	Entropy	Pref. MAE
0	0.51	0.52	0.50	0.08	0.91	0.049
1	0.59	0.57	0.50	0.14	0.84	0.066
2	0.67	0.64	0.51	0.23	0.73	0.083
3	0.75	0.71	0.51	0.37	0.60	0.101
4	0.83	0.78	0.52	0.46	0.49	0.092
5	0.88	0.84	0.52	0.55	0.40	0.078

Table 9: Geometry-side gate-coefficient diagnostic on ScanNetV2. Oracle, affected, and unaffected are mean routing coefficients; entropy is normalized by $\log M_a$. These coefficients are not calibrated reliability estimates or scale-invariant feature attributions.

Candidate-cost result. Table 8 separates the two non-redundant terms in the candidate cost. Removing the clean-risk term slightly raises TR but reduces clean and external-texture mIoU and increases candidate-family regret. Removing the intervention-regret term causes the larger robustness loss, including a 1.52-point reduction in external-texture mIoU and an increase in regret from 0.084 to 0.121. The full cost therefore provides the best joint clean/robustness trade-off among these variants; the table does not support a claim that every individual metric improves monotonically.

Preference response to intervention severity. We retain the same pre-generated target mask and construct severity $s \in \{0, 1, 2, 3, 4, 5\}$ by $T_s = (1 - \rho_s)T_{\mathrm{tgt}} + \rho_s T_{\mathrm{src}}$, where $\rho_s = s/5$. Severity zero changes no RGB value, while severity five gives complete replacement. Normalized candidate entropy is $H(\pi_i^{(t)})/\log M_a$.

As severity increases from 0 to 5, the oracle geometry-side coefficient rises from 0.51 to 0.88, while the student coefficient inside affected regions increases from 0.52 to 0.84. The coefficient outside the target mask remains nearly unchanged at 0.50–0.52. This localized coefficient shift is consistent with intervention-dependent routing rather than a scene-wide gate change; by itself, it does not measure the scale-normalized contribution of either raw feature stream. The mean candidate-comparison quality weight increases from 0.08 to 0.55 as normalized candidate entropy decreases from 0.91 to 0.40, indicating increasingly concentrated candidate rankings under stronger conflicts. CORD supervises only the channel-mean coefficient; channel-wise deviations remain governed by the segmentation objectives.

D.2 Internal Design of SA-TSI

SA-TSI contains a paired clean anchor and a graph residual-consistency term. We evaluate the two terms separately and report performance near intervention boundaries to determine whether the graph term improves local consistency without indiscriminate smoothing.

SA-TSI result. Table 10 shows that neither term provides a favorable overall trade-off in isolation. In particular, the graph-only variant improves boundary mIoU but reduces clean and external-texture mIoU relative to omitting SA-TSI. The combined objective improves clean mIoU from 67.73 to 69.29, boundary mIoU from 72.11 to 86.74, and external-texture mIoU from 52.86 to 55.40. We therefore interpret SA-TSI as a coupled objective: the clean anchor fixes the reference state, while the graph term controls how the anchored residual varies near compatible neighbors. The result supports their complementarity, not two independently monotonic gains; the corresponding overall conflict-region result is reported in Table 11.

Variant	Clean	TR↑	Boundary mIoU↑	Ext.-Tex.↑
Without SA-TSI	67.73	85.67	72.11	52.86
Anchor only	61.12	83.71	64.63	47.62
Graph residual only	63.05	84.74	75.19	45.55
Anchor + graph	**69.29**	**89.75**	**86.74**	**55.40**

Table 10: Ablation of SA-TSI on ScanNetV2. Boundary mIoU is evaluated on affected points within the fixed two-hop band defined in Appendix C.

Fusion	Params	GFLOPs	Latency	Clean	TR↑	$\mathrm{mIoU}_{\mathrm{conf}}$ ↑	Ext.-Tex.↑
Dual-stream mean	0	0.01	0.02 ms	**70.71**	61.48	55.38	45.07
PM-Concat	49.41K	1.97	0.18 ms	70.12	72.31	67.42	48.63
Static channel gate	0.13K	0.01	0.03 ms	69.94	69.76	63.95	47.54
Symmetric conflict gate	49.41K	1.99	0.21 ms	69.62	80.89	76.84	51.68
Ordered GT-IGU without CORD	49.34K	3.95	0.38 ms	69.56	83.84	82.16	49.73
Ordered GT-IGU + CORD	49.34K	3.95	0.38 ms	69.29	**89.75**	**88.64**	**55.40**

Table 11: Parameter-matched fusion comparison on ScanNetV2. Complexity is measured for the fusion module only; latency is measured under the same input shape and hardware.

D.3 Parameter-Matched Fusion Baselines

We compare GT-IGU with fixed mean fusion, parameter-matched concatenation (PM-Concat), a static channel gate, and a symmetric conflict gate. All variants share the modality encoders, feature dimension, decoder, segmentation head, point sampling, and optimization schedule. Fusion-head parameters are matched as closely as possible without adding inactive parameters. To isolate fusion structure, candidate exposure and CORD are disabled for the first five rows, and all variants use the same segmentation, STS-pair, and SA-TSI objectives. The final row adds candidate exposure and CORD to GT-IGU to report the complete framework.

Fusion comparison result. Table 11 isolates fusion structure under a shared STS-pair and SA-TSI training recipe. Without CORD, ordered GT-IGU improves conflict-region mIoU from 76.84 for the strongest parameter-matched conflict gate to 82.16, while using a comparable number of parameters. The symmetric gate is slightly better on external texture (51.68 versus 49.73), so ordered gating alone is not uniformly dominant. Adding the complete CORD pathway raises external-texture mIoU to 55.40 and gives the best TR and conflict-region mIoU. Under the shared STS+SA-TSI recipe, ordered GT-IGU provides the strongest conflict-region result among the CORD-free fusion designs, while the complete CORD pathway further improves held-out external-texture performance.

End-to-end training and inference cost. Clean and intervened EMA encoder features are reused across candidates, but the complete post-fusion decoder and prediction head are evaluated for all $M_a = 5$ candidates in both states, yielding $2M_a = 10$ EMA post-fusion evaluations per paired sample. The EMA teacher, candidate bank, graph refinement, and SA-TSI graph are removed at inference.

Relative to the corresponding backbone, the added training pathway increases time per epoch by $1.33\times$ for PointNeXt and $2.76\times$ for PTv3, and training memory by 28.4% and 47.7%, respectively. Inference latency increases by only 1.5% and 0.8%, because the EMA teacher, candidate bank, and graph objectives are training-only. The principal cost is therefore training, especially for PTv3, rather than deployment.

D.4 Semantic-Conflict Benchmark Validity

Controlled alternatives. Main-paper Table 2 compares the complete framework with geometry-only input, modality dropout, RSMix-style region mixing, a matched-training adaptation of MS-FER, and a class-agnostic donor ablation. CorrAug and RSMix-style use neither paired STS consistency nor reliability supervision.

Method	Params (M)	Train time/epoch	Train memory	Inference time	Infer. memory
PointNeXt baseline	7.10	4.8 min	11.6 GB	33.9 ms	12.4 GB
PointNeXt + Ours	7.15	6.4 min	14.9 GB	34.4 ms	12.5 GB
PTv3 baseline	46.20	12.6 min	15.1 GB	61.8 ms	17.2 GB
PTv3 + Ours	46.25	34.8 min	22.3 GB	62.3 ms	17.3 GB

Table 12: End-to-end efficiency on ScanNetV2. Values assume one RTX 4090 and identical batch size, point count, precision, and data loading. Total params counts the inference network; inference time is per scene.

Evaluation design. For every target region R, we construct a paired clean input X^{clean} and perturbed input $X^{\text{perturbed}}$. The paired inputs share identical geometry, semantic labels, point count, target instance, point indices, and evaluation mask; only the texture within R is modified. All models are evaluated using the same pre-generated target–source pairs. TR, TDR, OER, and conflict-region mIoU follow Appendix H; the additional paired diagnostics used only here are defined below.

Aggregate result. Main-paper Table 2 shows that geometry-only and modality-dropout training substantially improve robustness over the RGB+XYZ baseline, confirming that reducing texture dependence is already a strong strategy. The complete model nevertheless exceeds geometry-only by 7.38 TR points and 16.18 conflict-region mIoU points, so its result cannot be explained by discarding appearance altogether.

Validity-specific metrics. For the common target region R, the paired flip rate is

$$\text{PFR} = \frac{100}{|R|} \sum_{i \in R} \mathbb{I}\left(\hat{y}_i^{\text{perturbed}} \neq \hat{y}_i^{\text{clean}}\right). \tag{27}$$

We report the paired performance drop as:

$$\text{Drop}_R = \text{mIoU}_R^{\text{clean}} - \text{mIoU}_R^{\text{perturbed}}, \tag{28}$$

where a smaller value indicates stronger robustness. We additionally report within-model conditional diagnostics. For model m, let

$$C_m = \left\{ i \in R : \hat{y}_{m,i}^{\text{clean}} = y_i \right\}. \tag{29}$$

Conditional target retention and texture deception are defined as

$$\text{cTR}_m = \frac{100}{|C_m|} \sum_{i \in C_m} \mathbb{I}\left(\hat{y}_{m,i}^{\text{perturbed}} = y_i\right), \tag{30}$$

$$\text{cTDR}_m = \frac{100}{|C_m|} \sum_{i \in C_m} \mathbb{I}\left(\hat{y}_{m,i}^{\text{perturbed}} = c_{\text{src},i}\right). \tag{31}$$

Because C_m is model-specific, cTR and cTDR are interpreted as paired within-model diagnostics rather than the sole basis for cross-model ranking. cTDR is not defined for same-class or randomized texture controls.

Restricted-region evaluation. We compute restricted mIoU on the union of all evaluated target regions R. For each selected target category, true positives, false positives, and false negatives are accumulated only within this common evaluation support. IoU is macro-averaged over the selected target categories, preventing large regions from dominating the evaluation. Consequently, clean mIoU@R is not directly comparable with clean mIoU computed over the complete ScanNetV2 validation set. Aggregate cross-method comparisons are already reported in the main paper; the matched controls below focus on benchmark validity. The class-agnostic donor ablation in main-paper Table 2 is a training-time control that samples real donor patches. It is distinct from the source-free random-texture control below, which is applied only at evaluation and synthesizes RGB content without a donor patch or source-category label.

Method	Perturbation	mIoU Drop↓	cTR↑	cTDR↓	PFR↓
RGB+XYZ baseline	Same-class swap	5.92	92.48	–	7.84
	Random texture control	29.84	64.87	–	31.96
	Cross-class local swap	48.36	47.12	1.39	49.72
	Cross-class object swap	75.45	36.71	2.18	63.84
Ours	Same-class swap	0.71	98.31	–	1.94
	Random texture control	3.27	95.18	–	4.73
	Cross-class local swap	2.05	96.22	0.91	3.58
	Cross-class object swap	1.09	96.04	1.67	7.86

Table 13: Validity controls for semantic conflict. The random-texture control shares its target-region mask with the cross-class local swap and is directly matched only to that condition. Conditional metrics are computed over each model's own clean-correct subset. cTDR is defined only for cross-class replacement. mIoU Drop is measured in percentage points; the remaining metrics are percentages.

Matched control conditions. We consider four texture transformations. Same-class replacement changes instance appearance without introducing a cross-category texture conflict. The random-texture control uses the same target region as the corresponding cross-class local replacement but generates source-free randomized RGB content: it does not draw a texture from a semantic donor category, and hence has no source label for TDR. Cross-class local replacement modifies the same local target region, whereas cross-class object replacement modifies the complete target instance. Geometry and semantic labels remain unchanged in all conditions. Because the randomized control uses the same region mask only as the local replacement, it is treated as a matched control for the local condition but not for the object-level condition.

Analysis. Same-class replacement causes a smaller paired drop than cross-class replacement for the RGB+XYZ baseline, showing that sensitivity depends on the replacement condition. This comparison alone does not isolate semantic category identity from low-level texture distance. Random texture perturbations also cause a considerable degradation, showing that generic appearance shift constitutes a major part of the benchmark difficulty. Consistently, the baseline global-conflict result has a TDR of 2.01 and a derived OER of 64.43 $(100 - 33.56 - 2.01)$; most failures therefore correspond to other categories rather than direct prediction of the texture-source class. Under the matched local-region setting, cross-class replacement causes a larger mIoU drop than the randomized control for the RGB+XYZ baseline (48.36 vs. 29.84 points) and induces a non-zero conditional texture-deception rate. The larger drop shows only that the cross-class local swap is harder than this particular randomized control; the non-zero cTDR provides limited evidence of a source-category-directed component on top of the dominant generic appearance shift. Because the randomized content is not claimed to match every low-level statistic of a semantic donor texture, the difference is descriptive rather than a causal estimate of category identity alone. Under the same pre-generated masks and evaluation support, the full model obtains a maximum mIoU drop of 3.27 and a minimum cTR of 95.18 across the four conditions.

Coverage pointer. Appendix I reports baseline-versus-ours per-class sensor-corruption breakdowns and every target category used by the semantic-conflict protocol.

E Cross-Dataset Generalization

Protocol. Models are trained on ScanNetV2 and applied to S3DIS without fine-tuning or test-time adaptation, thereby changing both dataset and appearance statistics. The fixed evaluation support contains door, chair, table, and sofa, whose class names and semantic meanings are shared by the two taxonomies. Global metrics are computed on the union of affected points from these four target classes, and $\mathrm{mIoU}_{\mathrm{conf}}$ is macro-averaged over the same four classes. A prediction outside the target class counts as a target-retention failure. All methods use frozen checkpoints, identical pre-generated target–source pairs, and identical evaluation masks.

Method	Aggregate Metrics on Shared Support			Per-Class Robustness (TR %)			
	TR (%) ↑	**TDR** (%) ↓	$\mathrm{mIoU}_{\mathrm{conf}}$ (%) ↑	Door	Chair	Table	Sofa
PointNeXt	0.45	12.47	0.95	0.60	**1.75**	0.20	0.00
+ Ours	**6.68**	**5.21**	**10.85**	**32.01**	1.50	**11.34**	**0.25**
PTv3	47.89	15.92	48.74	36.77	86.38	36.88	39.13
+ Ours	**56.44**	**8.32**	**55.94**	**56.52**	**88.00**	**39.00**	**61.33**

Table 14: Semantic-conflict robustness under cross-domain transfer (ScanNet → S3DIS).

Result and scope. Table 14 shows improved aggregate metrics for both backbones, including a 7.20-point $\mathrm{mIoU}_{\mathrm{conf}}$ gain for PTv3; the PointNeXt chair TR is an exception at the class level. Absolute PointNeXt performance remains low (10.85 $\mathrm{mIoU}_{\mathrm{conf}}$ after improvement), whereas PTv3 retains substantially stronger transfer. We therefore treat this as secondary, backbone-dependent system-level transfer evidence.

F Generalization to Held-Out Intervention Operators

This section groups evaluation-only shifts into appearance-side tests and geometry-side or joint tests. All shared selection and frozen-evaluation rules follow Appendix H.

F.1 Appearance-Side Shifts

External-projection motivation. The global texture–geometry conflict benchmark in the main paper evaluates extrapolation from local training-time perturbations to global object-level texture replacement. Because both training and test protocols belong to the same broad texture-swapping family, the held-out external-projection benchmark evaluates transfer to a texture generator outside that family.

External texture projection. We construct a held-out material library that is never used during training, checkpoint selection, or hyperparameter tuning. External textures are mapped onto the input point cloud using tri-planar projection. In contrast to STS, this protocol uses neither intra-scene source textures, semantic source-category labels, nor STS region masks. Given the original point texture T and the externally projected texture T_{ext}, the shifted observation is constructed as:

$$T' = (1-\alpha)T + \alpha T_{\mathrm{ext}}, \tag{32}$$

where $\alpha \in \{0.25, 0.50, 0.75, 1.00\}$ controls the projection strength. We freeze all model parameters during evaluation and use the same projected point clouds for all compared methods. In Table 15, Avg. is the mean over the four shifted severities, Worst is the minimum over those severities, and Ret. is $100 \times \mathrm{mIoU}_{\alpha=1}/\mathrm{mIoU}_{\mathrm{clean}}$.

Compared backbones. Table 15 reports results for both backbones. All methods use the same held-out material library, projected point clouds, and frozen evaluation protocol.

Generic RGB-space corruptions. The fog, brightness, and hue families defined in Appendix H modify only RGB observations and are absent from STS training.

Appearance-side results. Table 15 shows higher mIoU than the reported RGB–geometry comparators at every external-texture strength, while Table 16 shows a higher RGB-C average than the listed RGB-using methods. These results support transfer relative to those comparators. Both protocols alter RGB alone. PointNeXt geometry-only reaches 68.04 clean full-scene mIoU on the same ScanNetV2 split (Table 2). If point sampling and preprocessing are RGB-independent, its performance under these shifts would be expected to remain near this value, potentially exceeding our measured full-scene mIoU of 55.40 under complete external projection and 63.88 on average under RGB-C (Tables 15–16). This is an unverified inference from the protocol, not a reported geometry-only result; a matched evaluation on the same shifted samples is required.

Method	Clean	$\alpha = .25$	$\alpha = .50$	$\alpha = .75$	$\alpha = 1$	Avg.↑	Worst↑	Ret.↑
PointNeXt	70.8	67.4	60.1	49.2	39.6	54.1	39.6	55.9
+ CorrAug	**71.0**	67.3	62.0	52.7	43.5	56.4	43.5	61.2
+ Ours	69.3	**67.9**	**64.8**	**60.7**	**55.4**	**62.2**	**55.4**	**80.0**
PTv3	77.2	74.8	69.8	62.7	56.0	65.8	56.0	72.5
+ CorrAug	77.4	75.2	71.0	64.8	58.5	67.4	58.5	75.6
+ Ours	**77.6**	**75.9**	**72.5**	**68.4**	**64.2**	**70.3**	**64.2**	**82.7**

Table 15: External-texture severity sweep on ScanNetV2.

Method	Fog-3	Fog-5	Bright.-3	Bright.-5	Hue-3	Hue-5	Avg.↑
PointNeXt baseline	56.81	42.64	61.72	49.81	58.96	44.73	56.84
+ CorrAug	57.42	43.87	62.17	50.63	59.75	45.99	57.42
+ RSMix-style	59.97	47.14	63.53	52.86	61.85	49.51	59.73
+ Ours	**64.73**	**56.34**	**66.87**	**60.21**	**65.95**	**57.67**	**63.88**

Table 16: Generalization to held-out RGB-space corruptions on ScanNetV2. Results are mIoU (%); Avg. covers all families and severities.

Method	Clean	N-3	N-5	J-3	J-5	D-3	D-5
PointNeXt	70.8	51.0	37.4	46.1	42.1	44.7	31.6
+ CorrAug	**71.0**	58.1	42.3	54.3	45.1	51.5	38.6
+ Ours	69.3	**62.5**	**50.8**	**68.2**	**54.1**	**68.8**	**45.9**
PTv3	77.2	65.0	58.9	62.6	61.8	67.8	53.7
+ CorrAug	77.4	67.1	61.2	65.9	62.1	68.3	57.8
+ Ours	**77.6**	**75.2**	**68.8**	**70.1**	**68.9**	**72.2**	**64.2**

Table 17: Out-of-intervention generalization to geometry corruptions on ScanNetV2. N, J, and D denote noise, jitter, and dropout.

Method	Ext.+N-3	Ext.+J-3	Ext.+D-3	Avg.↑
PointNeXt baseline	28.43	30.17	25.75	28.12
+ CorrAug	34.84	36.11	32.92	34.62
+ Ours	**47.60**	**49.10**	**44.80**	**47.17**

Table 18: Joint appearance and geometry degradation on ScanNetV2. Results are mIoU (%).

F.2 Geometry-Side and Joint Degradation

Geometry-side corruptions. The complete framework is trained without the coordinate-noise, jitter, or point-dropout families defined in Appendix H; its intervention-specific supervision is generated only from texture-side STS pairs. These tests measure system-level out-of-intervention robustness and do not isolate CORD or imply bidirectional reliability supervision.

Joint composition. We combine complete external texture projection ($\alpha = 1.0$) with each severity-3 geometry corruption. This composition is absent from training and tests whether the model remains usable when neither modality is fully clean.

Geometry-side and joint results. Table 17 shows that the complete model attains higher mIoU than both comparators for every reported severity-3 and severity-5 geometry corruption, with gain magnitude varying by backbone and corruption. Table 18 evaluates the stricter composition of external texture and severity-3 geometry corruption; average mIoU is 47.17, compared with 34.62 for CorrAug and 28.12 for the baseline. These results characterize the complete framework under the tested geometry-side and joint degradations.

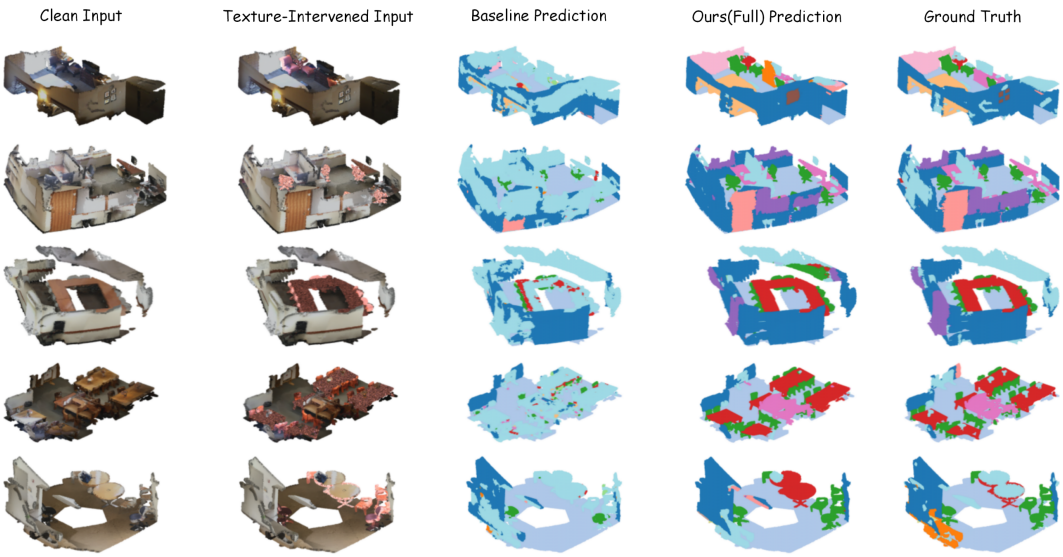

Figure 4: Additional scene-level comparisons under global semantic texture conflicts. Each row shows the clean input, texture-intervened input, baseline prediction, full-model prediction, and ground truth. Both models receive only the texture-intervened input at inference; the clean input is shown solely as a visual reference. All comparisons use identical point sampling and camera viewpoints.

G Additional Qualitative and Feature-Space Analysis

This appendix provides three complementary qualitative analyses: scene-level segmentation results under global semantic texture conflicts, object-level target-retention cases, and t-SNE projections of post-fusion representations. Within each prediction comparison, the baseline and the full framework use identical inputs, intervention settings, point sampling, and camera viewpoints. Clean inputs are shown only as visual references and are not provided to either model at inference.

G.1 Scene-Level Semantic-Conflict Results

Figure 4 provides five matched scene-level comparisons under global semantic texture conflicts.

G.2 Object-Level Texture-Conflict Cases

Figure 5 provides object-level examples in which sofa geometry is paired with wall appearance. These examples isolate the target object and illustrate the difference in target retention between the baseline and the full framework. They are selected qualitative cases; aggregate source-directed and non-source-directed errors are quantified by TDR and OER in Appendix D.4.

G.3 Fused-Feature Visualization

Figure 6 visualizes post-fusion point features using t-SNE. The semantic-class view examines within-model class organization, while the texture-condition view provides a qualitative diagnostic of how original and texture-intervened samples are distributed in the learned feature space. The projections are used only for qualitative inspection; quantitative conclusions are based on the segmentation and robustness metrics.

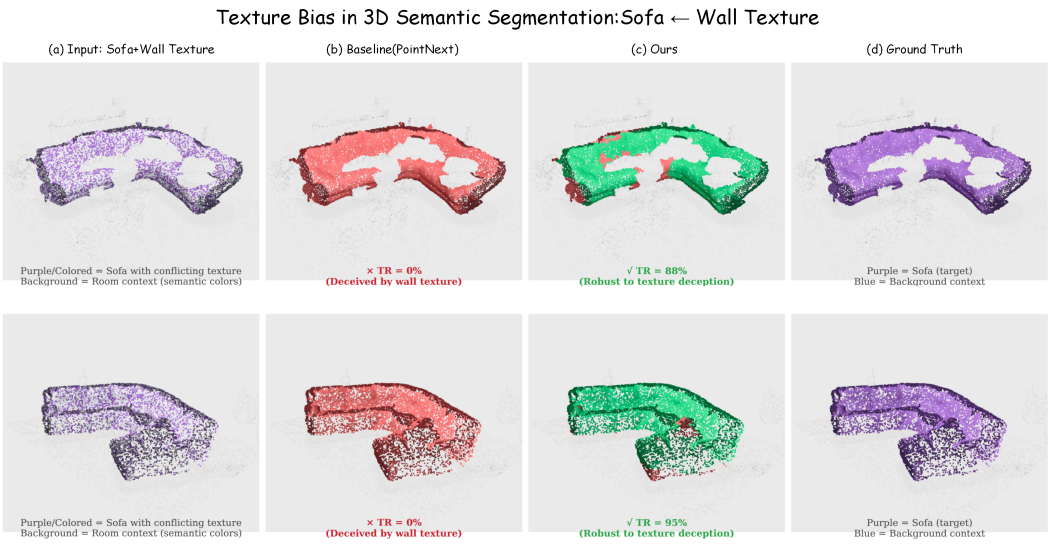

Figure 5: Object-level semantic texture conflicts for sofa instances.

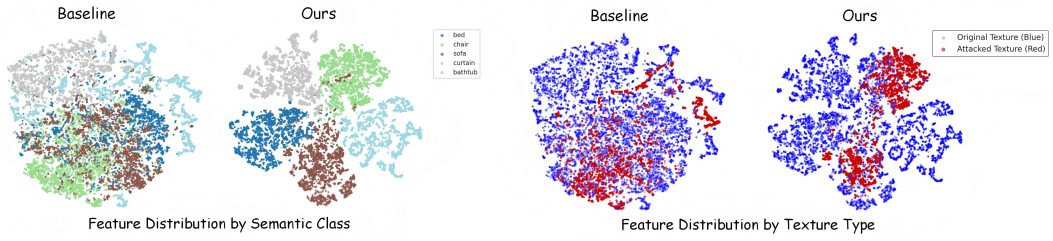

Figure 6: t-SNE visualization of post-fusion point features. Left: features colored by semantic category. Right: features colored by texture condition. Baseline and full-model panels use matched evaluation samples and identical t-SNE hyperparameters. The projections are computed separately for the two models; therefore, absolute coordinates and cross-panel distances are not directly comparable.

H Consolidated Experimental and Implementation Protocol

This section consolidates the fixed dataset splits, training configuration, baseline definitions, and evaluation metrics shared by all experiments.

H.1 Datasets and Evaluation Splits

For S3DIS, we follow the standard Area 5 protocol. For ScanNetV2, models are trained on the official training split and evaluated on the validation split. Unless otherwise stated, all methods use identical point sampling, RGB–geometry inputs, preprocessing, optimization schedules, and checkpoint-selection criteria. Corrupted validation samples are not used for model selection, hyperparameter selection, or early stopping.

H.2 Backbones and Implementation

We instantiate the framework with PointNeXt and Point Transformer V3 (PTv3), representing convolutional and Transformer-based point-cloud backbones, respectively. CORD uses the candidate family $\mathcal{A} = \{0, 0.25, 0.5, 0.75, 1\}$. Candidate exposure is applied only to clean observations with

Component	Setting	Value
STS	neighborhood size K_{nn}	128
STS	intervention probability / affected-point ratio	0.5 / 5–10% target
Candidate bank	$\mathcal{A}$ and M_a	$\{0, 0.25, 0.5, 0.75, 1\}$; 5
Candidate exposure	sampled candidates per clean batch	1; encoder-detached
EMA teacher	momentum β_{ema} / CORD schedule	0.99 / 10-epoch burn-in + 10-epoch linear ramp
Candidate oracle	λ_{clean}, η, τ, τ_c	0.25 / 0.50 / 0.20 / 1.00
CORD graph	K_{C}, $\sigma_{g,c}$, $\sigma_{n,c}$	16 / median KNN edge length / 0.30
CORD refinement	μ, ϵ_a, Jacobi iterations	0.20 / 10^{-3} / 5
SA-TSI graph	K_{S}, σ_g, σ_n	16 / median KNN edge length / 0.30
SA-TSI	residual weight β_{m} / boundary-band width	0.50 / two graph hops
Training objective	λ_{pair}, λ_{cand}, $\lambda_{\mathrm{C}}^{\max}$, λ_{TSI}	0.50 / 0.10 / 0.50 / 0.10
Optimization	optimizer / epochs / effective batch size / base learning rate	AdamW / 100 / 8 / 10^{-3}

Table 19: Implementation settings used in the reported experiments.

detached modality features, while candidate costs are evaluated by the EMA teacher through the complete post-fusion decoder and prediction head.

All experiments are conducted on a single NVIDIA RTX 4090 GPU running Ubuntu 22.04. PointNeXt uses PyTorch 2.5.1 with CUDA 12.1, while PTv3 uses PyTorch 2.8.0 with CUDA 12.6. The EMA momentum, CORD warm-up schedule, candidate temperatures, graph construction, normal-estimation settings, and loss coefficients are fixed across controlled comparisons. All compared methods use the same data splits, training schedules, fixed random seed within each paired comparison, and clean-validation checkpoint-selection criterion. Reported entries are point estimates under this paired protocol; sub-point differences are not interpreted as statistical separation.

Encoder features are reused across candidates. Nevertheless, evaluating both the clean and intervened risks requires $2M_a = 10$ complete EMA post-fusion decoder/head passes per paired scene during training. The EMA teacher, candidate bank, and graph solver are discarded at inference. Measured end-to-end training time, memory, and inference cost are reported in Table 12.

H.3 Compared Methods

In addition to the original RGB–geometry backbone, we compare against geometry-only and texture-only models, modality dropout, geometry-corruption augmentation under a matched training budget (CorrAug), and an adapted RSMix-style region-mixing baseline. CorrAug follows the same training schedule as our method and applies Gaussian coordinate noise, bounded coordinate jitter, and random point dropout, but does not use STS, CORD, or SA-TSI. The RSMix-style baseline uses the same number and size of mixed regions as STS but does not construct paired texture interventions or reliability targets. MS-FER + STS augments the fixed-fusion baseline with the published prediction-level and four-scale feature-level MS-FER terms. It uses the same structured STS clean–intervened pairs, paired-consistency loss, optimization steps, and clean-validation selection as our method, while excluding GT-IGU, candidate exposure, CORD, and SA-TSI. Ours w/ class-agnostic donor retains the full framework and changes only the training donor sampler as defined in Appendix D.4. Parameter-matched fusion baselines are reported with the detailed ablations.

H.4 Clean and Semantic-Conflict Metrics

We report overall accuracy (OA), mean class accuracy (mAcc), and mean intersection-over-union (mIoU) on clean scenes. The Clean column in main-paper Table 2 denotes full-scene ScanNetV2 validation mIoU. OER is algebraically determined by TR and TDR and is retained only as a derived diagnostic; it is not displayed as an independent column in main-paper Table 2. For semantic-conflict evaluation, let Ω_{eval} be the affected point set, $\widehat{y}_i$ the predicted category, y_i the original target category, and $c_{\mathrm{src},i}$ the category from which the replacement texture is obtained. Target Retention

(TR) and Texture Deception Rate (TDR) are

$$\begin{aligned} \mathrm{TR} &= \frac{100}{|\Omega_{\mathrm{eval}}|} \sum_{i \in \Omega_{\mathrm{eval}}} \mathbb{I}[\widehat{y}_i = y_i], \\ \mathrm{TDR} &= \frac{100}{|\Omega_{\mathrm{eval}}|} \sum_{i \in \Omega_{\mathrm{eval}}} \mathbb{I}[\widehat{y}_i = c_{\mathrm{src},i}]. \end{aligned} \tag{33}$$

For cross-class replacement, where $c_{\mathrm{src},i} \neq y_i$, we additionally report

$$\mathrm{OER} = 100 - \mathrm{TR} - \mathrm{TDR}. \tag{34}$$

OER is not defined for same-class or source-free randomized controls. We also report conflict-region mIoU ($\mathrm{mIoU}_{\mathrm{conf}}$), computed only over the affected points and evaluated target categories; it is therefore not directly comparable with full-scene clean mIoU. Target–source category pairs and intervention masks are generated before evaluation and shared by all methods.

Evaluation support. The global-conflict protocol uses the five target categories per dataset listed in Table 23. TR, TDR, derived OER, and conflict-region mIoU are restricted to this fixed support, whereas Clean is evaluated over the complete validation split. Consequently, claims concern the selected categories and pre-generated target–source pairs, not unlisted categories.

H.5 Corruption Metrics

After coordinate normalization, Gaussian coordinate noise uses standard deviation $0.05s$, bounded coordinate jitter uses amplitude $0.10s$, and random point dropout uses probability $0.10s$, where $s \in \{1, 2, 3, 4, 5\}$. The same corrupted samples are shared by all methods. Dropped points are handled by the fixed-size sampling strategy used by the corresponding backbone. The main paper reports severity 3, and Table 17 reports severities 3 and 5. These are representative points from the same five-level sweep rather than separately tuned conditions. All compared methods receive the same pre-generated corrupted samples at every severity.

H.6 Appearance Corruption Protocols

Fog, brightness shift, and hue shift are applied only to the RGB observations; point coordinates and semantic labels remain unchanged. None of these operators is used during STS training. The six displayed columns in Table 16 show severities 3 and 5, whereas Avg. averages all five severities of all three families. Every method is evaluated on the same pre-generated observations with a frozen checkpoint; no appearance-corruption result is used for selection or tuning.

I Per-Class Results

To avoid duplicating the main paper, we report only the per-class quantities here. The sensor-corruption settings are identical to those defined in Appendix H. Within each backbone and evaluation condition, bold values indicate the better result; ties are left unbolded.

I.1 Per-Class Sensor-Corruption Results

Tables 20– 22 show that sensor-corruption gains span multiple categories, while retaining near-zero baseline IoUs and occasional negative changes. Table 23 lists all global-conflict targets; TR improves for every target and both backbones, with category-dependent magnitude. Thus, neither result is driven by a single class, although conclusions remain restricted to the listed support.

Corr.	Method	Ceil	Floor	Wall	Beam	Col.	Win.	Door	Chair	Table	Bkc.	Sofa	Board	Clut.
Clean	PointNeXt	94.0	98.6	84.9	0.0	36.8	61.4	73.7	83.2	**91.7**	76.0	**77.5**	77.4	59.7
	+ Ours	**95.1**	98.6	**85.5**	0.0	**47.3**	**74.6**	**78.8**	**85.0**	91.0	**77.3**	77.1	**78.6**	**73.1**
	PTv3	91.7	97.8	**85.8**	0.0	**27.6**	63.6	70.2	92.4	82.3	77.5	**81.4**	83.7	54.9
	+ Ours	**93.3**	**98.2**	85.3	0.0	25.0	**64.9**	**82.5**	**93.3**	**82.9**	**80.2**	79.8	**85.4**	**65.4**
Noise	PointNeXt	87.8	84.5	60.1	0.0	0.0	34.2	16.5	44.3	70.0	47.4	39.9	20.4	44.4
	+ Ours	**94.7**	**98.2**	**83.9**	0.0	**41.6**	**61.9**	**71.4**	**64.9**	**88.6**	**73.2**	**72.8**	**67.6**	**71.2**
	PTv3	82.4	86.3	23.2	0.0	0.0	14.3	50.6	91.1	64.6	40.0	59.3	11.1	46.1
	+ Ours	**93.4**	**97.8**	**81.8**	0.0	**23.4**	**46.7**	**74.0**	**91.4**	**82.7**	**79.3**	**75.2**	**80.5**	**64.0**
Jitter	PointNeXt	91.6	97.3	71.4	0.0	0.0	35.8	21.1	51.5	74.0	50.8	47.0	34.5	49.2
	+ Ours	**95.0**	**98.5**	**84.8**	0.0	**46.8**	**63.5**	**74.3**	**70.5**	**89.3**	**76.5**	**74.9**	**76.3**	**72.2**
	PTv3	81.9	85.0	9.5	0.0	0.0	12.7	42.1	**90.6**	60.3	34.8	49.5	0.4	44.8
	+ Ours	**92.6**	**97.1**	**72.5**	0.0	**16.7**	**27.3**	**70.6**	89.4	**82.5**	**77.0**	**72.3**	**60.1**	**61.6**
Dropout	PointNeXt	88.0	83.3	63.5	0.0	2.9	37.1	20.3	45.5	71.0	34.3	43.2	39.5	46.5
	+ Ours	**94.6**	**98.1**	**83.7**	0.0	**39.7**	**62.1**	**71.3**	**62.9**	**88.7**	**73.1**	**71.6**	**68.6**	**70.9**
	PTv3	86.0	89.8	24.0	0.0	1.8	57.0	24.7	90.3	66.3	30.9	49.3	44.8	**32.6**
	+ Ours	**93.3**	**98.0**	**30.5**	0.0	**14.4**	**65.2**	**54.5**	**91.6**	**82.0**	**43.3**	**82.0**	**73.7**	31.7

Table 20: Per-class IoU (%) under clean and severity-3 corrupted observations on S3DIS.

Corr.	Method	Wall	Floor	Cab.	Bed	Chair	Sofa	Table	Door	Wind.	BkShf
Clean	PointNeXt	**81.5**	93.9	59.6	**82.9**	**86.1**	**85.4**	69.2	56.1	**68.0**	69.8
	+ Ours	81.3	93.9	**60.2**	81.1	85.0	84.4	**69.6**	**56.6**	64.6	**72.0**
	PTv3	87.6	95.8	**74.2**	83.1	**92.7**	83.1	**77.9**	76.1	73.0	82.0
	+ Ours	87.6	**95.9**	74.1	**83.2**	92.6	**83.3**	77.7	76.1	**73.1**	**82.6**
Noise	PointNeXt	45.1	92.9	38.3	75.0	82.1	**79.5**	62.1	37.4	46.9	59.5
	+ Ours	**72.0**	**93.6**	**50.3**	**77.1**	**83.7**	79.4	**67.7**	**44.5**	**49.0**	**66.5**
	PTv3	77.4	95.2	53.9	78.8	89.8	**85.1**	70.3	52.9	54.4	78.1
	+ Ours	**85.1**	**95.3**	**70.2**	**82.2**	**90.5**	85.0	**76.1**	**68.7**	**70.5**	**83.0**
Jitter	PointNeXt	35.4	92.7	34.9	72.9	80.9	77.6	60.8	33.9	42.6	56.8
	+ Ours	**80.1**	**93.9**	**57.3**	**80.9**	**84.6**	**84.0**	**69.6**	**55.2**	**61.8**	**70.5**
	PTv3	76.0	95.0	50.2	77.8	89.5	83.6	69.2	50.2	51.6	77.8
	+ Ours	**78.1**	95.0	**63.0**	77.8	89.5	83.6	**69.5**	**61.8**	**63.6**	77.8
Dropout	PointNeXt	37.2	91.9	30.0	64.9	78.9	74.0	59.4	23.6	34.6	56.8
	+ Ours	**80.5**	**93.9**	**59.2**	**80.4**	**84.8**	**83.8**	**69.5**	**55.3**	**63.0**	**70.9**
	PTv3	79.1	95.1	59.3	78.4	90.3	**85.8**	71.9	55.8	54.9	79.8
	+ Ours	**80.3**	95.1	**65.6**	**79.5**	**90.7**	85.7	71.9	**62.7**	**65.9**	79.8

Table 21: Per-class IoU (%) under clean and severity-3 corrupted observations on ScanNetV2 (wall through bookshelf).

Corr.	Method	Pic.	Cntr	Desk	Curt.	Frid.	Show.	Toil.	Sink	Bath	Other
Clean	PointNeXt	**47.0**	**50.9**	**57.8**	**77.2**	71.9	**55.4**	**87.4**	**63.3**	**89.1**	**63.4**
	+ Ours	43.3	48.5	57.5	72.3	**74.3**	51.1	81.8	60.9	87.2	60.2
	PTv3	42.0	71.1	70.6	77.5	**72.3**	68.2	91.9	69.8	88.8	66.3
	+ Ours	42.0	71.1	**71.0**	**77.8**	72.0	**72.8**	**92.0**	**71.9**	**88.9**	66.3
Noise	PointNeXt	3.8	35.0	51.3	55.6	58.5	27.5	30.5	37.6	60.1	41.0
	+ Ours	**20.5**	**45.8**	**55.2**	**72.2**	**69.2**	**40.0**	**74.4**	**54.6**	**81.6**	**52.8**
	PTv3	10.8	69.8	66.1	33.2	39.4	56.2	91.8	67.5	82.9	47.3
	+ Ours	**30.2**	69.8	**71.0**	**75.9**	**67.5**	**72.3**	91.8	**71.8**	**88.7**	**58.4**
Jitter	PointNeXt	3.4	29.3	50.3	48.9	52.0	24.0	14.9	26.6	47.0	36.4
	+ Ours	**41.4**	**48.2**	**57.8**	**72.7**	**72.8**	**49.8**	**78.8**	**59.7**	**86.8**	**58.6**
	PTv3	7.3	69.5	64.6	21.3	34.9	53.9	90.1	67.0	77.7	45.3
	+ Ours	**23.2**	69.5	64.6	**69.0**	**60.4**	**65.4**	90.1	67.0	**81.7**	**51.4**
Dropout	PointNeXt	3.5	32.5	46.0	45.1	56.9	23.6	21.0	22.6	56.9	34.5
	+ Ours	**41.5**	**47.8**	**57.9**	**74.8**	**73.5**	**49.0**	**83.3**	**60.3**	**86.5**	**59.8**
	PTv3	16.3	70.4	67.5	51.3	42.8	60.6	93.9	68.4	85.9	47.9
	+ Ours	**25.0**	**70.8**	**67.7**	**70.6**	**62.9**	**67.8**	93.9	**68.9**	**86.4**	**52.8**

Table 22: Per-class IoU (%) under clean and severity-3 corrupted observations on ScanNetV2 (picture through other; continuation of Table 21).

S3DIS						ScanNetV2					
Method	**Window**	**Door**	**Chair**	**Table**	**Sofa**	**Method**	**Chair**	**Sofa**	**Table**	**Toilet**	**Bathtub**
PointNeXt	3.75	33.81	80.15	93.67	55.51	PointNeXt	58.77	3.48	2.58	9.92	1.35
+ Ours	**82.15**	**80.51**	**87.90**	**93.73**	**87.80**	+ Ours	**93.67**	**92.80**	**83.05**	**89.78**	**90.36**
Gain	+78.40	+46.70	+7.75	+0.06	+32.29	*Gain*	+34.90	+89.32	+80.47	+79.86	+89.01
PTv3	61.99	22.92	29.66	84.57	64.97	PTv3	84.31	35.96	60.35	65.40	26.19
+ Ours	**71.00**	**46.00**	**54.58**	**91.00**	**73.45**	+ Ours	**90.80**	**81.00**	**76.00**	**86.00**	**43.00**
Gain	+9.01	+23.08	+24.92	+6.43	+8.48	*Gain*	+6.49	+45.04	+15.65	+20.60	+16.81

Table 23: Per-class TR (%) under global semantic texture conflicts on S3DIS and ScanNetV2. Gains are relative to the corresponding backbone.