

LiEqualizer: Seeing the Lost Spectrum of Sparse LiDARs via Spectral-Invariant Distillation

Zikun Xu^{1,2}, Shuocheng Yang¹, Jianqiang Wang^{1*}, Keqiang Li^{1*},
Gim Hee Lee^{2*}, Shaobing Xu^{1*}

¹School of Vehicle and Mobility, Tsinghua University, Haidian, Beijing,
100084, China.

²Department of Computer Science, National University of Singapore,
117418, Singapore.

*Corresponding author(s). E-mail(s): wjqlws@tsinghua.edu.cn;
wjqlws@tsinghua.edu.cn; gimhee.lee@nus.edu.sg;
shaobxu@tsinghua.edu.cn;

Contributing authors: xuzk23@mails.tsinghua.edu.cn;
ysc24@mails.tsinghua.edu.cn;

Code, Data, Video: <https://easonxu-01.github.io/LiEQ-Simple/>

Abstract

Engineers and researchers pay twice for LiDAR perception: a cheap sensor with few beams sees less of the scene, and every new sensor means new data, labels and training, team after team. Recent work cuts that cost by randomizing the sampling pattern during training [1], but ignoring a pattern does not return the measurements it never took, and putting them back has meant reconstructing the scan before the network reads it. *Different LiDARs do not see different worlds; they see different spectra of the same world.* Read as frequencies, sparse sampling keeps a scan’s coarse shape but weakens real detail and invents false detail, which a filter cannot do but a sampler can. We prove sparse scans alone cannot bring the weakened detail back, but a teacher trained on densified scenes can. Based on this, we propose **LiEqualizer**, a distillation method that equalizes the student’s spectrum against that teacher during training. A low-frequency branch restores what sampling weakened, in the encoder’s compressed code, returning structure its beams missed; a high-frequency branch removes what sampling faked, in the decoder’s voxel logits, so false detail cannot pass as real; each acts only where the teacher’s response marks something to restore or remove. Across SemanticKITTI, Waymo and nuScenes, LiEqualizer delivers consistent zero-shot gains, raising cross-sensor mIoU by up to **11.5** points without altering the network, its parameter count, or the cost of training it. We also propose

an open protocol for measuring such spectra. Its losses act only in training, so the deployed model keeps the baseline’s architecture and inference cost, and the unchanged model transfers to three vehicles carrying unseen LiDARs. Deploying a new LiDAR therefore becomes a software change, not another round of data collection.

Keywords: Cross-LiDAR generalization, LiDAR semantic segmentation, Knowledge distillation, Graph spectral analysis, Frequency-domain learning

1 Main

LiDAR provides autonomous vehicles and mobile robots with precise three-dimensional measurements of their surroundings, and semantic segmentation, which assigns a class label to every point, is one of its core perception tasks. Benchmark accuracy on this task has improved steadily. However, the improvement is tied to the sensor that collected the training data. When the same model is applied to a different LiDAR, with fewer beams, another scan pattern or a different mounting height, its zero-shot accuracy drops sharply [1]. In practice every new sensor then requires its own collection, annotation and retraining, which hinders cross-vehicle deployment, slows the adoption of newer sensors, and increases development cost. Supporting each of them with freshly annotated data is expensive, since annotating LiDAR point clouds takes roughly $30\times$ the effort of annotating images [2]. A method that transfers to an unseen LiDAR without that effort would therefore be of substantial practical value.

Most LiDAR segmentation models are trained and tested on scans from the same sensor. Such a setting lets a model lean on the sampling pattern it has been shown instead of the scene behind it. As a result, accuracy falls as soon as that pattern changes. Existing research usually relieves this issue in three ways, and each stops at a different point. Domain adaptation aligns features towards one known target sensor [3–5], which is effective for that sensor but requires target-domain data and a further round of training for every additional one. Domain generalization instead regularizes features against restricted forms of sensor variation such as beam dropping or density change [6, 7], or reconstructs a denser scene and re-renders it under predefined configurations [8, 9]. Either way it improves robustness against the variations it enumerates, but not beyond them. A third group transfers knowledge from a denser observation to a sparser one by distillation [10, 11]. This is the closest setting to ours, but the transfer is carried out by matching features or predictions point by point. Yet these approaches remain tied to what they were given: a particular target sensor, a list of variations, or a point-to-point correspondence. More fundamentally, none of them provides an explicit mechanism to say what sparse sampling takes out of the signal. A transfer that does not know what was removed cannot know what is worth putting back. A sparser sensor does return less of the scene, and that loss is ordinarily left to the hardware to fix. Software has put some of it back, but so far only by reconstructing the scan before the network reads it, which the deployed system then pays for on

every frame. A natural question then arises: *what does a sparse LiDAR lose, and can it be put back without paying for it at inference?*

LiEqualizer

Seeing the Lost Spectrum: Recover Structure, Suppress Artifacts

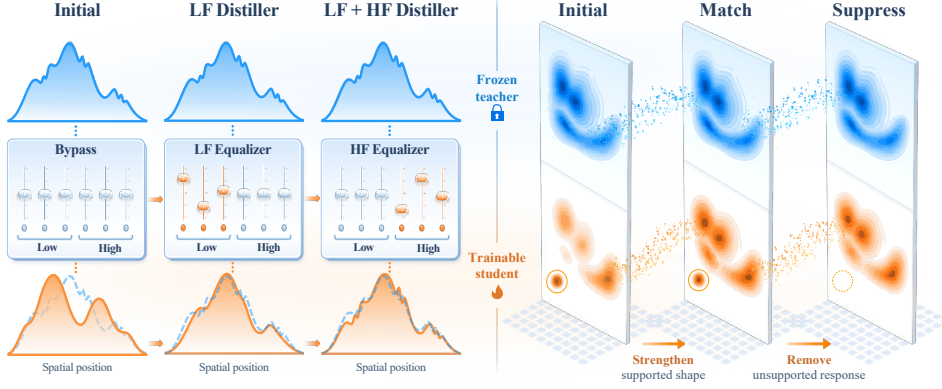


Fig. 1: Conceptual illustration of LiEqualizer.

Note. Different LiDARs observe different spectra of the same world. Sparse sampling attenuates real high frequencies in the point domain and, unlike a filter, creates near-white fake high frequencies on the voxel grid. LiEqualizer distills the spectrum from a teacher trained on densified scenes and equalizes the student’s spectral response, matching detail where the teacher has it and suppressing ripple where it does not.

This work answers this question from a new perspective. Part of the vocabulary for it is already in place [1]. A recent study views every LiDAR as a sparse sampler of the same underlying 3D world, and splits a scan into two factors. *Scene-intrinsic geometry* (“SIG”) is the set of surfaces that exist in the scene whatever the sensor, and *sampling geometry* (“SG”) is which points on those surfaces a particular sensor returns. Under that view, cross-LiDAR generalization reduces to learning “SIG” without overfitting “SG”, and the practical recipe is to randomize “SG” during training. Randomization, however, teaches a model which variations to ignore, and ignoring a sampling pattern is not the same as recovering what it left out. To say what is lost, rather than which variations to ignore, we adopt the language of graph spectra, in which each feature channel is treated as a signal on a graph. Low graph frequencies describe smooth, large-scale structure such as the road surface, terrain and building facades, and high graph frequencies describe fine detail such as curbs, lane markings, poles and object boundaries. In this language our finding is simple:

Different LiDARs do not see different worlds; they see different spectra of the same world.

“SG” is a sparse sampling operator acting on “SIG”: the sensor decides which frequencies of the scene survive in the scan and which are corrupted, and the difference between two LiDARs is the difference between their spectral curves. From this perspective, the essence of cross-sensor transfer is not to rebuild the scan but to repair the spectrum: to put back the bands that sampling took away, and to take out the bands it added. If the gap is spectral, then so is the fix: part of it can be closed by reshaping the student’s spectrum against a reference that still holds the whole spectrum, and

99 whether that reshaping has happened can be measured, band by band, before and
100 after training.

101 We measure this directly, by comparing dense and sparse observations of the same
102 scenes. The result has to be read in two domains, because the damage runs in opposite
103 directions in each. On the raw point cloud, sparsity removes real high frequencies,
104 and the loss is not uniform across classes. Smooth expanses such as road surfaces lose
105 the most, whereas porous vegetation instead gains spurious high-frequency energy as
106 sampling fragments its support. On the voxel grid that segmentation networks operate
107 on, the sparse branch shows the opposite, carrying near-white high-frequency energy
108 while its low bands still track the dense branch closely. The two observations are
109 not in conflict. A filter can only attenuate frequencies that already exist, whereas a
110 sampler also creates components the scene never contained. Voxelization makes this
111 visible by turning the sampling pattern itself into signal content: isolated occupied
112 voxels and staircase boundaries are high-frequency structures on the grid, and a dense
113 scan does not produce them. One undersampling therefore produces two symptoms,
114 and a single per-frequency model captures both: the student spectrum is the teacher
115 spectrum multiplied by a decaying attenuation, plus a flat aliasing noise floor.

116 That model makes the question tractable, and the answer has two parts. The atten-
117 uated content is not recoverable from sparse observations alone, because consistency
118 between two sparse views cannot invert the attenuation when both views have lost
119 the same frequencies. It is recoverable from a teacher trained on the densified scenes,
120 which has observed the full spectrum. Two obstacles stand in the way. Spectra cannot
121 be computed explicitly at the sizes involved, and matching the student to the teacher
122 point by point at high frequencies would copy the aliasing we measured rather than the
123 detail we want. Our theorem removes both. The energies surviving repeated smoothing
124 identify the entire spectral-energy distribution without any eigendecomposition,
125 and a gate driven by the teacher keeps the unreliable channel from being copied.

126 Based on this insight we propose **LiEqualizer**, a distillation method for zero-shot
127 cross-sensor segmentation that works as its name suggests. An audio equalizer does
128 not re-record a signal. It corrects the balance between the frequency bands that the
129 channel has distorted, and it is set against a reference response. LiEqualizer takes a
130 teacher trained on densified scenes as that reference: during training a sparse scan
131 passes through the student and a dense observation of the same scene through the
132 frozen teacher, and the comparison is made band by band instead of point by point. In
133 this way the teacher supplies the spectrum a sparse scan can no longer carry, and its
134 own response says where that spectrum can be trusted. To equalize the student against
135 it, LiEqualizer introduces two complementary branches. The first is the low-frequency
136 branch, which acts as the boost: on the encoder’s projected feature planes, where
137 sparse sampling has attenuated the bands that carry large-scale structure, it restores
138 them, so that the student recovers layout its own beams never resolved. The second is
139 the high-frequency branch, which acts as the cut: on the decoder’s voxel logits, where
140 sampling has instead injected a band the scene never contained, it removes that band,
141 so that the student’s own return pattern stops being read as detail. Each reads where
142 to act off the teacher and never off the student, so it acts only where the teacher’s own
143 response says there is something to restore or to remove. Together, these two branches

realize the core insight: because the gap between sensors is a gap between spectra, the student is brought towards the teacher by equalizing its spectral response rather than by reconstructing its scan. What is restored is a teacher-consistent estimate of the reliable spectral *energy*, not a unique dense scene and not the measurements the sensor never returned. All of these terms act during training only. In practice the distillation modules are never invoked at inference and contribute no trainable parameters, so the deployed student is the baseline network itself, and the part of the sampling gap this estimate closes is a property of how the network was trained rather than of the sensor it runs on.

Experiments on SemanticKITTI, Waymo and nuScenes under a source-to-unseen protocol, together with qualitative deployment on three vehicle platforms carrying previously unseen 32-, 80- and 128-beam LiDARs, ask whether the equalized spectrum buys accuracy on sensors the model has never seen. On the benchmarks, LiEqualizer adds up to 11.5 zero-shot cross-sensor mIoU on top of the pipeline of [1], and it does so without changing the architecture, the parameter count or the training budget. Two further tests address the mechanism rather than the outcome: a comparison against prior distillation objectives under one pipeline, with the same teacher and student throughout, separates the contribution of the spectral objective from that of the dense teacher itself, and a re-measurement of the student’s voxel spectrum after distillation asks whether the excess high-frequency energy has in fact been removed. The practical reading is simple: because the losses act at training time only, part of what a sparse LiDAR fails to see can be put back in software, and a new sensor is served without paying for that recovery on every frame. The main contributions of this paper are as follows:

- **A spectral account of sampling geometry.** We measure, in the graph-frequency domain, what “*SG*” does to the signal. Real high frequencies are attenuated on raw point clouds, and the attenuation follows a class law set by density and geometry, while near-white fake high frequencies appear on the voxel grids the network operates on. Both symptoms follow from one property, that “*SG*” acts as a sampler rather than a filter, and both are unified in a single per-frequency transfer model.
- **A unified theorem for spectral distillation under sampling.** One self-contained theorem establishes four properties: identifiability, in that the responses surviving repeated smoothing determine the spectral-energy distribution; finite-scale stability, in that a short prefix of responses controls every Lipschitz spectral statistic, with an explicit bound; denoising under sampling, in that each smoothing step provably improves the reliability of the comparison and ungated residuals are never more reliable than the corresponding low-pass channel; and the optimality of complementary teacher gates for low- and high-frequency transfer. A companion proposition shows that consistency between sparse views cannot invert the attenuation, which is why a spectrally complete teacher is necessary rather than merely convenient.
- **LiEqualizer.** A distillation method in which every loss term is dictated by one part of the theorem. Added to the pipeline of [1] at training time only, it improves zero-shot cross-sensor mIoU by up to 11.5 across SemanticKITTI, Waymo and nuScenes,

188 and it does so without a single change to the network, to its parameter count or to
 189 what training it costs.
 190 • **An open spectral measurement protocol.** Paired comparison suites for raw
 191 point clouds and voxel logits, covering retention, band divergence, energy ratios and
 192 flatness, released so that sampling degradation and learned representations can be
 193 analyzed by others under the same definitions.

194 2 Results

195 2.1 Experimental Settings

196 Every experiment below shares one optimization setting, one label space and one evalu-
 197 ation script, and no configuration is tuned per experiment. Additional implementation
 198 details are provided in Appendix G.5.

199 2.1.1 Baseline Model Selection

200 LiDAR segmentation models broadly fall into point-based, projection-based, and
 201 voxel-based families, with the latter two dominating large-scale outdoor benchmarks.
 202 Most existing studies [8, 9, 12] evaluate cross-LiDAR generalization on relatively early
 203 segmentation models [13, 14]. Those models give a common benchmark, but their
 204 performance is limited, so they are less representative of current LiDAR perception sys-
 205 tems. To test LiEqualizer against current practice, we adopt one recent, state-of-the-art
 206 and representative segmentation baseline: the projection-based method, PointSeg [15],
 207 called P-baseline, which employs the Swin Transformer [16] backbone.

208 For fair comparison, we do not modify the architecture or hyperparameters of
 209 any baseline. All methods are trained under a unified optimization setting, and no
 210 experiment is individually tuned. Since official pretrained checkpoints are unavailable
 211 for these baselines, all reported results are obtained from our own implementations.

212 2.1.2 Dataset Selection

213 Given the scarcity of large-scale outdoor LiDAR datasets, existing benchmarks pre-
 214 dominantly employ 32-beam or 64-beam LiDAR sensors. We select three representative
 215 datasets widely used in autonomous driving research: nuScenes, which employs a
 216 32-beam LiDAR, and SemanticKITTI and Waymo, both equipped with 64-beam
 217 LiDARs. These datasets cover the most common LiDAR configurations used in modern
 218 autonomous driving benchmarks.

219 To further evaluate the practical value of LiEqualizer beyond standard benchmarks,
 220 we additionally deploy it on three real vehicles equipped with completely unseen 32-,
 221 80- and 128-beam LiDARs, and provide qualitative results in Appendix G.3.

222 2.1.3 Coarse Labels for Cross-Domain Evaluation

223 Label definitions and annotation granularity are not fully aligned across datasets, so
 224 a direct mIoU comparison does not make sense.

225 To address this issue, we adopt a *coarse label* strategy following COLA [17]. Instead
 226 of requiring a one-to-one mapping matrix $\mathcal{T}_{ij} \in \{0, 1\}^{|\mathcal{L}_i| \times |\mathcal{L}_j|}$ between datasets, we

project all dataset-specific labels into a unified semantic space. For the three datasets considered, the unified coarse label set is defined as $\mathcal{L}_{\text{unified}} = \{\text{'Other'}, \text{'Driveable Ground'}, \text{'Structure'}, \text{'Vehicle'}, \text{'Nature'}, \text{'Living Being'}, \text{'Objects'}, \text{'Other Ground'}\}$. This abstraction merges fine-grained categories into broader semantic groups. Detailed mapping rules are provided in Section F. Since all comparisons here share the same coarse label space, they remain directly comparable.

2.1.4 Teacher Construction and Distillation Configuration

The dense teacher shares the student’s architecture and is trained on the densified scenes produced by the data factory of [1]; the student consumes randomized sparse re-samplings of the same scenes, so every training pair observes the same “*SIG*” under different “*SG*”. Distillation follows Sec. 3.4: the teacher is frozen, $\mathcal{L}_{\text{LF}}$ acts on the three TPV planes, $\mathcal{L}_{\text{HF}}$ acts on the common-support voxel logits, and $\mathcal{L}_{\text{KD}}$ compares the corresponding temperature-softened class distributions on that same voxel support. The optimizer, schedule, framework and hardware are the baselines’ own and are listed in Appendix G.5.

2.1.5 What the Spectrum Shows, Before and After

Two spectral readings sit behind the design and behind the claim, and both are reported in full in Appendix D. The first measures what sampling does to the signal, on 600 dense and sparse frame pairs across the three datasets. It is the reason LiEqualizer has two branches rather than one: on the raw point cloud sparsity attenuates real high frequencies, and on the voxel grid the network operates on it instead injects near-white ones, which a filter cannot do but a sampler can. One undersampling, two opposite symptoms, and a single per-frequency model that covers both. The second asks whether distillation repaired what the first measured. It reads the student’s spectrum band by band against the teacher’s, before and after training, and it is the direct test of the claim this paper makes; the segmentation results below are the indirect one. Where the two readings disagree would be where a design rule failed, so we state both. The rest of this section reports segmentation accuracy.

2.2 Source-to-Unseen Generalization Experiments

We focus on the more challenging *source-to-unseen* ($S \rightarrow U$) setting, which serves as the primary evaluation protocol of this work. In this setting, the model is trained exclusively on diverse generated data from a single source dataset and is then evaluated on real scans of completely unseen LiDAR, without any target-domain access or fine-tuning. Compared with standard cross-dataset evaluation, this protocol more directly reflects the practical objective of zero-shot cross-LiDAR semantic segmentation.

2.2.1 Benchmark Generalization Results

The results are summarized in Table 1, whose three rows per source are the single-dataset baseline, the randomized-sampling baseline of [1], and LiEqualizer. In the second of these, R denotes training on randomly sampled observations, while S and CL denote that method’s two modules.

Table 1: $S \rightarrow U$ domain generalization results.



Trained on	Method	Test on SK mIoU_{SK}	Test on NU mIoU_{NU}	Test on Waymo mIoU_{Waymo}
only SK	P-baseline	71.6	19.4 (−52.2)	17.3 (−54.3)
	LiSeer-P-RSCL	68.1 (−3.5)	56.8 (+37.4)	54.6 (+37.3)
	LiEqualizer-P(ours)	75.5 (+3.9)	58.3 (+38.9)	64.4 (+47.1)
only NU	P-baseline	20.5 (−49.9)	70.4	18.1 (−52.3)
	LiSeer-P-RSCL	54.4 (+33.9)	66.5 (−3.9)	53.8 (+35.7)
	LiEqualizer-P(ours)	55.3 (+34.8)	76.1 (+5.7)	61.4 (+43.3)
only Waymo	P-baseline	24.7 (−36.2)	5.4 (−55.5)	60.9
	LiSeer-P-RSCL	52.3 (+27.6)	49.0 (+43.6)	62.2 (+1.3)
	LiEqualizer-P(ours)	63.8 (+39.1)	58.5 (+53.1)	79.2 (+18.3)

Note. Cell shading encodes mIoU magnitude (Blue monochrome: light = low, saturated = high). **Green** values show performance drop on the target domain relative to the source domain for baseline methods. **Red** values show the improvement of LiSeer over the corresponding baseline on the target domain. The baseline and LiSeer rows are quoted from the original publication [1] under that work’s own evaluation protocol.

Changing the sensor is enough to undo most of what the first baseline learned. For example, trained on Waymo alone it scores 60.9 on its own sensor and 5.4 on nuScenes, a drop of 55.5 points, and the pattern repeats on every other pair. A model trained under one sampling geometry has therefore largely fitted that geometry rather than the scene. Randomizing the geometry recovers a large part of the loss, by up to 43.6 points on the same pair, which is what invariance to the sampler buys.

It does not, however, buy back what the sampler removed, and that is where the third row begins. LiEqualizer keeps the randomization and adds the spectrum the teacher still holds, raising cross-sensor mIoU by as much as 11.5 over the same backbone, the same schedule and the same generated data, and leaving the network and its parameter count exactly as they were. Because the two act on different parts of the problem, the gain is not a reallocation of the first: the baseline is already invariant to the sampler, so what remains can only come from restoring content rather than from ignoring the sampling pattern more thoroughly. These results separate the two halves of the problem, in the terms this work started from: randomization makes the student indifferent to which spectrum its sensor delivers, and distillation gives it the part of the spectrum that no sensor in training delivered.

2.2.2 Zero-shot Deployment on Real Vehicles

Benchmarks share their collection conventions, so we also deploy LiEqualizer-P on three real vehicle platforms, *TH-Courier*, *TH-AVT* and *TH-CS55*, carrying previously unseen 32-, 80- and 128-beam LiDARs. These platforms span sensor densities and mounting heights that appear in none of the training sets, which makes them a test of the spectral account rather than of the benchmarks. Segmentation stays accurate and stable on all three (Fig. 2), and further qualitative results are given in Appendix G.3. We draw no numerical claim from these deployments; what they show is that a spectrum corrected during training survives sensors whose sampling geometry was never observed.

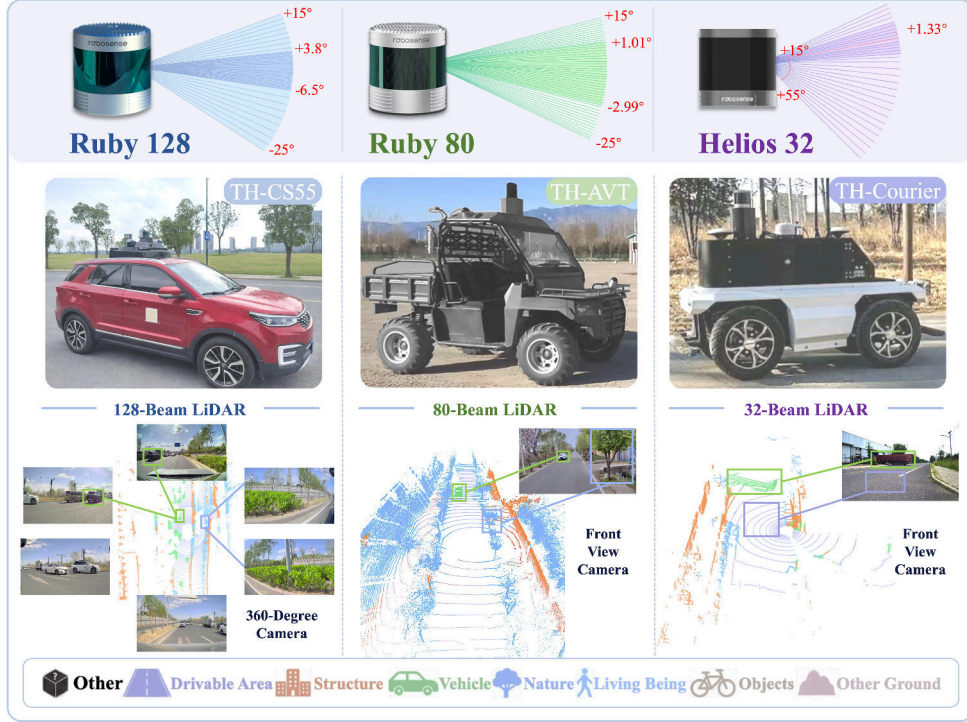


Fig. 2: Zero-shot cross-LiDAR generalization results on real-world vehicles.

Note. Camera images are time-synchronized with the corresponding LiDAR scans on all three platforms. For visualization, we establish camera-point-cloud correspondences for selected classes: “vehicle” and “nature” on TH-CS55 (360° camera) and TH-AVT (front-view camera), and “vehicle” and “drivable area” on TH-Courier (front-view camera).

2.3 Comparison with Previous Methods

Because LiEqualizer is itself a distillation objective, we further compare against the objectives of prior dense-to-sparse and frequency-domain KD, re-implemented under one pipeline with the same teacher, the same student, the same signals (TPV planes and common-support voxel logits) and the same schedule, so that only the objective differs. This comparison separates two possible sources of the gain. Every row has access to the same dense teacher, so whatever distinguishes the last row from the rest is attributable to the spectral objective rather than to the teacher itself. The spatial-domain representatives reduce to the classical two-term form

$$\mathcal{L}_{\text{pw}} = \lambda_{\text{kd}} T^2 \text{KL}(\text{softmax}(z_t/T) \parallel \text{softmax}(z_s/T)) + \lambda_{\text{feat}} \|F_s - F_t\|^2, \quad (1)$$

where z_t, z_s are the teacher and student logits, F_t, F_s the corresponding features and T the temperature. Theorem 1(iii) predicts the ordering. Objectives that match spatial pointwise detail inherit the aliasing floor and should gain little on high-frequency classes, whereas spectrum-level alignment should not.

Table 2: Distillation objectives compared under one pipeline (source-domain and zero-shot mIoU; numbers to be filled).

Objective	Attached to	SK→SK	SK→Waymo	SK→nuScenes
student (no distillation)	—	68.1	54.6	56.8
+ feature mimicking [10]	TPV planes	73.7 (+5.6)	64.8 (+10.2)	54.0 (−2.8)
+ soft-label KL [11, 18]	Voxel logits	74.1 (+6.0)	64.1 (+9.5)	53.6 (−3.2)
+ both, Eq. (1)	Both	73.6 (+5.5)	64.0 (+9.4)	53.9 (−2.9)
+ PVKD [19]	Both	73.8 (+5.7)	64.5 (+9.9)	54.1 (−2.7)
+ CWD [20]	TPV planes	74.2 (+6.1)	64.2 (+9.6)	53.7 (−3.1)
+ LiEqualizer (ours)*	Both	75.5 (+7.4)	64.4 (+9.8)	58.3 (+1.5)

Note. All competing objectives are implemented by us within this pipeline. Where the released code runs on this architecture we follow it and its default hyper-parameters; the substitutions we made are listed in Appendix E.1. The first row is the same student without any distillation, and the bracketed figures are differences against that row.

2.4 Ablation Studies

Back to the reading this work started from: the two branches exist because sparse sampling does two different things to one spectrum. It takes broad structure away at the encoder’s code, and at the decoder’s output it additionally leaves behind structure the scene never had. If that reading is right, each branch should earn its place on the fault it sits on, and the gate rather than the mere presence of a dense teacher should be what makes the transfer safe. We put that to the test by building the objective up rather than taking it apart, starting from the undistilled student and adding one part at a time, with everything else, the teacher included, held fixed.

The first two additions are alternatives rather than steps. One gives the student the low-frequency branch alone, which aligns the smoothed responses of the projected feature planes and nothing on the grid. The other gives it the high-frequency branch alone, which acts on the grid where the aliasing floor appears and carries the soft-label term that shares its switch. The next three hold both branches and differ in how much of design rule D3 of Sec. 3.3 they carry: no gate at all, the gate used to decide where to match but nothing suppressed, and the gate used to decide where to suppress but nothing matched. The last row has both halves of the gate and is the full objective. One further variant changes neither branch nor gate but the target of the low-frequency suppression, charging the student only for the band-pass energy by which it exceeds the teacher, as the high-frequency branch already does, instead of driving it to zero.

Table 3 is read downwards. Each row adds a part of the objective to the undistilled student, and the bracketed figure beside it is what that part is worth on the sensor named at the head of the column.

Three readings matter. The first two rows say which branch carries the transfer, because they add one and withhold the other. The middle three say how much of the gain belongs to the gate rather than to the dense teacher alone, and they separate the gate’s two halves: deciding where to match and deciding where to suppress are different instructions, and the table charges them separately. The last row says whether the two halves together are worth more than either of them alone, which is what design rule D3 predicts.

Table 3: What each part of the objective adds, under the standard single-pass protocol.

Objective	SK→SK	SK→Waymo	SK→nuScenes
LiSeer-P-SK, no distillation	68.1	54.6	<u>56.8</u>
+ LF alone	74.5 (+6.4)	65.4 (+10.8)	54.4 (−2.4)
+ HF and soft label alone	73.5 (+5.4)	63.7 (+9.1)	53.0 (−3.8)
+ both branches, ungated	73.8 (+5.7)	64.0 (+9.4)	53.3 (−3.5)
+ gates, matching only	74.6 (+6.5)	64.1 (+9.5)	53.9 (−2.9)
+ gates, suppressing only	74.2 (+6.1)	64.2 (+9.6)	53.1 (−3.7)
with LF suppressed towards the teacher	74.6 (+6.5)	64.0 (+9.4)	53.7 (−3.1)
LiEqualizer (gates, matching and suppressing)	75.5 (+7.4)	<u>64.4 (+9.8)</u>	58.3 (+1.5)

Note. The table is read downwards, from the undistilled student to the full objective, and each bracketed figure is the difference from the first row. That row is the randomized-sampling student of Table 1 trained on SemanticKITTI, quoted from the original publication [1]. The two single-branch rows are alternatives rather than steps: one keeps the low-frequency branch alone, the other the high-frequency branch that shares its switch. The three rows after them hold both branches and differ in how much of the gate they carry.

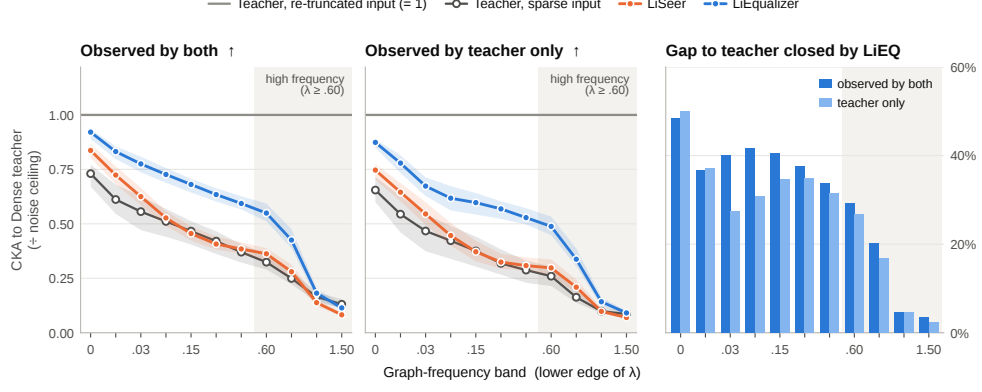
Two limits on what the table can be asked. The rows marked † were trained with one sample per GPU and the last two with four, so comparisons are read inside one of those groups and not across them; a difference between the groups also carries a change of batch-normalization statistics and of the iteration-counted schedules. And the spectral diagnosis of Sec. 2.5 was run on the full method only, so the prediction that an ungated objective re-inflates the flatness gap is untested here: segmentation accuracy is the only quantity these rows report, and it is consistent both with a spectral change the task metric does not resolve and with no spectral change at all.

2.5 Spectral Diagnosis: Is the Lost Spectrum Actually Restored?

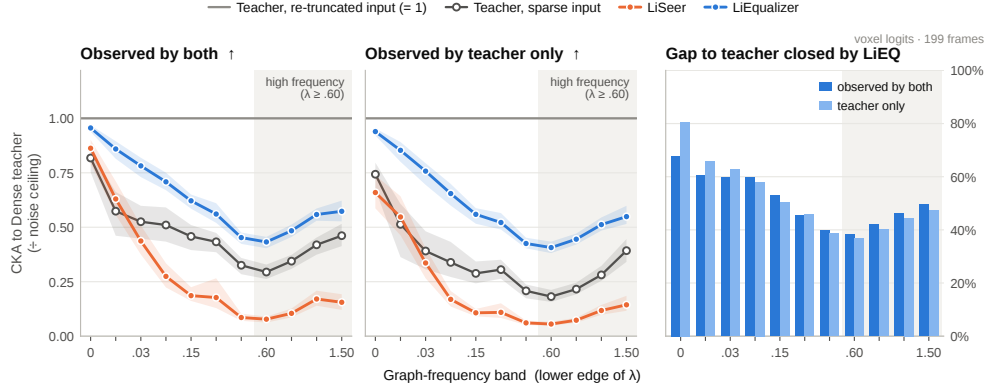
The diagnosis above says what sampling does to the signal. Whether distillation undoes it is a separate question, and it is the direct test of our claim. We apply the protocol of Sec. C.1 on held-out scenes, comparing the student before and after distillation against the same frozen teacher, in both domains. Three readings are expected if the mechanism works as Theorem 1 prescribes.

First, the aliasing floor of the masked field should not move. Section D.2 located that floor in the support geometry rather than in the values the cells carry, and no loss applied during training can change which cells the sensor returned. Measurement bears this out in the direction that matters: the floor does not fall. It stands at 32.3 before distillation and 33.3 after it, as means over frames, the small difference going upwards on 88% of them. A reading that fell here would not be evidence of repair; it would mean the diagnosis of Sec. D.2 had misattributed the excess.

The first reading is about what distillation cannot touch. What it can repair is the field on the teacher’s own support, and there the band distribution is the thing to watch. The frequency-band discrepancy should fall, and it falls from 0.43 to 0.11 on every frame, roughly a quarter of its undistilled value. The student’s high-frequency



(a) on the feature planes



(b) on the voxel logits

Fig. 3: Alignment, band by band.

Note. Agreement between the student’s field and the teacher’s within each band, scaled so that one is the agreement the teacher reaches with itself when its input is re-drawn under the same budget. That ceiling is the right unit here, because high-frequency detail depends on which points happened to be sampled and the teacher does not reproduce it exactly either. Each row carries three columns: the left two read the band-wise agreement with the dense teacher, normalized by that ceiling, on the cells both sensors observe and on those the teacher alone observes, drawn as the median over frames with the quartiles shaded; the right one gives the fraction of the undistilled student’s distance to the ceiling that distillation closes, as the median of the per-frame ratio rather than a ratio of medians. Higher is better in all three.

share, read against the teacher’s on that support, should not exceed it, and under every channel convention we examined it does not: the distilled student stays between 0.67 and 0.77 of the teacher, whereas the undistilled student crosses from one side of unity to the other as the convention changes. Its field inside the support is therefore smoother than the teacher’s while its band distribution is the closest to the teacher’s of any model we measured. Second, the flatness gap $\text{Flat}(\hat{\mathbf{q}}^{\text{H},s}) - \text{Flat}(\hat{\mathbf{q}}^{\text{H},t})$ should become non-positive, so that the student is never noisier than the teacher. We read it

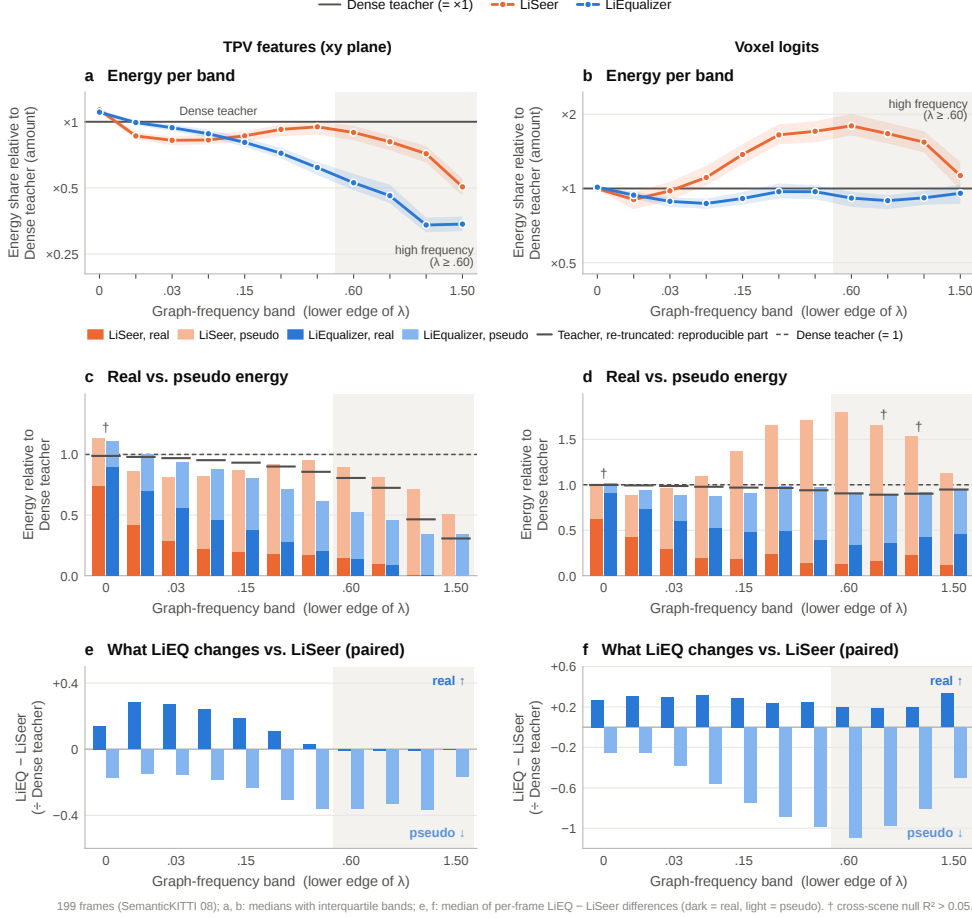


Fig. 4: How much energy a band carries, and how much of it the teacher accounts for.

Note. **a, b:** the share of the total energy that falls in each band, as a multiple of the teacher's share in that band, on the feature planes and on the voxel logits. The axis is logarithmic because the quantity is a ratio, so that twice the teacher and half the teacher sit equally far from it. **c, d:** the same energy split into its shared and invented parts as defined in Sec. 2.5.2, drawn dark and light. The grey marks give the teacher's own ceiling, the part of each band it can account for when its input is re-truncated. **e, f:** what distillation changed, as the median over frames of the per-frame difference between the distilled and the undistilled student, each part first divided by the teacher's energy in that band. A dark bar upwards is shared energy the distilled student holds and the undistilled one does not; a light bar downwards is invented energy it no longer holds. The two bars are not stacked, and their difference is not the change in total energy. The comparison is against the undistilled baseline, so it says what the objective as a whole changed and not which branch changed it.

on the channel-aggregated residual profile of Eq. (24), the profile the one-sided hinge
 constrains during training, and it crosses from $+0.0052$ to -0.0013 , non-positive on
 87% of frames. The smoothed profile behind the remaining readings answers a different
 question, and a floor of injected high frequencies is visible only in the residual one.
 Theorem 1(iv) gives the direction of that gap, but on a different profile and only at
 the ideal gated optimum: there the optimum is the teacher itself, so the student's

377 *coordinate*-energy profile is the teacher’s own and its flatness cannot exceed it. What
 378 the hinge constrains, and what we report here, is flatness along the *diffusion-scale*
 379 axis instead. The two axes agree in direction: at that optimum the coordinate-axis
 380 gap is zero, and a near-white floor along the scale axis is exactly what the residual
 381 profile registers at the coarser scales. We therefore check the scale-axis gap empirically
 382 rather than inherit it from the theorem. Third, high-frequency recovery should rise
 383 most where the structure is finest, showing that genuine detail rather than broadband
 384 energy was transferred. We read it as the share of each model’s own local energy that
 385 falls above the band, so that the comparison cannot be carried by one model’s larger
 386 logit scale. It does rise: on poles the share goes from 0.17 to 0.66 of the teacher’s, and
 387 on traffic signs from 0.11 to 0.55, while on road surfaces, which carry almost no fine
 388 structure, it rises only from 0.60 to 0.74. The contrast is the reading that matters, a
 389 fourfold to fivefold gain against a fifth, because an objective that simply added high-
 390 frequency energy everywhere would have moved them together. Pedestrians, on the
 391 44 frames that contain them, rise from about 0.29 to roughly the teacher’s own level.
 392 Occupancy boundaries, which start high at 0.54 and reach 0.79, gain under half again
 393 although they too are edges, far short of the fourfold and fivefold gains above, so
 394 what separates the classes is the fineness of the structure and not the presence of a
 395 boundary. Holding the network fixed and changing only the observation, so that the
 396 dense-trained teacher itself reads the sparse scan, leaves 0.58 on poles and 0.47 on
 397 signs, well above the undistilled student’s 0.17 and 0.11: what that student had lost
 398 on fine upright structure exceeded what the sampling alone removes. Because these
 399 are the same statistics the losses optimize (Table 4), agreement here closes the loop
 400 between what is trained and what is measured; disagreement would localize which rule
 401 failed.

402 2.5.1 Alignment, Band by Band

403 The statistics above summarise a spectrum by a few scalars. A stronger test asks, for
 404 each band separately, whether the student’s field is *aligned* with the teacher’s rather
 405 than merely as energetic. We measure alignment by the agreement between the two
 406 fields within a band, scaled so that one is the agreement the teacher reaches with itself
 407 when its input is re-drawn under the same budget. That ceiling is the right unit here,
 408 because high-frequency detail depends on which points happened to be sampled and
 409 the teacher does not reproduce it exactly either.

410 On the feature planes, where the low-frequency branch acts, the distilled student is
 411 closer to the teacher than the undistilled one in every band and in every frame, and at
 412 high frequency it closes between a sixth and a third of the undistilled student’s distance
 413 to that ceiling (Fig. 3a). Energy alone would have hidden this, and Appendix D reports
 414 what energy alone says.

415 On the voxel logits, where the high-frequency branch acts, the separation is wider.
 416 The undistilled student holds 1.37 to 1.79 times the teacher’s share between $\lambda = 0.15$
 417 and $\lambda = 1.15$, whereas the distilled student holds 0.87 to 1.01 of it in every band, is
 418 closer to the teacher in every frame, and closes between a third and two thirds of the
 419 undistilled student’s distance to the ceiling (Fig. 3b and Fig. 4b).

On the planes the amount itself falls, and that is the design working rather than failing. In the bands above $\lambda = 0.6$ the distilled student carries 0.34 to 0.53 of the teacher’s energy share, against 0.51 to 0.89 for the undistilled one (Fig. 4a). The low-frequency objective drives the student’s band-pass response towards zero wherever the teacher’s class-probability gate is closed, so a deficit there is what it was built to produce; the two vertical planes, which are matched without a gate, show no such deficit. What matters is which part of the energy went. Splitting each band into the part a linear reading of the teacher’s energy in that same band accounts for and the remainder (Fig. 4c) shows that above $\lambda = 0.6$ the distilled student retains as much teacher-explained energy as the undistilled one, and that the reduction falls on the unexplained part, which drops from 2.4% to 1.2% of the total feature energy; the teacher itself, re-run on a re-truncated input, leaves 0.9% unexplained. Across scenes the more high-frequency energy the distilled student keeps, the better it aligns with the teacher, at a rank correlation of 0.61, while no such relation holds for the undistilled one, at 0.07. The feature planes are therefore where teacher-consistent structure is restored in the low and middle bands and where sampling artefacts are removed in all of them, and the voxel logits are where the high-frequency recovery itself happens: there the distilled student matches the teacher’s energy in every band and holds 1.4 to 3.8 times the undistilled student’s teacher-explained energy, the highest bands included (Fig. 4d).

2.5.2 Which Part of the Gain Is Real

Alignment and amount still leave one question open. A student could become better aligned by discarding what it had rather than by acquiring what the teacher has. We therefore split the student’s energy in each band in two, and the split needs stating precisely because everything in this subsection rests on it. Inside the teacher’s support, and band by band, we ask how much of the student’s field a linear predictor built from the teacher’s field *in that same band* can account for. That predictor is fitted on one half of the cells and evaluated on the other, so that what it explains is explained out of sample; the estimator is given in Appendix D.1. The part it accounts for we call *shared* energy, the remainder *invented*, and both are divided by the teacher’s energy in the band, so that the two add up to the ratio plotted in Fig. 4a. Two things follow from the definition and bound how it may be read. Structure that the student has moved into a neighbouring band is not explained by the teacher’s energy in this one, so it is counted as invented: shared energy is a lower bound on what is genuinely held in common and invented energy an upper bound on what is not. And the ceiling is not one. Re-running the teacher on a re-truncated input leaves part of each band unexplained by its own earlier field, and that residual is the most any student could be asked to share; it is below 0.86 only above $\lambda = 0.6$, and falls to 0.46 and 0.31 in the two highest bands, where what a field holds depends on which points happened to be sampled. Both students are read under one convention.

The split falls differently at the two levels, and it falls the way the two branches were built. On the planes the distilled student gains shared energy between $\lambda = 0.01$ and 0.45 and none above it, where the change is between -0.01 and 0, while invented energy drops in every band (Fig. 4e). On the logits it gains shared energy in every

band, the highest included, by 0.19 to 0.33 of the teacher’s energy in that band, while
invented energy falls by as much as 1.09 (Fig. 4f). The preregistered high-frequency
recovery test, which asks instead how much of the teacher’s high-frequency structure a
linear reading of the student recovers over the four bands above $\lambda = 0.6$ taken together,
improves by 0.31, and improves in all 199 frames. These readings show that the gain
is a repair of the spectrum rather than a trade within it: the low-frequency branch
restores broad structure and removes ripple wherever it finds it, and the high-frequency
branch is what puts genuine detail back.

3 Methods

3.1 Preliminary: Task, Goal, and a Graph-Spectral View of LiDAR Features

We study *cross-LiDAR generalizable semantic segmentation*. Given a labelled source
dataset

$$\mathcal{D}_s = \{(\mathcal{P}_i, \mathcal{Y}_i, \mathbf{s}_i)\}_{i=1}^N, \quad (2)$$

where $\mathcal{P}_i$ is a sparse scan, $\mathcal{Y}_i$ its per-point labels and $\mathbf{s}_i$ the sensor configuration that
produced it, the goal is a model $f_\Theta : \mathcal{P} \mapsto \hat{\mathcal{Y}}$ that remains accurate on scans acquired
by sensors never observed during training, without any target-domain data or fine-
tuning. Following [1], each observed scene is first densified and then re-sampled under
controllable sensor configurations,

$$\mathcal{P}^d = \Phi_{\text{dens}}(\mathcal{P}), \quad \tilde{\mathcal{P}}^{(k)} = \Phi_{\text{samp}}(\mathcal{P}^d; \mathbf{s}^{(k)}), \quad \mathbf{s}^{(k)} \sim p(\mathbf{s}), \quad (3)$$

so that every training scene exists in two forms, a dense observation and a family of
sparse re-samplings that share the same “SIG” and differ only in “SG”.

A central difficulty remains after this randomization. Exposing the model to many
sampling geometries teaches it which variations to ignore, but every sparse re-sampling
has lost the same content. Moreover, Proposition 2 shows that no objective defined
on those observations alone can recover it. The dense observation, in contrast, was
never subjected to that loss. We therefore train a *teacher* on the dense observations
and a *student* on the sparse ones, freeze the teacher, and add a distillation objective
that equalizes the student’s spectral response to the teacher’s. The objective acts only
during training, so the deployed student keeps the architecture, parameter count and
inference cost of the baseline.

The quantities this objective aligns do not live on the point cloud. The backbone
produces three tri-perspective-view (TPV) feature planes, each channel of which is a
two-dimensional grid signal, and a voxel logit volume, each class channel of which is
a three-dimensional grid signal. We write either as $x \in \mathbb{R}^N$ on the d -dimensional grid
graph $G = (V, E)$ with $d \in \{2, 3\}$. Here N counts grid locations, HW or HWZ and
of order 10^4 to 10^6 , rather than returned points. Empty cells therefore remain nodes
of the ambient grid. The two fields sit at opposite ends of one network. The planes
are what the encoder outputs, a compressed code that keeps what the scene means
rather than how large it is. The logits are what the decoder puts back on the grid
at the scene’s own scale, and they are where the task loss is applied and where the

aliasing floor appears. The low-frequency branch of Sec. 3.4 accordingly acts on the planes and the high-frequency branch on the logits. It is the position in the network that fixes which band a field carries, not the form the features take.

Comparing two such signals spectrally requires a graph operator. A natural choice is to build a neighbourhood graph on the occupied cells of each branch, as the point-domain suite of Sec. C.1 does. However, teacher and student occupy different cells, so the two graphs would differ and their eigenbases with them. A mode-by-mode comparison would then lose its meaning. Both branches are therefore read with one operator, defined by the ambient grid and identical for the two of them. It is a depth-wise stencil with centre weight 1, weight $-\frac{1}{2d}$ on each axis neighbour, and replicate padding. This stencil is a graph Laplacian, since it equals $L = I - A$ with symmetric A and satisfies $L = L^\top \succeq 0$, $x^\top Lx = \frac{1}{2d} \sum_{(u,v) \in E} (x_u - x_v)^2 \geq 0$, $\text{spec}(L) \subset [0, 2]$ and $\ker L = \text{span}\{\mathbf{1}\}$ (Lemma 1, Appendix B.2). Replicate padding is what folds the boundary ghosts back and makes A symmetric; zero padding would break every spectral statement that follows. The eigenpairs (λ_i, φ_i) then play the role of frequencies, with small- λ modes encoding smooth, large-scale structure such as the road surface, terrain and building facades, and large- λ modes encoding fine detail such as curbs, lane markings, poles and object boundaries.

The quantity we would like to compare between teacher and student is the spectral energy distribution of each channel,

$$\mu_x = \sum_{i=1}^N \frac{\hat{x}_i^2}{\|x\|^2} \delta_{\lambda_i}, \quad \hat{x}_i = \langle x, \varphi_i \rangle, \quad (4)$$

which records how the channel’s energy is distributed over frequencies. Forming μ_x inside the training loop, however, is not an option. First, it would require diagonalizing L on 10^4 to 10^6 nodes for every sample at every step. Second, on the irregular k NN graph of a point cloud there is no closed-form decomposition to fall back on. It is therefore never computed during training. Instead we probe it through a family of smoothing operators. Fix a step $\alpha \in (0, \frac{1}{2})$ and let $D_\alpha = I - \alpha L$, which acts on frequency λ by $\rho(\lambda) = 1 - \alpha\lambda \in (0, 1]$. The upper limit on α is not cosmetic, and Theorem 1(i) analyses what fails beyond it. All probes are then built from the diffusion pyramid $\{D_\alpha^b x\}_{b=1}^B$,

$$e_b(x) = \frac{\|D_\alpha^b x\|^2}{\|x\|^2} = \int_0^2 \rho(\lambda)^{2b} d\mu_x(\lambda), \quad h^{(b)}(x) = x - D_\alpha^b x, \quad q(x) = \frac{(e_1, \dots, e_B)}{\sum_b e_b}, \quad (5)$$

where e_b is the energy that survives b smoothing steps, $h^{(b)}$ is the high-pass residual and q is the multi-scale response profile. These operators are *spectral-invariant*, in that they commute with L , act on each frequency independently through ρ^b , and create no new frequency components. As a result they can probe μ_x without reshaping it. “SG” does the opposite, relocating energy across the spectrum. Every statistic below is finally normalized per sample, and where stated per channel. Several of them are individually invariant to a positive rescaling of their own input, and we say so where that holds; the assembled objective is not, because the high-frequency comparison

divides both sides by one teacher-derived scale and the gates are read from the teacher's own class probabilities.

3.2 Spectral Measurement Protocol

Nothing in the argument that follows can be stated without a way of measuring a spectrum, and the spectra at issue cannot be formed at these sizes. We therefore built a protocol that reads them through a family of smoothing operators instead, on the raw point cloud and on the voxel logits alike, and we report it in full in Appendix C. It fixes the operator, the scales, the bands, the statistics and the two ways a branch's logits can be placed on the lattice, and it is the source of the measurement model and of the assumption the theorem below rests on. The protocol is released with the paper, so that sampling degradation and learned representations can be analysed by others under the same definitions.

3.3 Spectral-Invariant Distillation under Sampling Degradation: A Unified Theorem

The measurements of Sec. D.2 say what sampling does to the signal, but they do not say what can be done about it. Two questions have to be settled before an objective can be written down. The first is which quantities a student can be asked to match, given that the spectrum itself cannot be formed at this size. The second is which of those quantities remain trustworthy once the student's own copy carries an aliasing floor. The theorem below answers both, and in doing so fixes what to align, how many scales suffice, which channels to trust, and how to transfer frequency information safely. Every symbol it uses is defined at its first appearance. Complete proofs with all constants are in Appendix B.2, and a self-contained glossary of the spectral notions is provided in Appendix B.1.

Theorem 1 (Spectral-invariant distillation) *Fix $\alpha \in (0, \frac{1}{2})$ and adopt Secs. 3.1–C.1; part (iii) additionally assumes model (C18) and Assumption 2.*

(i) **Identifiability.** *For any two nonzero signals x, y on the same grid,*

$$e_b(x) = e_b(y) \quad \forall b \geq 1 \quad \Longleftrightarrow \quad \mu_x = \mu_y : \quad (6)$$

the diffusion responses determine the entire spectral energy distribution. They determine energy only, not signs, phases, or a unique spatial signal, and the premise $\alpha < \frac{1}{2}$ is what makes the change of variables $u = \rho(\lambda)^2$ injective; beyond it two frequencies collapse onto one response and identifiability fails.

(ii) **Finite-scale stability.** *Suppose only B responses are matched, $|e_b(x) - e_b(y)| \leq \varepsilon$ for $b = 1, \dots, B$, and let $\psi : [0, 2] \rightarrow \mathbb{R}$ be any frequency statistic with Lipschitz constant $\text{Lip}(\psi) = \sup_{\lambda \neq \lambda'} |\psi(\lambda) - \psi(\lambda')|/|\lambda - \lambda'|$. Writing $p_B^*(u) = \sum_{b=0}^B c_b^* u^b$ for a best degree- B polynomial approximant of ψ after the change of variables $u = \rho(\lambda)^2$, with coefficient size $\|c^*\|_1 = \sum_b |c_b^*|$,*

$$\left| \int \psi d\mu_x - \int \psi d\mu_y \right| \leq \frac{C_\alpha \text{Lip}(\psi)}{B+1} + \|c^*\|_1 \varepsilon, \quad C_\alpha = \frac{C_J(1 - u_{\min})}{\alpha(1 - 2\alpha)}, \quad u_{\min} = (1 - 2\alpha)^2, \quad (7)$$

where C_J is the universal constant of Jackson's approximation theorem. Band energies and, through their finite profiles, the smoothed flatness of Eqs. (C16)–(C17) are such statistics. The default $B = 8$ is motivated by this $O(1/B)$ rate and validated empirically; the bound alone does not certify a task-independent tolerance, since C_α , $\|c^*\|_1$ and ε enter as well.

(iii) **Denoising under sampling.** With the distinct eigenvalues $\theta_1 < \dots < \theta_K$ of L , their multiplicities d_k , and $\rho_k = \rho(\theta_k)$, define two probability weights over frequency bands,

$$W_b(k) = \frac{d_k \rho_k^{2b}}{\sum_\ell d_\ell \rho_\ell^{2b}}, \quad V_b(k) = \frac{d_k (1 - \rho_k^b)^2}{\sum_\ell d_\ell (1 - \rho_\ell^b)^2}, \quad (8)$$

the frequency weightings of the b -step low-pass channel D_α^b and of its high-pass residual $I - D_\alpha^b$, and write $\mathbb{E}_P[\text{snr}] = \sum_k P(k) \text{snr}_k$ for the average reliability a channel perceives. Then

$$\mathbb{E}_{W_{b+1}}[\text{snr}] \geq \mathbb{E}_{W_b}[\text{snr}] \quad \text{and} \quad \mathbb{E}_{V_b}[\text{snr}] \leq \mathbb{E}_{W_b}[\text{snr}] \quad \forall b \geq 1 : \quad (9)$$

each smoothing step can only raise the reliability of the aligned statistic, and the ungated pointwise residual is the less reliable channel at every scale. The same ordering selects the moment system. For an integer raw-moment order $p \geq 1$, the normalized raw Laplacian moment and the depth- b contracted moment read, on the spectrum,

$$\frac{x^\top L^p x}{\|x\|^2} = \int_0^2 \lambda^p d\mu_x(\lambda) \quad \text{vs.} \quad e_b(x) = \int_0^2 \rho(\lambda)^{2b} d\mu_x(\lambda) : \quad (10)$$

the weight λ^p concentrates geometrically on the corrupted top band as p grows ($a \sim 10^2$ reweighting already at $p = 8$, $\alpha = \frac{1}{4}$), while ρ^{2b} suppresses it geometrically in b ; both infinite systems identify μ_x , so the contracted choice costs no information.

(iv) **Complementary teacher-gated transfer.** At any fixed diffusion depth (omitted here to keep the notation compact), for a branch $r \in \{L, H\}$ let $z_r^t, z_r^s \in \mathbb{R}^J$ be its detached teacher and trainable student responses on the same J coordinates, and define the branch operator $T_L = L$ and $T_H = I$. Thus $T_L z_L$ is LF curvature and $T_H z_H$ is the HF response itself. For a floor $\tau_r > 0$ and suppression weight $\lambda_{\text{sup}}^r > 0$, let

$$c_r^t = T_r z_r^t, \quad a_{r,i} = \frac{(c_{r,i}^t)^2}{(c_{r,i}^t)^2 + \tau_r}, \quad A_r = \text{diag}(a_{r,1}, \dots, a_{r,J}), \quad (11)$$

where i indexes response coordinates and A_r is the diagonal matrix of detached gate values. The ideal complementary-gate objective is

$$\mathcal{J}_r(z_r^s) = \|A_r^{1/2}(z_r^s - z_r^t)\|_2^2 + \lambda_{\text{sup}}^r \|(I - A_r)^{1/2} T_r z_r^s\|_2^2. \quad (12)$$

Every minimizer $z_r^{s,*}$ satisfies

$$\|(I - A_r)^{1/2} T_r z_r^{s,*}\|_2^2 \leq \|(I - A_r)^{1/2} T_r z_r^t\|_2^2 : \quad (13)$$

the student cannot reduce the objective by adding branch-specific structure on the teacher-complement coordinates. The two branches suppress towards different references, and the optimum accordingly takes two forms. For HF , $T_H = I$ makes (12) separable, and because the complementary term charges only an excess over the teacher, each coordinate minimizes $a(z^s - z^t)^2 + \lambda_{\text{sup}}^H(1 - a)[(z^s)^2 - (z^t)^2]_+$, whose minimizer is

$$(z_{H,i}^s)^* = z_{H,i}^t, \quad (14)$$

for every $a_{H,i} \in [0, 1]$ and uniquely so wherever the gate is open. The gate therefore selects which deviation is charged rather than where the branch comes to rest: an open coordinate is penalized on both sides of the teacher, a fully closed one only above it, and is then free below it. At the teacher minimizer the student's coordinate-energy profile coincides with the teacher's,

621 so its Shannon entropy and Wiener flatness are the teacher's own and the one-sided hinge
622 of Eq. (24) is inactive. One consequence is worth naming, since the gated objective and the
623 dark-knowledge term act on the same pair of logits: the teacher is a stationary point of both,
624 so neither can pull the student away from what the other prescribes. The continuity term is
625 the one part of the branch that does not take the teacher as its optimum, and deliberately
626 so, since the teacher's own variation inside a class is confidence ripple rather than structure
627 worth copying.

628 For LF , the complementary term still charges the student's curvature against zero, and
629 $T_L = L$ couples neighbouring coordinates. Pointwise in the suppressed quantity the optimum
630 is accordingly shrunk towards zero by a factor $a_L/(a_L + \lambda_{\text{sup}}^L(1 - a_L))$, and at the level of the
631 field it is (13), rather than any coordinatewise formula, that guarantees student curvature in
632 teacher-flat regions cannot exceed the teacher's complementary-gate curvature energy at the
633 ideal optimum. On the two planes that carry no gate there is no complementary term at all,
634 so there the optimum is again the teacher itself.

635 *Scope.* The theory identifies and stabilizes spectral energy, not every Fourier coefficient
636 and not a unique dense scene. *Seeing the lost spectrum* is therefore a claim about
637 energy and nothing more: it means learning a teacher-consistent estimate of the reliable
638 spectral-energy structure from sparse observations. The dense teacher supplies the
639 missing prior, and the spectral invariants supply a supervision channel that stays
640 stable.

641 **Proposition 2** (Consistency cannot invert attenuation) *Under model (C18), let $\hat{s}^{(1)}, \hat{s}^{(2)}$ be*
642 *two sparse views of the same scene, $\hat{s}_i^{(v)} = m_i \hat{t}_i + \hat{n}_i^{(v)}$ with independent noise. For any per-*
643 *mode gains $c_i > 0$, the parameters $(m_i, \hat{t}_i)$ and $(m_i/c_i, c_i \hat{t}_i)$ induce the same joint distribution*
644 *of the views. Hence every objective that depends only on sparse observations, including all*
645 *cross-view consistency losses, has the same optima under both, so $|\hat{t}_i|$ is not identifiable; on*
646 *modes with $m_i \rightarrow 0$ the observations carry no information about $\hat{t}_i$ at all. Access to the*
647 *teacher coefficients $\hat{t}$ removes the ambiguity.*

648 The proposition draws the line that no sparse-only objective can cross. Randomiz-
649 ing the input and making sparse views of the same scene agree can teach a model what
650 to ignore, but it cannot give back what the sampler attenuated: that content is reach-
651 able only through a reference that carries it. This is why LiEqualizer is built around
652 a teacher at all, rather than around more views of the same sparse scan. The theorem
653 then condenses into four design rules, which Sec. 3.4 implements term by term: **D1**
654 compare responses of the diffusion family, rather than raw features (from (9)); **D2** use
655 contracted moments rather than raw Laplacian powers (from (10)); **D3** split trans-
656 fer into teacher-supported matching and teacher-complement suppression (from (iv));
657 **D4** describe the response-energy matrix along its two complementary axes, channels
658 at a fixed scale and scales after aggregating channels, and constrain high-frequency
659 flatness only in the harmful direction.

3.4 LiEqualizer: Unified Spectral Equalization

We return to the observation that organizes this work: different LiDARs see different spectra of the same world. Under that reading, the job of a distiller is not to copy the teacher’s field but to correct the student’s spectrum against it. Sparse sampling damages the scan before the network reads it, and the two stages of a segmentation network do not pass that damage on in the same way. Models of this family squeeze the scan into a compact semantic code and then decode it back onto the grid. The code keeps what the scene means rather than how big it is, so what a sparse scan leaves out of it is macroscopic structure the teacher’s code still holds. The decoded output sits on the scene’s own scale and carries its detail, and the student’s own pattern of returns is written onto the grid along with it. As a result the student is short of content at one end of the network and has content to spare at the other. However, a single objective that matches the two stages in the same way cannot do both.

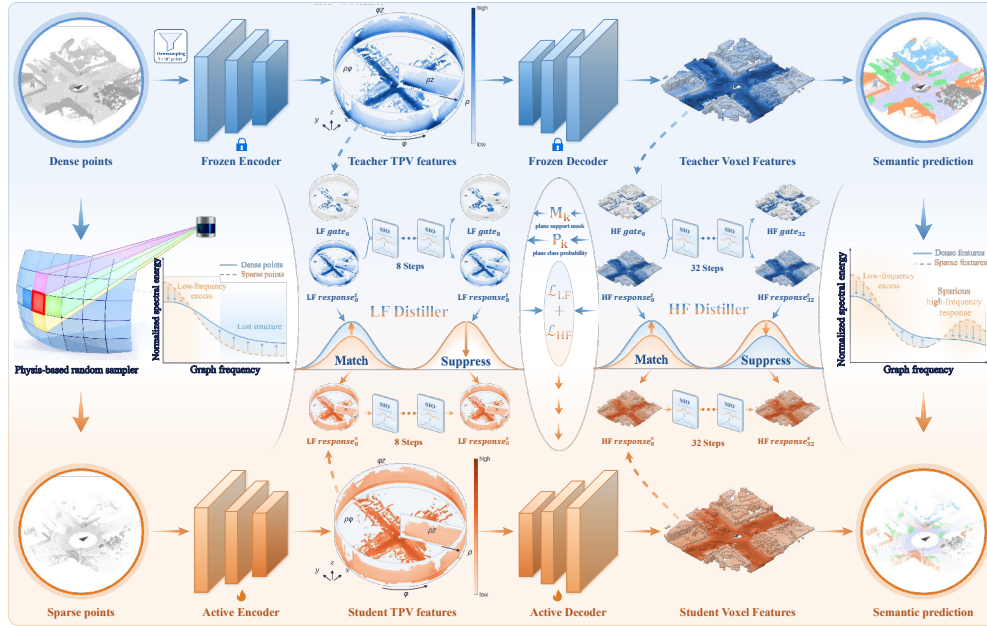


Fig. 5: LiEqualizer, and where its two branches act.

Note. A sparse re-sampling of a scene passes through the student and a dense observation of the same scene through the frozen teacher, and each network yields its three tri-perspective-view planes and its voxel logit volume. The low-frequency branch acts on the planes, which are the encoder’s output: it smooths both sides with the diffusion pyramid of Sec. 3.1 and compares the smoothed responses under a gate read off the teacher’s class field. The high-frequency branch acts on the voxel logits, which the decoder returns to the grid: it places both volumes on the teacher’s support, forms their high-pass residuals with the operator that support induces, and compares them under a gate read off the teacher’s class probabilities. Both gates are read from the teacher and never from the student. All of it runs during training only; at inference the student is the baseline network.

A distiller could simply match the student’s field to the teacher’s wherever the two can be compared. However, matching everywhere is not a solution, for three reasons.

First, nothing in the student’s own copy distinguishes detail it has observed from detail its sampling fabricated, so where to match cannot be decided from the student at all; only the teacher’s copy is free of the aliasing floor. Second, part of what the teacher carries is content a sparse student cannot predict, and copying its magnitude there produces structure of the right size and the wrong content. Third, the spectrum itself cannot be formed at these sizes, and the channel a naive comparison would reach for, the raw pointwise residual, is precisely the one in which a sparse student carries ripple rather than contours.

To this end we introduce **LiEqualizer**, which realizes the four design rules of Sec. 3.3 as two branches placed where the network itself separates the two bands. The low-frequency (LF) branch acts on the encoder’s code, where sampling takes macroscopic structure away. The high-frequency (HF) branch acts on the decoder’s output, where sampling additionally leaves aliased structure behind. In this architecture those two stages are the three tri-perspective-view planes and the common-support voxel logits, so the branches are built for those two representations, but the band each one addresses follows from where it sits and not from the form the features take. During training, a sparse re-sampling of a scene passes through the student and the dense observation of the same scene through the frozen teacher, and each network yields its three TPV planes and its voxel logit volume. The LF branch smooths each TPV plane with the diffusion pyramid of Sec. 3.1, reads a gate off the teacher’s class field projected onto that plane, and compares the two smoothed responses under that gate. The HF branch places both logit volumes on the teacher’s support, forms their high-pass residuals there with the operator that support induces, reads a gate off the teacher’s class probabilities, and compares the two residuals under that gate. The three parts of each branch answer those three obstacles in turn. A gate read from the teacher, and never from the student, decides where teacher-supported structure is *matched* and where student-only structure is *suppressed*. What that suppression aims at is set by what the student could have predicted. Both halves are then computed on smoothed responses rather than on raw residuals, together with a small set of auxiliary spectral invariants that a finite prefix of those responses controls.

The two suppressions do not aim at the same place. On the planes the student’s structure in a teacher-flat region is driven to zero; on the logits only its excess over the teacher is charged. Both follow from one criterion: a part of the teacher’s field is worth copying only where the student can predict it from its own sparse input. Under squared error the best estimate of that part is its conditional expectation given the input. Where the information is there to be recovered, that estimate is the teacher; where it is not, it is zero. The second case is the one that is easy to get wrong. Detail of the right magnitude but of unrelated content leaves the student twice as far from the teacher as learning nothing at all, since two uncorrelated fields of equal energy differ by twice that energy. Copying the teacher is therefore not a safe default. Theorem 1(iv) separates the two regimes exactly: the gated low-frequency problem is solved by shrinkage, whose gain falls to zero with the recoverable share, while the high-frequency problem has the teacher as its stationary point. Which regime a branch is in follows from the level it acts on, and the two subsections below settle it for each.

Let $\nu \in \{s, t\}$ index the student or teacher side and let x_ν denote its input field for the current loss branch. For diffusion depth $b \geq 1$, the LF diffusion response and the HF residual response are $g_b^\nu = D_\alpha^b x_\nu$ and $h_b^\nu = (I - D_\alpha^b)x_\nu$. Throughout, $r \in \{L, H\}$ identifies the branch, $\mathcal{M}_r$ its response scales, $\mathcal{S}_r \subseteq \mathcal{M}_r$ its suppression scales and $\bar{\eta}_b^r = \eta_b^r / \sum_{j \in \mathcal{M}_r} \eta_j^r$ its normalized scale weight, while i denotes a spatial or voxel-class coordinate and c a feature or semantic channel. At each scale, the matched response $z_b^{r,\nu}$, detached teacher gate a_b^r , and student quantity $u_b^{r,s}$ being suppressed are

$$\begin{array}{c|ccc} r & z_b^{r,\nu} & a_b^r & u_b^{r,s} \\ \hline L & g_b^\nu & v_b & Lg_b^s \\ H & h_b^\nu & w_b & hg_b^s \end{array} \quad (15)$$

With $T_L = L$ and $T_H = I$ from Theorem 1(iv), this table is exactly $u_b^{r,s} = T_r z_b^{r,s}$, with the generic gate a_b^r instantiated as v_b in LF and w_b in HF. The two branches then share the single form

$$\mathcal{L}_r = \sum_{b \in \mathcal{M}_r} \bar{\eta}_b^r \left[\text{Match}_r(z_b^{r,s}, z_b^{r,t}; a_b^r) + \lambda_{\text{sup}}^r \mathbf{1}\{b \in \mathcal{S}_r\} \text{Suppress}_r(u_b^{r,s}; a_b^r) \right] + \mathcal{R}_r, \quad r \in \{L, H\}. \quad (16)$$

Here Match_r is the branch-specific normalized response distance, Suppress_r is the complementary-gate energy penalty, $\lambda_{\text{sup}}^r \geq 0$ sets its strength, $\mathbf{1}\{\cdot\}$ is the indicator function, and $\mathcal{R}_r$ collects the auxiliary spectral invariants. The gate is soft rather than binary. Here a_b^r weights teacher-supported response matching, while its complement $1 - a_b^r$ weights student-only response energy. Thus the match term compares normalized response *shape* where the teacher exhibits structure, whereas the suppress term penalizes the branch-specific student quantity in teacher-flat regions. Their exact weighted forms, including the LF block construction, are given in Appendix E.1; Eq. (16) is the conceptual objective used throughout the paper.

3.4.1 Low-Frequency Gated Spectral-Invariant Distillation

On the tri-perspective-view planes the same question returns in a different form. A plane is dense, so no support has to be built, but the gate still has to tell structure from mere presence. Reading it off the plane’s own variation does not do that, because a smoothed feature plane varies wherever the scene is occupied at all, and a gate driven by the field itself then marks the outline of the observed region instead of the boundaries inside it.

What the gate needs on the planes, then, is a quantity that changes with the class rather than with the presence of returns; only then does an open gate mark the macro structure a sparse student has lost.

The branch therefore reads the LF gate from the teacher’s class field projected onto the plane, as Fig. 6 lays out. That map is formed from the teacher’s distillation field by averaging its class probabilities along the axis the plane collapses, over the occupied cells alone, and it carries no gradient. It is then diffused on the support subgraph of

the plane, under the same schedule of scales the loss uses, so that the gate is read where the teacher has returns and nowhere else,

$$v_{b,i} = \frac{c_{b,i}^t}{c_{b,i}^t + \tau_{L,b}}, \quad \tau_{L,b} = \kappa_L \text{mean}_{i \in R} c_{b,i}^t, \quad (17)$$

where c_b^t is the band-pass energy of that class map at scale b and R is the teacher's support on the plane. The loss reads the features with the ordinary plane operator, so gate and loss share the step size and the schedule of scales but not the graph on which they are computed. The dimensionless multiplier κ_L controls selectivity, a smaller value opening the gate more broadly, and $v_{b,i} = 1/2$ exactly when $c_{b,i}^t = \tau_{L,b}$.

The three planes do not carry the same kind of information, and the branch treats them accordingly. On the bird's-eye plane, which holds the horizontal layout of roads, footprints and the geometry between objects, the gate is used in full. Its open region is matched block by block under the gate, on the diffusion responses g_b^t themselves rather than on their curvature, each block normalized and matched on its own so that a large smooth background cannot dominate a global comparison, and the block losses are then averaged. On its closed region the student's band-pass energy is driven towards zero, and the level is what makes zero right. The teacher's class-interior detail on these planes comes from the texture of a dense return, and a sparse scan does not carry it. The teacher itself recovers only part of that detail when the same scene is sampled again under its own budget, down to about a third in the highest bands. Of the high-frequency energy an undistilled student puts there, some seven eighths is unrelated to the teacher's (Sec. 2.5.2). The conditional expectation is zero in that regime. Matching the teacher's magnitude would grow structure of the right size and the wrong content. Nor does the segmentation head give anything up for it. A closed gate marks a region the teacher reads as one class, and variation inside it does not change the prediction. On the two vertical planes the features vary mainly with how the scene occupies height rather than with what class occupies it, so a gate read there would again select occupancy. They receive the support-restricted match alone, without a gate and without suppression, which is what the plane geometry supports and no more.

Suppression towards zero has no floor of its own. Zero is the term's fixed point by design, and the student's interior curvature keeps falling for as long as training runs. Some of the teacher's genuine interior texture goes with it, most visibly on vegetation and building facades. That loss does not reach the outputs. A distilled student's planes hold as much teacher-aligned high frequency as an undistilled one, with fewer spurious boundaries and fewer missed ones (Sec. 2.5.2).

The exact block sizes follow each plane's own shape and are deferred, with the suppression reference, to Appendix E.1. The two halves of the branch therefore act on different functionals of the same response, the match on g_b^s itself and the suppression on its band-pass energy. Suppression is restricted to the smoother scales $b \in \mathcal{S}_L$, where the teacher gate is a reliable anti-aliasing cue; later scales receive larger η_b , following the first inequality of (9).

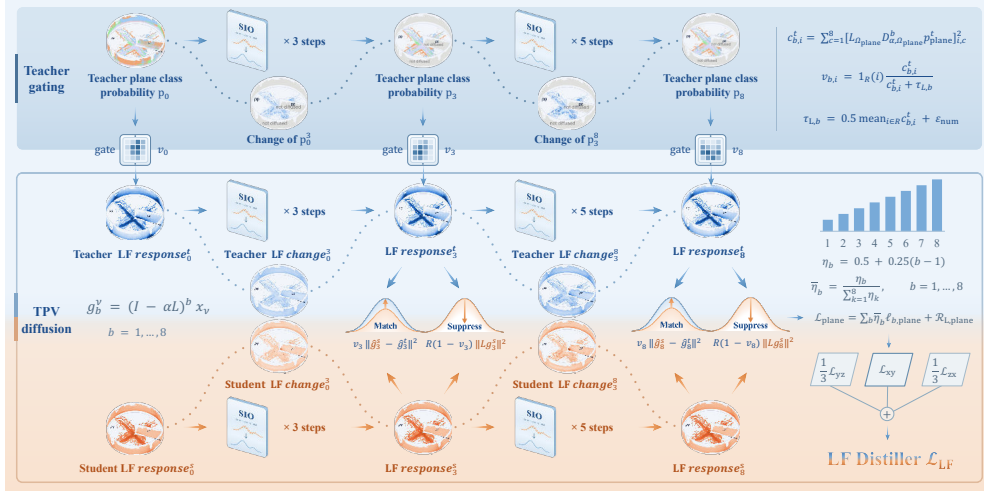


Fig. 6: The low-frequency branch.

Note. The three tri-perspective-view planes of the student and of the frozen teacher are smoothed by the same diffusion pyramid, and the branch reads the smoothed responses rather than the raw planes. Its gate is not read from the planes themselves: the teacher’s voxel class probabilities are averaged along the axis each plane collapses, over occupied cells alone, and that projected class map is diffused on the plane’s support subgraph to give the gate. On the bird’s-eye plane the gate is used in full, matching block by block where it is open and driving the student’s band-pass energy towards zero where it is closed. The two vertical planes receive the support-restricted match alone, with no gate and no suppression. Gate and loss share the step size and the schedule of scales but not the graph on which they are computed. All teacher quantities are detached, so gradients flow to the student only.

A spectrum is a shape, not a list of places, so the branch carries a further term. The LF regularizer $\mathcal{R}_L$ records how response energy is organized along two complementary axes; $\epsilon_{\text{num}} > 0$ below is a fixed numerical stabilizer, distinct from the response error ϵ in Theorem 1(ii). Let

$$E_{b,c}^{L,\nu} = \|g_{b,c}^\nu\|_2^2, \quad \mathbf{e}^{L,\nu} = \frac{E^{L,\nu}}{\|E^{L,\nu}\|_F + \epsilon_{\text{num}}}, \quad (18)$$

$$q_b^{L,\nu} = \sum_i \sum_c \omega_{b,i} (g_{b,i,c}^\nu)^2, \quad \mathbf{q}^{L,\nu} = \frac{\mathbf{q}^{L,\nu}}{\|\mathbf{q}^{L,\nu}\|_1 + \epsilon_{\text{num}}}, \quad \nu \in \{s, t\}.$$

The two statistics are not two marginals of one matrix, and the difference matters. The moments $E^{L,\nu}$ are taken over the whole plane with no weight, and their normalization is joint over channels and scales, so it divides out one overall gain per sample and nothing else; $\mathbf{e}^{L,\nu}$ therefore retains both how energy is allocated across channels and how it decays along the diffusion scales, whereas normalizing each scale separately would divide out that decay as well. The profile $q^{L,\nu}$ is instead read under a teacher-supplied spatial weight ω_b , which is the gate v_b at that scale on the bird’s-eye plane and the teacher’s non-isolated support on the two vertical planes. Both sides carry the same teacher weight, so the profile compares where the teacher says there is something to compare while the moments compare the field as a whole. After the weighted spatial

810 sum, $\hat{\mathbf{q}}^{L,\nu}$ describes the scale allocation alone. We align both:

$$811 \quad \mathcal{R}_L = \lambda_{\text{mom}}^L \frac{1}{K} \sum_{b=1}^K \|\hat{\mathbf{e}}_b^{L,s} - \hat{\mathbf{e}}_b^{L,t}\|_2^2 + \frac{\lambda_{\text{prof}}^L}{2} \|\hat{\mathbf{q}}^{L,s} - \hat{\mathbf{q}}^{L,t}\|_1. \quad (19)$$

812 These two summaries are complementary marginals of the same scale-channel energy
 813 matrix, rather than two unrelated penalties. The nonnegative coefficients λ_{mom}^L and
 814 λ_{prof}^L weight the two terms. The first is the stable contracted-moment statistic of
 815 Theorem 1(iii); the second is a first-order profile discrepancy that retains a non-
 816 vanishing gradient near alignment. The three TPV losses are combined as $\mathcal{L}_{\text{LF}} =$
 817 $\mathcal{L}_{xy} + \frac{1}{3}\mathcal{L}_{yz} + \frac{1}{3}\mathcal{L}_{zx}$. The xy plane receives the largest weight because the bird’s-eye
 818 view carries the primary horizontal scene layout, including roads, object footprints
 819 and inter-object geometry, while the two vertical planes provide complementary height
 820 structure.

821 3.4.2 High-Frequency Gated Distillation

822 On the output logits the band at issue is how the class field changes from one cell
 823 to the next along a surface. Sparse sampling does two things to that band. It loses
 824 changes the scene has, and, once the scan is put on a grid, it adds changes the scene
 825 never had. The job of the high-frequency branch is to tell the two apart. It works on
 826 the voxel logits, compared on the common support (C15), where both branches carry
 827 a value and the student carries a gradient. The teacher is frozen throughout, so what
 828 the common support buys is a differentiable student output at every coordinate the
 829 teacher queries.

830 However, comparing the two logit fields on the grid is not a trivial solution. First,
 831 the grid operator treats the empty cells as part of the domain. A residual formed across
 832 the boundary between an observed cell and an empty one is dominated by the step
 833 itself, so the response reports where the sensor happened to return rather than what
 834 the scene holds there. However large such a residual is, it describes occupancy, and no
 835 teacher can supply what is missing from it. Second, the gate has to say whether the
 836 teacher has content at a cell, and the teacher’s logit field answers that badly: most
 837 of its variation along a surface is confidence ripple inside a single class rather than a
 838 change of class. Third, zero would be the wrong place for the term to come to rest
 839 here, since part of what varies in the teacher is real and a sparse student can predict
 840 it. The branch answers the three in that order: it builds its operator from the teacher
 841 support, reads its gate from the teacher’s class probabilities, and charges the student
 842 only for the response that exceeds the teacher’s. In contrast to the planes, then, the
 843 first thing to fix here is the domain. The comparison has to be confined to the cells
 844 the teacher observed, and the operator must not reach across the edge of that region.
 845 A large response then means that the class field varies *along* the observed surface,
 846 which is the variation a sparse student loses.

847 To stop the operator from reaching into the empty cells the branch replaces the
 848 grid operator by the one the teacher support induces, and Fig. 7 lays out the path.
 849 Let each cell of Ω_T be adjacent to those of its six grid neighbours that also lie in Ω_T ,

and let d_i count them. Then

$$(L_\Omega x)_i = \frac{1}{6} \left(d_i x_i - \sum_{j \sim i} x_j \right), \quad i \in \Omega_T, \quad (20)$$

where neighbours outside Ω_T or outside the grid are simply absent, which is the same Neumann convention already used at the grid border. The operator is symmetric and positive semidefinite. Its quadratic form is $\frac{1}{6} \sum_{(u,v)} (x_u - x_v)^2$ over the retained edges, and its kernel consists of the fields that are constant on each connected component of the support. Every retained edge keeps the weight $1/6$, so a pattern lying within a surface has the eigenvalue it would have on the full grid and the scale axis keeps the spatial meaning it carries throughout the paper. The operator depends on the frame. But teacher and student share Ω_T , the same operator filters both, and spectral quantities are compared only within one frame. No residual the branch reports now straddles the edge of the scan.

To keep the gate off the teacher's confidence ripple we read it from the teacher's class probabilities $p^t = \text{softmax}(z^t)$ rather than from its logits,

$$e_{b,i} = \sum_c (p^t - D_\alpha^b p^t)_{i,c}^2, \quad w_{b,i} = \frac{e_{b,i}}{e_{b,i} + \tau_b}, \quad \tau_b = \kappa_H \text{mean}_{i \in R} e_{b,i}, \quad (21)$$

with $D_\alpha = I - \alpha L_\Omega$, one gate per response scale, shared across channels and detached from the student. The averages are taken over R , the support with its isolated cells removed, since a cell with no retained neighbour has an identically vanishing residual and would only dilute the threshold. The gate opens where the teacher's class assignment changes across the surface, at object boundaries and where the teacher is itself uncertain, and closes on the smooth interior of a class. The diffusion, the gated terms and their normalizers, τ and the weights included, are taken over Ω_T alone. One auxiliary statistic is not, and Appendix E.1 says which and why.

Match and suppress then divide that support between them. Where the gate is open the student's high-pass residual is aligned to the teacher's, both sides taken with their channel mean removed and divided by the same teacher-derived residual scale, so that the comparison is made in a common unit set by the teacher. Where the gate is closed the teacher is smooth along the surface. Zero would be the wrong resting place there so the student is charged only for the part of its residual that exceeds the teacher's,

$$\text{Suppress}_H = \frac{\sum_{i \in R} \sum_c c_c (1 - w_{b,i}) [(n_{i,c}^s)^2 - (n_{i,c}^t)^2]_+}{\sum_{i \in R} \sum_c c_c (1 - w_{b,i}) + \epsilon_{\text{num}}}, \quad (22)$$

where c_c are the per-class weights of Appendix E.1, so that the teacher, and not zero, is where the term is at rest.

Because the matched and suppressed halves share that reference, the pointwise minimizer of $w(n - n^t)^2 + (1 - w)[n^2 - (n^t)^2]_+$ is $n = n^t$ whatever the gate value. The gate therefore decides which deviation is charged rather than what the branch converges to: where it is open the student is drawn to the teacher from either side, and where it is closed only an excess over the teacher is charged.

At this level the teacher is the right target and zero would be the wrong one. What varies in the teacher’s logits inside a class is not texture but calibrated confidence, highest deep inside an object and falling as a boundary approaches. A sparse student can predict it: distillation recovers teacher-aligned energy in every band of the logit spectrum, the highest included (Sec. 2.5.2). A target of zero would also work against the logit distillation the objective already carries, which holds the teacher’s field as its own. The two terms would pull on one quantity in opposite directions, and the student would settle between them. Charging the excess alone avoids that. It leaves the teacher’s own ripple untouched and removes the part of the student’s that the sampler put there. At this level that is not a cosmetic difference: it is a flipped class and a boundary absent from the scene.

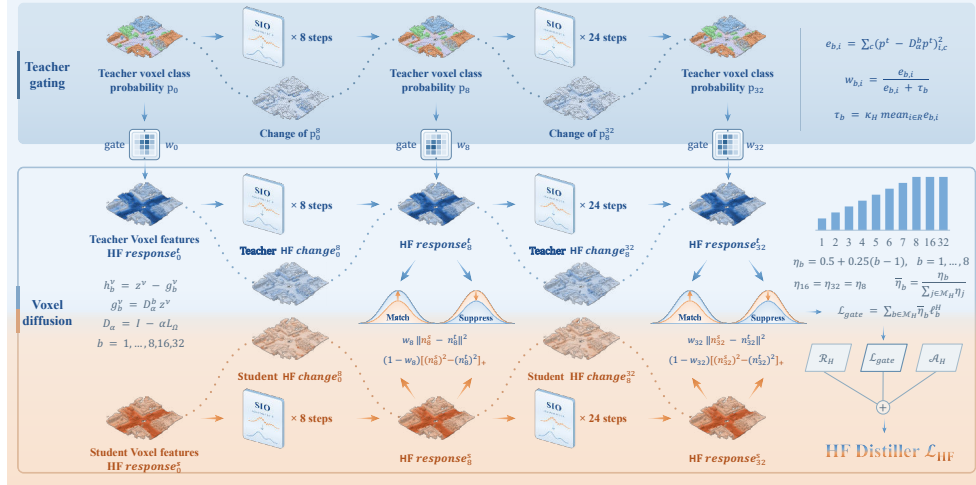


Fig. 7: The high-frequency branch.

Note. Both logit volumes are placed on the teacher’s support, and the operator that support induces replaces the grid operator, so that a large response means the class field varies along an observed surface rather than at the edge of the observed region. High-pass residuals are formed there, centred across channels and divided by one and the same teacher-derived scale, so that the two sides are compared in a common unit. The gate is read from the teacher’s class probabilities rather than from its logits, because most of the variation a logit field shows along a surface is confidence ripple inside one class. Where the gate is open the student is drawn to the teacher from either side; where it is closed only the part of its residual that exceeds the teacher’s is charged, so the teacher, and not zero, is where the term comes to rest. Cells the student returns but the teacher does not lie outside the common support and are neither matched nor suppressed. All teacher quantities are detached, so gradients flow to the student only.

The distinction the branch has to draw is between detail the student has observed and detail that sampling has fabricated, and the support draws it. Cells the student returns but the teacher does not are real surface; they are absent from the common-support field by construction, so nothing there is suppressed and nothing there is matched either. Roughly an eighth of the student’s occupied cells fall in that category, and for them the branch has no teacher reference to offer, which is the honest limit of what a distillation on the teacher’s support can reach. What is left inside Ω_T is the

case that matters, where the teacher observes a cell more densely and finds it smooth while the student's field varies, and it is that variation the sampler created.

A student could meet the response comparison in ways that change nothing semantic, and two further terms rule those out. Both act inside $\mathcal{L}_{\text{HF}}$ and are written $\mathcal{L}_V$ and $\mathcal{L}_C$; neither is a separate path of Eq. (26), and neither is a temperature-softened comparison, which is what $\mathcal{L}_{\text{KD}}$ alone provides. The first compares the logit values themselves, both sides with their channel mean removed and normalized by the root mean square over Ω_T , since a comparison of residuals is blind to a constant added across a whole component of the support. The second charges the student for the variation of its class probabilities between adjacent cells at which the gate is closed at both ends, weighted by $(1 - w_u)(1 - w_v)$, which asks it to hold one class wherever the teacher reads one class. Both act inside $\mathcal{L}_{\text{HF}}$ and leave the objective of Eq. (26) with its four terms. Besides the base pyramid $b = 1, \dots, K$, the gated terms reuse cumulative scales $b \in \{16, 32\}$ so their high-pass response reaches the measured aliasing band below $\lambda \approx 0.4$; these extra scales do not enter the auxiliary statistics.

The HF auxiliary regularizer has the same two-axis form as Eq. (19), but is evaluated on high-pass responses. Let $\hat{\mathbf{e}}_k^{\text{H},\nu}$ denote the contracted moments of the terminal base residual h_K^ν under the same joint channel-scale normalization, and let $\hat{\mathbf{q}}^{\text{H},\nu}$ be the ℓ_1 -normalized vector of channel-aggregated residual energies $q_b^{\text{H},\nu} = \sum_c \|h_{b,c}^\nu\|_2^2$ over the base scales, where $\nu \in \{s, t\}$. For a normalized K -scale response profile q , its Wiener flatness is

$$\text{Flat}(q) = \frac{\exp\left(K^{-1} \sum_{b=1}^K \log(q_b + \epsilon_{\text{num}})\right)}{K^{-1} \sum_{b=1}^K q_b + \epsilon_{\text{num}}} \in (0, 1]. \quad (23)$$

A value near one means that residual energy is spread nearly uniformly across response scales, which is a broadband, noise-like profile, whereas a small value indicates concentration on fewer scales. Then

$$\begin{aligned} \mathcal{R}_{\text{H}} = & \lambda_{\text{mom}}^{\text{H}} \frac{1}{K} \sum_{k=1}^K \|\hat{\mathbf{e}}_k^{\text{H},s} - \hat{\mathbf{e}}_k^{\text{H},t}\|_2^2 + \frac{\lambda_{\text{prof}}^{\text{H}}}{2} \|\hat{\mathbf{q}}^{\text{H},s} - \hat{\mathbf{q}}^{\text{H},t}\|_1 \\ & + \lambda_{\text{flat}} [\text{Flat}(\hat{\mathbf{q}}^{\text{H},s}) - \text{Flat}(\hat{\mathbf{q}}^{\text{H},t})]_+. \end{aligned} \quad (24)$$

Here $[u]_+ = \max(u, 0)$ and the three nonnegative λ coefficients weight the moment, profile, and flatness terms. The moment and profile terms again constrain channel allocation and scale allocation, respectively. Only HF needs the final one-sided flatness hinge. It acts when the student's residual profile is more broadband and noise-like than the teacher's, but never rewards making an already more concentrated student artificially flatter. This global scale-profile constraint complements the spatial gate; it is motivated by the measured broadband aliasing and is not identical to the coordinate-energy flatness analyzed for the ideal gate in Theorem 1(iv).

Taken together, the branch equalizes the student's response in this band against the teacher's: it matches the variation the teacher reads along a surface, charges the

student for the excess where the teacher reads none, and leaves the teacher’s own ripple where it is.

3.4.3 Overall Objective

The two branches equalize the spectrum where the network itself splits the bands, the LF branch on the tri-perspective-view planes and the HF branch on the common-support voxel logits, each matching teacher-supported response, suppressing student-only response, and holding its auxiliary spectral invariants. What is left is to make one training objective out of them. However, summing the two branch losses is not a trivial solution. First, each branch treats its own channels as spatial fields, the low-frequency branch the feature channels of the projected planes and the high-frequency branch the semantic-logit channels. Both align where and at which scales the teacher’s response varies, and neither constrains the relative class probabilities at an individual voxel: what they normalize and aggregate are not those relations. Those cross-class relations carry the teacher’s “dark knowledge”, for example that a rare pedestrian-like voxel is more easily confused with a cyclist than with road, and become especially informative when the teacher is strong [18]. Second, the two kinds of knowledge do not sit in the same part of the signal: the spectral comparisons are normalized per sample, while the cross-class information sits in the relative values of the softened class probabilities. The first gap is closed by a per-voxel softened-logit term, the second by keeping the terms separate, each with its own weight. Let $\mathcal{V}$ be the common teacher-support voxel set, let $\ell_u^\nu \in \mathbb{R}^C$ be the C -class logit vector at voxel $u \in \mathcal{V}$ on side $\nu \in \{s, t\}$, and let $T > 1$ be the distillation temperature. With $\mathbf{p}_u^\nu(T) = \text{softmax}(\ell_u^\nu/T)$, the dark-knowledge term is

$$\mathcal{L}_{\text{KD}} = \frac{T^2}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \text{KL}(\mathbf{p}_u^t(T) \parallel \mathbf{p}_u^s(T)). \quad (25)$$

Here the factor T^2 keeps the gradient scale comparable across temperatures. This term answers the first gap. It adds no spatial comparison; it only asks the student to weigh the classes at a voxel as the teacher weighs them, runners-up included. The two signals are therefore complementary rather than redundant. $\mathcal{L}_{\text{LF}}$ and $\mathcal{L}_{\text{HF}}$ specify *where and how* class-response fields should vary in space, whereas $\mathcal{L}_{\text{KD}}$ specifies *how classes compete* at each voxel through the teacher’s softened secondary-class probabilities.

LiEqualizer augments the task loss with both forms of knowledge,

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{LF}} \mathcal{L}_{\text{LF}} + \lambda_{\text{HF}} \mathcal{L}_{\text{HF}} + \lambda_{\text{KD}} \mathcal{L}_{\text{KD}}. \quad (26)$$

where $\lambda_{\text{LF}}, \lambda_{\text{HF}}, \lambda_{\text{KD}} \geq 0$ balance the LF, HF, and dark-knowledge terms against the supervised task loss $\mathcal{L}_{\text{task}}$. Read term by term, the sum asks the student to do the task, to restore the structure sampling took out of the code, to drop the detail sampling wrote into the logits, and to match the teacher’s class probabilities at each voxel. Giving each term its own weight is what answers the second gap. The spectral response comparisons are normalized per sample and are therefore insensitive to branch gain; in contrast, $\mathcal{L}_{\text{KD}}$ deliberately preserves relative cross-class information through the temperature- T probability vectors. Robustness to sampling degradation comes from the response family and teacher gates, not from normalization alone. Exact spectral

Table 4: The LiEqualizer training objective. LF and HF equalize complementary spatial response fields, while soft-label KD transfers per-voxel cross-class knowledge.

Component	Role	Branches
$\mathcal{L}_{\text{match}}$	match teacher-supported response	LF and HF
$\mathcal{L}_{\text{sup}}$	suppress student-only response	LF and HF
$\mathcal{R}_{\text{mom}}$	energy across channels at fixed scale	LF and HF
$\mathcal{R}_{\text{prof}}$ (profile ℓ_1)	energy across scales after channel aggregation	LF and HF
$\mathcal{R}_{\text{flat}}$	one-sided broadband-noise penalty	HF only
$\mathcal{L}_{\text{KD}}$	per-voxel softened class distribution	voxel logits

normalizers, scale sets, implementation-only stabilizers, and branch hyper-parameters are given in Appendix E.1.

In this way the sampling gap is no longer an unexplained difference between two feature maps. It is a spectral imbalance, and the objective corrects it scale by scale, matching the detail the teacher has and suppressing the response it does not, under the teacher’s own judgement of where content exists.

4 Discussion and Conclusion

This work studies zero-shot cross-LiDAR semantic segmentation from a spectral perspective. Building on the view that a LiDAR is a sparse sampler of one shared 3D world, prior work separates scene-intrinsic geometry (“SIG”) from sampling geometry (“SG”) and randomizes “SG” so the model learns “SIG” [1]. We ask what “SG” is as a property of the signal, and answer that it is spectral: different LiDARs see different spectra of the same world. **LiEqualizer** turns this observation into a method. It measures the spectral damage of sampling in two domains and models it with a per-frequency transfer law. A theorem then fixes which spectral statistics to align and which channels to trust. The missing structure is distilled from a teacher that has seen the full spectrum, equalizing the student’s spectral response through teacher-gated low- and high-frequency branches. Experiments on three benchmarks, together with zero-shot deployment on three full-size vehicle platforms carrying unseen LiDARs, show that the restored spectrum translates into better cross-sensor accuracy.

The main findings of this study include:

- **The cross-LiDAR gap has a measurable spectral signature, with opposite symptoms in the two representation domains.** Sparse sampling does two distinct things to the signal, and they appear in opposite directions depending on where one looks: on raw point clouds it *attenuates* real high frequencies (the density signal retains 24–30% of the dense value, driveable ground 0.42–0.56), while on the voxel grids the network operates on it *injects* near-white high-frequency energy (32× the teacher’s, with the low-frequency shape left intact). The reversal is not a contradiction but a consequence of rasterization turning the sampling pattern itself into signal content; both symptoms follow from a single model once “SG” is treated

as a sampler rather than a filter, that is, as an operator that creates frequencies rather than only removing them.

- **The lost spectrum cannot be recovered from sparse observations alone; it must be distilled.** Input-side randomization controls what the model ignores; it cannot restore measurements the sensor never returned. We make this precise: consistency between two sparse views cannot invert the attenuation, because both views have lost the same frequencies, so the attenuated magnitudes are simply not identifiable from sparse data. A teacher trained on densified scenes has observed the full spectrum and supplies the missing information, which is why an input-side treatment and a distillation treatment are complementary rather than redundant, and why the two combine in our experiments.
- **What to align is dictated by theory, not chosen by tuning.** A single theorem fixes the design: the complete diffusion-response sequence identifies spectral energy, while any finite prefix controls Lipschitz spectral statistics with an explicit $O(1/B)$ approximation term. Every smoothing step provably raises the signal-to-noise ratio of the comparison, so alignment happens after smoothing and never on an ungated raw pointwise residual, the channel that in a sparse student carries sampling ripple rather than real contours. Each loss term in LiEqualizer traces back to one part of this theorem, and the ablations follow the directions it predicts.
- **Both branches require teacher-complement gating.** Because sparse sampling also folds artifacts into low and middle frequencies, LF uses teacher curvature to distinguish supported structure from student-only ripple; HF uses teacher residual amplitude to distinguish real detail from broadband aliasing. The ideal LF optimum obeys a complementary curvature-energy bound, while the ideal HF optimum is provably never coordinate-profile flatter than the teacher, so fabricated structure cannot reduce either ideal objective. Empirically, the distilled student’s spectrum moves toward the teacher’s in exactly the statistics the objective optimizes, closing the loop between what is trained and what is measured.

Taken together, these findings extend the understanding of cross-LiDAR generalization in two ways. First, the domain gap should be read not only as a mismatch of density or sensor configuration, but as a difference between the *spectra* that different samplers produce from the same scene. That difference appears as attenuation in one representation and as aliasing in another. Second, closing it requires both **invariance** to the sampler, which randomization provides, and **restoration** of what the sampler removed, which only a spectrally complete teacher can provide; invariance to a sampler does not make the sampler harmless. Together these define the principle behind LiEqualizer: equalizing the spectrum that sparse sampling reshapes.

We also note the boundaries of the approach. By design, the teacher is where the high-frequency branch comes to rest, and the gate decides only which departure from it is charged: information the teacher itself does not carry is deliberately not synthesized, so LiEqualizer restores transferable spectral structure rather than a hypothetical “complete” signal. The teacher is a relative reference, so teacher quality upper-bounds what can be transferred. Natural next steps are to extend the spectral account to detection and occupancy prediction, to model other degradations such as

adverse weather and motion distortion in the same per-frequency language, and to explore spectrum-aware data generation.

Overall, LiEqualizer offers a spectral route toward sensor-agnostic LiDAR perception. By characterizing sampling degradation in the frequency domain and repairing it through theory-guided distillation that acts at training time only and leaves the deployed model unchanged, it complements input-side generalization and further lowers the cost of bringing existing perception systems to new LiDARs. Part of what a sparser sensor fails to see is therefore a limit of how its perception model was trained, not of the sensor itself.

Data availability. The datasets used in this article, SemanticKITTI, nuScenes and Waymo, are publicly accessible. The real-vehicle data and the segmentation results of LiEqualizer on the three platforms, together with the outputs of the spectral analyses reported here, can be obtained through the [project page](#).

Code availability. We provide a streamlined version of this project on the [project page](#), which allows reproduction of the best-performing LiEqualizer-P model. The release includes the spectral measurement protocol, so that sampling degradation and learned representations can be analysed under the same definitions.

Supplementary information. The appendix provides additional analyses and implementation details that complement the main paper:

- Appendix A: extended related work on LiDAR semantic segmentation, cross-sensor generalization, knowledge distillation, and graph spectral analysis.
- Appendix B: the vocabulary and the proofs. A stand-alone glossary of every spectral term used in this paper (Appendix B.1), for readers without graph signal-processing background, and the unified theory in full (Appendix B.2): setting, three lemmas, complete proofs of Theorem 1 and Proposition 2, and scope remarks.
- Appendix C: the spectral measurement protocol, which the main text cites as the basis of the design. How a spectrum is read when it cannot be formed, the operator and the scales, what each statistic measures, the two ways a branch’s logits can be placed on the lattice, and the operational details of the suite (Appendix C.2).
- Appendix D: the two spectral readings behind the design and behind the claim. What sampling does to the signal, measured on dense and sparse frame pairs across the three datasets (Appendix D.2), and what distillation does to that damage, read band by band against the teacher (Appendix 2.5).
- Appendix E: every loss term in full, with its weights, ablation switches and cost (Appendix E.1).
- Appendix F: coarse-label mapping rules for nuScenes, SemanticKITTI, and Waymo.
- Appendix G: the remaining empirical material. Band-wise spectral figures for the two representations (Appendix G.1), the attribution of the SemanticKITTI-to-nuScenes drop (Appendix G.2), qualitative results from on-vehicle testing on unseen LiDAR platforms (Appendix G.3), the student’s multi-scale profile beside the teacher’s before and after distillation, and the gate maps that show where the low-frequency gate opens (Appendix G.4), and additional experimental settings with class-wise results (Appendix G.5).

1100 **Acknowledgements.** This work is supported by the National Natural Science
 1101 Foundation of China (Grants 52372415 and 52221005), the Fundamental and Interdis-
 1102 ciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant
 1103 JYB2025XDXM123) and the Beijing Natural Science Foundation (Grant L259013).

1104 References

- 1105 [1] Anonymous for review: Zero-shot Generalizable LiDAR Semantic Segmentation
 1106 via Scene–Sensor Disentanglement. arXiv preprint, number TBD (2026)
- 1107 [2] Xiao, A., Zhang, X., Shao, L., Lu, S.: A Survey of Label-Efficient Deep Learning
 1108 for 3D Point Clouds. arXiv (2023)
- 1109 [3] Yi, L., Gong, B., Funkhouser, T.: Complete & label: A domain adaptation
 1110 approach to semantic segmentation of lidar point clouds. In: Proceedings of
 1111 the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.
 1112 15363–15373 (2021)
- 1113 [4] Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: Squeezesegv2: Improved
 1114 model structure and unsupervised domain adaptation for road-object segmenta-
 1115 tion from a lidar point cloud. In: 2019 International Conference on Robotics and
 1116 Automation (ICRA), pp. 4376–4382 (2019). [https://doi.org/10.1109/ICRA.2019.](https://doi.org/10.1109/ICRA.2019.8793495)
 1117 [8793495](https://doi.org/10.1109/ICRA.2019.8793495)
- 1118 [5] Peng, X., Zhu, X., Ma, Y.: Cl3d: Unsupervised domain adaptation for cross-lidar
 1119 3d detection. Proceedings of the AAAI Conference on Artificial Intelligence **37**(2),
 1120 2047–2055 (2023) <https://doi.org/10.1609/aaai.v37i2.25297>
- 1121 [6] Kim, H., Kang, Y., Oh, C., Yoon, K.-J.: Single domain generalization for
 1122 lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on
 1123 Computer Vision and Pattern Recognition, pp. 17587–17598 (2023)
- 1124 [7] Saltori, C., Osep, A., Ricci, E., Leal-Taixé, L.: Walking your lidog: A journey
 1125 through multiple domains for lidar semantic segmentation. In: Proceedings of the
 1126 IEEE/CVF International Conference on Computer Vision, pp. 196–206 (2023)
- 1127 [8] Sanchez, J., Deschaud, J.-E., Goulette, F.: Domain generalization of 3d seman-
 1128 tic segmentation in autonomous driving. In: Proceedings of the IEEE/CVF
 1129 International Conference on Computer Vision, pp. 18077–18087 (2023)
- 1130 [9] Ryu, K., Hwang, S., Park, J.: Instant domain augmentation for lidar semantic
 1131 segmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pat-
 1132 tern Recognition (CVPR), pp. 9350–9360. IEEE, Vancouver, BC, Canada (2023).
 1133 <https://doi.org/10.1109/CVPR52729.2023.00902>
- 1134 [10] Wei, Y., Wei, Z., Rao, Y., Li, J., Zhou, J., Lu, J.: LiDAR distillation: Bridging
 1135 the beam-induced domain gap for 3D object detection. In: ECCV (2022)

- [11] Liu, M., Zhou, Y., Qi, C.R., *et al.*: LESS: Label-efficient semantic segmentation for LiDAR point clouds. In: ECCV (2022)
- [12] Kim, J., Woo, J., Kim, J., Im, S.: Rethinking lidar domain generalization: Single source as multiple density domains. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, pp. 310–327. Springer, Cham (2025)
- [13] Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019)
- [14] Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6411–6420 (2019)
- [15] Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: PointOcc: Cylindrical Tri-Perspective View for Point-Based 3D Semantic Occupancy Prediction. arXiv (2023). <https://doi.org/10.48550/arXiv.2308.16896>
- [16] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
- [17] Sanchez, J., Deschaud, J.-E., Goulette, F.: COLA: COarse LAbel Pre-Training for 3D Semantic Segmentation of Sparse LiDAR Datasets. arXiv (2023). <https://doi.org/10.48550/arXiv.2202.06884>
- [18] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
- [19] Hou, Y., Zhu, X., Ma, Y., Loy, C.C., Li, Y.: Point-to-voxel knowledge distillation for lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8479–8488 (2022)
- [20] Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5311–5320 (2021)
- [21] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)
- [22] Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet plus plus : Deep hierarchical feature learning on point sets in a metric space. In: Guyon, I., Luxburg, U.V., Bengio, S.,

- 1172 Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neu-
 1173 ral Information Processing Systems 30 (Nips 2017), vol. 30. Neural Information
 1174 Processing Systems (nips), La Jolla (2017)
- 1175 [23] Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham,
 1176 A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In:
 1177 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 1178 Recognition, pp. 11108–11117 (2020)
- 1179 [24] Zhao, H., Jiang, L., Jia, J., Torr, P.H.S., Koltun, V.: Point transformer. In: Pro-
 1180 ceedings of the IEEE/CVF International Conference on Computer Vision, pp.
 1181 16259–16268 (2021)
- 1182 [25] Milioto, A., Vizzo, I., Chley, J., Stachniss, C.: Rangenet plus plus : Fast and
 1183 accurate lidar semantic segmentation. In: 2019 Ieee/Rsj International Conference
 1184 on Intelligent Robots and Systems (Iros), pp. 4213–4220. Ieee, New York (2019)
- 1185 [26] Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z.:
 1186 Rethinking Range View Representation for LiDAR Segmentation. arXiv (2023).
 1187 <https://doi.org/10.48550/arXiv.2303.05367>
- 1188 [27] Ando, A., Gidaris, S., Bursuc, A., Puy, G., Boulch, A., Marlet, R.: RangeViT:
 1189 Towards Vision Transformers for 3D Semantic Segmentation in Autonomous
 1190 Driving. arXiv (2023). <https://doi.org/10.48550/arXiv.2301.10222>
- 1191 [28] Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation
 1192 with submanifold sparse convolutional networks. In: Proceedings of the IEEE
 1193 Conference on Computer Vision and Pattern Recognition, pp. 9224–9232 (2018)
- 1194 [29] Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection.
 1195 Sensors **18**(10), 3337 (2018) <https://doi.org/10.3390/s18103337>
- 1196 [30] Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical
 1197 and asymmetrical 3d convolution networks for lidar segmentation. In: Proceedings
 1198 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.
 1199 9939–9948 (2021)
- 1200 [31] Kong, L., Quader, N., Liong, V.E.: Conda: Unsupervised domain adaptation for
 1201 lidar segmentation via regularized domain concatenation. In: 2023 IEEE Interna-
 1202 tional Conference on Robotics and Automation (ICRA), pp. 9338–9345 (2023).
 1203 <https://doi.org/10.1109/ICRA48891.2023.10160410>
- 1204 [32] Shaban, A., Lee, J., Jung, S., Meng, X., Boots, B.: Lidar-uda: Self-ensembling
 1205 through time for unsupervised lidar domain adaptation. In: Proceedings of
 1206 the IEEE/CVF International Conference on Computer Vision, pp. 19784–19794
 1207 (2023)

- [33] Bešić, B., Gosala, N., Cattaneo, D., Valada, A.: Unsupervised domain adaptation for lidar panoptic segmentation. *IEEE Robotics and Automation Letters* **7**(2), 3404–3411 (2022) <https://doi.org/10.1109/LRA.2022.3147326>
- [34] Zhang, B., Yuan, J., Shi, B., Chen, T., Li, Y., Qiao, Y.: Uni3d: A unified baseline for multi-dataset 3d object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9253–9262 (2023)
- [35] Xu, Z., Xu, S.: Mergeocc: Bridge the domain gap between different lidars for robust occupancy prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26539–26548 (2025)
- [36] Li, M., Zhang, Y., Ma, X., Qu, Y., Fu, Y.: Bev-dg: Cross-modal learning under bird’s-eye view for domain generalization of 3d semantic segmentation. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11598–11608 (2023). <https://doi.org/10.1109/ICCV51070.2023.01068>
- [37] Yan, X., Gao, J., Zheng, C., *et al.*: 2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds. In: *ECCV* (2022)
- [38] Cen, J., Zhang, S., *et al.*: CMDFusion: Bidirectional fusion network with cross-modality knowledge distillation. *arXiv:2307.04091* (2023)
- [39] Kang, J., Wang, J., Luo, D.: Domain adaptation-based crossmodal knowledge distillation for 3D semantic segmentation. *arXiv:2509.00379* (2025)
- [40] Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: *ICCV* (2021)
- [41] Pham, C., *et al.*: Frequency attention for knowledge distillation. In: *WACV* (2024)
- [42] Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine* **30**(3), 83–98 (2013)
- [43] Ortega, A., Frossard, P., *et al.*: Graph signal processing: Overview, challenges, and applications. *Proc. IEEE* **106**(5), 808–828 (2018)
- [44] Dinesh, C., Cheung, G., Bajić, I.V.: Point cloud denoising via feature graph laplacian regularization. *IEEE TIP* **29**, 4143–4158 (2020)
- [45] Shao, Y., *et al.*: Attribute compression of 3D point clouds using laplacian sparsity optimized graph transform. In: *IEEE VCIP* (2017)
- [46] Chen, S., Tian, D., Feng, C., Vetro, A., Kovačević, J.: Fast resampling of three-dimensional point clouds via graphs. *IEEE TSP* **66**(3), 666–681 (2018)
- [47] Fu, Z., *et al.*: GraphReg: Dynamical point cloud registration with geometry-aware

- graph signal processing. arXiv:2302.01109 (2023)
- [48] Ding, X., Wu, H.-T.: Impact of signal-to-noise ratio and bandwidth on graph Laplacian spectrum from high-dimensional noisy point cloud. arXiv:2011.10725 (2021)
- [49] Rahaman, N., Baratin, A., *et al.*: On the spectral bias of neural networks. In: ICML (2019)
- [50] Xu, Z.-Q.J., Zhang, Y., Luo, T., Xiao, Y., Ma, Z.: Frequency principle: Fourier analysis sheds light on deep neural networks. *Commun. Comput. Phys.* **28**(5), 1746–1767 (2020)
- [51] Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421 (2018)

Appendix A Extended Related Work

This appendix expands the positioning sketched in the Main text. Prior work is grouped by the mechanism it relies on rather than by publication order, and each group closes with the boundary that motivates the present paper.

A.1 LiDAR Semantic Segmentation and Cross-Sensor Generalization

Architectures for LiDAR semantic segmentation are commonly grouped by representation. Point-based networks operate on raw coordinates [14, 21–24]; projection-based networks rasterize the scan into range or bird’s-eye views to reuse 2D machinery [25–27]; voxel-based networks discretize space and apply sparse 3D convolutions, the strongest family on current benchmarks [28–30]. All of them, however, are trained and evaluated on data from a single sensor, and inherit that sensor’s sampling pattern as an implicit prior.

When the sensor changes, three lines of work address the resulting gap. *Domain adaptation* [3–5, 31–33] aligns feature distributions toward one known target LiDAR; it is effective for that target but needs target-domain data and retraining for every new sensor. *Multi-dataset training* [34, 35] unifies several LiDARs during training, but its coverage is fixed by the chosen datasets. *Domain generalization* either regularizes features against restricted variations such as beam dropping or density change [6, 7, 12, 36], or reconstructs denser geometry before learning, via multi-frame accumulation with label propagation [8] or re-rendering under predefined sensor configurations [9]. These methods improve robustness against the variations they enumerate, but none of them describes what sampling does to the signal itself. LiSeer [1] goes one step further: it treats every LiDAR as a sparse sampler of one shared world, separates

scene-intrinsic geometry (“SIG”) from sampling geometry (“SG”), and randomizes “SG” through a densify-then-resample data factory so that the model learns “SIG”.

Our work is orthogonal to all of the above. We modify neither the backbone nor the input pipeline; we characterize, in the graph-frequency domain, the degradation that “SG” imprints on the signal, and restore the affected spectral structure by distillation. What these methods enumerate as a list of sensor variations to be robust against, we treat as one measurable change of spectrum, which is what makes it possible to say not only that the gap exists, but which part of the signal it consists of.

A.2 Knowledge Distillation for LiDAR Perception

Knowledge distillation (KD) transfers the behavior of a teacher network to a student [18]. For LiDAR segmentation, existing work falls into three groups by the nature of the teacher.

Compression-oriented KD keeps the input fixed and shrinks the model. PVKD [19] is the representative for LiDAR segmentation: it distills both point-level and voxel-level outputs and adds supervoxel-level affinity terms, explicitly designed around the sparsity and varying density of point clouds. Here teacher and student observe the same scan; the knowledge gap comes from capacity, not from the data.

Cross-modal KD imports knowledge from another modality. 2DPASS [37] injects 2D priors from images during training while keeping a LiDAR-only network at inference; CMDFusion [38] distills bidirectionally between camera and LiDAR branches; recent work combines cross-modal distillation with domain adaptation to reduce annotation cost [39]. The knowledge source is a different modality of the same scene, not a different sampling density of the same modality.

Dense-to-sparse KD is the closest setting to ours, and two representatives define it. LiDAR Distillation [10] targets the beam-induced gap in 3D detection: it progressively downsamples high-beam scans into pseudo low-beam data and matches the student’s BEV feature maps to the teacher’s by pointwise regression. LESS [11] targets label efficiency in segmentation: a teacher trained on temporally accumulated multi-scan point clouds supervises a single-scan student through its soft predictions. Both confirm the premise that a denser observation makes a stronger teacher, yet neither shares our setting. First, the dense reference itself differs in kind: pseudo low-beam pairs are downsampled from a single high-beam sensor, multi-scan accumulation depends on ego-motion and leaves trails on dynamic objects, whereas we distill from per-scan densified scenes [1]. Second, and more fundamentally, their analyses never enter the frequency domain: knowledge is transferred by spatial pointwise matching of features or predictions, with no account of *what* the density gap removes from the signal, which spectral components remain trustworthy, or why pointwise high-frequency matching might backfire. Our measurements and theory answer exactly these questions, and Sec. 3.3 shows that the spatial pointwise channel these methods rely on is, under sampling degradation, the least reliable one.

A separate line explores *frequency-domain losses* in 2D vision. The focal frequency loss [40] narrows spectral gaps in image reconstruction; frequency-attention distillation [41] matches teacher and student features through a learnable global filter in the Fourier domain, exploiting the fact that each frequency aggregates global context.

1323 These methods use the frequency domain as a better coordinate system for matching:
 1324 teacher and student still observe the same input, and their spectral differences stem
 1325 from model capacity. Our situation is qualitatively different: the student’s *input* is
 1326 systematically degraded by sampling, so its spectrum is attenuated and aliased in a
 1327 measurable way. Handling this requires two ingredients that existing frequency-domain
 1328 KD does not provide: an explicit spectral model of the sampling degradation, and a
 1329 theory of which spectral statistics are worth aligning and which channels cannot be
 1330 trusted. This paper supplies both.

1331 A.3 Graph Spectral Analysis of Point Clouds

1332 Graph signal processing (GSP) extends harmonic analysis to irregular domains [42, 43],
 1333 and has long served point-cloud *processing*: graph-Laplacian regularizers for denois-
 1334 ing [44], graph transforms for attribute compression [45], graph-filter-based resampling
 1335 that preserves high-frequency content [46], and spectral embeddings for registra-
 1336 tion [47]. In these works the spectrum is a tool for low-level reconstruction quality;
 1337 the question of cross-sensor generalization does not arise. On the theory side, the
 1338 spectral behavior of graph Laplacians built from noisy point clouds has been character-
 1339 ized [48], and the spectral bias of neural networks, the tendency to fit low frequencies
 1340 first [49, 50], implies that whatever high-frequency evidence sparse sampling removes is
 1341 also the part a network recovers most slowly. This compounds the problem we measure
 1342 and strengthens the case for transferring high frequencies explicitly from a teacher. To
 1343 our knowledge, no prior work describes the cross-LiDAR domain gap in graph-spectral
 1344 terms, and none uses spectral statistics as distillation targets; this paper does both,
 1345 with a measured degradation model and a theorem that dictates the objective.

1346 Appendix B Spectral Terminology and Proofs

1347 This appendix collects the vocabulary and the proofs behind the statements of the
 1348 main text. Readers without a graph signal-processing background may read the
 1349 glossary first; the proofs that follow are self-contained.

1350 B.1 Glossary of Spectral Notions: Terminology, Formulas, 1351 and Elementary Proofs

1352 This part is a stand-alone reference for readers without prior graph signal-processing
 1353 background. It separates mathematical definitions, proved properties, modeling
 1354 assumptions, and empirical diagnostics. Unless explicitly stated otherwise, all signals
 1355 are real-valued and finite-dimensional.

1356 B.1.1 Domains, samples, grids, and supports

1357 *Raw samples and representation grids.*

1358 A point cloud is a finite set of measured locations and attributes, $\mathcal{P} = \{(p_\ell, a_\ell)\}_{\ell=1}^{n_{\text{pts}}}$.
 1359 Its cardinality n_{pts} is unrelated to the symbol N used below. A TPV, BEV, or voxel
 1360 representation is a tensor on a fixed ambient grid $V = \{1, \dots, N\}$. After fixing one

sample and one channel, its values form a graph signal $x : V \rightarrow \mathbb{R}$, identified with a vector $x \in \mathbb{R}^N$. The domain is discrete even when the values are continuous and differentiable logits.

Ambient, algebraic, observed, and query supports.

The ambient domain is the complete node set V . The algebraic support of a particular tensor is

$$\text{supp}(x) = \{u \in V : x_u \neq 0\}.$$

This set is not automatically the set of observed coordinates: a legitimately observed logit can equal zero, while an unobserved scatter coordinate can be filled by a structural zero. We therefore reserve Ω_{obs} for measurement support and Ω_{query} for the coordinates at which a decoder is evaluated. For a coordinate subset $\Omega \subseteq V$, define the diagonal mask M_Ω by

$$(M_\Omega x)_u = \begin{cases} x_u, & u \in \Omega, \\ 0, & u \notin \Omega. \end{cases}$$

Teacher and student raw point sets may differ, and their algebraic or observed supports may differ, but a pointwise loss requires a common coordinate system. In the corrected high-frequency path, both are evaluated on Ω_T , the teacher query set. The graph-theoretic statements in Part II use the common ambient graph; they do not require identical raw point sets.

Why gradient support is a separate issue.

If a scatter operation writes trainable updates only on Ω_S and fills the rest with constants, then

$$\frac{\partial (M_{\Omega_S} F_\theta)_u}{\partial \theta} = 0 \quad \text{for } u \notin \Omega_S.$$

A mathematically defined loss can therefore contain terms that are optimization-inert. Querying the trainable student field at Ω_T restores a gradient path on every teacher coordinate.

B.1.2 Weighted graphs and graph Laplacians

Weighted graph.

A finite undirected weighted graph is $G = (V, E, W)$, where $W = W^\top \in \mathbb{R}^{N \times N}$, $W_{uv} \geq 0$, and $W_{uv} > 0$ only when u and v are adjacent. Its degree matrix is diagonal, $D_{uu} = \sum_v W_{uv}$. Two common Laplacians are

$$L_{\text{comb}} = D - W, \quad L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}.$$

The symmetric normalized form requires $D_{uu} > 0$ (no isolated node), or an explicit convention for zero degrees. The implemented regular-grid operator is written $L = I - A$, where A is a symmetric averaging matrix with boundary self-loops induced by replicate padding. It is a valid weighted graph Laplacian but should not be confused blindly with every textbook normalization.

1397 **Positive semidefinite matrix.**

1398 A real symmetric matrix L is positive semidefinite, written $L \succeq 0$, when $z^\top L z \geq 0$ for
 1399 all $z \in \mathbb{R}^N$. It is positive definite if the inequality is strict for every nonzero z .

1400 **Proposition 3** (PSD property of an undirected weighted Laplacian) *For $L_{\text{comb}} = D - W$
 1401 with $W = W^\top \geq 0$,*

$$1402 \quad z^\top L_{\text{comb}} z = \frac{1}{2} \sum_{u,v} W_{uv} (z_u - z_v)^2 \geq 0.$$

1403 *Hence L_{comb} is symmetric PSD.*

1404 *Proof* Expand the right-hand side and use $\sum_v W_{uv} = D_{uu}$:

$$1405 \quad \frac{1}{2} \sum_{u,v} W_{uv} (z_u^2 + z_v^2 - 2z_u z_v) = \sum_u D_{uu} z_u^2 - \sum_{u,v} W_{uv} z_u z_v = z^\top (D - W) z.$$

1406 □

1407 Lemma 1 is the implementation-specific version with the exact coefficient $1/(2d)$
 1408 and replicate boundary.

1409 **Kernel and connectedness.**

1410 For L_{comb} , every row sums to zero and $L_{\text{comb}} \mathbf{1} = 0$. On a connected graph, the Dirichlet
 1411 form can vanish only when $z_u = z_v$ along every edge, hence $\ker L_{\text{comb}} = \text{span}\{\mathbf{1}\}$. The
 1412 implemented $I - A$ operator has the same property because $A\mathbf{1} = \mathbf{1}$. By contrast, for a
 1413 connected graph with positive degrees, $L_{\text{sym}} D^{1/2} \mathbf{1} = 0$; its row sums need not vanish.
 1414 These are the corresponding zero-frequency modes.

1415 B.1.3 Spectral theorem and graph Fourier transform

1416 **Spectral theorem.**

1417 Every real symmetric L has an orthogonal eigendecomposition

$$1418 \quad L = U \Lambda U^\top, \quad U = [\varphi_1, \dots, \varphi_N], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N),$$

1419 with $U^\top U = U U^\top = I$. If $L \succeq 0$, then $\lambda_i \geq 0$. The graph-Fourier transform (GFT)
 1420 and inverse GFT are

$$1421 \quad \hat{x} = U^\top x, \quad x = U \hat{x}, \quad \hat{x}_i = \langle x, \varphi_i \rangle.$$

1422 The eigenvalue, not the eigenvector index by itself, is the graph frequency.

1423 **Proposition 4** (Parseval identity) *The GFT preserves inner products and energy:*

$$1424 \quad \langle x, y \rangle = \langle \hat{x}, \hat{y} \rangle, \quad \|x\|^2 = \sum_i |\hat{x}_i|^2.$$

Proof $\langle \hat{x}, \hat{y} \rangle = x^\top U U^\top y = x^\top y$ because U is orthogonal. Taking $y = x$ gives the energy identity. $\square$

Repeated eigenvalues and spectral projectors.

If λ has multiplicity greater than one, its individual eigenvectors are not unique. Let

$$P_\lambda = \sum_{i: \lambda_i = \lambda} \varphi_i \varphi_i^\top$$

be the orthogonal projector onto the full eigenspace. The invariant energy at frequency λ is $\|P_\lambda x\|^2$, not the energy of an arbitrarily selected basis vector.

Spectral energy measure.

For $x \neq 0$,

$$\mu_x = \sum_{\lambda \in \text{spec}(L)} \frac{\|P_\lambda x\|^2}{\|x\|^2} \delta_\lambda.$$

$\text{spec}(L)$ denotes the set of eigenvalues of L , and δ_λ is the unit Dirac point mass located at λ . This is a probability measure because Parseval gives total mass one. For any function ψ on the spectrum,

$$\int \psi(\lambda) d\mu_x(\lambda) = \frac{\sum_i \psi(\lambda_i) |\hat{x}_i|^2}{\|x\|^2}.$$

Proposition 5 (Basis and gain invariance) μ_x is unchanged by (i) any orthogonal change of basis within a repeated eigenspace and (ii) a nonzero global rescaling $x \mapsto cx$.

Proof An orthogonal basis change leaves the projector P_λ and its norm $\|P_\lambda x\|$ unchanged. Under $x \mapsto cx$, both the numerator and denominator of every mass are multiplied by c^2 . $\square$

Neither invariance removes a sampling mask or permits comparison on misregistered coordinates.

Meaning of “spectral invariant” in this work.

A spectral invariant is a statistic that depends on x only through μ_x , for example $\int \psi d\mu_x$ or the diffusion-response sequence. It is invariant to basis rotations inside repeated eigenspaces and, after energy normalization, to nonzero global gain. The phrase does not mean invariance to arbitrary subsampling, voxelization, masking, or coordinate changes.

B.1.4 Graph filters, diffusion, and high-pass residuals

Spectral graph filter.

For a scalar response h , functional calculus defines

$$h(L) = U h(\Lambda) U^\top, \quad h(\Lambda) = \text{diag}(h(\lambda_1), \dots, h(\lambda_N)).$$

Thus $\widehat{h(L)x_i} = h(\lambda_i)\hat{x}_i$. A polynomial $h(\lambda) = \sum_{k=0}^K a_k \lambda^k$ yields $h(L)x = \sum_{k=0}^K a_k L^k x$ and can be evaluated without eigendecomposition.

Low-pass and high-pass responses.

On the ordered graph spectrum, a nonnegative response $h(\lambda)$ is called low-pass when it is nonincreasing in λ , and high-pass when it is nondecreasing. These terms describe spectral weighting, not whether an output is noise-free: a high-pass channel contains both true boundaries and high-frequency artifacts.

Proposition 6 (Filtered-energy identity) *For real h and $x \neq 0$,*

$$\frac{\|h(L)x\|^2}{\|x\|^2} = \int h(\lambda)^2 d\mu_x(\lambda).$$

Proof By Parseval, $\|h(L)x\|^2 = \sum_i h(\lambda_i)^2 |\hat{x}_i|^2$; divide by $\|x\|^2$ and use the definition of μ_x . $\square$

Dirichlet energy and graph variation.

The unnormalized Dirichlet energy is

$$x^\top Lx = \sum_i \lambda_i |\hat{x}_i|^2.$$

Its normalized version $x^\top Lx / \|x\|^2 = \int \lambda d\mu_x$ is the first raw spectral moment. It weights rapidly varying modes more heavily and is the graph analogue of squared-gradient energy. It is a global scalar; the implementation computes it separately for each batch item and feature channel.

Diffusion and heat kernel.

The exact heat kernel is e^{-tL} with response $e^{-t\lambda}$. The implementation uses an explicit-Euler diffusion step

$$D_\alpha = I - \alpha L, \quad D_\alpha^b = (I - \alpha L)^b, \quad \rho(\lambda) = 1 - \alpha\lambda, \quad \rho_i = \rho(\lambda_i), \quad \rho_b(\lambda) = \rho(\lambda)^b.$$

For $\text{spec}(L) \subset [0, 2]$ and $0 < \alpha < 1/2$, $0 < 1 - 2\alpha \leq \rho(\lambda) \leq 1$ and ρ is strictly decreasing. It is therefore a monotone low-pass response with no sign reversal. At $\alpha > 1/2$, ρ changes sign and the energy response ρ^2 folds two different frequencies together. At $\alpha = 1/2$, injectivity remains but the inverse map has an unbounded derivative at the top endpoint, so the finite-scale Lipschitz constant used in Theorem 1(ii) is not finite.

High-pass residual.

The b -scale residual is

$$H_b x = (I - D_\alpha^b)x, \quad \widehat{H_b x}_i = (1 - \rho_i^b)\hat{x}_i.$$

It increasingly emphasizes modes for which ρ_i is small. This is why it is sensitive both to true boundaries and to mask/quantization artifacts.

B.1.5 Raw and contracted spectral moments

Raw Laplacian moments.

For raw-moment order $p \geq 0$,

$$m_p^{\text{raw}}(x) = \frac{x^\top L^p x}{\|x\|^2} = \int \lambda^p d\mu_x(\lambda).$$

Because the ratio $(\lambda_{\text{hi}}/\lambda_{\text{lo}})^p$ grows with p , high orders emphasize top-band contributions relative to low-band ones. This does not assert that every unnormalized moment grows.

Contracted moments.

For $b \geq 0$,

$$m_b^{\text{con}}(x) = e_b(x) = \frac{\|D_\alpha^b x\|^2}{\|x\|^2} = \int (1 - \alpha\lambda)^{2b} d\mu_x(\lambda).$$

They are called contracted because $0 < (1 - \alpha\lambda)^{2b} \leq 1$ and the weight decreases with frequency; for every $\lambda > 0$ it also decreases with depth b .

Proposition 7 (Geometric relative band reweighting) *Let π be a probability measure on $[0, 2]$, let $0 < \lambda_0 < \lambda_1 \leq 2$, and assume both the low band $A_0 = [0, \lambda_0]$ and high band $A_1 = [\lambda_1, 2]$ have positive π -mass. Then*

$$\frac{\int_{A_1} \rho(\lambda)^{2b} d\pi / \pi(A_1)}{\int_{A_0} \rho(\lambda)^{2b} d\pi / \pi(A_0)} \leq \left(\frac{\rho(\lambda_1)}{\rho(\lambda_0)} \right)^{2b},$$

whereas

$$\frac{\int_{A_0} \lambda^p d\pi / \pi(A_0)}{\int_{A_1} \lambda^p d\pi / \pi(A_1)} \leq \left(\frac{\lambda_0}{\lambda_1} \right)^p.$$

Thus contracted and raw moments move the high-to-low relative weighting in opposite geometric directions.

Proof On A_1 , $\rho(\lambda)^{2b} \leq \rho(\lambda_1)^{2b}$, while on A_0 it is at least $\rho(\lambda_0)^{2b}$. Divide the corresponding band averages. The second inequality follows identically from $\lambda^p \leq \lambda_0^p$ on A_0 and $\lambda^p \geq \lambda_1^p$ on A_1 . $\square$

Moment determinacy.

Set $u = (1 - \alpha\lambda)^2$ and let $\nu_x = u_\# \mu_x$ be the pushforward measure, defined by $\nu_x(A) = \mu_x(u^{-1}(A))$. Then

$$e_b(x) = \int u^b d\nu_x(u).$$

1515 On a compact interval, equality of all polynomial moments implies equality of prob-
 1516 ability measures: equality first holds for all polynomials; polynomials are uniformly
 1517 dense in continuous functions by Stone–Weierstrass; equality then extends to all con-
 1518 tinuous test functions, which determines the measure. This is the Hausdorff moment
 1519 argument used in Theorem 1(i).

1520 ***Finite moments and Lipschitz statistics.***

1521 A function ψ is K -Lipschitz if $|\psi(\lambda) - \psi(\lambda')| \leq K|\lambda - \lambda'|$. Jackson’s theorem is applied
 1522 after the change of variable $u = (1 - \alpha\lambda)^2$. Define

$$1523 \quad \tilde{\psi}(u) = \psi\left(\frac{1 - \sqrt{u}}{\alpha}\right), \quad u \in [u_{\min}, 1].$$

1524 It supplies a universal constant C_J such that a degree- B polynomial p_B can be chosen
 1525 with

$$1526 \quad \|\tilde{\psi} - p_B\|_{\infty} \leq \frac{C_J \text{Lip}(\tilde{\psi})(1 - u_{\min})}{B + 1}.$$

1527 Because $p_B(u) = \sum_{b=0}^B c_b u^b$ and $\int u^b d\nu_x = e_b(x)$, its monomials are exactly the
 1528 diffusion responses. The triangle inequality then gives Theorem 1(ii). A hard band
 1529 indicator is discontinuous and is not covered directly; a smoothed band selector is
 1530 Lipschitz.

1531 **B.1.6 Power spectra, bands, and diagnostic profiles**

1532 ***Graph spectral energy versus Fourier PSD.***

1533 On an irregular point graph, frequency is indexed by Laplacian eigenvalues and energy
 1534 is $|\hat{x}_i|^2$. On a rectangular lattice, a discrete Fourier transform (DFT) may instead be
 1535 used:

$$1536 \quad \mathcal{F}[x](\omega) = \frac{1}{\sqrt{N}} \sum_{u \in V} x_u e^{-i\langle \omega, r_u \rangle}, \quad \text{PSD}(\omega) = |\mathcal{F}[x](\omega)|^2.$$

1537 Here $r_u \in \mathbb{Z}^d$ is the lattice coordinate of node u and ω is a DFT frequency vector.
 1538 For grid spacing Δ_j on axis j , the Nyquist interval is $[-\pi/\Delta_j, \pi/\Delta_j]$: frequencies out-
 1539 side it cannot be distinguished from an in-band alias after uniform sampling. These
 1540 are related for periodic regular grids but are not identical in general: the implemented
 1541 graph Laplacian uses replicate boundaries, whereas the standard DFT assumes peri-
 1542 odic extension. A radial PSD averages DFT energy over frequency vectors with similar
 1543 $\|\omega\|$.

1544 ***Normalized band energy, HFER, and HF retention.***

1545 Given J nonnegative spectral-bin energies S_j , let $p_j = S_j / \sum_{\ell=1}^J S_{\ell}$. Let f_j be the
 1546 frequency coordinate of bin j , let f_{hf} be a declared high-frequency threshold, and let
 1547 $\epsilon_{\text{num}} > 0$ be a numerical stabilizer. For a band $\mathcal{B} \subseteq \{1, \dots, J\}$,

$$1548 \quad E_{\mathcal{B}} = \sum_{j \in \mathcal{B}} p_j, \quad \text{HFER} = \sum_{j: f_j \geq f_{\text{hf}}} p_j.$$

HFER is a high-frequency energy *fraction*, not a reconstruction accuracy by itself. The recovery ratio used diagnostically is

$$\text{HFER}_{\text{recovery}} = \frac{\text{HFER}_S}{\text{HFER}_T + \epsilon_{\text{num}}}. \quad (1551)$$

Equation (C17) writes this recovery ratio as HFER, which is the quantity reported throughout; the bare fraction above is an intermediate and is never quoted on its own. A value above one can indicate useful extra detail or pseudo-high frequency; flatness and teacher-relative band discrepancies are needed to distinguish them.

For the native point-graph diagnostic, let $R_{\text{hf}}(x)$ be the high-band fraction within the computed graph eigenmodes. Its reported retention is

$$\text{HFret} = \frac{R_{\text{hf}}(x^{\text{sparse}})}{R_{\text{hf}}(x^{\text{dense}}) + \epsilon_{\text{num}}}. \quad (1558)$$

Because sparsification rebuilds the k NN graph, the two R_{hf} values use different Laplacians and bases. HFret is therefore a cross-graph band-level diagnostic, not a same-mode estimate of $m(\lambda)$.

Frequency-band discrepancy and correlation.

For teacher and student normalized energies in band $\mathcal{B}$, a common relative discrepancy is

$$\text{FBD}_{\mathcal{B}} = \frac{|E_{\mathcal{B}}^S - E_{\mathcal{B}}^T|}{E_{\mathcal{B}}^T + \epsilon_{\text{num}}}. \quad (1564)$$

The spectral correlation coefficient (SCC) is the Pearson correlation between two sampled log-PSD curves:

$$\text{SCC} = \frac{\sum_j (a_j - \bar{a})(b_j - \bar{b})}{\sqrt{\sum_j (a_j - \bar{a})^2} \sqrt{\sum_j (b_j - \bar{b})^2}}, \quad a_j = \log(S_j^S + \epsilon_{\text{num}}), \quad b_j = \log(S_j^T + \epsilon_{\text{num}}). \quad (1568)$$

$\bar{a}$ and $\bar{b}$ are sample means over bins. SCC is undefined when either log-PSD curve has zero variance; code should return a declared fallback or mark that case invalid. High SCC can coexist with a large top-band floor because correlation tracks shape rather than absolute band agreement.

Diffusion response distribution and residual profile.

For scales $b = 1, \dots, B$, define

$$q_b(x) = \frac{\|D_{\alpha}^b x\|^2}{\sum_{\ell=1}^B \|D_{\alpha}^{\ell} x\|^2}, \quad q_b^{\text{res}}(x) = \frac{\|(I - D_{\alpha}^b)x\|^2}{\sum_{\ell=1}^B \|(I - D_{\alpha}^{\ell})x\|^2 + \epsilon_{\text{num}}}. \quad (1575)$$

The first is the diffusion response distribution (DRD) used by the low-frequency branch; the second is its high-pass analogue. For $x \in \ker L$, the residual vector is identically zero and is skipped with the stated ϵ_{num} guard. The training loss compares

1579 either normalized profile using first-order total variation,

$$1580 \quad \mathcal{L}_{\text{profile}}^r = \frac{1}{2} \|\hat{q}^{S,r} - \hat{q}^{T,r}\|_1, \quad r \in \{\text{L}, \text{H}\}.$$

1581 This must not be conflated with the evaluation pipeline’s *Spectral Information*
 1582 *Distance*, which is a radial-PSD distance: after aligning normalized radial PSD
 1583 distributions p^S, p^T on the declared relative-frequency/CDF axis, it reports

$$1584 \quad \text{SID}_{\text{PSD}}(p^S, p^T) = \sqrt{\text{JS}(p^S, p^T) / \log 2} \in [0, 1].$$

1585 Thus the training profile ℓ_1 and evaluation SID operate on different summaries and
 1586 use different discrepancies.

1587 **B.1.7 White noise, attenuation, aliasing, and mask mixing**

1588 ***White noise.***

1589 A zero-mean vector $\xi \in \mathbb{R}^N$ is white with variance σ^2 when

$$1590 \quad \mathbb{E}[\xi] = 0, \quad \mathbb{E}[\xi\xi^\top] = \sigma^2 I.$$

1591 White means equal second-order power in every orthonormal direction, not that one
 1592 finite realization has an exactly flat empirical curve.

1593 **Proposition 8** (Whiteness is invariant under an orthogonal basis) *If ξ is white and U is*
 1594 *orthogonal, then $\hat{\xi} = U^\top \xi$ is also white with variance σ^2 .*

1595 *Proof* $\mathbb{E}[\hat{\xi}] = U^\top \mathbb{E}[\xi] = 0$, and $\mathbb{E}[\hat{\xi}\hat{\xi}^\top] = U^\top \mathbb{E}[\xi\xi^\top]U = \sigma^2 U^\top U = \sigma^2 I$. □

1596 Near-white is an empirical approximation: a top-band PSD floor that is approxi-
 1597 mately constant after accounting for the teacher-correlated component.

1598 ***Attenuation and effective transfer.***

1599 The model $\hat{x}_i^S = m(\lambda_i)\hat{x}_i^T + \xi_i$ decomposes the student into a teacher-correlated compo-
 1600 nent and a residual. The constraint $0 \leq m \leq 1$ and monotonicity are assumptions. An
 1601 energy retention ratio approximates m^2 only when residual energy and cross terms are
 1602 small and both signals use the same spectral basis. The denoising theorem ultimately
 1603 requires the basis-invariant eigenspace SNRs snr_k from (C19) to be nonincreasing, not
 1604 merely monotone m .

1605 ***Aliasing.***

1606 Aliasing is the non-injective ambiguity created when a sampling or representation
 1607 operator maps distinct input patterns to the same observation. Spectral leakage or
 1608 mode mixing is the redistribution of one mode into several output modes; it can occur
 1609 even when the operator is injective. On a regular sampled signal, frequencies separated

by a sampling frequency can coincide. On a graph or voxel pipeline, node removal, quantization, interpolation, and masks can mix Laplacian modes even without classical periodic aliasing; non-invertible subsampling then adds genuine ambiguity.

Proposition 9 (A spatial mask mixes graph modes) *Let M_Ω be a nonconstant diagonal mask and $L = U\Lambda U^\top$. In the graph-Fourier basis,*

$$\widehat{M_\Omega}x = U^\top M_\Omega U \hat{x}.$$

Unless M_Ω commutes with L , the matrix $U^\top M_\Omega U$ is not block diagonal across distinct eigenspaces, so one input graph frequency contributes to multiple output frequencies.

Proof Substitute $x = U\hat{x}$. A symmetric matrix commutes with L exactly when it preserves every distinct eigenspace of L , which is equivalent to block diagonality in the eigenspace decomposition. The contrapositive gives cross-frequency mixing whenever commutation fails. $\square$

For a periodic regular lattice, this is the familiar rule that pointwise multiplication in space becomes convolution in frequency. Whether the resulting tail is broad or flat depends on the mask and signal; it is not guaranteed for an arbitrary deterministic mask.

Proposition 10 (Bernoulli masking gives a flat DFT residual in expectation) *Let $x \in \mathbb{R}^N$ be fixed on a periodic lattice, let $B_u \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$, and define $y_u = B_u x_u$. Write $y = px + \eta$ and let F be the unitary DFT, so $|F_{ku}|^2 = 1/N$. Then for every DFT bin k ,*

$$\mathbb{E} |(F\eta)_k|^2 = \frac{p(1-p)}{N} \|x\|^2.$$

Thus the mask-induced residual has an exactly flat DFT power floor in expectation.

Proof $\eta_u = (B_u - p)x_u$ has zero mean, independent coordinates, and $\mathbb{E}[\eta_u \eta_v] = p(1-p)x_u^2 \delta_{uv}$. Therefore

$$\mathbb{E} |(F\eta)_k|^2 = \sum_{u,v} F_{ku} \overline{F_{kv}} \mathbb{E}[\eta_u \eta_v] = p(1-p) \sum_u |F_{ku}|^2 x_u^2 = \frac{p(1-p)}{N} \|x\|^2.$$

$\square$ 1634

For a general graph eigenvector φ_k , the analogous expectation is $p(1-p) \sum_u \varphi_k(u)^2 x_u^2$, which need not be constant in k . Near-flat graph power therefore needs additional energy-uniformity or eigenvector-delocalization assumptions, or direct empirical support.

1639 ***Why point and voxel diagnostics differ.***

1640 Rebuilding a point graph after node removal does not insert a zero/nonzero boundary
 1641 on a fixed lattice. A zero-filled voxel scatter does. In addition, the current point diag-
 1642 nostic retains only the lowest 128 graph eigenmodes, while the voxel DFT diagnostic
 1643 observes the full lattice Nyquist range. The observed difference therefore combines a
 1644 real representation effect with a measurement-bandwidth difference; it is not evidence
 1645 that point clouds are discrete while voxels are continuous.

1646 **B.1.8 Signal-to-noise ratios and monotone reweighting**

1647 ***Eigenspace and channel SNR.***

1648 Let θ_k , P_k , and d_k denote a distinct graph frequency, its projector, and its multiplicity;
 1649 set $m_k = m(\theta_k)$, $\rho_k = 1 - \alpha\theta_k$, and $T_k = \|P_k t\|^2$. Under (B3), the residual has expected
 1650 power $d_k \sigma^2$ in this eigenspace, so

$$1651 \quad \text{snr}_k = \frac{m_k^2 T_k}{d_k \sigma^2}.$$

1652 If a channel applies nonnegative squared response weights a_k with $\sum_k d_k a_k > 0$, its
 1653 aggregate SNR is

$$1654 \quad \text{SNR}(a) = \frac{\sum_k a_k m_k^2 T_k}{\sigma^2 \sum_k d_k a_k} = \sum_k \frac{d_k a_k}{\sum_\ell d_\ell a_\ell} \text{snr}_k.$$

1655 Thus channel SNR is a weighted average of eigenspace-average SNR. The weights W_b
 1656 and V_b in Theorem 1(iii) are obtained by setting $a_k = \rho_k^{2b}$ and $a_k = (1 - \rho_k^b)^2$.

1657 ***Monotone reweighting.***

1658 If snr_k decreases with k , multiplying a probability distribution by another decreas-
 1659 ing sequence and renormalizing shifts mass toward larger SNR. Lemma 3 proves the
 1660 resulting average cannot decrease. This is a deterministic ordering statement; the
 1661 monotone-SNR premise remains an empirical/model assumption.

1662 **B.1.9 Probability-profile divergences and concentration**

1663 ***Jensen–Shannon divergence.***

1664 For probability vectors p, q and $m = (p + q)/2$,

$$1665 \quad \text{JS}(p, q) = \frac{1}{2} \sum_i p_i \log \frac{p_i}{m_i} + \frac{1}{2} \sum_i q_i \log \frac{q_i}{m_i}.$$

1666 It is symmetric, nonnegative, and bounded above by $\log 2$ when natural logarithms are
 1667 used. Nonnegativity follows from nonnegativity of each Kullback–Leibler divergence.
 1668 The code adds ϵ_{num} before logarithms and normalization for numerical stability; this
 1669 is the ϵ_{num} used above.

Entropy.

For a probability profile p ,

$$\text{Ent}(p) = - \sum_i p_i \log p_i.$$

It is zero for a point mass and maximal at the uniform distribution. For $K > 1$ bins, the normalized graph spectral entropy reported as GSE is commonly $\text{GSE}(p) = \text{Ent}(p)/\log K \in [0, 1]$; the binning and set of computed eigenmodes must be reported with it.

Wiener flatness.

For a nonnegative profile $a \in \mathbb{R}_+^K$ with positive arithmetic mean,

$$\text{Flat}(a) = \frac{(\prod_{k=1}^K a_k)^{1/K}}{\frac{1}{K} \sum_{k=1}^K a_k}.$$

It lies in $[0, 1]$ by the arithmetic–geometric mean inequality, equals one exactly for a constant positive profile, and equals zero if any bin vanishes while the arithmetic mean is positive. It approaches zero for increasingly concentrated positive profiles. The implementation uses $a_k + \epsilon_{\text{num}}$, so its flatness is a smoothed finite-profile statistic.

Majorization and Schur concavity.

Let $p^\downarrow$ denote p sorted in decreasing order. A probability vector q majorizes p , written $q \succ p$, if

$$\sum_{i=1}^k q_i^\downarrow \geq \sum_{i=1}^k p_i^\downarrow \quad (k = 1, \dots, K-1), \quad \sum_i q_i = \sum_i p_i = 1.$$

It means q is more concentrated. Shannon entropy and, on the positive probability simplex, the logarithm of Wiener flatness are symmetric concave functions and are therefore Schur-concave. Extending flatness continuously to zero components gives: $q \succ p$ implies $\text{Ent}(q) \leq \text{Ent}(p)$ and $\text{Flat}(q) \leq \text{Flat}(p)$.

Coordinate profile versus graph spectrum.

If a gate is computed pointwise on voxel coordinates, its direct majorization result concerns the coordinate-energy profile. It does not automatically imply majorization of the graph-Fourier energy profile because a spatially varying gate mixes modes (Proposition 9). The implemented residual-profile flatness hinge is therefore a separate, explicit training term.

1698 B.1.10 Taxonomy of claims and compact symbol index

1699 *Proved structural facts.*

1700 The implemented L is symmetric PSD with spectrum in $[0, 2]$; diffusion responses are
 1701 compact-interval moments; the full response sequence identifies spectral energy; the
 1702 finite response bound follows from polynomial approximation; and the ideal coordinate
 1703 gate has the two-case optimum stated in Theorem 1, a shrinkage towards zero for the
 1704 branch that suppresses against zero and the teacher itself for the branch that charges
 1705 only an excess over it.

1706 *Modeling assumptions.*

1707 The effective correlated response satisfies $0 \leq m(\lambda) \leq 1$; its band trend is nonincreas-
 1708 ing; the residual is approximately white; and eigenspace-average SNR is nonincreasing.
 1709 These assumptions can be supported or falsified by data but do not follow from PSD
 1710 of L .

1711 *Empirical diagnostics.*

1712 HF retention, HFER, FBD, SCC, DRD, profile ℓ_1 , PSD-SID, entropy, and flatness
 1713 summarize particular measured representations. Their interpretation depends on the
 1714 graph/grid, mask, boundary condition, frequency range, and normalization. Num-
 1715 bers measured before the common-query correction must be rerun before serving as
 1716 quantitative evidence for the corrected training signal.

1717 *Compact symbol index.*

1718 N Number of ambient grid nodes, not raw LiDAR points.
 1719 d Grid dimension, 2 for TPV/BEV and 3 for voxels.
 1720 L Symmetric PSD grid Laplacian; λ_i, φ_i are its eigenpairs.
 1721 θ_k, P_k, d_k A distinct eigenvalue, its eigenspace projector, and its multiplicity.
 1722 $\hat{x}_i$ Graph-Fourier coefficient $\langle x, \varphi_i \rangle$.
 1723 μ_x Normalized spectral energy measure.
 1724 D_α, ρ Diffusion operator and its one-step response.
 1725 b, B One diffusion depth and the number of matched depths.
 1726 p Nonnegative integer order of a raw Laplacian moment.
 1727 e_b Normalized energy surviving b diffusion steps.
 1728 $m(\lambda_i)$ Assumed effective amplitude retention.
 1729 ξ_i, σ^2 Residual mode and idealized white-noise power.
 1730 snr_k Teacher-correlated power divided by residual power, averaged within eigenspace k .
 1731 W_b, V_b Normalized low-pass and residual mode weights.
 1732 ε Uniform response-matching error for the first B moments.
 1733 ϵ_{num} Numerical stabilizer for ratios, logs, and profiles; distinct from ε .
 1734 $\psi, \text{Lip}(\psi)$ A spectral test function and its Lipschitz constant.
 1735 c^* Coefficients of the degree- B Jackson polynomial approximating the transported test
 1736 function.
 1737 C_J, C_α Universal Jackson constant and its operator-dependent stability prefactor.

r, ν, b, i, c Loss branch (L or H), network side (s or t), diffusion depth, response coordinate, and channel. 1738
 z_r^ν, T_r, c_r^t Branch response, branch suppression operator ($T_L = L, T_H = I$), and teacher saliency $c_r^t = T_r z_r^t$. 1739
 $a_{r,i}, A_r, \tau_r, \lambda_{\text{sup}}^r$ Generic detached teacher gate, its diagonal matrix, gate floor, and complementary suppression weight. 1740
 π Normalized residual-noise spectral measure. 1741
 λ_0, λ_1 Low/high band endpoints used in relative reweighting bounds. 1742
JS, Ent, Flat Profile divergence, entropy, and Wiener flatness. 1743
 M_Ω Coordinate mask or scatter support operator. 1744

B.2 Proofs: The Unified Theory in Full 1748

This appendix is self-contained: it restates the setting with all constants, proves the three supporting lemmas, gives the complete proof of Theorem 1 part by part, proves Proposition 2, and collects the remarks that delimit the scope of the theory. Notation mirrors Sec. 3; equation numbers prefixed here are internal to this appendix. 1749
1750
1751
1752

B.2.1 Setup 1753

Each channel of a feature plane or logit volume is a signal $x \in \mathbb{R}^N$ on a connected graph $G = (V, E)$ ($N = |V| \geq 2$; in practice the d -dimensional grid, $d \in \{2, 3\}$). We assume a symmetric positive semidefinite (PSD) graph Laplacian $L = L^\top \succeq 0$ with $\text{spec}(L) \subset [0, 2]$; Lemma 1 shows the grid operator we use is exactly of this form, so the hypothesis is discharged rather than posited. Write $L = \sum_i \lambda_i \varphi_i \varphi_i^\top$ with orthonormal $\{\varphi_i\}$ and $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_N \leq 2$, and let $\hat{x}_i = \langle x, \varphi_i \rangle$. The object we align is the *spectral energy distribution* of $x \neq 0$, the probability measure 1754
1755
1756
1757
1758
1759
1760

$$\mu_x = \sum_{i=1}^N \frac{\hat{x}_i^2}{\|x\|^2} \delta_{\lambda_i} \quad \text{on } [0, 2]. \quad (\text{B1}) \quad 1761$$

Let $\theta_1 < \dots < \theta_K$ be the K distinct eigenvalues of L , let P_k be the orthogonal projector onto the θ_k eigenspace, and let $d_k = \text{rank}(P_k)$ be its multiplicity. Grouping by P_k makes every sampling statement below invariant to rotations inside repeated eigenspaces. 1762
1763
1764
1765

Fix a diffusion step $\alpha \in (0, \frac{1}{2})$ and set $D_\alpha = I - \alpha L$, acting as $D_\alpha \varphi_i = \rho(\lambda_i) \varphi_i$ with low-pass response $\rho(\lambda) = 1 - \alpha \lambda \in (0, 1]$. The statistics we compute are the normalized *diffusion responses* 1766
1767
1768

$$e_b(x) = \frac{\|D_\alpha^b x\|^2}{\|x\|^2} = \int_0^2 \rho(\lambda)^{2b} d\mu_x(\lambda), \quad b \in \mathbb{N}. \quad (\text{B2}) \quad 1769$$

These are spatial-domain quantities — apply the smoothing operator b times, then measure the remaining energy — that need no eigendecomposition; (B2) records that they nonetheless read out μ_x . 1770
1771
1772

1773 **Effective sampling transfer model (assumed).**

1774 Modewise, the sparse student branch is modeled as

$$1775 \quad \hat{s}_i = m(\lambda_i) \hat{t}_i + \hat{n}_i, \quad \mathbb{E}[\hat{n}_i] = 0, \quad \mathbb{E}[\hat{n}_i \hat{n}_j] = \sigma^2 \delta_{ij}, \quad (\text{B3})$$

1776 where the teacher coefficients are treated as fixed, with nonincreasing attenuation $m : [0, 2] \rightarrow [0, 1]$ (high frequencies damped) and white noise floor $\sigma^2 > 0$ (aliasing). The
 1777 range, monotonicity, and whiteness are modeling assumptions. In the special case $S = m(L)$, the positive-contraction condition $0 \preceq S \preceq I$ proves only the range $0 \leq m(\lambda_i) \leq$
 1778 1; monotonicity still requires the additional low-pass assumption, and whiteness is a
 1779 separate property of the residual. A general sampling/voxelization operator need not
 1780 have this form. The basis-invariant, eigenspace-average signal-to-noise ratio is
 1781
 1782

$$1783 \quad \text{snr}_k = \frac{m(\theta_k)^2 \|P_k t\|^2}{d_k \sigma^2}. \quad (\text{B4})$$

1784 *Assumption 1* (Low-frequency reliability trend) Ordering the distinct frequencies increas-
 1785 ingly, with the low-pass weights $W_b(k) \propto d_k \rho_k^{2b}$ of part (iii) below, for every $b \geq 1$ and every
 1786 j ,

$$1787 \quad G_j^{(b)} = \sum_{k \leq j} W_b(k) (\text{snr}_k - \mathbb{E}_{W_b}[\text{snr}]) \geq 0. \quad (\text{B5})$$

1788 The reliability accumulated from the low-frequency end never falls below its overall aver-
 1789 age: the teacher’s true structure is low-frequency dominated while the residual noise floor is
 1790 isotropic.

1791 Monotone reliability, $\text{snr}_1 \geq \dots \geq \text{snr}_K$, implies (B5) but is strictly stronger, and
 1792 (B5) is in turn the weakest form that does not depend on which reweighting is applied: by
 1793 Lemma 3(iv) it holds at a given baseline weighting if and only if the conclusion of Lemma 3(ii)
 1794 holds at every nonincreasing reweighting of that baseline, and that conclusion is the only way
 1795 the proof of part (iii) uses the hypothesis. A snr that rebounds in its top bands can therefore
 1796 satisfy it. For the two reweightings part (iii) actually applies, the sharp condition is weaker
 1797 still—the sign of the corresponding weighted covariance of ρ^2 , respectively $(\rho^{-b} - 1)^2$, with
 1798 snr—so (B5) remains a sufficient condition there. A coarser banded version follows by group-
 1799 ing adjacent eigenspaces first and assuming the condition for the resulting band averages;
 1800 merely replacing a nonmonotone sequence by an envelope would not prove the claim for the
 1801 original sequence.

1802 B.2.2 Three lemmas

1803 **Lemma 1** (Operator identification) *Let $\mathcal{S}$ be the depthwise grid stencil with center weight*
 1804 *1, weight $-\frac{1}{2d}$ on each axis neighbor, and replicate padding. Then $\mathcal{S} = L := I - A$ with*
 1805 *$A = A^\top$, $A_{uv} = \frac{1}{2d}$ for $\{u, v\} \in E$ and $A_{uu} = \frac{2d - \deg u}{2d}$. Moreover (i) $L = L^\top \succeq 0$ with*
 1806 *$x^\top L x = \frac{1}{2d} \sum_{\{u, v\} \in E} (x_u - x_v)^2$; (ii) $\text{spec}(L) \subset [0, 2]$; (iii) $L\mathbf{1} = 0$ and $\ker L = \text{span}\{\mathbf{1}\}$.*

1807 *Proof* Replicate padding maps each out-of-domain neighbor of a boundary node back to that
 1808 node, so the weight $-\frac{1}{2d}$ attaches only to genuine edges (symmetrically), and the $2d - \deg u$

folded ghost weights accumulate on the diagonal; hence A is symmetric with the stated entries. 1809
Every row of A sums to $\frac{2d - \deg u}{2d} + \deg u \cdot \frac{1}{2d} = 1$, so with $\sum_u \deg u x_u^2 = \sum_{\{u,v\} \in E} (x_u^2 + x_v^2)$, 1810

$$x^\top Lx = \sum_u x_u^2 - \sum_u A_{uu} x_u^2 - \frac{1}{d} \sum_{\{u,v\} \in E} x_u x_v = \frac{1}{2d} \sum_{\{u,v\} \in E} (x_u - x_v)^2 \geq 0, \quad 1811$$

giving (i) and, by connectedness, (iii). Gershgorin discs place $\text{spec}(A) \subset [-1, 1]$, so $\text{spec}(L) = 1 - \text{spec}(A) \subset [0, 2]$, proving (ii). $\square$ 1812

Lemma 1 legitimizes every spectral statement that follows; it also shows $x^\top L^p x = \|L^{p/2} x\|^2 \geq 0$ for integer $p \geq 0$, so these quantities never need clamping. 1814

Lemma 2 (Diffusion responses are moments on a compact interval) *For $\alpha \in (0, \frac{1}{2})$ the map $u(\lambda) = \rho(\lambda)^2 = (1 - \alpha\lambda)^2$ is a strictly decreasing homeomorphism of $[0, 2]$ onto $[u_{\min}, 1]$, $u_{\min} = (1 - 2\alpha)^2 > 0$. With $\nu_x = u_{\#} \mu_x$ the pushforward,* 1816

$$e_b(x) = \int_{u_{\min}}^1 u^b d\nu_x(u), \quad b \geq 0, \quad (\text{B6}) \quad 1819$$

so $\{e_b(x)\}_{b \geq 0}$ is the moment sequence of the probability measure ν_x on $[u_{\min}, 1]$ (with $e_0 = 1$). 1820

Proof ρ is affine with $\rho' = -\alpha < 0$ and, for $\alpha < \frac{1}{2}$, $\rho(\lambda) \in [1 - 2\alpha, 1] \subset (0, 1]$ on $[0, 2]$; 1821
hence $u = \rho^2$ is continuous, strictly decreasing and positive, a homeomorphism onto $[u_{\min}, 1]$. 1822
Substituting $u = \rho(\lambda)^2$ in (B2) gives (B6). $\square$ 1823

Remark 1 (Why $\alpha \leq \frac{1}{2}$ is not cosmetic) If $\alpha \in (\frac{1}{2}, 1)$ then $\|D_\alpha\| \leq 1$ still holds, but ρ changes 1824
sign on $[0, 2]$ and $u = \rho^2$ ceases to be injective: two distinct frequencies collapse to the same 1825
response, and no family of response energies can separate them. A particular finite graph can 1826
still be identifiable if no such collision occurs on its actual spectrum, but the graph-uniform 1827
guarantee in the Main Theorem(i) can fail. We keep $\alpha \in (0, \frac{1}{2})$; the strict inequality also 1828
keeps the stability constant in (ii) finite. 1829

Lemma 3 (Monotone reweighting) *Let p be a probability vector on ordered indices and let $a_1 \geq a_2 \geq \dots \geq a_n \geq 0$, not all zero. Define $q_i = a_i p_i / \sum_j a_j p_j$. Then (i) $\sum_{i \leq k} q_i \geq \sum_{i \leq k} p_i$ for all k ; (ii) $\sum_i q_i f_i \geq \sum_i p_i f_i$ for every nonincreasing sequence f ; (iii) if p is itself nonincreasing then so is q , and q majorizes p ; (iv) if f is arbitrary and $G_j = \sum_{i \leq j} p_i (f_i - \sum_k p_k f_k) \geq 0$ for every j , the conclusion of (ii) holds all the same, and it reverses if instead $0 \leq a_1 \leq a_2 \leq \dots \leq a_n$, not all zero. Conversely, if the conclusion of (ii) holds at every nonincreasing a , then $G_j \geq 0$ for every j .* 1830

Proof (i) The sign of $\sum_{i \leq k} q_i - \sum_{i \leq k} p_i$ equals that of $\sum_{i \leq k} \sum_j (a_i - a_j) p_i p_j = \sum_{i \leq k} \sum_{j > k} (a_i - a_j) p_i p_j \geq 0$, the diagonal block canceling by antisymmetry. (ii) Abel summation, $\sum_i (q_i - p_i) f_i = \sum_{k=1}^{n-1} (f_k - f_{k+1}) \sum_{i \leq k} (q_i - p_i)$, each factor nonnegative by (i). (iii) $q_i \propto a_i p_i$ is a product of nonincreasing nonnegative sequences; then (i) is the definition of majorization. (iv) In every case $\sum_i (q_i - p_i) f_i = \text{Cov}_p(a, f) / \sum_j a_j p_j$ with $\text{Cov}_p(a, f) = \sum_i p_i a_i (f_i - \sum_k p_k f_k)$. For a nonincreasing, $a_i = \sum_{j \geq i} \Delta_j$ with $\Delta_j = a_j - a_{j+1} \geq 0$ and 1837

1843 $a_{n+1} = 0$, so $\text{Cov}_p(a, f) = \sum_j \Delta_j G_j \geq 0$; for a nondecreasing, $a_i = a_1 + \sum_{j < i} \delta_j$ with
 1844 $\delta_j = a_{j+1} - a_j \geq 0$, so $\text{Cov}_p(a, f) = -\sum_j \delta_j G_j \leq 0$, the constant a_1 contributing nothing.
 1845 For the converse take $a_i = \mathbf{1}\{i \leq j\}$, which reduces the conclusion of (ii) to $G_j \geq 0$. $\square$

1846 This single lemma is the engine behind every monotonicity in the Main Theorem: it
 1847 is invoked once for scale monotonicity, once for the channel comparison, and once for
 1848 the gated-optimum profile.

1849 B.2.3 Main theorem

1850 *Theorem 1 (restated, with all constants)* Fix $\alpha \in (0, \frac{1}{2})$ and adopt the setup of Section B.2.1.
 1851 Let x, y be signals with responses e_b from (B2).

1852 **(i) Identifiability.** If $e_b(x) = e_b(y)$ for all $b \geq 1$, then $\mu_x = \mu_y$.

1853 **(ii) Finite-scale stability.** If $|e_b(x) - e_b(y)| \leq \varepsilon$ for $b = 1, \dots, B$, where $B \geq 1$, then for
 1854 every Lipschitz statistic $\psi : [0, 2] \rightarrow \mathbb{R}$,

$$1855 \quad \left| \int \psi d\mu_x - \int \psi d\mu_y \right| \leq \frac{C_\alpha \text{Lip}(\psi)}{B+1} + \|c^*\|_1 \varepsilon, \quad C_\alpha = \frac{C_J(1-u_{\min})}{\alpha(1-2\alpha)}, \quad (\text{B7})$$

1856 where C_J is the universal Jackson constant defined in Part I, and $c^* = (c_0^*, \dots, c_B^*)$ are the
 1857 coefficients of a degree- B Jackson approximant $p_B^*(u) = \sum_{b=0}^B c_b^* u^b$ to $\tilde{\psi}(u) = \psi((1-\sqrt{u})/\alpha)$
 1858 on $[u_{\min}, 1]$.

1859 **(iii) Denoising under sampling.** Adopt (B3) and Assumption 1; write $\rho_k = \rho(\theta_k)$ and
 1860 $m_k = m(\theta_k)$. Set the low-pass and residual spectral weights $W_b(k) \propto d_k \rho_k^{2b}$ and $V_b(k) \propto$
 1861 $d_k(1 - \rho_k^b)^2$, each normalized to sum to one, and write $\mathbb{E}_P[\text{snr}] = \sum_k P(k) \text{snr}_k$. Then

$$1862 \quad \underbrace{\mathbb{E}_{W_{b+1}}[\text{snr}] \geq \mathbb{E}_{W_b}[\text{snr}]}_{\text{smoothing cannot hurt}} \quad \text{and} \quad \underbrace{\mathbb{E}_{V_b}[\text{snr}] \leq \mathbb{E}_{W_b}[\text{snr}]}_{\text{residual SNR is no higher}} \quad \forall b \geq 1. \quad (\text{B8})$$

1863 Moreover, let π be a probability measure on $[0, 2]$ representing the normalized spectral energy
 1864 of the residual noise. For any $0 < \lambda_0 < \lambda_1 \leq 2$ with $\pi([\lambda_1, 2]) > 0$ and $\pi([0, \lambda_0]) > 0$, and
 1865 any integer raw-moment order $p \geq 1$,

$$1866 \quad \int_{[\lambda_1, 2]} \rho^{2b} d\pi \leq \left(\frac{\rho(\lambda_1)}{\rho(\lambda_0)} \right)^{2b} \frac{\int_{[0, \lambda_0]} \rho^{2b} d\pi}{\pi([0, \lambda_0])}, \quad \int_{[0, \lambda_0]} \lambda^p d\pi \leq \left(\frac{\lambda_0}{\lambda_1} \right)^p \frac{\int_{[\lambda_1, 2]} \lambda^p d\pi}{\pi([\lambda_1, 2])} : \quad (\text{B9})$$

1867 the noise mass carried by the contracted moments e_b concentrates on the clean low band geo-
 1868 metrically in b , while that carried by the raw moments $\int \lambda^p d\mu$ concentrates on the corrupted
 1869 top band geometrically in the raw-moment order p . Both $\{e_b\}$ and $\{\int \lambda^p d\mu\}$, indexed over
 1870 all nonnegative integer orders, are complete for μ .

1871 **(iv) Complementary teacher-gated transfer.** For $r \in \{L, H\}$, let $z_r^t, z_r^s \in \mathbb{R}^J$ be the
 1872 detached teacher and trainable student responses on the same J coordinates. Set $T_L = L$ and
 1873 $T_H = I$, define $c_r^t = T_r z_r^t$, and let

$$1874 \quad a_{r,i} = \frac{(c_{r,i}^t)^2}{(c_{r,i}^t)^2 + \tau_r}, \quad A_r = \text{diag}(a_{r,1}, \dots, a_{r,J}), \quad \tau_r, \lambda_{\text{sup}}^r > 0. \quad (\text{B10})$$

1875 Here i indexes response coordinates, $a_{r,i}$ is a detached teacher gate, A_r is its diagonal matrix,
 1876 τ_r is the gate floor, and λ_{sup}^r is the suppression weight. Consider

$$1877 \quad \mathcal{J}_r(z_r^s) = \|A_r^{1/2}(z_r^s - z_r^t)\|_2^2 + \lambda_{\text{sup}}^r \|(I - A_r)^{1/2} T_r z_r^s\|_2^2. \quad (\text{B11})$$

1878 Every minimizer $z_r^{s,*}$ obeys

$$1879 \quad \|(I - A_r)^{1/2} T_r z_r^{s,*}\|_2^2 \leq \|(I - A_r)^{1/2} T_r z_r^t\|_2^2. \quad (\text{B12})$$

If $H_r = A_r + \lambda_{\text{sup}}^r T_r^\top (I - A_r) T_r$ is positive definite, the minimizer is unique and satisfies $z_r^{s,*} = H_r^{-1} A_r z_r^t$. For HF, $T_H = I$ is separable, and the complementary term charges only an excess over the teacher, so each coordinate minimizes

$$f(z) = a(z - z^t)^2 + \lambda_{\text{sup}}^H (1 - a)[z^2 - (z^t)^2]_+. \quad (\text{B13})$$

On $|z| \leq |z^t|$ the hinge vanishes, so $f(z) = a(z - z^t)^2$ and $f(z^t) = 0$. On $|z| > |z^t|$ the hinge is strictly positive and the quadratic is nonnegative, so $f(z) > 0$ whenever $\lambda_{\text{sup}}^H > 0$. Hence $z = z^t$ minimizes f for every $a \in [0, 1]$, and it is the unique minimizer when $a > 0$. At $a = 0$ the coordinate is only forbidden to exceed the teacher and every $|z| \leq |z^t|$ is optimal; in the implementation the gate $w = e/(e + \tau)$ with $\tau > 0$ vanishes only where the teacher residual does. The gate therefore affects which deviation is charged, and whether the minimizer is unique, but not that the teacher attains it. Consequently, at that minimizer and with $p_i = (z_{H,i}^t)^2 / \sum_\ell (z_{H,\ell}^t)^2$ and $q_i = ((z_{H,i}^s)^*)^2 / \sum_\ell ((z_{H,\ell}^s)^*)^2$ one has $q = p$, and therefore

$$\text{Flat}(q) = \text{Flat}(p) \quad \text{and} \quad \text{Ent}(q) = \text{Ent}(p). \quad (\text{B14})$$

Thus the LF specialization controls complementary curvature energy and retains a shrinkage optimum in its suppressed quantity, whereas for HF the teacher is the optimum itself and the coordinate-profile conclusions hold with equality. A spatially varying gate need not commute with L , so graph-spectral flatness is handled separately by the implemented residual-profile hinge.

Proof (i). By Lemma 2 the e_b are the moments of ν_x on the compact interval $[u_{\min}, 1]$. Equal moments for all $b \geq 1$ (and $e_0 = 1$ automatically) force, by determinacy of the Hausdorff moment problem — polynomials are uniformly dense in $C([u_{\min}, 1])$ (Stone–Weierstrass) — the equality $\nu_x = \nu_y$; since u is injective, $\mu_x = u_{\#}^{-1} \nu_x = u_{\#}^{-1} \nu_y = \mu_y$.

(ii). Transport ψ to $\tilde{\psi}(u) = \psi((1 - \sqrt{u})/\alpha)$ on $[u_{\min}, 1]$; since $|\frac{d}{du}(1 - \sqrt{u})/\alpha| \leq \frac{1}{2\alpha(1-2\alpha)}$, $\text{Lip}(\tilde{\psi}) \leq \text{Lip}(\psi)/(2\alpha(1-2\alpha))$. For any polynomial $p = \sum_{b=0}^B c_b u^b$,

$$\left| \int \psi d\mu_x - \int \psi d\mu_y \right| = \left| \int \tilde{\psi} d\nu_x - \int \tilde{\psi} d\nu_y \right| \leq 2\|\tilde{\psi} - p\|_\infty + \sum_{b=1}^B |c_b| |e_b(x) - e_b(y)|,$$

the $b = 0$ term canceling as both measures are probability measures. Taking $p = p_B^*$ and applying the defining Jackson bound, $\|\tilde{\psi} - p_B^*\|_\infty \leq C_J \text{Lip}(\tilde{\psi})(1 - u_{\min})/(B + 1)$, gives (B7).

(iii). Under (B3), $\mathbb{E}\|D_\alpha^b s\|^2 = \sum_k \rho_k^{2b} m_k^2 \|P_k t\|^2 + \sigma^2 \sum_k d_k \rho_k^{2b}$, so $\mathbb{E}_{W_b}[\text{snr}]$ is exactly the aligned SNR of the low-pass channel, and likewise $\mathbb{E}_{V_b}[\text{snr}]$ for the residual channel. Both inequalities read off Lemma 3(iv) with $f = \text{snr}$ and baseline W_b , whose hypothesis $G_j^{(b)} \geq 0$ is Assumption 1 at that b . First: $W_{b+1} \propto \rho^2 W_b$ with ρ^2 nonincreasing in k , so the lemma gives $\mathbb{E}_{W_{b+1}}[\text{snr}] \geq \mathbb{E}_{W_b}[\text{snr}]$. Second: $V_b \propto a W_b$ with $a_k = (\rho_k^{-b} - 1)^2$ nondecreasing in k , so the reversed half of the lemma gives $\mathbb{E}_{V_b}[\text{snr}] \leq \mathbb{E}_{W_b}[\text{snr}]$. One cumulative condition therefore yields both, the two reweightings being monotone in opposite directions. The bounds (B9) follow from monotonicity of ρ and of $\lambda \mapsto \lambda^p$: $\int_{[\lambda_1, 2]} \rho^{2b} d\pi \leq \rho(\lambda_1)^{2b}$, $\int_{[0, \lambda_0]} \rho^{2b} d\pi \geq \rho(\lambda_0)^{2b} \pi([0, \lambda_0])$, symmetrically for λ^p . Completeness of $\{\int \lambda^p d\mu\}$ is the Hausdorff argument on $[0, 2]$; of $\{e_b\}$, it is part (i).

(iv). Equation (B11) is a sum of squared linear functions and is therefore convex. A minimizer exists in finite dimensions. Comparing any minimizer $z_r^{s,*}$ with the feasible choice $z_r^s = z_r^t$ gives

$$\mathcal{J}_r(z_r^{s,*}) \leq \mathcal{J}_r(z_r^t) = \lambda_{\text{sup}}^r \|(I - A_r)^{1/2} T_r z_r^t\|_2^2.$$

1921 Dropping the nonnegative match term from the left and dividing by $\lambda_{\text{sup}}^r > 0$ proves (B12).
 1922 Differentiation yields the normal equation

$$1923 [A_r + \lambda_{\text{sup}}^r T_r^\top (I - A_r) T_r] z_r^{s,*} = A_r z_r^t.$$

1924 If the bracketed Hessian H_r is positive definite, it is invertible and gives the stated unique
 1925 solution.

1926 For HF, $T_H = I$ makes the normal equation coordinatewise: $a_{H,i}(z_{H,i}^s - z_{H,i}^t) + \lambda_{\text{sup}}^H(1 -$
 1927 $a_{H,i})z_{H,i}^s = 0$, which gives (B13). The map $a \mapsto a/[a + \lambda_{\text{sup}}^H(1 - a)]$ is increasing on $[0, 1]$, and
 1928 $a_{H,i}$ increases with $|z_{H,i}^t|$, so β_i is nondecreasing in teacher amplitude. Sort p_i in descending
 1929 order; then $q_i \propto \beta_i^2 p_i$ reweights it by a nonincreasing sequence. Lemma 3(iii) gives $q \succ p$.
 1930 Both Shannon entropy and Wiener flatness $\text{Flat}(q) = J(\prod_i q_i)^{1/J}$ are Schur-concave on the
 1931 probability simplex (with continuous extension at zeros), so they cannot increase under this
 1932 majorization, proving (B14). $\square$

1933 B.2.4 Design rules

1934 **Corollary 1** (The four rules motivated by the theorem) *the Main Theorem motivates*
 1935 *the following auditable design choices; the indicated theorem part supplies the mathematical*
 1936 *rationale.*

1937 **D1** (iii, first inequality) *Compare the two branches only after at least one diffusion step,*
 1938 *and optionally upweight later scales while checking oversmoothing empirically; (iii,*
 1939 *second inequality) do not align pointwise high-pass residuals uniformly or without*
 1940 *a reliability gate, since they have no higher SNR than the corresponding low-pass*
 1941 *channel under the model.*

1942 **D2** (iii, (B9)) *Use the contracted moments $e_b = \|D_\alpha^b x\|^2 / \|x\|^2$. The code builds a multi-*
 1943 *channel relative surrogate from the same contracted energies by normalizing the*
 1944 *vector of channel energies; this is related but not identical to per-channel e_b . Use*
 1945 *these quantities rather than normalized raw Laplacian moments $x^\top L^p x / \|x\|^2$.*

1946 **D3** (iv) *Apply a detached teacher gate in both branches: match the LF diffusion response*
 1947 *or HF residual where the teacher supports structure, and use the complementary*
 1948 *gate to suppress student LF curvature or HF amplitude where the teacher is flat. The*
 1949 *auxiliary moment, profile, and HF-flatness terms remain global spectral summaries*
 1950 *rather than pointwise gated quantities.*

1951 **D4** (i–ii) *Preserve both axes of the finite scale–channel response-energy matrix: con-*
 1952 *tracted moments compare channel allocation at a fixed diffusion depth, while the*
 1953 *response profile compares scale allocation after channels are aggregated. The nor-*
 1954 *malized local response match supplies the spatial response anchor, and HF adds*
 1955 *only a one-sided scale-profile flatness penalty. These finite normalized quantities*
 1956 *are implementation surrogates for a prefix of $\{e_b\}$; the formal Lipschitz bound in*
 1957 *part (ii) applies whenever their induced response errors meet its stated premise, and*
 1958 *is not claimed from the profile alone.*

1959 B.2.5 Remarks

1960 **Scope of each part.**

1961 Parts (i)–(ii) require nonzero signals on one shared grid with the operator of Sec. 3.1;
 1962 part (iii) is a statement about the effective model (C18) under Assumption 2; part

(iv) analyzes an ideal unnormalized quadratic objective for both gates. Its HF specialization is separable and yields the coordinate-profile entropy/flatness result; its LF specialization is spatially coupled through L and yields the complementary curvature-energy bound. The blockwise and gate-normalized implementation preserves these roles but is not claimed to share the ideal objective’s exact minimizer; its numerical behaviour is verified empirically (Sec. 2.5).

Remark 2 (Contracted moments are the estimator we adopt) Part (iii) shows both moment systems are complete for μ_x , so the choice is one of robustness, not information. Because the top-to-low noise weighting ratio of the normalized raw moments $x^\top L^p x / \|x\|^2$ grows geometrically with order p , while the same ratio for $\|D_\alpha^b x\|^2 / \|x\|^2$ decays geometrically, the method uses the contracted moments throughout; raw Laplacian powers are not used. (At $\alpha = \frac{1}{4}$, $\lambda_0 = 1$, $\lambda_1 = 1.8$, the top-band reweighting of an order-8 raw moment is $\sim 10^2$, versus $(0.55/0.75)^{2b}$ damping for the contracted one.)

Remark 3 (The linear-observation viewpoint, subsumed) Stacking M responses gives $\mathbf{E} = \tilde{H} \tilde{\mathbf{s}}$ with $\tilde{H}_{m,k} = \rho(\lambda_k)^{2m}$ over the K distinct eigenvalues. When $M \geq K$ this Vandermonde-type $\tilde{H}$ has rank K and the binned spectral energies are *exactly* recovered — the finite, full-rank shadow of the Main Theorem(i). But $\tilde{H}$ has nodes in $(0, 1)$ and is exponentially ill-conditioned, so exact recovery is unusable at the operative regime $M \ll K$; parts (i)–(ii) instead control a Lipschitz spectral statistic ψ to $O(\text{Lip}(\psi)/B) + O(\|c^*\|_1 \varepsilon)$ without explicitly inverting $\tilde{H}$. This avoids a Vandermonde solve but does not erase conditioning: growth of $\|c^*\|_1$ records possible noise amplification in the finite moment coordinates.

Remark 4 (Scope of invariance) Norm normalization gives exact invariance only to a per-sample positive gain $x \mapsto cx$ (network temperature), and $\{e_b\}$ is invariant to orthogonal maps commuting with L . Neither removes the sampling action (B3); robustness to sampling comes from estimator design — scale scheduling and contracted moments (iii), the gate (iv) — not from a symmetry argument. “Sampling invariance” should be read in this precise sense.

Remark 5 (Gates on continuous responses; low-contrast regions) The generic variables in (B11) are continuous network responses, so gradients are well defined. In HF they are pre-softmax voxel logits; in LF they are TPV feature responses. On a large smooth region, the teacher saliency $c_r^t = T_r z_r^t$ is small and the generic gate a_r closes: matching is downweighted and complementary suppression removes student-only HF amplitude ($T_H = I$) or LF curvature ($T_L = L$). At curbs, lane markings, and object boundaries the teacher saliency grows and the gate re-opens. Lowering τ_r opens branch r more broadly; increasing it makes the gate more selective. The threshold is a fraction of the mean saliency over the whole grid rather than over the teacher’s returns alone, so the separation the gate achieves in practice is chiefly between those returns and the empty cells around them, and only secondarily between detail and smooth interior within them.

B.2.6 Proof of Proposition 2

Proof Fix per-mode gains $c_i > 0$ and set $m_i' = m_i/c_i$, $\hat{t}_i' = c_i \hat{t}_i$. Then $m_i' \hat{t}_i' = m_i \hat{t}_i$ for every i , so for each view $\hat{s}_i^{(v)} = m_i \hat{t}_i + \hat{n}_i^{(v)} = m_i' \hat{t}_i' + \hat{n}_i^{(v)}$: the two parameterizations induce

the identical joint law of $(\hat{s}^{(1)}, \dots, \hat{s}^{(V)})$ for any number of views. Any objective that is a functional of the sparse observations alone therefore takes the same values, and has the same optima, under both parameterizations; no such objective can determine $|\hat{t}_i|$ beyond the product $m_i|\hat{t}_i|$. On a mode with $m_i = 0$ the observations equal pure noise and are independent of $\hat{t}_i$, so no information about $\hat{t}_i$ is available at all; by continuity, modes with small m_i carry vanishing information. A teacher branch supplies $\hat{t}_i$ itself, which selects a single member of each equivalence class and removes the ambiguity. $\square$

Appendix C The Spectral Measurement Protocol

This appendix states the protocol the main text relies on: how a spectrum is read when it cannot be formed, which operator and which scales are used, what each statistic measures, and how a branch’s logits are placed on the voxel lattice. The main text cites it as the basis of the design.

C.1 Reading a spectrum without forming it

The degradation originates on the point cloud, but the network learns and is supervised on the voxel grid. A protocol that measured only one of the two would therefore describe only half of the effect, and Sec. D.2 showed that the same undersampling looks opposite in the two. The protocol accordingly has two suites. The *native suite* builds a k NN graph over the points of a scan and compares dense and sparse observations of the same scene. The *voxel suite* applies the grid Laplacian of Sec. 3.1 and its diffusion pyramid to the class-logit volumes of the two branches. Graph construction, band boundaries and aggregation rules are specified in Appendix C.2.

Placing a branch’s logits onto the voxel lattice admits two constructions, and they measure different things. Let F denote a branch’s latent logit field and M_Ω the operator that keeps values on a coordinate set Ω and writes zeros elsewhere. The *masked field* $M_\Omega F$ uses each branch’s own hit set, Ω_S for sparse and Ω_T for dense. The two are not nested. The sparse scan is re-sampled rather than decimated, and the dense reference is itself truncated to the point budget before it is rasterized, so between 8% and 19% of Ω_S has no counterpart in Ω_T , about two thirds of that because the beams land where the teacher has no return and the remainder because of the truncation. It exposes the rasterized sampling pattern itself, which is exactly what a diagnosis of the raw degradation needs. It cannot be used for distillation, however, because zero-filled cells carry no gradient and the difference between the two hit sets contaminates any teacher–student comparison. We therefore also define the *common-support field*, which evaluates both branches on the teacher’s coordinates,

$$X^S = M_{\Omega_T} F^S, \quad X^T = M_{\Omega_T} F^T, \quad (\text{C15})$$

The two spectra then live on identical coordinates and every compared cell has a gradient path to the student. Diagnosis of the raw degradation accordingly uses the masked field, while distillation and post-training diagnosis use the common-support field (C15).

On the point cloud, let x_t, x_s be paired dense/sparse signals of one scene (global geometric signals, and per-class indicator signals on each branch’s own k NN graph),

and let μ_x be the spectral energy measure of Sec. 3.1. The suite reports four quantities. The *band-energy profile* partitions the full spectrum $[0, 2]$ into 11 bands β and estimates each band’s share $E_\beta(x) = \mu_x(\beta)$ by a Chebyshev–Jackson expansion of order 1024, giving a spectral resolution of about 0.006. As a result the entire axis is covered, and $\sum_\beta E_\beta = 1$ serves as a built-in capture check. The *band gap* $G_\beta = E_\beta(x_t) - E_\beta(x_s)$ localizes where energy is missing or spurious. Two scalars then summarize the profile,

$$\begin{aligned} \text{HFR} &= \frac{\sum_{\lambda \geq 0.6} E_\beta(x_s)}{\sum_{\lambda \geq 0.6} E_\beta(x_t)}, & \lambda_{85}(x) &= \inf \left\{ \lambda : \int_0^\lambda d\mu_x \geq 0.85 \right\}, \\ \text{DRD} &= \text{JS}(q(x_s) \parallel q(x_t)), \end{aligned} \quad (\text{C16})$$

the *true high-frequency retention* (share of energy above $\lambda = 0.6$, sparse over dense), the *spectral roll-off* (where 85% of the energy is accumulated), and the diffusion-profile distance of the multi-scale profile q from (5). A second and independent estimator of the same retention is the diffusion residual $\text{HFret}(b_0) = [E_{\text{hi}}^{(b_0)}/E_{\text{tot}}]_s / [E_{\text{hi}}^{(b_0)}/E_{\text{tot}}]_t$ with $E_{\text{hi}}^{(b_0)}(x) = \|x - D_\alpha^{b_0} x\|^2$ is computed for $b_0 \in \{2, 8, 16, 32\}$ entirely through sparse diffusion, never touching an eigensolver. Three conventions keep these numbers comparable across datasets. The dense reference is subsampled to the same 5×10^5 -point budget with the same uniform sampler the teacher consumes in training, so that the protocol measures the point clouds the models actually see. Per-class statistics are reported as medians with interquartile ranges together with the fraction of frames above parity, which is robust to the heavy right tail that ratio statistics develop. Every conclusion is finally cross-checked between the two estimator families, the Chebyshev spectral one and the diffusion residual one, whose Spearman correlation, pooled over frame–class pairs, is 0.94 to 0.99 across the reference scales b_0 .

On the voxel grid, with E_β the energy in band $\beta \in \{\text{low}, \text{mid}, \text{high}\}$ and q the profile as above,

$$\begin{aligned} \text{HFER} &= \frac{[E_{\text{hi}}/E_{\text{tot}}]_s}{[E_{\text{hi}}/E_{\text{tot}}]_t}, & \text{FBD}_\beta &= \frac{|E_\beta(x_s) - E_\beta(x_t)|}{E_\beta(x_t) + \epsilon_{\text{num}}}, \\ \text{SCC} &= \text{corr}(\log q(x_s), \log q(x_t)), & \text{Flat}(q) &= \frac{\text{GM}(q)}{\text{AM}(q)}. \end{aligned} \quad (\text{C17})$$

The high-frequency energy ratio HFER compares the student’s share of high-frequency energy with the teacher’s, so that 1 is ideal and a value above 1 means the student carries high-frequency energy the teacher does not have, which is excess rather than detail. Throughout, both this ratio and the band discrepancies below are formed from energies summed over semantic channels, as (C17) is written; averaging the per-channel ratios instead puts equal weight on the smallest classes and is reported as a robustness check, since for HFER the two conventions can fall on opposite sides of 1. The frequency-band discrepancy FBD_β is the per-band energy difference relative to the teacher’s energy in that band and localizes where the two spectra disagree, a value of 1 meaning that the student’s band energy differs from the teacher’s by 100%. The spectral curve correlation SCC is the Pearson correlation of the log-profiles, so

that 1 means the spectral shapes agree. Both profiles decay monotonically, however, so this correlation saturates close to one; we report it for completeness and let the band discrepancies carry the low-frequency reading. The spectral flatness Flat is the ratio of geometric to arithmetic mean of the profile, approaching 1 for a white, noise-like spectrum and small for a structured one; the flatness gap $F_s - F_t$ serves as the single-number health check used throughout training and diagnosis (Secs. 3.4, 2.5). A symmetric spectral information divergence over the profiles is reported alongside for completeness (Appendix C.2).

Both symptoms are summarized by one effective per-frequency model on the common grid. Writing $\hat{t}_i, \hat{s}_i$ for the teacher and student coefficients *in the same eigenmode* φ_i ,

$$\hat{s}_i = m(\lambda_i) \hat{t}_i + \hat{n}_i, \quad \mathbb{E}[\hat{n}_i] = 0, \quad \mathbb{E}[\hat{n}_i \hat{n}_j] = \sigma^2 \delta_{ij}, \quad \hat{n} \perp \hat{t}, \quad (\text{C18})$$

where δ_{ij} is the Kronecker delta, $m : [0, 2] \rightarrow [0, 1]$ is the amplitude retention of the teacher-correlated component, and $\hat{n}$ collects aliasing, mode mixing, and student-specific error with idealized white power $\sigma^2 > 0$. The range $[0, 1]$ and the monotone decrease of m are modelling assumptions rather than derived facts, since real beam subsampling and voxelization do not commute with L . They are calibrated by the measurements above, with the native suite fixing the low-pass trend of m and the masked-field excess fixing the floor σ^2 . The quantity the theory actually needs is basis-invariant. Let $\theta_1 < \dots < \theta_K$ be the distinct eigenvalues of L , P_k the orthogonal projector onto the θ_k -eigenspace and d_k its multiplicity, and define the eigenspace-averaged signal-to-noise ratio

$$\text{snr}_k = \frac{m(\theta_k)^2 \|P_k t\|^2}{d_k \sigma^2}, \quad (\text{C19})$$

the reliability of frequency band θ_k in the student's copy.

Assumption 2 (Low-frequency reliability trend) With the low-pass weights W_b of (8), for every $b \geq 1$ and every j ,

$$\sum_{k \leq j} W_b(k) (\text{snr}_k - \mathbb{E}_{W_b}[\text{snr}]) \geq 0 : \quad (\text{C20})$$

accumulated from the low-frequency end, reliability never falls below its own overall average. The monotone reading $\text{snr}_1 \geq \text{snr}_2 \geq \dots \geq \text{snr}_K$ implies (C20) and is the easier one to picture, but only the cumulative form enters the proof, so a band that locally breaks the order is admissible. A decreasing m alone implies neither, since the teacher's band energies $\|P_k t\|^2$ also enter (C19); what either form encodes is that the teacher's structure is low-frequency dominated. The diagnosis of Sec. 2.5 reports the monotone reading, which holds exactly on the planes the low-frequency branch acts on, and checks (C20) on the common-support field, where the top bands rebound.

C.2 Spectral Measurement Protocol: Details

This appendix specifies the operational details behind Secs. C.1 and 2.5; the suite is released with our code.

Native suite (point clouds).

For each scene, the dense and sparse observations are paired by the data factory; per semantic class, the class’s points carry a local geometric feature signal. The dense branch is the densified cloud the teacher is trained on, subsampled without replacement to the 5×10^5 -point budget by the same uniform sampler; the sparse branch is the sensor’s own scan, so the comparison is between a real sparse observation and a dense reconstruction of the same scene rather than between two simulated samplings. A k NN graph is built over the points ($k = 16$, reduced to $N - 1$ on frames with fewer points, symmetrized), with a normalized adjacency so that the resulting Laplacian satisfies $L = L^\top \succeq 0$, $\text{spec}(L) \subset [0, 2]$ —the same standing hypotheses as the grid operator, so every statement of Appendix B.2 applies verbatim. The diffusion pyramid, the profiles q , the high-pass energies at the reference scales $b_0 \in \{2, 8, 16, 32\}$ and the roll-off scale are computed exactly as in Eq. (5). The spectrum is partitioned into eleven bands with edges $\{-0.05, 0.01, 0.03, 0.08, 0.15, 0.25, 0.40, 0.60, 0.85, 1.15, 1.50, 2.02\}$; the true high-frequency segment is $\lambda \geq 0.60$ and the roll-off is read at 85% of the cumulative energy. Per-class values are aggregated over scenes by the median, reported with the interquartile range and the fraction of frames above parity. Degenerate cases (fewer points than $k+1$; zero-energy signals) are skipped with an ϵ_{num} guard.

Voxel suite (class logits).

Signals are the pre-softmax class-logit channels on the voxel grid, analysed with the grid Laplacian and diffusion pyramid of Sec. 3.1. E_{hi} , band energies, log-profiles, SCC (Pearson over $\log q$), flatness, and the symmetric spectral information divergence $\text{SID} = \text{KL}(q_s \| q_t) + \text{KL}(q_t \| q_s)$ are computed per channel and averaged. The three bands used for FBD are defined by quantiles of the reference (teacher) spectrum’s cumulative energy at $(0, 0.5, 0.9, 1)$: the low band carries the teacher’s first 50% of energy, the mid band the next 40%, and the high band its last 10%. Defining the bands this way keeps each of them non-empty for the reference however steeply its spectrum decays, which a fixed split of the frequency axis does not guarantee. Where a single aggregate is quoted, the three band values are combined with weights $(0.2, 0.3, 0.5)$. HFER is not computed on this split: it uses a fixed threshold on the absolute frequency axis, so HFER and FBD_{high} report the same qualitative effect under two different definitions of “high”. Regional high-frequency recovery is reported as the share of each model’s own local energy that falls above the band, which is the form fixed when the regional statistic was specified and is free of the models’ differing logit scales by construction. A variant that instead divides each model’s local energies by its own mean on the teacher’s support was computed after those results were seen; it preserves the ordering and gives nearly the same multiples, the one difference being that boundaries and road surfaces then gain equally rather than boundaries gaining somewhat more. That fixed threshold is $\lambda \geq 0.60$ on both suites, and $\epsilon_{\text{num}} = 10^{-8}$ throughout. The quantile construction keeps $E_\beta(x_t) \geq 0.1$ in every band, so the stabilizer never affects a reported value.

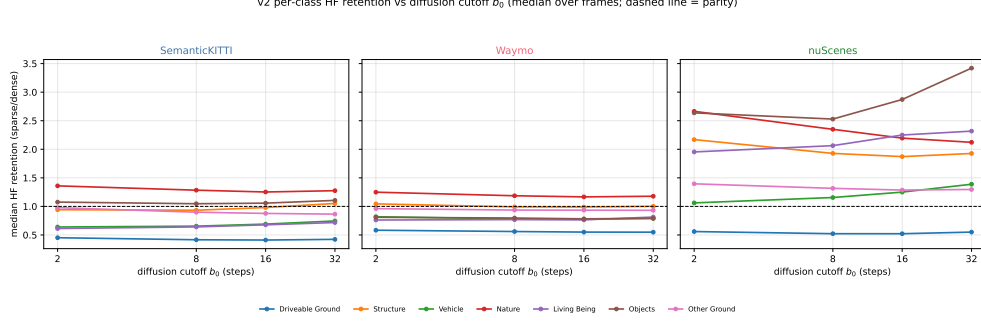


Fig. C1: Per-class retention vs. diffusion cutoff b_0 (medians). Conclusions are stable across b_0 : no (dataset, class) median crosses parity between $b_0 = 2$ and 32. The direction of the residual trend also localizes the pathology: ratios that rise with b_0 indicate spurious energy in the lower high-pass range, ratios that fall indicate energy at the very top.

Outputs.

The native suite emits per-class records (band-energy profile and gap, HFR with its two independent estimators, λ_{85} roll-off, DRD); the voxel suite emits per-model tables (HFER, $\text{FBD}_{\text{low/mid/high}}$, SCC, Flat, SID). Two robustness views are archived with the release: the per-class retention as a function of the diffusion cutoff $b_0 \in \{2, 8, 16, 32\}$ (Fig. C1; medians move by 6–16%, no class crosses parity), and the complete 11-band energy profiles per signal and dataset (Fig. C2), whose first two bands are merged as “quasi-DC” in the main text because they lie inside the Jackson kernel-width soft edge.

The anchors quoted in the paper come from these outputs.

Appendix D Spectral Analysis of the Sampling Gap and of Its Repair

This appendix holds the two readings the main text cites as the basis of the design and as the direct test of its claim: what sampling does to the signal, measured across three datasets, and what distillation does to that damage, measured band by band against the teacher.

One bookkeeping note, so that two parts of the paper can be read against each other. The band-wise analyses of Figs. 3 and 4 were computed on the epoch-21 checkpoint of the undistilled student, whereas the scalar readings of Sec. 2.5 use the epoch-24 checkpoint, which is the one the tables report. The two are the same training run. Recomputing the scalar readings on the earlier checkpoint moves every channel-merged quantity by less than 0.01 and changes neither the direction of any effect nor the fraction of frames carrying it, so the figures and the readings describe the same student; we state the difference rather than silently averaging over it.

v2 KPM band profiles: dense vs sparse mean normalized band energy (11 bands, N=200 frames/dataset)
dotted line: first two bands lie inside the Jackson kernel-width soft edge < 0.03

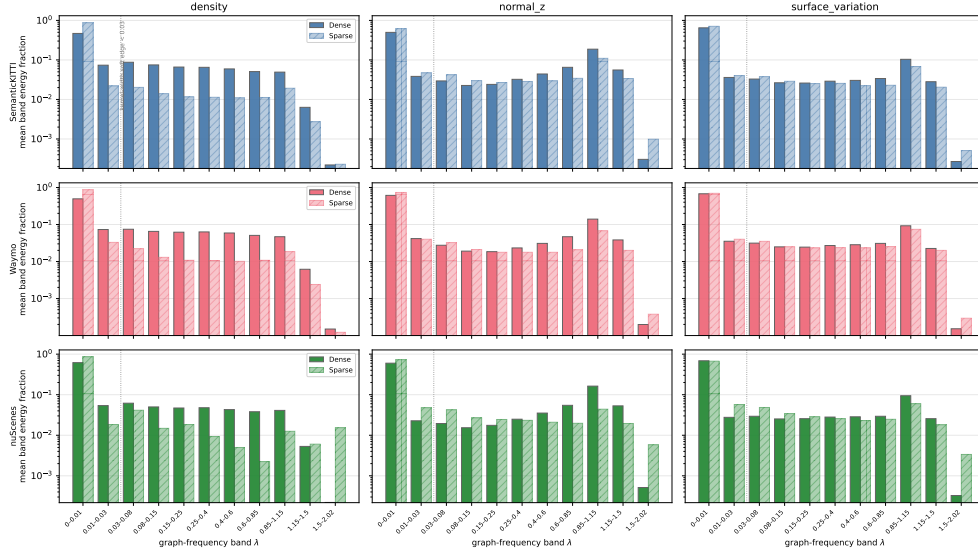


Fig. C2: Complete 11-band energy profiles (log scale) per signal and dataset. The sparse branch (hatched) collapses onto the quasi-DC bands on all datasets and all signals; the dense branch spreads across the full axis, with the normal_z peak at $\lambda \in [0.85, 1.15]$.

D.1 What Energy Alone Says, and What It Hides

The main text reads alignment band by band and reports how much of each band the teacher accounts for. Three further readings sit behind those statements, and Fig. D3 collects them, the feature planes on the left and the voxel logits on the right.

Carrying more is not carrying the right thing.

The first row plots, for each scene, the high-frequency energy a student carries against how well that energy aligns with the teacher's. On the feature planes the undistilled student carries almost as much high-frequency energy as the teacher, yet across scenes the amount it carries and the alignment it reaches are unrelated, at a rank correlation of 0.07; for the distilled student the two move together at 0.61. On the voxel logits the undistilled student's excess is *anti*-correlated with alignment, at -0.59 : the more high frequency it carries there, the less of it belongs to the scene. Amount and alignment come apart in one student and move together in the other, which is what a spectrum corrected against a reference looks like rather than a spectrum merely made larger.

How much is added where the teacher is flat.

The second row counts, on the cells where the teacher's own field is smooth, how far a student's high-frequency energy rises above the teacher's, and counts only the

2199 excess. The unit is the teacher’s own run-to-run fluctuation: re-running the teacher
 2200 on a re-truncated input leaves an excess of its own, and that is set to one. On the
 2201 feature planes the undistilled student sits at 3.04 and the distilled one at 0.19; on the
 2202 voxel logits the two are 10.52 and 2.05. The distilled student therefore falls below the
 2203 teacher’s own fluctuation on the planes, which is as much as this measurement can
 2204 say; it does not say that the student is cleaner than the teacher.

2205 *Where the surviving high frequency sits.*

2206 The third row divides the high-frequency share of boundary cells by that of class-
 2207 interior cells. It is a ratio, and it describes how the high frequency that survives is
 2208 distributed between the two regions rather than how much of it there is. On the feature
 2209 planes the teacher stands at 1.85, the undistilled student at 2.03 and the distilled one
 2210 at 2.41; on the voxel logits the three are 1.65, 1.98 and 2.13. Read in absolute terms,
 2211 both students carry less high frequency than the teacher in both regions, and the
 2212 distilled one carries the least: on the planes the boundary shares are 5.0, 4.5 and 3.0
 2213 per cent and the interior shares 2.7, 2.3 and 1.3 per cent. The distilled student’s ratio
 2214 is the highest because its interiors are suppressed harder than its boundaries, retaining
 2215 0.48 of the teacher inside a class against 0.61 at its edges. Part of what is removed
 2216 inside a class is the fabricated high frequency the objective was built to remove, and
 2217 part is interior texture the teacher does carry; the two cannot be separated by region
 2218 from these measurements. Read across the whole plane they do not add up to a loss:
 2219 above $\lambda = 0.6$ the distilled student retains as much teacher-explained energy as the
 2220 undistilled one, to within 0.01. What this row supports is that the high frequency the
 2221 distilled student keeps leans more towards boundaries, and not that it carries more
 2222 high frequency at boundaries.

2223 *How the shared and invented parts are estimated.*

2224 Inside the teacher’s support, for each band separately, a ridge regression predicts
 2225 the student’s band- k field from the teacher’s band- k field. The cells are divided into
 2226 alternating 60° azimuthal sectors; the regression is fitted on one set of sectors and
 2227 evaluated on the other, and the ridge parameter is chosen by a further cross-validation
 2228 inside the fitting half alone. What the predictor accounts for out of sample is the shared
 2229 part and the residual is the invented part, each divided by the teacher’s energy in
 2230 that band. The teacher’s own ceiling is obtained by running the same procedure with
 2231 a re-truncated copy of the teacher in place of the student. One further check guards
 2232 the shared part: the same predictor is evaluated against the teacher of an unrelated
 2233 frame, where it should explain nothing, and a band in which it still reaches $R^2 > 0.05$
 2234 is marked, because its shared part may be slightly overstated. On the feature planes
 2235 only the lowest band is marked, by 0.06 and 0.10 for the two students.

2236 **D.2 Spectral Structure of the Sampling Gap Across Datasets**

2237 Before asking whether distillation restores the lost spectrum, we first establish what is
 2238 lost. The native-domain suite of Sec. C.1 is applied to 200 dense/sparse frame pairs on
 2239 each of SemanticKITTI, Waymo and nuScenes, 600 frames in total, and all per-class
 2240 numbers below are medians over frames.

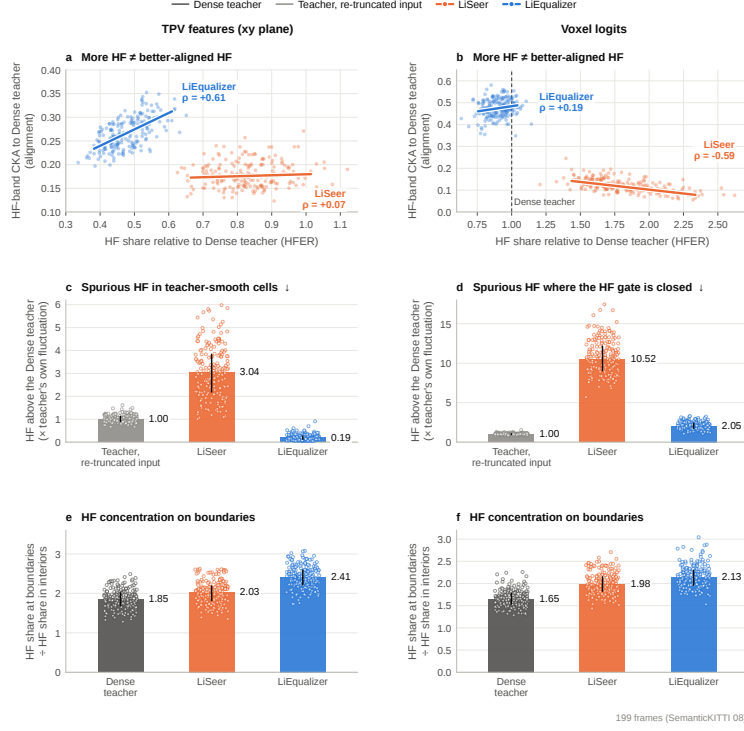


Fig. D3: The remaining energy readings, feature planes on the left and voxel logits on the right.

Note. Rows, from the top: whether carrying more high frequency goes with carrying the right high frequency, across scenes; how much high frequency a student adds where the teacher's own field is flat, in units of the teacher's own run-to-run fluctuation; and how the high frequency that survives is divided between class boundaries and class interiors, as a ratio rather than an amount.

At the level of whole scans, sampling acts as a low-pass over the full spectrum. Figure D4 shows the multi-scale diffusion profiles. On every dataset the dense curve decays fast while the sparse curve stays flat, because the sparse cloud has little high-frequency energy left for diffusion to remove. The full-spectrum view (Fig. D5) quantifies the same collapse on the true frequency axis. Between 62% and 88% of the sparse branch's energy sits below $\lambda < 0.01$, the λ_{85} roll-off separates by a factor of 35 to 47 (median 0.33–0.46 dense vs. 0.01 sparse for the density signal), and true high-frequency energy ($\lambda \geq 0.6$) retains only 24–30% of it. No SemanticKITTI frame exceeds parity on any of the three signals. Elsewhere the exceptions stay rare, at most 14% of frames on any one signal, and the normal_z signal stays at zero on all three datasets. The direction of the effect is therefore not a property of one dataset.

Per class, however, attenuation is not the whole story, and Fig. D6 separates two pathologies. Driveable ground attenuates everywhere (median 0.42–0.56; at most 8% of frames above parity), since a smooth expanse has only fine texture to lose. Nature instead exceeds parity everywhere (median 1.22–2.59; 87–100% of frames), because subsampling a porous volume fragments its support into salt-and-pepper islands whose

v2 dataset-mean multi-scale diffusion profiles (per-frame normalized, averaged over N=200 frames)

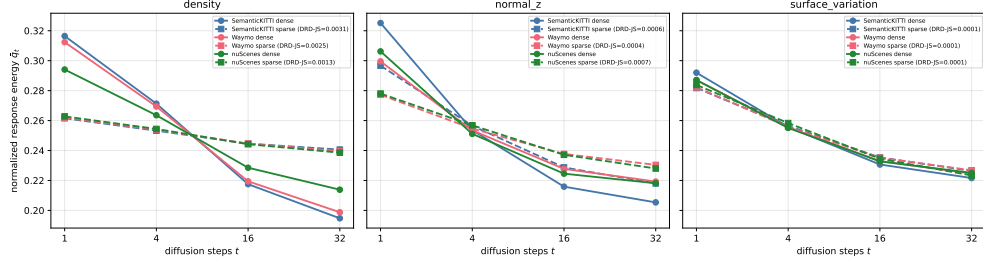


Fig. D4: Dataset-mean multi-scale diffusion profiles $\bar{q}_t$ (per-frame normalized, $N=200$ frames each). Dense profiles (solid) decay fast—there is high-frequency energy for diffusion to remove—while sparse profiles (dashed) are already flat. The ordering is identical on all three datasets and all three signals, with sensitivity density $\gg$ normal_z $>$ surface variation.

v2 KPM global-signal spectral concentration (N=200 frames/dataset)

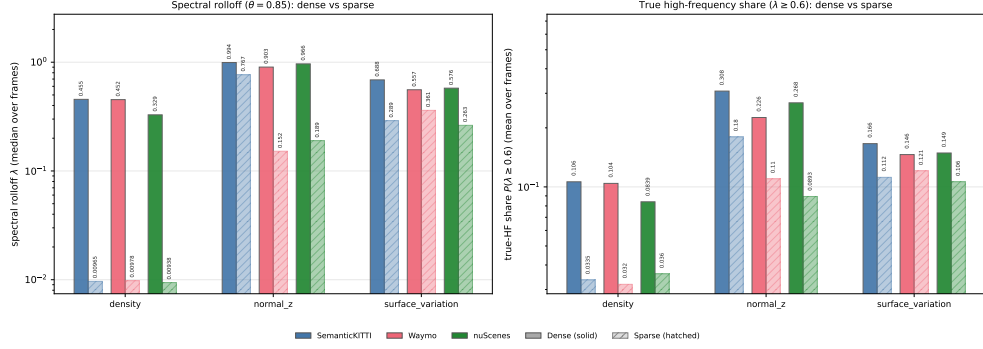


Fig. D5: Full-spectrum concentration statistics (medians/means over 200 frames per dataset). Left: spectral roll-off λ_{85} —the sparse branch accumulates 85% of its energy below $\lambda \approx 0.01$, a 35 to 47 fold separation from dense on the density signal. Right: true high-frequency share ($\lambda \geq 0.6$)—the sparse branch retains 24 to 30% of the dense reference.

boundaries are themselves high-frequency content. This is the native-domain precursor of the white floor that the voxel suite reports below. Every other class sits between these poles, and its side of parity is set by the density of the sparse branch. On 32-beam nuScenes (34k points) fragmentation dominates and all classes except ground exceed parity, whereas on 64-beam SemanticKITTI and Waymo (121k/162k points) attenuation dominates and most classes sit below. Attenuation and aliasing are therefore not attributes of a dataset but two competing effects of one sampling operation, balanced at a working point set by density and geometry. The black diamonds overlay

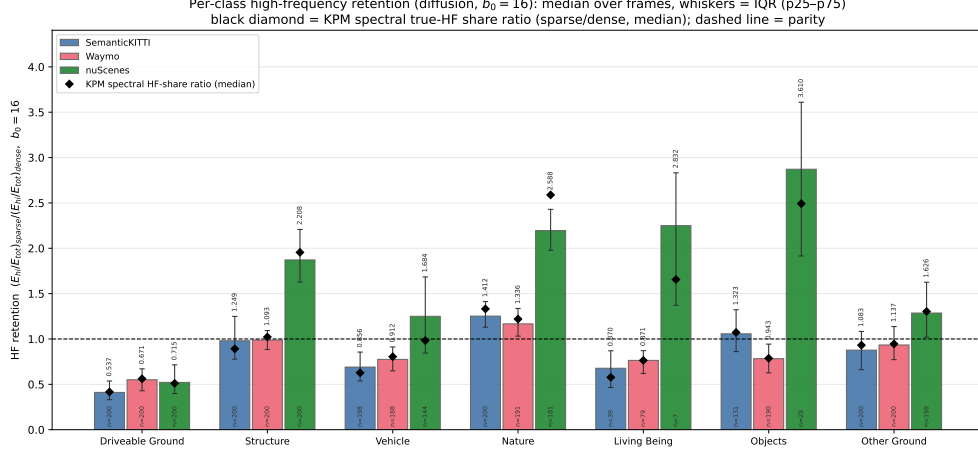


Fig. D6: Per-class high-frequency retention (diffusion estimator, $b_0=16$; bars = medians, whiskers = IQR, n = frames with the class on both sides). Black diamonds: the independent full-spectrum estimate (share of energy at $\lambda \geq 0.6$, sparse/dense). Ground attenuates on every dataset; Nature exceeds parity on every dataset; the remaining classes cross parity as the sparse branch gets sparser (SemanticKITTI/Waymo $\rightarrow$ nuScenes).

the independent Chebyshev-spectral estimate of the same ratio, and the two estimator families agree class by class and dataset by dataset.

The band-gap heatmap (Fig. D7) localizes both pathologies on the frequency axis. The sparse excess of the global signals is confined to the quasi-DC bands ($G_{<0.03}$ between -0.02 and -0.36), while the dense excess is spread over every band of $\lambda \in [0.03, 1.15]$, which is the complete shape of the low-pass. The excess of geometric orientation detail peaks at $\lambda \in [0.85, 1.15]$ on all three datasets (normal_z row), so the loss of fine geometry has a definite spectral address. The most consequential reading concerns the fragmenting classes, whose spurious energy is not confined to the top of the spectrum. For the nuScenes minority classes the sparse excess is largest in $\lambda \in [0.03, 0.4)$ and decays toward the top band, and on Waymo several classes show both pathologies at once, with dense excess in the mid bands and sparse excess at the very top. Aliasing injected by sampling is therefore broadband, and a mechanism that reasons about it only at the top of the spectrum sees a fraction of it.

On the voxel grid the same undersampling reverses sign, and the voxel suite, which compares both branches on one common grid, supplies this complementary view. Measured on the masked field with the operator the teacher support induces, the sparse branch carries between 24 and 34 times the high-frequency energy of the dense one, the middle half of frames spanning that interval. The band above $\lambda = 0.6$ holds only a small part of the teacher’s own energy on this operator, so the ratio is large where the grid operator, which counts the step at the support boundary as high frequency, would report a figure near two. The excess is concentrated in the top band: the relative band-energy discrepancy is $FBD_{\text{high}} = 5.35$ there, against $FBD_{\text{low}} = 0.72$ in the

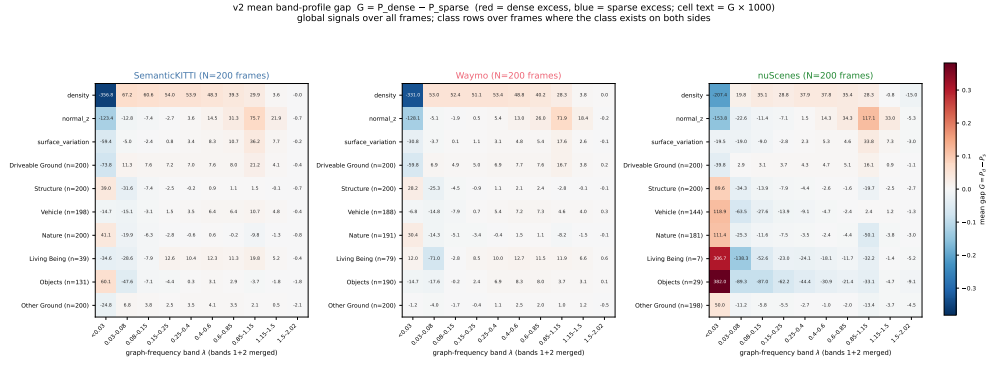


Fig. D7: Band-profile gap $G_\beta = E_\beta(\text{dense}) - E_\beta(\text{sparse})$ (red = dense excess, blue = sparse excess; cell text = $G \times 1000$). Rows: three global signals and seven class indicators; columns: frequency bands. The sparse surplus of the global rows lives entirely in the quasi-DC bands; the dense surplus spans $\lambda \in [0.03, 1.15]$, peaking at $[0.85, 1.15]$ for geometric orientation. For fragmenting classes (Nature everywhere; most nuScenes classes) the sparse excess is broadband, largest well below the top band.

low band, as means over frames. The low-frequency shape therefore still tracks the teacher where the top band no longer does.

One control separates this excess from any difference between the two networks. We hold the network fixed and change only the observation, so that the same dense-trained model reads the sparse scan; its field still carries between 21 and 31 times that energy. That accounts for 0.91 of the excess measured above, so the injection is a property of the sampling and not of the student’s weights. The mechanism is the sampling geometry itself. Multiplication by an irregular hit pattern already becomes, in any Fourier basis, a convolution that spreads scene energy across all bands, and for an independent random mask the expected residual floor is exactly flat (Prop. 10, Appendix B.1). That idealization does not capture everything the sensor does here, since the returns are also re-positioned rather than merely dropped. Isolated occupied voxels and staircase boundaries are the visible form. The excess is a property of which cells are present rather than of the values they carry. The student’s logits at a given voxel do not depend on which support set the bookkeeping uses, since both branches query the same voxel centre and read the same features through the same decoder, so the ratio above is not an artifact of evaluating the student differently in the two cases. What produces it is absence rather than addition. The student holds no value on the 84% of the teacher’s support it never observes, so its field steps at every boundary between a cell the two share and a cell only the teacher reaches, and those steps lie inside the teacher’s support. The cells the student holds and the teacher does not are a thin rim around that support, about two per cent of the grid the operator sees, and the boundaries there, where the teacher’s field steps instead, number about a fifth as many. A filter cannot produce this symptom, since a filter only attenuates existing frequencies. A sampler can, which is why “SG” must be modeled as one. That ratio quantifies the aliasing floor of the masked field. On the common-support field (C15),

where distillation operates, the corresponding excess is remeasured by the diagnosis of Sec. 2.5.

Taken together, these measurements fix the empirical shape of the transfer model (C18), a monotone attenuation m with a class- and density-dependent aliasing floor, and they are re-verified on the common-support voxel field in Sec. 2.5.

Appendix E Full Loss Definitions and Implementation Details

This appendix gives the loss terms with their weights, their switches and their cost. Besides the matched, suppressed and auxiliary terms named in the main text, the implementation carries three further stabilizers that are small but not zero, and they are listed here so that the objective can be reproduced: a per-scale Dirichlet-energy comparison on each branch, which aligns the channel-normalized distribution of $g_b^\top L g_b$ on the planes and of $h_b^\top L_\Omega h_b$ on the logits; and a soft Euler-characteristic curve matched on the logits. Their weights in the configuration used here are 0.5 and 2 for the two Dirichlet terms and 4 for the Euler-characteristic term, against 20 for each of $\mathcal{L}_V$ and $\mathcal{L}_C$. They are spectral quantities of the same fields, so they are not independent of the contracted moments, but neither are they redundant with them.

The Euler-characteristic term is the one statistic of the high-frequency branch that is not confined to the teacher’s support. It is read on the whole grid, so that cells with no return carry the uniform class distribution that a zero logit vector produces rather than being excluded. Everything else in the branch, the diffusion, the gates, the matched and suppressed terms and all their normalizers, is taken over Ω_T alone.

The nominal block for the low-frequency match is 30×30 , fitted to each plane’s own shape: 30×30 on the bird’s-eye plane, which is 240×180 and therefore divides into 8×6 blocks; 30×16 on the 180×16 plane, giving 6×1 ; and 16×30 on the 16×240 plane, giving 1×8 . Each block is normalized jointly over its channels and its two spatial extents rather than per channel, and the set $\mathcal{P}$ of Eq. (E21) is the blocks that contain teacher support, which are averaged with equal weight.

E.1 Overview

This appendix expands the compact objective of Sec. 3.4. Indices below absorb batch and channel dimensions whenever this keeps the notation readable; every teacher quantity, including each gate and threshold, is detached.

E.1.1 Normalized Match and Suppression

Let $\mathcal{P}$ be a partition of the response domain and let a be a nonnegative gate. The LF block match uses gate-weighted unit-norm responses,

$$\begin{aligned} \bar{u}_{a,P}^\Sigma &= \frac{u_P}{\left(\sum_{i \in P} a_i u_i^2 + \epsilon_{\text{num}}\right)^{1/2}}, \\ \text{Match}_L(u_s, u_t; a) &= \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \sum_{i \in P} a_i \left(\bar{u}_{s,a,P,i}^\Sigma - \bar{u}_{t,a,P,i}^\Sigma\right)^2. \end{aligned} \tag{E21}$$

2350 The HF branch does not normalize each side on its own. Both residuals are first centred
 2351 across channels and then divided by one and the same teacher-derived scale, so that
 2352 what is compared is expressed in the teacher’s own units,

$$2353 \quad \tilde{h}_b^\nu = h_b^\nu - \text{mean}_c h_b^\nu, \quad n_b^s = \frac{\tilde{h}_b^s}{r_b^t}, \quad n_b^t = \frac{\tilde{h}_b^t}{r_b^t}, \quad (\text{E22})$$

$$2354 \quad r_b^t = \left(\max\{\text{mean}_{R,c}(\tilde{h}_b^t)^2, (0.01 a_t)^2, \epsilon_{\text{num}}\} \right)^{1/2}, \quad a_t = \text{RMS}_{\Omega_T,c}(z^t - \text{mean}_c z^t), \quad (\text{E23})$$

$$2357 \quad \text{Match}_H(n_b^s, n_b^t; a) = \frac{\sum_{i \in R} \sum_c a_i (n_{b,i,c}^s - n_{b,i,c}^t)^2}{\sum_{i \in R} a_i + \epsilon_{\text{num}}}. \quad (\text{E24})$$

2358 Every teacher quantity here is detached. The term a_t only places a relative floor
 2359 under the residual scale; the logits themselves are not divided by it before diffusion,
 2360 and the raw residuals h_b^ν are what the moment, profile and flatness statistics are
 2361 computed from. Because the two sides share one denominator, the HF match is a
 2362 comparison in a common unit rather than a comparison of two separately normalized
 2363 shapes: it is invariant to a positive rescaling of the *teacher*, which moves numerator
 2364 and denominator together, but not to one of the student.

2365 The two suppressions share the complement-of-gate mechanism but are two dif-
 2366 ferent functions, and they are written separately here because they do not have the
 2367 same target. On the voxel logits the student is charged only for the residual energy by
 2368 which it exceeds the teacher,

$$2369 \quad \text{Suppress}_H^{(b)} = \frac{\sum_{i \in R} \sum_c c_c (1 - w_{b,i}) [(n_{b,i,c}^s)^2 - (n_{b,i,c}^t)^2]_+}{\sum_{i \in R} \sum_c c_c (1 - w_{b,i}) + \epsilon_{\text{num}}}, \quad (\text{E25})$$

2370 with n^s, n^t from Eq. (E22) and c_c the per-class weights, which on SemanticKITTI
 2371 raise the two dynamic classes by a factor of two. The teacher’s own residual therefore
 2372 enters the suppression as its resting point, which is the property Sec. 3.4.2 relies on.

2373 On the feature planes the target is zero, and the quantity driven there is the gate-
 2374 complement-weighted curvature energy measured against a teacher-derived reference,

$$2376 \quad T_b = \max\left\{\text{mean}_{R,c}(Lg_b^t)^2, 10^{-4} \text{mean}_{R,c}(g_b^t)^2\right\}, \quad u_b = \frac{\sum_{i \in R} \sum_c (1 - v_{b,i}) (Lg_b^s)_{i,c}^2}{T_b \sum_{i \in R} (1 - v_{b,i}) + \epsilon_{\text{num}}}, \quad (\text{E26})$$

$$2377 \quad \text{Suppress}_L^{(b)} = f(u_b), \quad f(u) = \begin{cases} u, & u \leq 1, \\ 1 + \log u, & u > 1, \end{cases} \quad (\text{E27})$$

2378 averaged over the batch. Both teacher terms in T_b are detached, and the second is
 2379 a floor that keeps the reference from collapsing where the teacher’s plane is flat; the
 2380 logarithmic tail bounds the gradient when the student’s curvature is far above the
 2381 reference. The target is zero in the sense that $u_b = 0$ minimizes the term, which is what
 2382 Theorem 1(iv) requires of the shrinkage case; it does not mean that the normalizer has

no lower bound. This suppression is applied on the bird’s-eye plane alone, and only
on the scales $b = 3, \dots, 8$.

With those two definitions in place the branch losses read

$$\mathcal{L}_{\text{match}}^{\text{L},b} = \text{Match}_{\text{L}}(g_b^s, g_b^t; v_b), \quad \mathcal{L}_{\text{sup}}^{\text{L},b} = \text{Suppress}_{\text{L}}^{(b)}, \quad (\text{E28})$$

$$\mathcal{L}_{\text{match}}^{\text{H},b} = \text{Match}_{\text{H}}(n_b^s, n_b^t; w_b), \quad \mathcal{L}_{\text{sup}}^{\text{H},b} = \text{Suppress}_{\text{H}}^{(b)}. \quad (\text{E29})$$

Here the partition $\mathcal{P}$ in Eq. (E21) is instantiated as non-overlapping 30×30 TPV blocks,
and Ω is the common teacher-support voxel–class domain. Equation (E28) suppresses
spurious *curvature* in a smoothed feature field; Eq. (E29) suppresses spurious *high-pass*
amplitude in voxel logits.

In one-based mathematical notation, the LF branch uses $\mathcal{M}_{\text{L}} = \{1, \dots, 8\}$ and
applies suppression only on $\mathcal{S}_{\text{L}} = \{3, \dots, 8\}$. The HF gate uses $\mathcal{M}_{\text{H}} = \{1, \dots, 8, 16, 32\}$
and $\mathcal{S}_{\text{H}} = \mathcal{M}_{\text{H}}$; only its first eight responses enter the auxiliary profiles. The TPV
view weights are $(\omega_{xy}, \omega_{yz}, \omega_{zx}) = (1, \frac{1}{3}, \frac{1}{3})$.

E.1.2 Two-Axis Spectral Regularization

For LF, the contracted-moment vectors are the jointly normalized energies of $u_b^{\text{L}} =$
 $D_{\alpha}^b x$. For HF, they are computed from $u_k^{\text{H}} = D_{\alpha}^k h_8$, i.e., a contracted-moment pyramid
of the terminal base residual. In either case,

$$E_{k,c}^r = \|u_{k,c}^r\|_2^2, \quad \widehat{\mathbf{e}}^r = \frac{E^r}{\|E^r\|_F + \epsilon_{\text{num}}}, \quad r \in \{\text{L}, \text{H}\}. \quad (\text{E30})$$

The normalization is taken jointly over the channel and scale axes for both branches.
The profile instead aggregates channels at each branch response scale: $q_b^{\text{L}} = \sum_c \|g_{b,c}\|_2^2$
and $q_b^{\text{H}} = \sum_c \|h_{b,c}\|_2^2$, followed by $\widehat{\mathbf{q}}^r = \mathbf{q}^r / (\|\mathbf{q}^r\|_1 + \epsilon_{\text{num}})$. Consequently, $\mathcal{R}_{\text{mom}}$ com-
pares the joint allocation of energy over channels and contracted scales up to one
overall gain per sample, whereas $\mathcal{R}_{\text{prof}}$ compares scale allocation alone after channel
aggregation. Their explicit combination is Eq. (19) for LF and Eq. (24) for HF.

E.1.3 Defaults, Ablations, and Cost

What carries the high-frequency branch.

The branch combines a gated spectral comparison with a value match and a continuity
term, and the weights alone do not say which of them moves the student. Measured as
the share $\|\nabla \text{term}\| / \|\nabla \text{total}\|$ on student fields taken from training, the spectral terms
together account for 0.93 of the gradient, their cosine with the total being 0.97, with
the gated match at 0.90, the suppression at 0.25, the value match at 0.16 and the
continuity term at 0.20. That balance is not constant over training. On a perturbed
state that stands in for an early student, far from the teacher, the spectral share falls
to 0.52 and the value match rises to 0.56, above the gated match at 0.43. The reading
is therefore that the value match carries the coarse alignment while the student is still
distant, and the spectral comparison takes over once it is close, which is the regime in
which the branch is meant to act. Two cautions belong with these numbers. They must

2420 be read as gradient shares and not as loss values, since the value match contributes the
2421 largest weighted loss while supplying a sixth of the gradient. The Euler-characteristic
2422 term stays at or below 0.007 in both states and is reported here only for completeness.

2423 ***Student checkpoint selection.***

2424 Students are selected by the highest source-domain SSC mean over the schedule.
2425 SemanticKITTI labels only sequences 00 to 10, of which 08 is held out, so that one
2426 sequence serves both as the selection signal and as the source-domain column; the
2427 target-domain columns take no part in the selection. The two differ by at most 0.001
2428 on the runs reported here.

2429 ***Teacher checkpoint selection.***

2430 The teachers for SemanticKITTI and nuScenes are the checkpoints with the highest
2431 dense-validation SSC mean over the training schedule. The Waymo teacher is the
2432 checkpoint at the end of training. The difference follows from the data rather than
2433 from the protocol: the densified completion covers only the Waymo training split, so
2434 the validation split carries no dense annotation and no dense-validation curve exists
2435 to select against.

2436 ***Adaptations of the compared objectives.***

2437 Two adaptations are required to run the compared objectives on this architecture.
2438 The channel-wise objective is applied to the three tri-perspective-view planes rather
2439 than to a single image feature map; the channel-wise softmax remains well defined on
2440 a cylindrical projection, but the choice of spatial axis is ours. Its loss weight is the
2441 value the authors report for feature maps, $\lambda = 50$; the value of 3 that appears in their
2442 released example applies to the logits map and is not used here.

2443 ***Adaptations of the point-to-voxel objective.***

2444 This objective does not transfer unchanged, and its substitutions are larger than for
2445 the others. The released implementation carries two affinity terms, one over per-point
2446 features and one over sparse-decoder voxel features, and no form that acts on planes.
2447 At the stage where the objective is applied this network exposes no per-point feature
2448 of the first kind, so the point-level term is evaluated on the tri-perspective-view planes
2449 instead, over a grid of plane positions that is fixed across frames and includes unoc-
2450 cupied cells, drawn from four supervoxel regions at six thousand positions each. The
2451 voxel-level affinity is evaluated on the class logits rather than on decoder features,
2452 which is what the common-support construction makes available on both sides. The
2453 voxel divergence is restricted to that same common-support mask, since positions out-
2454 side it hold structural zeros produced by scattering rather than observations, and the
2455 conversion from voxel indices to metres uses the voxel size of this architecture while
2456 the supervoxel partition is kept as released. That divergence is normalised per batch
2457 and enters at unit weight, where the released code normalises by the number of pairs
2458 and classes and enters at a weight of 0.15, so the three terms together correspond
2459 to no single branch of the released implementation. We report the objective under

these substitutions rather than omit it, and the entry should be read as a comparison against this adaptation rather than against the published method.

Two properties of the released code are reproduced rather than repaired. The minority-class count that scales the class factor is formed so that the factor is constant across classes and cancels in the sampling normalization, which leaves the sampler driven by the distance term alone; and the affinity term is limited from above by an assignment rather than by a clamp, so it contributes no gradient beyond that limit. Both reduce rather than increase the strength of this baseline, and we keep them because the comparison is against the method its authors released; repairing them would compare against a variant that was never published. The objective is weighted by a fixed coefficient and is not routed through the gradient-norm controller, which matches the other compared objectives.

Table E1: Default hyper-parameters of the LF and HF spectral branches. One-based scale indices are used.

Quantity	LF	HF
Diffusion step α	0.25	0.25
Base scales / moments K	8	8
Scale weights η_b	$0.5 + 0.25(b - 1)$	$0.5 + 0.25(b - 1)$
Gate floor multiplier κ	0.5	0.5
Suppression scales	3:8	1:8, 16, 32
λ_{sup}	0.25	1.0
λ_{mom}	0.5	2.0
λ_{prof} (profile ℓ_1)	8.0	5.0
λ_{flat}	—	1.0
Spatial partition	30×30 blocks	common voxel domain

Each design rule can be reverted independently—ungated matching, removal of suppression, raw Laplacian moments, JS instead of profile ℓ_1 , removal of the extra HF scales, or removal of the flatness hinge—to obtain the one-factor ablations of Sec. 2.4. Distillation is training-time only, and its dominant training cost is the depthwise diffusion pyramid. At inference the distillation modules are not executed at all: none of the three inference entry points invokes them, so there is no variant of the deployed network that carries distillation and no comparison between the two can be drawn. The spectral branches also contribute no trainable parameters, only fixed operator kernels and scale weights, so the parameter count of the deployed model is identical to the baseline’s by construction rather than by measurement. For reference, the deployed model has 32,277,463 parameters and runs at 0.224 s per frame (median over 60 frames after a 10-frame warm-up, batch size 1, single NVIDIA RTX 4090 with the remaining GPUs idle), with a peak of 5.66 GiB. The two extra cumulative scales of the high-frequency branch cost about 3% of the training step. We measured this by running the final method twice with the extra scales on and twice with them off, alternating the two arms, and comparing per-iteration times after discarding the first hundred iterations of each run; the two pairings give 2.5% and 3.7%, and reading the middle run against the mean of the two that bracket it gives 2.4% and 3.1%. Peak memory per device rises by 0.7%. The arms differ in the extra scales alone, share a seed and a

2491 data order, and were run on eight RTX 4090 devices with one sample per device and
2492 gradient accumulation of four, so the figure is a relative cost under that layout rather
2493 than a property of the objective; data loading took about 0.02 s per iteration and was
2494 not the limiting stage.

2495 Appendix F More Details on Coarse Label

2496 We use the coarse labels for all cross-domain (domain-generalization) evaluations; the
2497 indices range from 0 to 7. The detailed mapping rules for the three datasets are given
2498 in Listings F, F, and F, each stated over that dataset’s own original label indices.
2499 Each mapping is total: every annotated class of every dataset is assigned to one of the
2500 eight coarse classes, and none is left unscored. Readers comparing against numbers
2501 reported on a dataset’s own benchmark should note that those benchmarks exclude
2502 some annotated classes from scoring, whereas the protocol here scores all of them.

2503 Three differences between the coarse classes of the three datasets follow from
2504 the granularity of the original annotations rather than from the mapping itself, and
2505 cannot be removed by any choice of rule. First, nuScenes has no separate class for
2506 poles, street signs or traffic lights: its `static.manmade` label is defined to cover build-
2507 ings, walls, guard rails, fences, poles, street signs and traffic lights within one atomic
2508 category, so these returns fall under *Structure* on nuScenes while the corresponding
2509 returns fall under *Objects* on SemanticKITTI and Waymo. Second, nuScenes anno-
2510 tates a rider together with the vehicle—its instructions state that a rider or passenger
2511 is included in the vehicle box, and only a person standing beside a vehicle is labelled
2512 as a pedestrian—so riders fall under *Objects* on nuScenes and under *Living Being* on
2513 SemanticKITTI and Waymo. Third, Waymo provides no fence category, so its *Struc-*
2514 *ture* class contains buildings only. These residual differences affect the same classes for
2515 every method evaluated here and therefore do not favour any of them.

Code F.1: Label mapping rules from nuScenes to coarse label.

```

1  NUSCENES_TO_CL = {
2  0: 0,    \# noise -> Other
3  1: 5,    \# animal -> Living Being
4  2: 5,    \# human.pedestrian.adult -> Living Being
5  3: 5,    \# human.pedestrian.child -> Living Being
6  4: 5,    \# human.pedestrian.construction\_worker -> Living
      Being
7  5: 5,    \# human.pedestrian.personal\_mobility -> Living Being
8  6: 5,    \# human.pedestrian.police\_officer -> Living Being
9  7: 5,    \# human.pedestrian.stroller -> Living Being
10 8: 5,    \# human.pedestrian.wheelchair -> Living Being
11 9: 6,    \# movable\_object.barrier -> Objects
12 10: 6,   \# movable\_object.debris -> Objects
13 11: 6,   \# movable\_object.pushable\_pullable -> Objects
14 12: 6,   \# movable\_object.trafficcone -> Objects
15 13: 6,   \# static\_object.bicycle\_rack -> Objects
16 14: 6,   \# vehicle.bicycle -> Objects

```

```

17 15: 3, \# vehicle.bus.bendy -> Vehicle
18 16: 3, \# vehicle.bus.rigid -> Vehicle
19 17: 3, \# vehicle.car -> Vehicle
20 18: 3, \# vehicle.construction -> Vehicle
21 19: 3, \# vehicle.emergency.ambulance -> Vehicle
22 20: 3, \# vehicle.emergency.police -> Vehicle
23 21: 6, \# vehicle.motorcycle -> Objects
24 22: 3, \# vehicle.trailer -> Vehicle
25 23: 3, \# vehicle.truck -> Vehicle
26 24: 1, \# flat.driveable\_surface -> Driveable Ground
27 25: 7, \# flat.other -> Other Ground
28 26: 7, \# flat.sidewalk -> Other Ground
29 27: 7, \# flat.terrain -> Other Ground
30 28: 2, \# static.manmade -> Structure
31 29: 0, \# static.other -> Other
32 30: 4, \# static.vegetation -> Nature
33 31: 0, \# vehicle.ego -> Other
34 }

```

2517

Code F.2: Label mapping rules from SemanticKITTI to coarse label.

```

1 SEMANTICKITTI_TO_CL ={
2     0: 0, \# unlabeled -> Other
3     1: 0, \# outlier -> Other
4     10: 3, \# car -> Vehicle
5     11: 6, \# bicycle -> Objects
6     13: 3, \# bus -> Vehicle
7     15: 6, \# motorcycle -> Objects
8     16: 3, \# on-rails -> Vehicle
9     18: 3, \# truck -> Vehicle
10    20: 3, \# other-vehicle -> Vehicle
11    30: 5, \# person -> Living Being
12    31: 5, \# bicyclist -> Living Being
13    32: 5, \# motorcyclist -> Living Being
14    40: 1, \# road -> Driveable Ground
15    44: 1, \# parking -> Driveable Ground
16    48: 7, \# sidewalk -> Other Ground
17    49: 7, \# other-ground -> Other Ground
18    50: 2, \# building -> Structure
19    51: 2, \# fence -> Structure
20    52: 2, \# other-structure -> Structure
21    60: 1, \# lane-marking -> Driveable Ground
22    70: 4, \# vegetation -> Nature
23    71: 4, \# trunk -> Nature
24    72: 7, \# terrain -> Other Ground
25    80: 6, \# pole -> Objects
26    81: 6, \# traffic-sign -> Objects

```

2518

G.1 Spectral Analysis of the Two Representations

The figures below complete the band-wise reading of Sec. 2.5, which the shared and invented parts defined there and estimated in Appendix D.1 do not exhaust. Every quantity is computed over the teacher’s support, on the feature planes with the plane operator and on the voxel logits with the operator that support induces, so that each level is read with the operator its own branch uses.

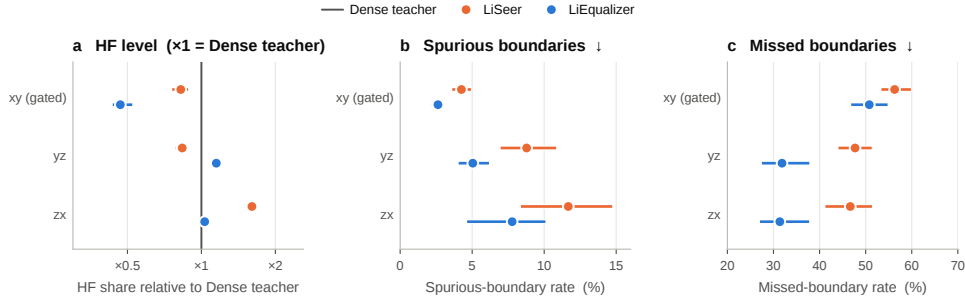


Fig. G8: The three planes do not agree.

Note. Band structure and boundary behaviour per tri-perspective plane. The two vertical planes vary mainly with how the scene occupies height, which is why they receive the support-restricted match alone.

G.2 Where the SemanticKITTI-to-nuScenes Drop Comes From

A model trained on SemanticKITTI loses 21.4 points of mIoU when it is evaluated on nuScenes, from 0.749 on its own sensor to 0.535. That is the largest transfer gap in this paper, and it is worth knowing what it consists of before reading any method into it. This appendix reports a diagnostic study of that single number. Quantities obtained with the help of nuScenes labels are marked as diagnostic: they say where the errors are and how many, and none of them is a deployable improvement.

The gap is generic, and it sits in two classes.

Eight SemanticKITTI-trained methods, evaluated under one protocol, land between 0.530 and 0.544, with Living Being between 0.090 and 0.118 and Objects between 0.033 and 0.041. The pattern does not depend on the distillation objective. Within it, Living Being falls from 0.511 to 0.118 and Objects from 0.384 to 0.036, and those two classes account for about half of the 21.4 points on their own.

nuScenes is not the harder dataset.

The decisive control replaces the source-trained model’s predictions by those of a same-architecture model trained on nuScenes, on the errors that are strictly specific to the domain change, that is the points the nuScenes-trained model reads correctly and a second SemanticKITTI-trained model also reads wrongly. Those points are 58% of

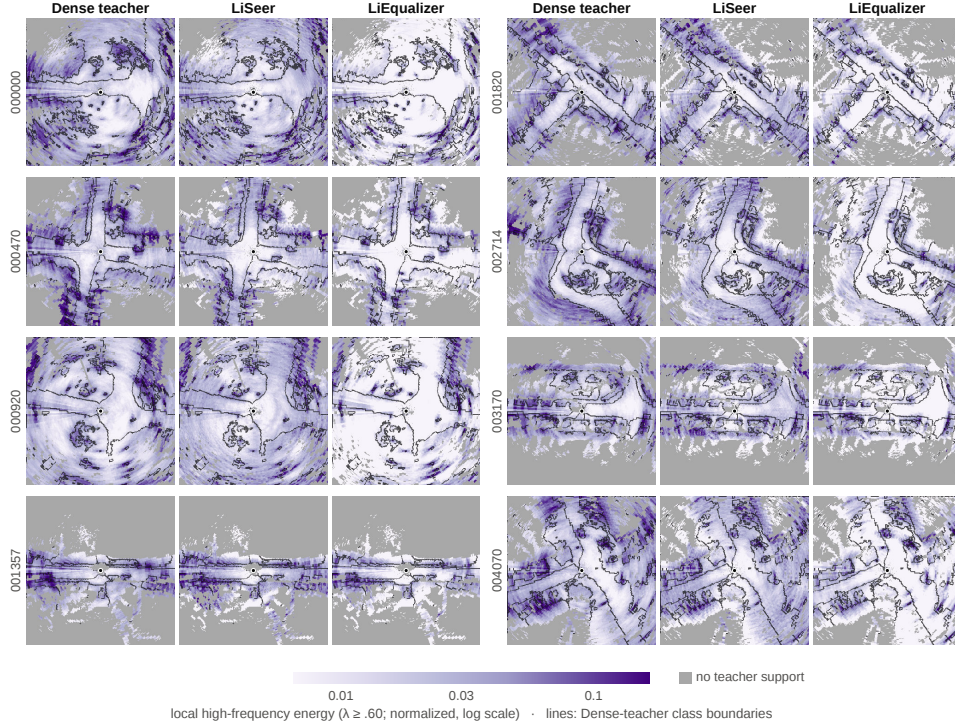


Fig. G9: Eight scenes, same reading, on the feature planes.

all errors. After that correction SemanticKITTI-to-nuScenes stands at 0.784, and the same correction applied within SemanticKITTI leaves it at 0.780. The two are level, so the validation set itself is not intrinsically more difficult; what is missing is in the model, not in the data. Both figures are diagnostic.

What the gap is made of.

The dominant failure is a whole-entity flip: an object type that SemanticKITTI does not contain, or contains only rarely, is read as one wrong class over its entire extent. Construction barriers are the clearest case. The source-trained model reads them at 0.9% accuracy against 88.7% for the nuScenes-trained model, and it stays at 0% even on the instances covered by fifteen scan lines or more, so this is not a matter of too few returns. Trailers and buses are read as Structure, manmade structure as Nature or Objects, and pedestrians as Objects, with the error growing with range. The errors are spatially coherent rather than scattered: 95.8% of wrong points agree with the majority of their eight nearest neighbours, and the best label-free smoothing we tried gains 0.03.

This failure is not created by the sensor change. The same confusion is already visible inside SemanticKITTI as a gap between instances the model has seen and instances it has not. Living Being is read as Objects on 1.5% of the training frames,

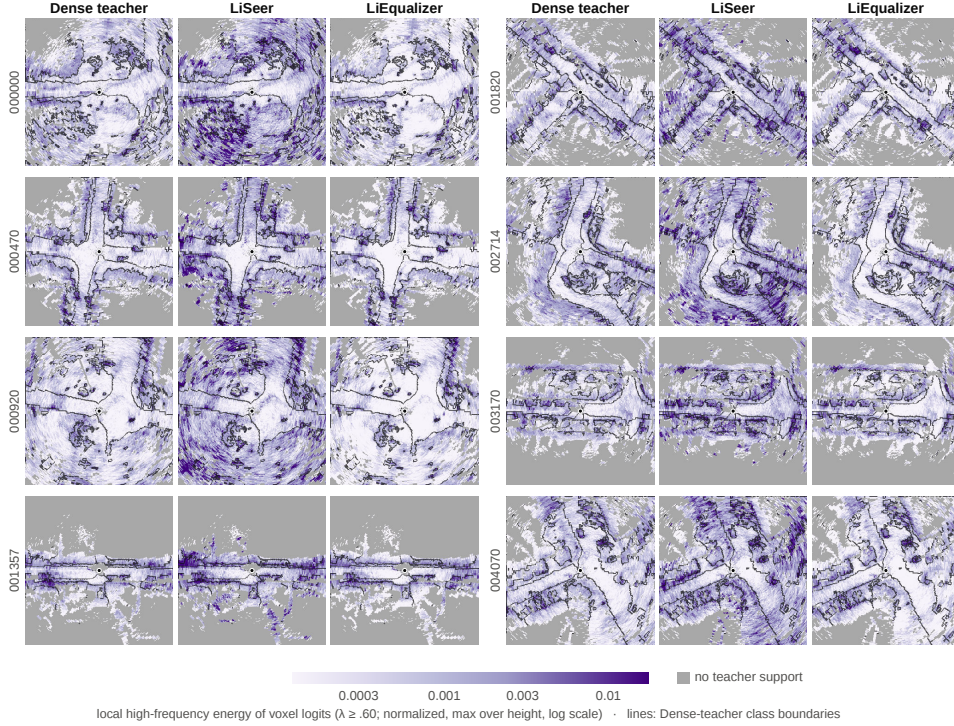


Fig. G10: Eight scenes, same reading, on the voxel logits.

22% of the held-out SemanticKITTI frames, and 45% of nuScenes. A nearest-neighbour test on the features, which involves no classifier at all, gives the same ordering at 0.2%, 13.5% and 33%, so the confusion is a property of the representation rather than of the loss used to train the head.

Sparsity is the second cause, and it cannot be undone after the fact.

Sub-sampling SemanticKITTI to the line density of nuScenes costs about 6 points, which is the share the sampling change can account for. The reverse intervention does not return it: densifying nuScenes by vertical interpolation recovers between 0.5 and 1.0 point, leaves barriers and trailers exactly as they were, and shows no dose response. The two measurements are not in conflict. Interpolation adds no measurement, and Proposition 2 is the statement that what sparse sampling attenuates cannot be recovered from sparse observations alone. Real sparsity costs about 6 points; filling the scan in afterwards is not a way of getting them back.

What the gap is not.

Table G2 lists the hypotheses that were tested and set aside. Label definitions deserve a note of their own, because the two datasets disagree on three of them. Barriers are a nuScenes class with no SemanticKITTI counterpart, poles and signs are inside the

2586 nuScenes *manmade* label and cannot be separated from it, and a rider is annotated
 2587 with the vehicle in one convention and apart from it in the other. Re-scoring the
 2588 nuScenes side under SemanticKITTI semantics recovers at most 0.8 points, and scoring
 2589 Structure and Objects as one class shrinks the gap by about 3.6, under a seven-class
 2590 to six-class change that is not directly comparable. These re-scorings are diagnostic.
 2591 The label mapping used throughout the paper is unchanged, and this analysis is not
 2592 an argument for changing it.

Table G2: Hypotheses tested against the SemanticKITTI-to-nuScenes gap.

Hypothesis	Magnitude	Verdict
Unseen or rare entity types	dominant	main cause
Scan sparsity	about 6	second cause
Class boundaries	4 to 7; 0 to 1 nuScenes-specific	not the cause
Label definitions	≤ 3.6 (diag.)	minor
Label prior, decision bias	≤ 2 each (diag.)	minor
Orientation	0.8	no
Scene conditions, weather, city	≈ 0	no
First-layer normalization, modulation	≤ 0.2	no
Head loss design	≈ 0	no

Note. Magnitudes are in points of mIoU and are not additive. Entries marked *diag.* use nuScenes labels and are upper bounds on what the corresponding error accounts for, not achievable gains.

2593 ***Where the missing capability sits, and what follows.***

2594 Re-fitting only the linear read-out on nuScenes features lifts the score by about eight
 2595 points, and the features of the nuScenes-trained model reach far higher under the same
 2596 protocol, so both the read-out and the representation are short. The split between the
 2597 two depends on where the read-out is cut and how much capacity the probe has, so
 2598 we report it as a magnitude rather than a fraction, and we state it as what a shallow
 2599 read-out can recover rather than as a claim about information being absent. Label-free
 2600 feature alignment made things worse in every form we tried.

2601 One conclusion carries the rest. Sampling is real and it is measurable: dropping
 2602 SemanticKITTI to the line density of nuScenes costs about 6 of the 21.4 points. What
 2603 remains once that share and the definition mismatches are set aside is a difference
 2604 between the scenes the two datasets contain rather than between the ways those scenes
 2605 are sampled. The difference is in what is present, and not in the conditions under
 2606 which the data were recorded, which the table above sets aside: SemanticKITTI covers
 2607 one city, largely suburban, with no construction barriers and few large vehicles, so a
 2608 model trained on it has never been asked to name those things. Sampling accounts for
 2609 about 6 points of the 21.4, and the rest is this.

2610 Two things follow for this paper. The residual gap is not a defect of the mapping,
 2611 of the evaluation or of any one module, and it is not something a distillation objective
 2612 is positioned to remove: changing what the student is taught does not change which
 2613 entities the source domain contains. The spectral account of this work addresses the

sampling component, and the measurements above are both the reason we say so and the bound on how much that component can be worth.

Limits of this study.

The nuScenes-trained control models differ from the source-trained model in more than their training domain: they use an earlier label mapping, an earlier code version and no distillation, so the representation part of the comparison is the feature space of another model rather than a counterfactual of the same model trained elsewhere. Every quantity obtained with nuScenes labels is diagnostic. The read-out and representation shares vary with the cut point and the probe, between about two fifths and one half for the read-out, and the geometry and within-group shares vary with the binning, so only magnitudes are reported. Several analyses use subsets, between 300 and 3,010 nuScenes frames and between 300 and 1,018 SemanticKITTI frames, and SemanticKITTI offers a single validation sequence. The sub-sampling and the interpolation measurements differ in direction and in method, and the second does not bound the first.

G.3 Qualitative Results and Analysis of On-Vehicle Testing on Unseen LiDAR Platforms

To further assess the practical value of LiEqualizer beyond standard benchmarks, we conduct real-vehicle validation on three in-house platforms equipped with completely unseen LiDARs. The goal of this experiment is to test whether the proposed zero-shot cross-LiDAR framework can transfer to real deployment platforms without any target-domain adaptation or fine-tuning. Since the collected real-world data do not provide dense point-wise ground truth, we evaluate qualitatively by comparing the projected LiDAR segmentation against synchronized camera images; to this end, each platform carries a forward-facing camera in addition to its LiDAR.

The real-world data are collected in three representative environments in Beijing: (1) the Tsinghua University campus in Haidian District, (2) the Beijing High-Level Autonomous Driving Demonstration Zone in Yizhuang, and (3) a rural road in Shunyi District. These scenes cover campus roads, urban roads, intersections, vegetation, parked vehicles, roadside structures, and open rural areas, providing diverse conditions for evaluating real-world cross-LiDAR transfer.

For each platform, we visualize the camera image alongside the predictions of the baseline PointSeg model and LiEqualizer-P. In every qualitative figure, the leftmost column shows the camera image, and the remaining six columns show Baseline-PointSeg and LiEqualizer-P trained on nuScenes, SemanticKITTI, and Waymo, respectively.

A consistent pattern emerges across all three real platforms. Models trained under the conventional SDT paradigm do not generalize reliably to unseen LiDARs, and their failure modes differ markedly with the source dataset. The nuScenes-trained baseline often preserves coarse road structure, yet its predictions are unstable on vehicles and roadside objects, especially under denser unseen scans. The SemanticKITTI-trained baseline degrades the most severely, frequently collapsing large portions of the scene into the *Other* category—a clear sign of strong sensitivity to sensor-dependent sampling. The Waymo-trained baseline is generally more stable than the

SemanticKITTI-trained one, but still shows dataset-specific bias, for example over-expanding certain structural or vegetation regions and producing inconsistent object boundaries across platforms. Under SDT training, then, the model does not fail uniformly; it inherits different dataset-specific biases and expresses them differently when facing unseen LiDAR configurations.

In contrast, LiSeer-P delivers markedly more consistent predictions across all three unseen sensors, regardless of whether it is trained on nuScenes, SemanticKITTI, or Waymo. Despite the substantial gap between the benchmark datasets and the real-world scenes collected in China, LiSeer-P still recovers coherent scene semantics on every platform. In particular, it identifies drivable areas, vehicles, vegetation/nature regions, and roadside objects more reliably than the corresponding baselines, while preserving a clearer large-scale scene layout. This cross-dataset consistency is especially telling: it indicates that the framework is not merely overfitting to a particular source dataset, but learning representations that remain stable across both unseen sensors and unseen geographic environments.

Some failure cases nonetheless remain. In extremely sparse or distant regions, predictions can become fragmented for lack of geometric support; fine-grained object boundaries may be imperfect when the camera-visible structure is only weakly sampled by the LiDAR; and in highly cluttered scenes, small movable objects are occasionally confused with surrounding structure or vegetation when their point support is limited. Even so, compared with the baselines, LiSeer-P exhibits far fewer catastrophic failures and maintains substantially stronger semantic consistency across platforms.

G.3.1 TH-CS55 platform

TH-CS55 is an in-house passenger-vehicle platform equipped with an unseen 128-beam LiDAR and a forward-facing camera. It represents a substantially denser LiDAR configuration than the benchmark sensors and is therefore particularly useful for testing whether the learned representations remain stable under previously unseen high-density sampling.

On TH-CS55, the gap between the baselines and LiSeer-P is particularly clear. The SemanticKITTI-trained baseline shows the strongest collapse, with many points incorrectly mapped to *Other*, especially in rows 1, 3, 5, and 8 of Fig. G12. The nuScenes-trained baseline is more stable on drivable areas but often under-segments vehicles and yields less coherent object regions. The Waymo-trained baseline captures large-scale structure more reliably, yet still shows inconsistent object extents and category transitions under the dense 128-beam observations. By contrast, LiSeer-P remains far more stable across all three source datasets, consistently recovering the drivable surface, nearby vehicles, roadside vegetation, and large roadside objects with better semantic coherence. Even in challenging scenes such as intersections and roadside parking areas, its predictions remain visually plausible and substantially more consistent than those of the corresponding baselines.

G.3.2 TH-AVT platform

TH-AVT is an in-house autonomous-vehicle platform equipped with an unseen 80-beam LiDAR and a forward-facing camera. Relative to the benchmark sensors used

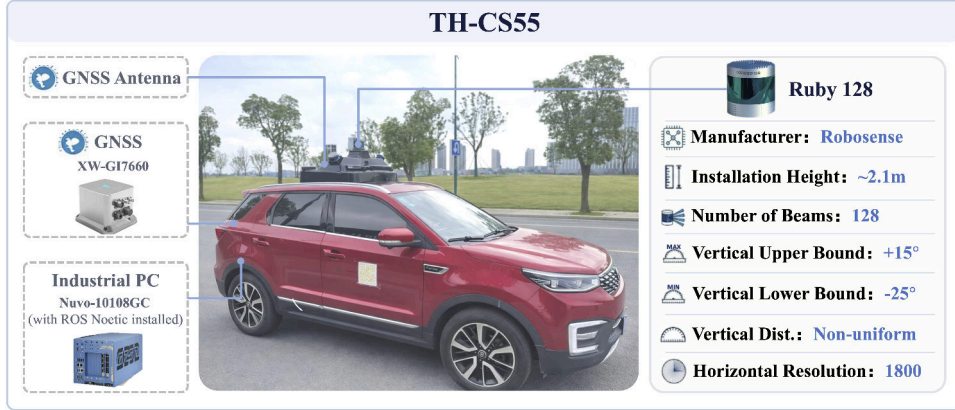


Fig. G11: Overview of the *TH-CS55* platform used for real-vehicle validation.

Note. TH-CS55 is equipped with a front-view camera, a GNSS system, an industrial PC, and an unseen 128-beam LiDAR mounted on the roof. This platform allows us to evaluate cross-LiDAR transfer under a substantially denser sampling configuration than those used in the benchmark datasets.

in training, it introduces a dense and previously unseen sampling pattern, making it a challenging test case for zero-shot cross-LiDAR generalization.

The TH-AVT results further highlight the dataset-specific bias of SDT-trained models. The nuScenes-trained baseline tends to preserve the drivable region and some roadside structure, but its vehicle predictions are often incomplete or unstable. The SemanticKITTI-trained baseline again over-predicts *Other*, especially in rows 1, 2, 4, 6, and 7 of Fig. G14, which substantially degrades semantic usability. The Waymo-trained baseline is visually cleaner, but still shows noticeable inconsistency in object boundaries and local semantic transitions. In comparison, LiSeer-P produces much more stable predictions across all three training sources, capturing road layout, building and wall structures, parked vehicles, and roadside vegetation with higher consistency—even though the scenes are collected in Beijing and differ significantly from the original benchmark domains. These results suggest that LiSeer-P is less sensitive to the specific source dataset and more robust to unseen platform geometry and installation conditions.

G.3.3 TH-Courier platform

TH-Courier is an in-house compact autonomous courier platform equipped with an unseen 32-beam LiDAR and a forward-facing camera. Unlike TH-AVT and TH-CS55, it uses a sparser LiDAR mounted at a lower height, providing a complementary deployment scenario for assessing the robustness of LiSeer across substantially different platform geometries.

The TH-Courier platform offers a complementary sparse-LiDAR setting. Here the nuScenes-trained baseline often retains a reasonable estimate of the drivable region, but its predictions on vehicles and roadside objects remain weak and fragmented. The SemanticKITTI-trained baseline again shows the most obvious failure mode, frequently assigning large portions of the scene to *Other*, particularly in rows 1, 3, 6,

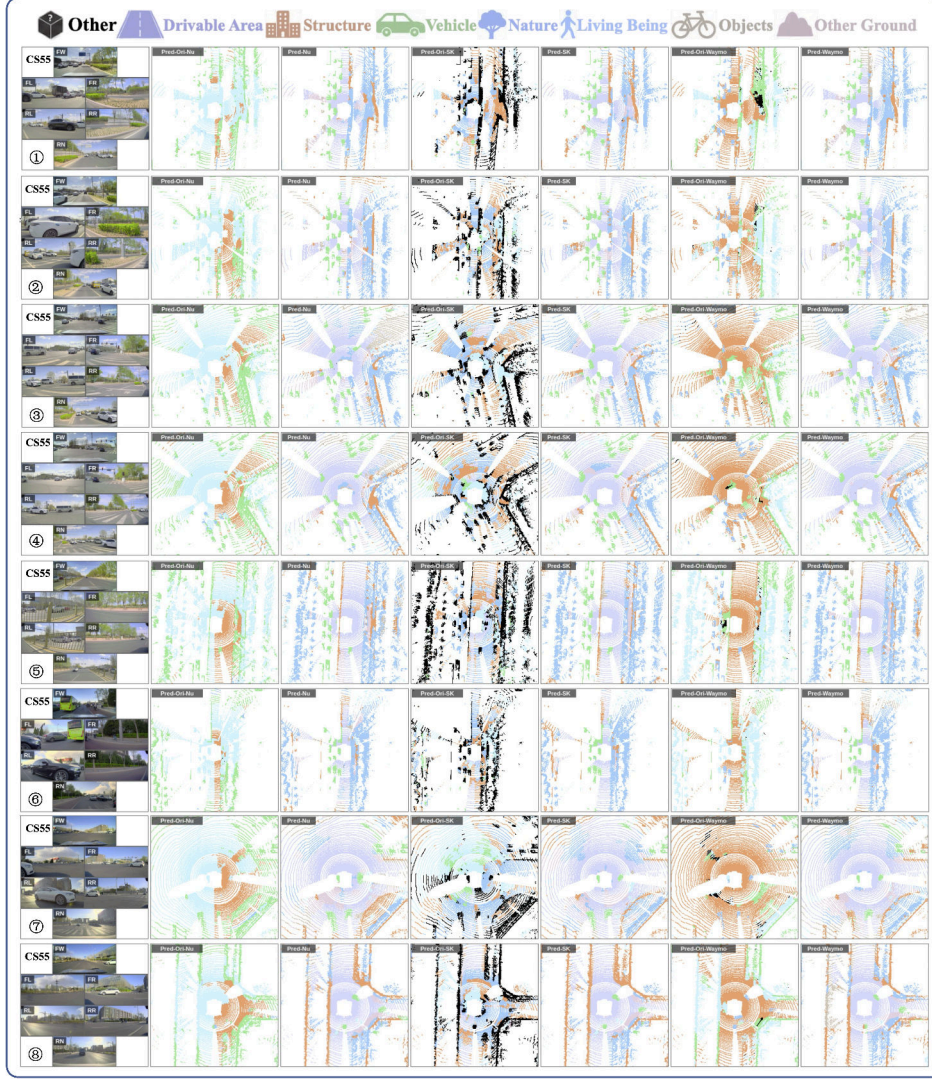


Fig. G12: Qualitative segmentation results on the *TH-CS55* platform.

Note. From left to right: camera image, Baseline-PointSeg and LiSeer-P trained on nuScenes, Baseline-PointSeg and LiSeer-P trained on SemanticKITTI, and Baseline-PointSeg and LiSeer-P trained on Waymo.

and 8 of Fig. G16. The Waymo-trained baseline is comparatively more structured, but still exhibits unstable object extents and category boundaries. In contrast, LiSeer-P remains consistently effective across all three source datasets despite the lower mounting height and sparser 32-beam observations, recovering the road surface, vehicles,

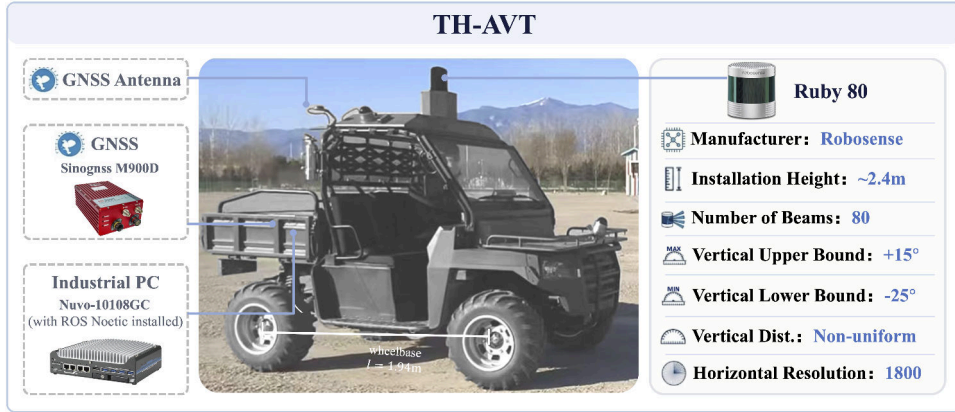


Fig. G13: Overview of the *TH-AVT* platform used for real-vehicle validation.

Note. TH-AVT is equipped with a front-view camera, a GNSS system, an industrial PC, and an unseen 80-beam LiDAR mounted at a relatively high position on the vehicle roof. This platform is used to evaluate transferability under a dense and previously unseen LiDAR configuration in real driving scenarios.

vegetation, and roadside objects more reliably and preserving a clearer large-scale scene organization in both urban-road and rural-road scenes.

Overall, the results on TH-CS55, TH-AVT, and TH-Courier support two observations. First, conventional SDT-trained models do not generalize reliably to unseen LiDARs, and their failure patterns are strongly source-dataset dependent. Second, LiSeer-P behaves far more consistently across all three unseen sensors and all three source datasets: despite being trained on benchmark data collected in other countries and under different sensor configurations, it transfers effectively to real-world scenes in China and continues to identify drivable surfaces, vehicles, vegetation, and roadside objects with substantially improved consistency. These findings provide direct qualitative evidence for the real-world value of LiSeer in zero-shot cross-LiDAR deployment.

G.4 Additional Visualizations

Two pictures are worth adding to the numbers of Sec. 2.5: what the student’s spectrum looks like beside the teacher’s once the band structure is taken away, and where on a scene the low-frequency gate actually opens.

Figure G17 is the shape behind the single-number flatness check. The two students differ most at the coarse scales, where a profile that falls below the teacher’s means the student has kept relatively more of its energy at the fine scales, which is the spurious high-frequency content the high-frequency branch exists to remove. That the distilled student tracks the teacher past $b = 8$, beyond the scales the loss constrains, is the part that was not fitted.

Figure G18 answers a question the equations leave open: the gate is defined from the teacher’s own curvature, but nothing in the definition says the result has to be interpretable. It is. The gate opens along class boundaries and closes over the interiors

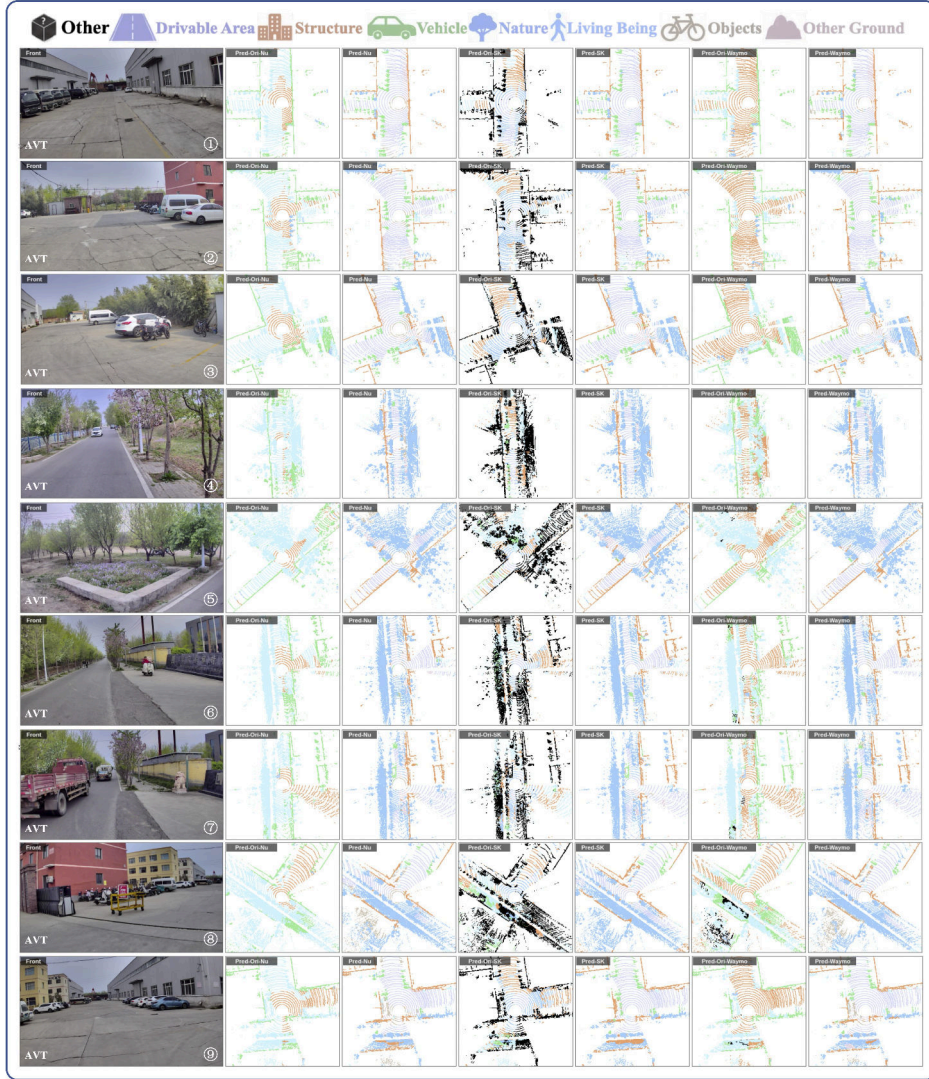


Fig. G14: Qualitative segmentation results on the *TH-AVT* platform.

Note. From left to right: camera image, Baseline-PointSeg and LiSeer-P trained on nuScenes, Baseline-PointSeg and LiSeer-P trained on SemanticKITTI, and Baseline-PointSeg and LiSeer-P trained on Waymo.

2755 of large smooth regions, so the suppression term is withheld exactly where the teacher
 2756 itself carries structure, and applied where it does not.

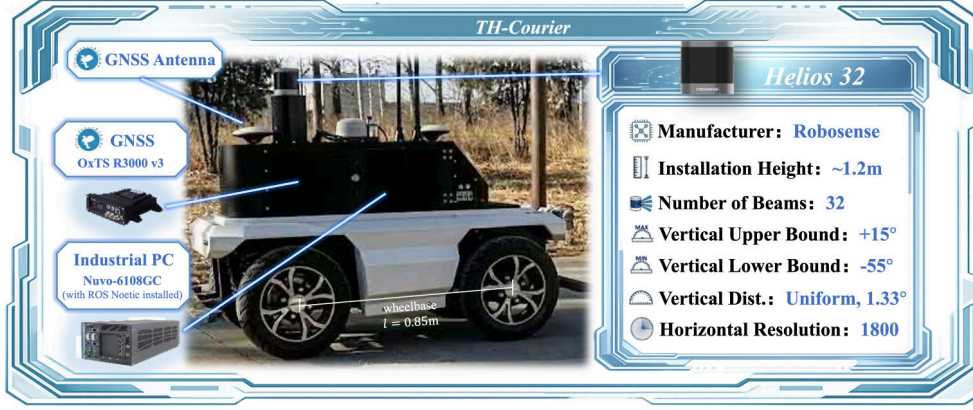


Fig. G15: Overview of the *TH-Courier* platform used for real-vehicle validation.

Note. TH-Courier is equipped with a front-view camera, a GNSS system, an industrial PC, and an unseen 32-beam LiDAR mounted at a relatively low position. Owing to its compact platform geometry and sparse LiDAR configuration, it provides a challenging real-world test case that differs substantially from the benchmark training setups.

G.5 More Settings and Results for Experiments

G.5.1 Experimental Settings

Model Architecture

As described in Section 2, we adopt the 2D projection-based method PointSeg as our baseline, which employs a Swin Transformer backbone. The input point clouds are partitioned cylindrically into $(\mathcal{H}_{\text{in}}, \mathcal{W}_{\text{in}}, \mathcal{D}_{\text{in}}) = (480, 360, 32)$ voxels, corresponding to radius, angle, and height, respectively.

To ensure a fair comparison, we do not modify any model parameters: all architectural designs and hyper-parameter settings are kept exactly as in the official baseline.

Metric for Semantic Segmentation

For 3D semantic segmentation, we evaluate performance using the mean Intersection-over-Union (**mIoU**) over the semantic classes:

$$\begin{aligned} \text{IoU} &= \frac{TP}{TP + FP + FN} \\ \text{mIoU} &= \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i + FN_i} \end{aligned} \quad (\text{G31})$$

where TP , FP , and FN denote the numbers of true positives, false positives and false negatives, and $C = 7$ is the number of semantic classes. Following common practice, class 0 (*Other*), which collects noise and returns that carry no semantic label, is excluded from the average.



Fig. G16: Qualitative segmentation results on the *TH-Courier* platform.

Note. From left to right: camera image, Baseline-PointSeg and LiSeer-P trained on nuScenes, Baseline-PointSeg and LiSeer-P trained on SemanticKITTI, and Baseline-PointSeg and LiSeer-P trained on Waymo.

2775 *Optimization, Training, and Testing*

2776 Across all experiments, we retain the original hyper-parameter configurations of
 2777 both baselines without any modification, which further highlights the robustness and
 2778 general applicability of our method.

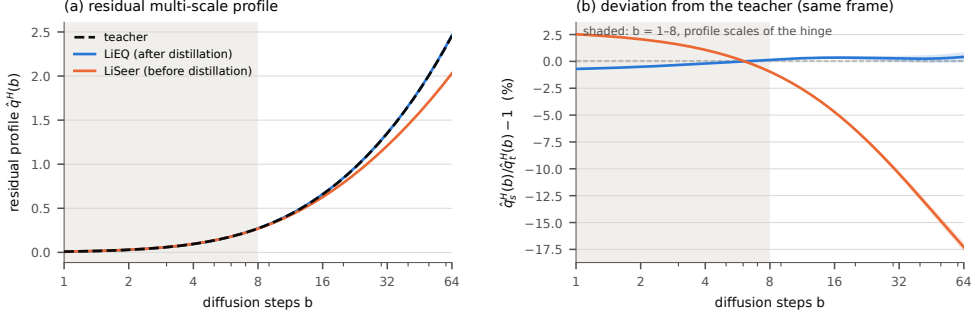


Fig. G17: The student’s multi-scale profile against the teacher’s, before and after distillation, on SemanticKITTI.

Note. The quantity is the residual profile $\hat{q}^H(b)$ of Eq. (24), the same profile the one-sided hinge constrains, read on the common-support field with the subgraph operator and pooled over 200 frames. The diffusion step b is taken continuously rather than at the eight scales the loss uses, and the shaded band marks $b = 1$ to 8 , the range the hinge sees. **a:** the three curves. Inside the shaded band they are hard to tell apart, which is the point of **b**, the per-frame ratio of each student’s profile to the teacher’s on the same frame. The distilled student stays within 0.7% of the teacher across the whole range, including the scales beyond those the loss constrains, whereas the undistilled one runs from 2.5% above the teacher at $b = 1$ to 17% below it at $b = 64$. Evaluated at the integer scales, the two curves reproduce the flatness gaps quoted in Sec. 2.5.

Training uses the Adam optimizer with a weight decay of 0.01. A cosine learning-rate schedule is adopted with a maximum learning rate of 3×10^{-4} , preceded by a linear warm-up over the first 500 iterations. The loss function combines standard cross-entropy with the Lovász-Softmax loss [51], applied jointly at the voxel and point levels.

All experiments are run within the mmdetection3d framework. Models are trained for 24 epochs, with a batch size of 4 for Cylinder3D and 1 for PointSeg, distributed across 4 NVIDIA A800 GPUs.

G.5.2 Class-wise Semantic Segmentation Results

For a more comprehensive view, the class-wise breakdown of the source-to-unseen results is presented in Table G3. As noted in Section 2, the coarse label set reduces category granularity, and both the $S \rightarrow S$ and $S \rightarrow U$ evaluations exhibit pronounced long-tail effects that lead to imbalanced per-class performance. For more intuitive comparison, we additionally plot line and bar charts for all twelve methods on each dataset, as shown in Fig. G19, Fig. G20, and Fig. G21.

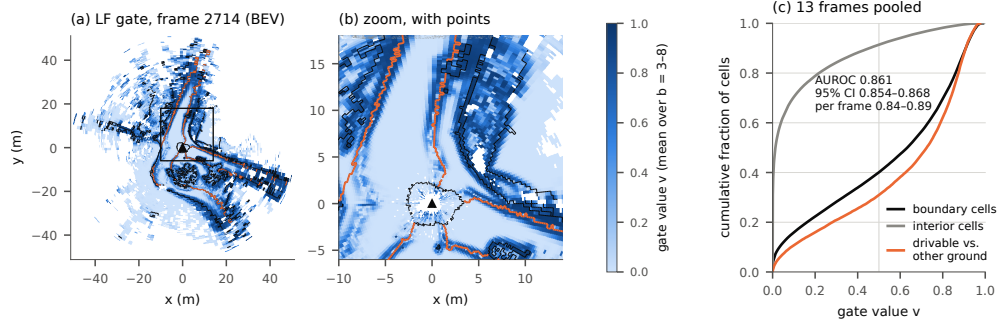


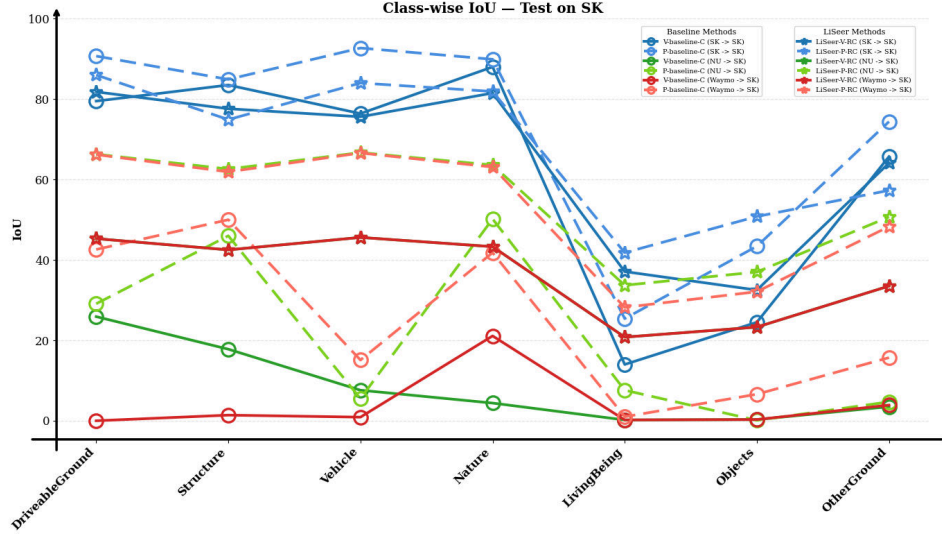
Fig. G18: Where the low-frequency gate opens.

Note. The gate is the teacher-derived probability gate of Sec. 3.4, averaged over the suppression scales $b = 3$ to 8 and computed from the frozen teacher. In the deployed method it acts on the bird's-eye plane alone, so that is the only plane there is to show. **a:** the gate over a full frame, with the junction between drivable and non-drivable ground drawn in orange. **b:** the boxed region, with the frame's own points underneath and class boundaries outlined. **c:** the test, fixed before the numbers were computed. Cells on a class boundary and cells in a class interior are separated by the gate value with a pooled AUROC of 0.861 (95% interval 0.854 to 0.868; 0.84 to 0.89 on each of the 13 frames, one-sided $p = 1.2 \times 10^{-4}$); the gate is open on 60% of boundary cells against 8.6% of interior cells. On the junction between drivable and non-drivable ground, which is where curbs fall, it reaches 0.907 and 69%. The gate is not open on boundaries alone—about a third of the plane's support is open—but it is on boundaries that it concentrates.

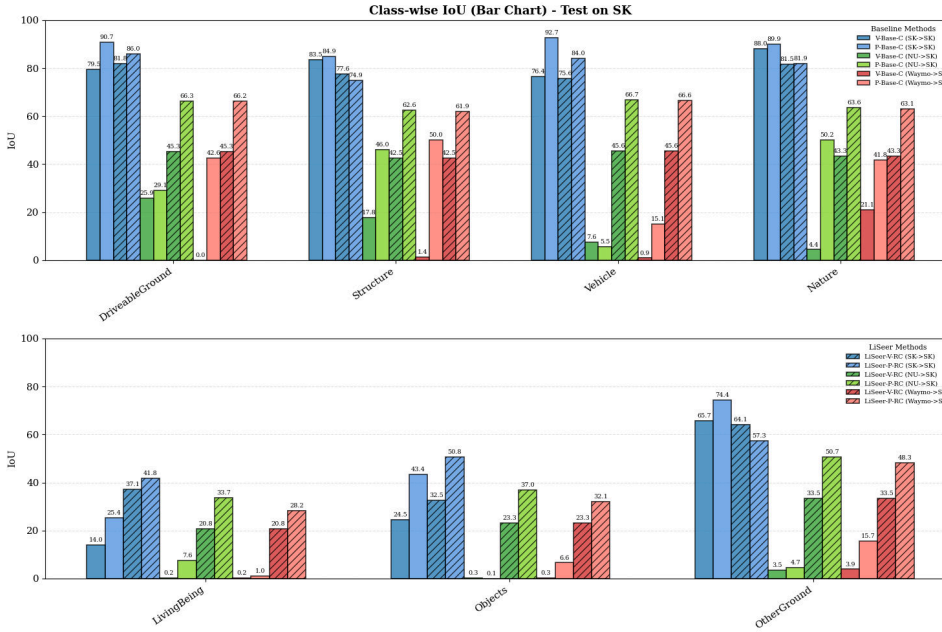
Table G3: Class-wise semantic segmentation results (mIoU, %).

Method	mIoU	Drivable Ground	Structure	Vehicle	Nature	Living Being	Objects	Other Ground
P – baseline – C [†] (SK → SK)	71.6	90.7	84.9	92.7	89.9	25.4	43.4	74.4
LiSeer-P-RSCL (SK → SK)	68.1	86.0	74.9	84.0	81.9	41.8	50.8	57.3
LiEqualizer-P (SK → SK)	75.5	91.1	82.6	95.0	83.9	52.4	39.5	84.1
P – baseline – C [†] (SK → NU)	19.4	43.3	40.3	2.0	34.7	0.6	0.1	15.0
LiSeer-P-RSCL (SK → NU)	56.8	70.8	67.7	70.7	63.2	32.8	40.4	51.7
LiEqualizer-P (SK → NU)	58.3	87.0	69.6	81.7	67.6	25.7	5.3	71.1
P – baseline – C [†] (SK → Waymo)	17.3	42.9	29.3	0.8	28.6	0.4	1.6	17.2
LiSeer-P-RSCL (SK → Waymo)	54.6	66.4	77.3	50.8	72.7	22.1	32.9	55.1
LiEqualizer-P (SK → Waymo)	64.4	81.3	81.7	87.8	70.8	25.7	37.4	66.1
P – baseline – C [†] (NU → SK)	20.5	29.1	46.0	5.5	50.2	7.6	0.1	4.7
LiSeer-P-RSCL (NU → SK)	54.4	66.3	62.6	66.7	63.6	33.7	37.0	50.7
LiEqualizer-P (NU → SK)	55.3	80.4	64.3	89.8	61.1	17.8	2.2	71.2
P – baseline – C [†] (NU → NU)	70.4	90.0	84.6	89.1	82.8	57.5	28.6	60.1
LiSeer-P-RSCL (NU → NU)	66.5	91.7	84.5	77.9	73.8	43.1	44.1	51.3
LiEqualizer-P (NU → NU)	76.1	94.3	83.4	90.0	77.3	56.3	48.7	82.9
P – baseline – C [†] (NU → Waymo)	18.1	39.3	50.8	4.5	24.0	0.1	0.4	7.8
LiSeer-P-RSCL (NU → Waymo)	53.8	65.3	61.7	65.7	62.7	33.8	37.0	50.5
LiEqualizer-P (NU → Waymo)	61.4	84.9	76.9	85.2	66.4	46.1	1.5	68.7
P – baseline – C [†] (Waymo → SK)	24.7	42.6	50.0	15.1	41.8	1.0	6.6	15.7
LiSeer-P-RSCL (Waymo → SK)	52.3	66.2	61.9	66.6	63.1	28.2	32.1	48.3
LiEqualizer-P (Waymo → SK)	63.8	72.9	73.5	90.3	75.0	34.5	30.4	70.0
P – baseline – C [†] (Waymo → NU)	5.4	9.3	19.3	0.7	6.1	0.0	0.0	2.3
LiSeer-P-RSCL (Waymo → NU)	49.0	61.4	57.5	61.7	58.6	26.4	30.9	65.3
LiEqualizer-P (Waymo → NU)	58.5	82.1	71.6	86.4	66.8	36.8	2.8	62.9
P – baseline – C [†] (Waymo → Waymo)	60.9	77.7	77.7	61.7	80.9	33.0	28.0	67.4
LiSeer-P-RSCL (Waymo → Waymo)	62.2	72.3	74.6	63.2	72.1	44.0	50.0	58.0
LiEqualizer-P (Waymo → Waymo)	79.2	92.4	89.6	94.4	81.9	63.2	50.7	82.1

Note. Results correspond to the source-to-unseen setting. †: reproduced by the authors. -V / -P: voxel-based / projection-based backbone. -R / -C: physics-based random sampling and label mapping. →: generalization from source to target domain. Highlighted cells encode mIoU magnitude as elsewhere in this paper.



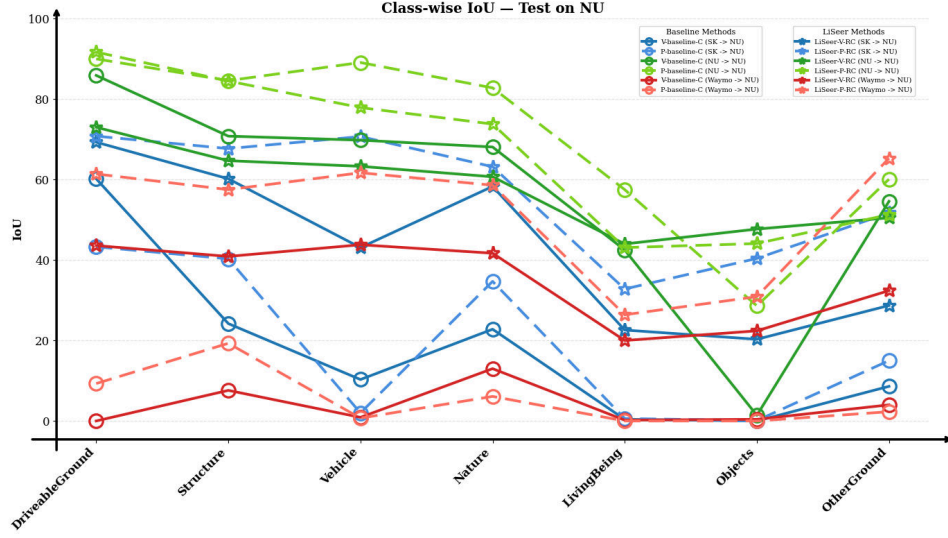
(a) SemanticKITTI (line chart).



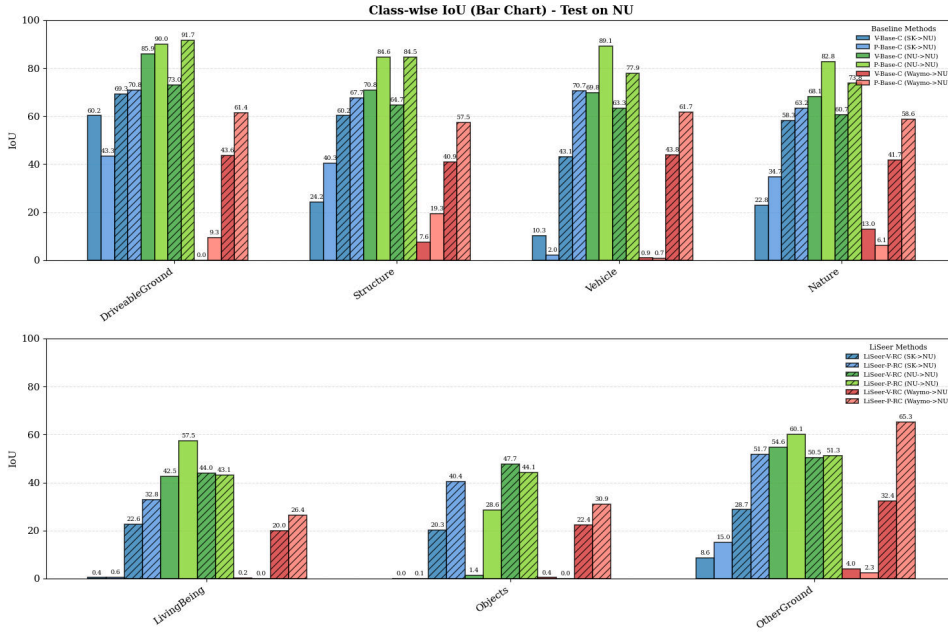
(b) SemanticKITTI (bar chart).

Fig. G19: Class-wise results on SemanticKITTI.

Note. Line and bar charts visualize the same class-wise IoU metrics in different styles.



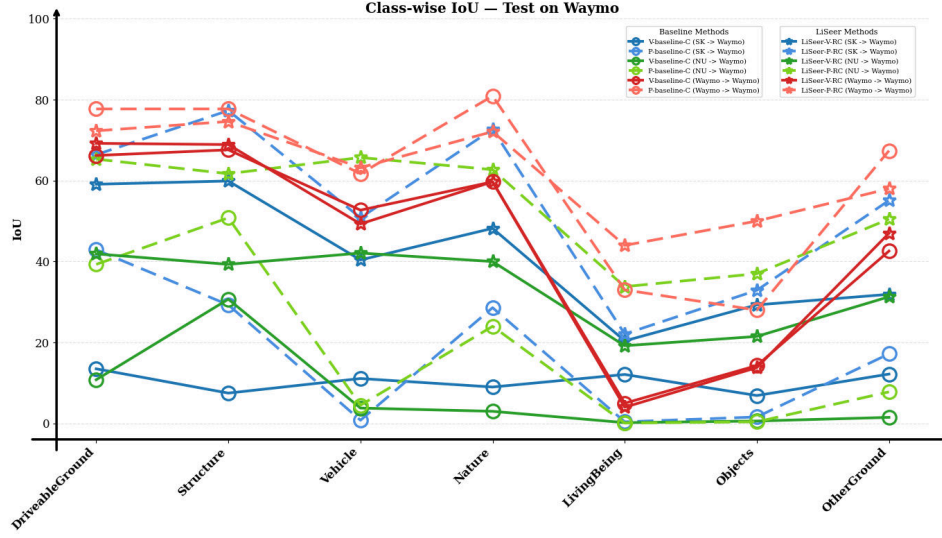
(a) nuScenes (line chart).



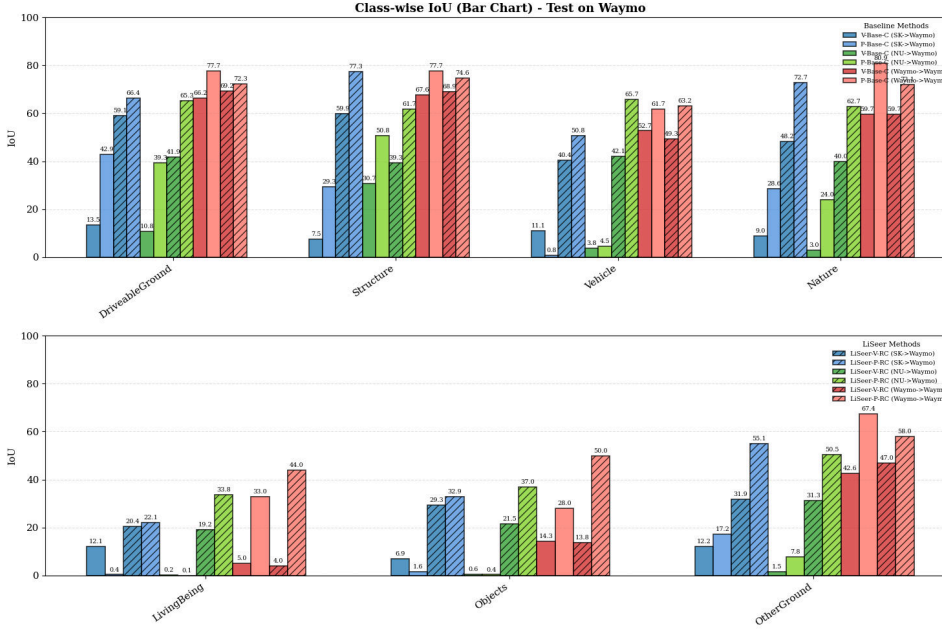
(b) nuScenes (bar chart).

Fig. G20: Class-wise results on nuScenes.

Note. Line and bar charts visualize the same class-wise IoU metrics in different styles.



(a) Waymo (line chart).



(b) Waymo (bar chart).

Fig. G21: Class-wise results on Waymo.

Note. Line and bar charts visualize the same class-wise IoU metrics in different styles.