Supplementary Materials for

UCRG: Reliability-Gated Multi-Source Prediction Integration for Forest Point Cloud Semantic Segmentation Across Sensing Platforms

Licheng Liu[1,†], Yu Li[1,†], Qian Zhou[2], Hua Zou[2,*], Xining Zhang[2], Xi Chen[3], and Bin Xie[3]

[1] School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430077, China

[2] School of Computer Science, Wuhan University, Wuhan, Hubei 430072, China

[3] Economic and Technical Research Institute, State Grid Shaanxi Electric Power Co., Ltd., Xi'an 710003, China

[*] Correspondence: zouhua@whu.edu.cn

[†] These authors contributed equally to this work and share first authorship.

Reproducibility status. All numerical results were cross-checked against the experiment scripts, frozen checkpoints, aligned probability outputs, and evaluation logs used for the manuscript. This supplement records the data roles, candidate-model provenance, parameter-selection rules, random seeds, aggregation procedures, and artifact identifiers required to reproduce the reported experiments. Environment and provenance details are provided in Section S7.

S1 Reproducibility Scope

The study separates candidate-model training, reliability fitting, parameter selection, and query evaluation. Candidate predictors are trained before integration and remain frozen thereafter. The reliability-fitting subset estimates source reliability and fits the PairMLP, whereas the disjoint parameter-selection subset determines gate parameters and refinement thresholds. Unless explicitly identified as a retrospective diagnostic, query labels are used only for final evaluation.

The canonical class orders are Ground, Low vegetation, Wood, and Crown for FOR-instance; Ground, Understorey, Wood, and Foliage for EvoMLS; and Ground, Lower objects, Wood, and Foliage for Lin3D. Lin3D uses one labeled support plot and three query plots per episode and is therefore a target-assisted transfer protocol rather than an unsupervised domain-generalization setting.

S1.1 Leakage-Controlled Shared-Parameter Selection

Shared UCRG selects its gate parameters using only the fixed and disjoint FOR-instance and EvoMLS calibration protocols. Lin3D is excluded because its annotated plots rotate between support and query roles across the four leave-one-plot episodes. The selected setting is therefore fixed before any Lin3D evaluation.

The predefined search grid contained 100 settings: $\mathcal{T} = \{0.01, 0.03, 0.05, 0.10, 0.15\}$, $\mathcal{B} = \{0, 0.25, 0.50, 0.75, 1.00\}$, and $\Gamma = \{0, 0.02, 0.05, 0.10\}$. Ties were resolved by mean relative regret, maximum relative regret, mean internal mIoU, and then smaller γ, τ, and β, in that order.

Table S1: Leakage-controlled shared-parameter selection audit.

Item	Value
Selection datasets	FOR-instance and EvoMLS
Held-out selection dataset	Lin3D
Selected (τ, β, γ)	$(0.01, 0, 0.02)$
Mean internal calibration mIoU (%)	73.7560
Mean relative regret	0.0006675
Maximum relative regret	0.0013350
Query-label access during selection	No

S2 Dataset-Specific Customized Reference Strategies

For completeness, Table S2 compares Shared UCRG with separately designed references: source-conditioned Height–AGF/SDEM fusion on FOR-instance, CSO+MSO+BR on EvoMLS, and target-plot branch selection on Lin3D. These references use different decision rules and are therefore reported separately from the shared-protocol comparison.

Table S2: Dataset-specific reference pipelines and Shared UCRG. Difference is computed as customized-reference mIoU minus Shared UCRG mIoU and is reported in percentage points. The Lin3D values are averaged over four leave-one-plot episodes, and the cross-dataset average is the arithmetic mean of the three reported protocol-level values.

Dataset	Reference strategy	Reference	UCRG	Difference
FOR-instance	Source-conditioned fusion	79.1913	78.8306	0.3607
EvoMLS	Complete CSO+MSO+BR pipeline	79.2342	77.9780	1.2562
Lin3D	Target-plot branch selection	55.4751	55.2329	0.2422
Cross-dataset average	Dataset-specific strategy for each protocol	71.3002	70.6805	0.6197

The comparison quantifies the modest performance cost of using one reusable decision protocol instead of separately optimized dataset-specific pipelines.

S3 FOR-instance Experiments

S3.1 Candidate Training Configuration

Table S3 reports the settings shared by the frozen Height and AGF/SDEM candidates. The models retain their respective feature configurations and native five-class training organization; their probability channels are mapped to the canonical four-class interface only after inference.

Table S3: Shared FOR-instance candidate-training configuration recorded from the training script.

Parameter	Value	Parameter	Value
Kernel points	15	Input radius	3.0 m
First subsampling grid	0.06 m	Convolution radius	2.5
Deformable radius	5.0	Kernel-point extent	1.2
First feature width	128	Input feature dimension	4
Class weights	$[1.5, 2.0, 2.0, 1.2, 2.0]$	Maximum epochs	40
Initial learning rate	10^{-3}	Momentum	0.98
Batch size	4	Steps per epoch	500
Validation steps	100	Vertical rotation	Enabled
Anisotropic scaling	Enabled	Scale range	$[0.9, 1.1]$
Coordinate noise	0.001	Color augmentation	0.8

S3.2 Source-Conditioned Reference Fusion

This audit documents the FOR-instance source-conditioned fusion reference reported in the main manuscript and in Table S2. The rows correspond to acquisition sources rather than candidate prediction models. CULS, NIBIO, and TUWIEN have labeled calibration support, whereas RMIT and SCION are absent from the calibration data used in this audit.

Let $\mathbf{q}_i^{(H)}$ and $\mathbf{q}_i^{(A)}$ denote the aligned probability vectors produced by the Height and AGF/SDEM candidates, respectively. For an acquisition source with calibration support, the fused probability vector is defined as

$$\mathbf{q}_i^{\text{fuse}}(\alpha_H) = \alpha_H \mathbf{q}_i^{(H)} + (1 - \alpha_H)\mathbf{q}_i^{(A)}, \qquad \alpha_H \in \Omega_H, \tag{S1}$$

where

$$\Omega_H = \{0, 0.05, \ldots, 1\}. \tag{S2}$$

The Height weight is selected by maximizing the calibration mIoU over the predefined grid:

$$\alpha_H^* = \arg\max_{\alpha_H \in \Omega_H} \text{mIoU}\left(\widehat{\mathbf{y}}^{\text{fuse}}(\alpha_H); \mathcal{D}_{\text{cal}}^{(a)}\right), \tag{S3}$$

where $\mathcal{D}_{\text{cal}}^{(a)}$ denotes the labeled calibration data associated with acquisition source a. For an acquisition source absent from calibration, the prespecified fallback $\alpha_H = 0.45$ is used without source-specific weight selection. The calibration mIoU values, selected weights, and support status are summarized in Table S4.

Table S4: Source-conditioned reference fusion on FOR-instance calibration data. Height, AGF/SDEM, and Fusion values are mIoU (%). The Height weight is selected from $\Omega_H = \{0, 0.05, \ldots, 1\}$ when labeled calibration support is available; otherwise, the prespecified fallback $\alpha_H = 0.45$ is used. "Support" indicates available calibration data, whereas "Fallback" indicates use of the fixed default weight.

Acquisition source	Height	AGF/SDEM	Height weight	Fusion	Status
CULS	69.43	68.62	0.70	69.56	Support
NIBIO	63.28	60.05	0.60	63.71	Support
TUWIEN	42.31	47.79	0.00	47.79	Support
RMIT	–	–	0.45	–	Fallback
SCION	–	–	0.45	–	Fallback

S3.3 Wood/Crown Refiner for the FOR-instance Reference Pipeline

The source-conditioned FOR-instance reference pipeline applies a lightweight binary refiner after the Height–AGF/SDEM fusion. The refiner is evaluated only on points whose base label is Wood

or Crown and whose fused Wood–Crown probability mass satisfies

$$q^{\text{fuse}}_{i,\text{Wood}} + q^{\text{fuse}}_{i,\text{Crown}} \geq \eta_{\text{WC}}, \qquad \eta_{\text{WC}} = 0.50. \tag{S4}$$

It predicts a correction between Wood and Crown for this restricted candidate set. The training and application settings are listed in Table S5.

Table S5: Wood/Crown refiner configuration for the FOR-instance source-conditioned reference pipeline.

Parameter	**Value**
Maximum samples per class	300,000
Hidden width	64
Training epochs	12
Batch size	8,192
Learning rate	10^{-3}
Weight decay	10^{-4}
Random seed	12,345
Candidate mass threshold η_{WC}	0.50
Evaluated confidence thresholds ξ_{WC}	0.90, 0.95, 0.97, 0.99
Output classes	Wood/Crown

Leave-one-validation-plot-out calibration selected $\xi_{\text{WC}} = 0.99$. With this threshold frozen, the final query evaluation changed 79 points and retained 79.1913% mIoU. This calibration-selected result is used throughout the reported comparison.

S4 EvoMLS Experiments

S4.1 Candidate Models and Class-Wise Calibration Reliability

Four candidate prediction sources were constructed by combining two training objectives with two random-seed realizations. The two objectives used the original Raw and Ground-weighted class-weight configurations, while the additional model pair was trained with seed 2027. All four models used the same KPConv architecture and optimization settings, including 11 input features, a first feature width of 128, an initial learning rate of 10^{-3}, momentum 0.98, batch size 4, 500 steps per epoch, and 80 training epochs. The entries in the class-weight vectors follow the canonical order Ground, Understorey, Wood, and Foliage.

The Raw and Ground-weighted objectives used class-weight vectors $[1.5, 1.2, 2.0, 1.0]$ and $[3.0, 1.2, 2.0, 1.0]$, respectively.

Figure S1 reports the class-wise IoU values measured on the calibration split. These values provide the empirical reliability evidence used by the subsequent decision procedures. The annotated heatmap shows that no single candidate is uniformly strongest across all classes: the base KPFCNN performs best for Ground and Understorey, whereas the seed-2027 base KPFCNN performs best for Wood and Foliage. This class-dependent source ordering motivates class-wise reliability weighting and explains why class-independent fusion or whole-source selection may fail to exploit the available complementarity.

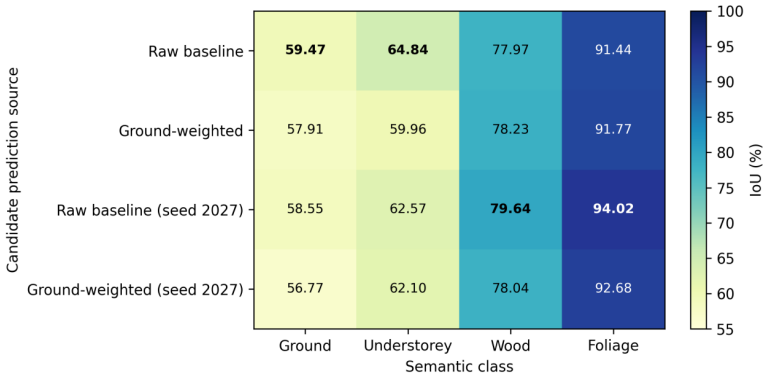


Figure S1: Class-wise calibration IoU of the four EvoMLS candidate prediction sources. The annotated values are IoU percentages, and darker colors indicate higher class-specific reliability.

S4.2 CSO/MSO Overrides and Ground/Understorey Refiner

The EvoMLS customized reference applies two calibration-driven selective overrides followed by a binary Ground/Understorey refiner. CSO allows the Ground-weighted candidate to replace Wood and Foliage predictions at thresholds of 0.80 and 0.50, respectively. MSO then applies the designated seed-2027 Foliage candidate and the seed-2027 Raw Understorey candidate, both at a threshold of 0.50. All thresholds are selected from calibration data before query evaluation.

The binary refiner is trained on Ground/Understorey calibration points and is applied only where the MSO prediction belongs to that class pair. Its configuration is reported in Table S6.

Table S6: Ground/Understorey refiner configuration for the EvoMLS CSO+MSO+BR reference pipeline.

Parameter	Value
Input dimension	71
Architecture	71–96–64–2
Dropout	0.05
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Maximum training samples	2,000,000
Sampling	Balanced Ground/Understorey
Training epochs	8
Training batch size	65,536
Evaluation batch size	524,288
Threshold grid	0.40–0.95, step 0.05
Selected Ground threshold	0.65
Selected Understorey threshold	0.40

The Ground and Understorey thresholds of 0.65 and 0.40 were selected from calibration data only. The frozen checkpoint, `test/evomls_gu_refiner_v4_multiseed_val.pt`, achieved 76.67% calibration mIoU and has SHA-256 value `2fc0ed5ec418554e995cc441b63311dd0eb9a87413f26d4db7fdbda675e27da1`. The complete CSO+MSO+BR pipeline changed 203,882 query points and achieved 79.2342% four-class mIoU. Its separate Wood/Foliage-only diagnostic result was 92.05% and is not compared directly with the four-class protocol.

S5 Cross-Dataset Confusion-Pair Selection Audit

This audit supports the optional UCRG confusion-pair experiment reported in the main manuscript and is distinct from the dataset-specific reference pipelines above. For each pair, the PairMLP

threshold is selected from calibration data and frozen before query evaluation. In both datasets, the manually specified pair coincides with the calibration-selected pair.

Table S7: Manual and seeded-random confusion-pair references. Refinement gains are mIoU points over unrefined Shared UCRG, reported as mean ± standard deviation over three PairMLP seeds. The seeded-random pair is drawn once before query evaluation and does not represent an average over repeated random pair selections.

Dataset	Selection rule	Class pair	mIoU gain
FOR-instance	Manual/automatic	Wood/Crown	0.5014 ± 0.0480
FOR-instance	Seeded random	Low vegetation/Crown	-0.0081 ± 0.0009
EvoMLS	Manual/automatic	Ground/Understorey	0.1059 ± 0.0622
EvoMLS	Seeded random	Understorey/Foliage	0.3142 ± 0.0376

On EvoMLS, the seeded-random pair provides the largest query gain, consistent with the weaker calibration-to-query rank correlation reported in the main manuscript.

S6 Lin3D Target-Plot Reliability Experiment

S6.1 Support/Query Episodes

Lin3D contains four leave-one-plot episodes. In each episode, one fully labeled plot supplies source-reliability evidence and the remaining three plots form the query set. Support labels are not used for model fine-tuning or feature adaptation, and query labels are used only for final evaluation. Because each plot appears in three query sets, episode results are not statistically independent.

Table S8: Lin3D leave-one-plot-out support/query episodes.

Fold	Support plot	Support points	Query plots	Query points
1	Plot 1	403,281	Plots 2, 3, 4	1,105,668
2	Plot 2	395,413	Plots 1, 3, 4	1,113,536
3	Plot 3	343,590	Plots 1, 2, 4	1,165,359
4	Plot 4	366,665	Plots 1, 2, 3	1,142,284

S6.2 Absolute Calibration-Label Budgets

Table S9 reports the per-class label counts used in the Lin3D calibration-label budget experiment. At each fraction below 100%, ten class-stratified samples were drawn per support plot. The class quotas, and therefore the total counts, are identical across repeats; the sampled point identities vary.

Table S9: Absolute calibration-label counts for each Lin3D leave-one-plot episode and budget fraction.

Fold	Budget	Total	Ground	Lower objects	Wood	Foliage
1	1%	4,033	231	40	324	3,438
1	5%	20,164	1,155	200	1,621	17,188
1	10%	40,328	2,311	400	3,242	34,375
1	25%	100,820	5,776	1,001	8,105	85,938
1	50%	201,641	11,553	2,002	16,210	171,876
1	100%	403,281	23,106	4,004	32,419	343,752
2	1%	3,954	204	12	361	3,377
2	5%	19,771	1,021	61	1,805	16,884
2	10%	39,541	2,042	121	3,610	33,768
2	25%	98,853	5,106	303	9,024	84,420
2	50%	197,706	10,212	606	18,048	168,840
2	100%	395,413	20,424	1,213	36,096	337,680
3	1%	3,436	118	38	222	3,058
3	5%	17,179	588	189	1,112	15,290
3	10%	34,358	1,176	378	2,224	30,580
3	25%	85,898	2,941	946	5,560	76,451
3	50%	171,795	5,882	1,892	11,119	152,902
3	100%	343,590	11,763	3,785	22,238	305,804
4	1%	3,667	167	57	264	3,179
4	5%	18,334	834	286	1,319	15,895
4	10%	36,667	1,668	571	2,638	31,790
4	25%	91,666	4,170	1,428	6,594	79,474
4	50%	183,333	8,341	2,855	13,188	158,949
4	100%	366,665	16,682	5,710	26,375	317,898

S6.3 Frozen Candidates and Label Mapping

The Height and AGF/SDEM branches were trained on FOR-instance and then transferred to Lin3D without target-domain fine-tuning. Their native five-channel outputs were mapped to the four-class Lin3D interface as

$$\text{Ground} \rightarrow \text{Ground}, \qquad \text{Low vegetation} \rightarrow \text{Lower objects},$$

$$\text{Stem} + \text{Woody branch} \rightarrow \text{Wood}, \qquad \text{Crown vegetation} \rightarrow \text{Foliage}.$$

Accordingly, the Wood probability channel was obtained by summing the Stem and Woody-branch probabilities before reliability estimation and query evaluation. The same mapping was applied to support and query predictions.

Table S10: Frozen candidate branches used in the FOR-instance-to-Lin3D transfer experiment.

Item	Height branch	AGF/SDEM branch
Source training dataset	FOR-instance	FOR-instance
Checkpoint identifier	`Log_2026-06-16_09-13-32`	`Log_2026-06-19_13-30-24`
Native output channels	5	5
Target-domain fine-tuning	No	No
Inference votes	20	20
Archived checkpoint variants	Current and best	Current and best

S6.4 Per-Episode Reliability and Query Evaluation

For episode j, the support-set reliability of candidate branch b is defined as its four-class mIoU:

$$R_j(b) = \mathrm{mIoU}\left(b; \mathcal{D}_{t,j}^{\mathrm{sup}}\right). \tag{S5}$$

The branch with the higher support-set reliability is then applied unchanged to the query plots of episode j. Query labels are used only to compute the evaluation results reported below.

Table S11: Per-episode Lin3D support reliability and selected branch. Height and AGF/SDEM reliability values are four-class mIoU (%) measured on the corresponding support plot.

Fold	Support	Height	AGF/SDEM	Selected branch
1	Plot 1	29.82	59.67	AGF/SDEM
2	Plot 2	26.15	54.09	AGF/SDEM
3	Plot 3	25.75	56.90	AGF/SDEM
4	Plot 4	22.44	53.53	AGF/SDEM

Table S12: Per-episode Lin3D query mIoU (%) from the frozen candidate probabilities. Calibration-based hard selection and Calibration-tuned UCRG coincide with the AGF/SDEM output in all four episodes. Shared UCRG uses the externally selected parameterization. The final row reports the episode mean and standard deviation.

Fold	Height	AGF/SDEM (hard/tuned)	Fixed fusion	Shared UCRG
1	24.62	53.90	41.76	53.90
2	26.02	56.72	44.28	55.75
3	26.10	55.40	43.24	55.40
4	27.11	55.88	44.09	55.88
Mean ± SD	25.96 ± 1.02	55.48 ± 1.18	43.34 ± 1.15	55.23 ± 0.91

Table S13: Pooled FOR-instance-to-Lin3D transfer results obtained by summing the four query confusion matrices.

Decision output	Pooled mIoU (%)
Height branch	25.95
AGF/SDEM branch	55.35
Fixed fusion	43.31
Calibration-based hard selection	55.35

Because the query plots overlap across the four leave-one-plot-out episodes, the reported standard deviation describes episode-level variation rather than uncertainty estimated from independent trials. Pooled metrics are computed from

$$\mathbf{C}_{\mathrm{pool}} = \sum_{j=1}^{4} \mathbf{C}_j^{\mathrm{qry}}. \tag{S6}$$

Each plot appears in the query set of three episodes and therefore contributes equally often to the pooled evaluation.

AGF/SDEM is selected in all four episodes, so the calibration-based hard-selection output matches this branch and substantially outperforms fixed fusion.

S6.5 Four-Architecture Candidate-Pool Extension

To evaluate UCRG across heterogeneous network architectures, we constructed a Lin3D candidate pool containing PointNet, KPFCNN, RandLA-Net, and Sen-net. All four predictors were trained on FOR-instance, frozen before Lin3D inference, and evaluated on the same four annotated Lin3D plots. In each leave-one-plot-out episode, one plot provided labeled calibration support and the remaining three plots formed the query set. Shared UCRG used the same calibration-supported decision procedure as in the primary protocol, but the four-architecture candidate pool was evaluated separately and was not included in the three-dataset aggregate results.

Table S14: Frozen four-architecture candidate-pool extension on Lin3D. Pooled OA (%) and pooled mIoU (%) are reported. Episode results are reported as mean mIoU ± standard deviation (%) over four leave-one-plot query episodes.

Method	Pooled OA	Pooled mIoU	Episode result
PointNet	12.55	24.26	24.26 ± 1.53
KPFCNN	46.19	25.95	25.96 ± 1.02
RandLA-Net	3.33	5.65	5.64 ± 0.26
Sen-net	91.92	47.40	47.35 ± 0.90
Shared UCRG	**93.48**	**57.44**	$\mathbf{57.49 \pm 2.07}$

No leave-one-plot episode falls below the strongest individual candidate.

S6.6 PTv3-Augmented Lin3D Candidate Pool

PTv3 was trained on the official FOR-instance training split using XYZ-only input, a 0.06 m grid, and a point limit of 131,072. Optimization used AdamW with an initial learning rate of 0.003 and weight decay of 0.05, a OneCycle learning-rate schedule, batch size 2, and cross-entropy plus Lovasz losses. The recorded seed was 56395375, and the configuration performed 200 effective training-data loops within one outer evaluation cycle. The checkpoint with the best FOR-instance validation mIoU of 45.3449% was frozen before Lin3D inference.

The native output order was Ground, Low vegetation, Stem, Crown, and Woody branch. Stem and Woody-branch probabilities were summed into Wood, while Crown was mapped to Foliage. No Lin3D label was used for PTv3 training or checkpoint selection.

Table S15: Frozen PTv3 transfer results on the canonical four-class Lin3D interface.

Metric	Value
OA (%)	91.9022
mIoU (%)	53.2326
Ground IoU (%)	49.8831
Lower objects IoU (%)	12.3565
Wood IoU (%)	58.2004
Foliage IoU (%)	92.4906
Evaluated points	1,508,949

The frozen PTv3 probabilities were subsequently added to the Height and AGF/SDEM candidate pool. Shared UCRG uses the externally selected setting $(0.01, 0, 0.02)$, whereas Calibration-tuned UCRG selects its gate parameters from the support plot of each episode.

Table S16: Fold-level PTv3-augmented results and calibration-tuned gate parameters.

Support plot	Calibration-tuned	Shared UCRG	Θ_{tuned}
Plot 1	60.8674	54.3181	(0.05, 0.75, 0.02)
Plot 2	54.5319	53.6178	(0.10, 0.00, 0.10)
Plot 3	59.2263	55.4048	(0.10, 0.00, 0.10)
Plot 4	62.1601	57.6653	(0.05, 0.75, 0.02)
Mean ± SD	59.1964 ± 3.3334	55.2515 ± 1.7692	–

S7 Artifact Provenance and Environment

Hardware and software environment. All experiments were conducted on a Linux 5.15.0-67-generic server equipped with eight NVIDIA GeForce RTX 3090 GPUs. The primary KPConv experiments used Python 3.8.20, NumPy 1.21.6, PyTorch 1.7.0, CUDA 11.0, and cuDNN 8.0.3. The Height and AGF/SDEM Lin3D probability outputs were generated using 20 inference votes, as verified from the archived evaluation logs.

The PTv3 experiments were conducted in a separate Pointcept environment using Python 3.10.21, NumPy 2.2.6, PyTorch 2.5.0+cu124, CUDA 12.4, and cuDNN 9.1.0 on an NVIDIA GeForce RTX 3090 GPU. The associated configuration, checkpoint, probability, and result-file checksums are recorded in the provenance manifests listed below.

Artifact verification. Complete artifact directories, file-level SHA-256 checksums, checkpoint metadata, and experiment records are stored in the machine-readable provenance manifests listed in Table S17. These records cover the labeled Lin3D plots, frozen candidate probabilities, archived checkpoints, PTv3 outputs, and UCRG result files. All paths are relative to the accompanying reproducibility archive.

Table S17: Machine-readable provenance records included in the reproducibility archive.

Artifact group	Verification record
Primary Lin3D labels, Height/AGF/SDEM probabilities, and checkpoints	`test/lin3d_server_provenance.json`
PTv3 configuration, checkpoint, native and mapped probabilities	`test/ptv3_reproducibility_bundle/ptv3_artifact_sha256.txt`
PTv3 environment, configuration, and checkpoint metadata	`test/ptv3_reproducibility_bundle/ptv3_environment_config_checkpoint.json`
PTv3 probability preflight and three-source UCRG results	`test/ptv3_reproducibility_bundle/ptv3_probability_and_ucrg_results.json`
Leakage-controlled shared-parameter selection	`test/ucrg_shared_selection_for_evo/selection_protocol.json`
Shared-parameter and component-ablation protocol	`test/ucrg_leakage_controlled_ablations/component_ablation_protocol.json`

S8 Additional Validation and Audit Analyses

S8.1 UCRG Gate-Parameter Seed Audit

Table S18: Ten-seed audit of Shared UCRG under point-level 50:50 class-stratified calibration fit–selection partitions. The parameter columns show the shared triple selected jointly from FOR-instance and EvoMLS and applied to all query protocols. Reliability is recomputed from complete calibration support before query evaluation. Values are mean ± SD across seeds; Lin3D metrics are first averaged over four episodes within each seed. The primary aggregate averages the three protocol-level mIoU values, and negative transfer is counted over six episodes per seed.

Scope	τ	β	γ	Query mIoU	Query OA	Neg.
FOR-instance	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	78.8306 ± 0.0000	93.0437 ± 0.0000	1/1
EvoMLS	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	77.9780 ± 0.0000	91.7402 ± 0.0000	0/1
Lin3D	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	55.2329 ± 0.0000	84.7541 ± 0.0000	1/4
Primary aggregate	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	70.6805 ± 0.0000	—	2/6

S8.2 Calibration-Based Ensemble Baselines

Calibration-optimized class-wise fusion fits one source-weight vector per class on the complete calibration support. With $z_{i,c} = \sum_s w_{s,c} q_{i,c}^{(s)}$, it minimizes the mean loss $-\log[z_{i,y_i} / \sum_k z_{i,k}]$, subject to $w_{s,c} \geq 0$ and $\sum_s w_{s,c} = 1$. SciPy SLSQP starts from uniform weights, with a maximum of 300 iterations and function tolerance 10^{-9}. Scores are clipped at 10^{-12} for loss evaluation. Query predictions are $\arg\max_c z_{i,c}$.

KNORA-E uses seven Euclidean nearest neighbors in the space of concatenated frozen source-probability vectors. For each episode, its competence set contains 100,000 labeled calibration points sampled without replacement in proportion to class and scene counts. The run uses base seed 7, with episode sampling seeds derived as $7 + 3001 + 7919f$, where f is the zero-based fold index. Sources correct on all neighbors are retained; if none qualifies, the farthest neighbor is removed until a source qualifies. If none qualifies even at one neighbor, sources with maximal accuracy over the original neighborhood are retained. Selected sources predict by majority vote, with ties resolved by summed source probabilities and then canonical class order. Query labels are used only for final evaluation.

Run settings, competence-set class counts, and query results are recorded in `test/reviewer1_calibration_baselines_with_knorae_20261009_185106/`.

S8.3 Source Probability Calibration Audit

Table S19: Source-level calibration metrics before and after probability-space temperature recalibration. Temperatures were fitted by minimizing NLL on the calibration fit subset; ECE, Brier score, and NLL are reported on the disjoint calibration selection subset. Paired values are raw/recalibrated.

Dataset/fold	Source	T	ECE	Brier	NLL
FOR-instance	Height	1.0996	0.0216/0.0207	0.1819/0.1811	0.3182/0.3178
FOR-instance	AGF/SDEM	0.8437	0.0270/0.0060	0.1764/0.1749	0.3130/0.3098
EvoMLS	Raw baseline	0.7071	0.0559/0.0166	0.1992/0.1912	0.3404/0.3243
EvoMLS	Ground-weighted	0.9216	0.0131/0.0122	0.2018/0.2017	0.3470/0.3458
EvoMLS	Raw seed 2027	0.7724	0.0415/0.0055	0.1888/0.1848	0.3342/0.3189
EvoMLS	Ground-weighted seed 2027	0.7724	0.0396/0.0050	0.1983/0.1945	0.3381/0.3259
Lin3D, fold 1	Height	1.0067	0.1247/0.1254	0.5410/0.5408	0.8667/0.8666
Lin3D, fold 1	AGF/SDEM	0.4163	0.1749/0.0055	0.1970/0.1400	0.3929/0.2630
Lin3D, fold 2	Height	1.5655	0.2518/0.2088	0.7039/0.6395	1.0722/1.0146
Lin3D, fold 2	AGF/SDEM	0.7071	0.0996/0.0486	0.3846/0.3747	0.6362/0.6088
Lin3D, fold 3	Height	1.2011	0.2020/0.1950	0.6390/0.6204	0.9734/0.9631
Lin3D, fold 3	AGF/SDEM	0.3489	0.1945/0.0059	0.1936/0.1251	0.3843/0.2174
Lin3D, fold 4	Height	1.7100	0.2946/0.2127	0.7454/0.6602	1.1294/1.0632
Lin3D, fold 4	AGF/SDEM	0.4547	0.1660/0.0100	0.2733/0.2233	0.4920/0.3764

For each metric, the first value is before and the second is after recalibration. Lower values indicate better calibration.

S8.4 Class-Wise Query Audit

Table S20: Fold-level Lin3D query class-wise IoU (%) for the primary integration protocols. Class order is Ground, Lower objects, Wood, and Foliage.

Fold	Method	Ground	Lower objects	Wood	Foliage
1	Fixed-weight fusion	49.7282	3.7041	44.6079	68.9939
1	Calibration-optimized class-wise fusion	71.1483	5.6747	56.8230	81.9695
1	KNORA-E dynamic ensemble selection	68.5447	6.0796	55.5203	82.1091
1	Calibration-based hard selection	71.1483	5.6747	56.8230	81.9695
1	Calibration-tuned UCRG	71.1483	5.6747	56.8230	81.9695
1	Shared UCRG	71.1483	5.6747	56.8230	81.9695
2	Fixed-weight fusion	49.6343	6.8636	42.8404	77.7628
2	Calibration-optimized class-wise fusion	69.4264	12.5040	55.9330	89.0026
2	KNORA-E dynamic ensemble selection	65.9428	12.2891	54.2604	88.6990
2	Calibration-based hard selection	69.4264	12.5040	55.9330	89.0026
2	Calibration-tuned UCRG	69.4264	12.5040	55.9330	89.0026
2	Shared UCRG	66.9898	11.6472	55.9130	88.4406
3	Fixed-weight fusion	50.8889	3.4453	48.4129	70.2254
3	Calibration-optimized class-wise fusion	73.9168	5.0073	60.7876	81.9076
3	KNORA-E dynamic ensemble selection	73.3427	5.3626	59.3325	81.9895
3	Calibration-based hard selection	73.9168	5.0073	60.7876	81.9076
3	Calibration-tuned UCRG	73.9168	5.0073	60.7876	81.9076
3	Shared UCRG	73.9168	5.0073	60.7876	81.9076
4	Fixed-weight fusion	51.3229	3.4042	47.5018	74.1237
4	Calibration-optimized class-wise fusion	74.3319	4.6426	60.3602	84.1654
4	KNORA-E dynamic ensemble selection	73.3051	5.0594	58.4776	84.2025
4	Calibration-based hard selection	74.3319	4.6426	60.3602	84.1654
4	Calibration-tuned UCRG	74.3319	4.6426	60.3602	84.1654
4	Shared UCRG	74.3319	4.6426	60.3602	84.1654

Ground-truth query support, in the class order above:
Fold 1: (48,869, 10,708, 84,709, 961,382).
Fold 2: (51,551, 13,499, 81,032, 967,454).
Fold 3: (60,212, 10,927, 94,890, 999,330).
Fold 4: (55,293, 9,002, 90,753, 987,236).

S8.5 Calibration-Budget Gate Audit

Table S21: Repeat-level audit of the Lin3D calibration-label budget experiment. The hard-selection mask and selected source are summarized across ten repeats for each support plot at fractions below 100%. Source-weight ranges and complete repeat-level records are provided in the accompanying CSV.

Fraction	Repeats/plot	Hard mask	Selected source	Height soft-weight range	AGF/SDEM soft-weight range
1%	10	(1, 1, 1, 1)	AGF/SDEM for all classes	See CSV	See CSV
5%	10	(1, 1, 1, 1)	AGF/SDEM for all classes	See CSV	See CSV
10%	10	(1, 1, 1, 1)	AGF/SDEM for all classes	See CSV	See CSV
25%	10	(1, 1, 1, 1)	AGF/SDEM for all classes	See CSV	See CSV
50%	10	(1, 1, 1, 1)	AGF/SDEM for all classes	See CSV	See CSV
100%	1	(1, 1, 1, 1) in three plots; (1, 0, 1, 1) in one plot	AGF/SDEM where hard-selected; soft fusion for Lower objects in one plot	See CSV	See CSV

The four mask entries correspond to Ground, Lower objects, Wood, and Foliage. A value of 1 denotes hard selection. At 100%, the Lower objects class in one support plot uses soft weights instead of the hard-selection branch, although AGF/SDEM remains the higher-reliability source. Per-class calibration counts are reported in Table S9. The complete repeat-level records are provided in `test/reviewer2_lin3d_budget_audit/repeat_level_gate_audit.csv`.

S8.6 Limited-Label Parameter Sensitivity

Figures S2 and S3 examine the sensitivity of the Lin3D results to the joint variation of τ and β under different calibration-label fractions. The reliability-gap threshold is fixed at $\gamma = 0.02$. Each cell reports the mean over the four Lin3D leave-one-plot episodes, using one fixed calibration sample for each fraction and episode across all parameter combinations.

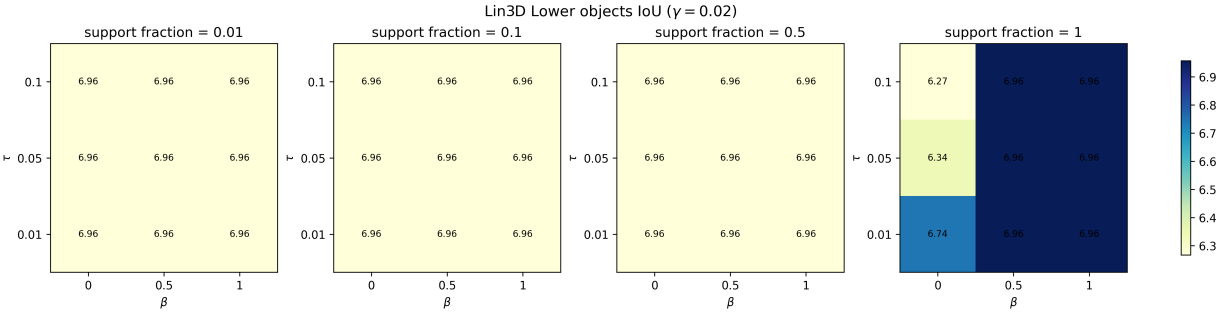


Figure S2: Limited-label sensitivity of the Lin3D Lower objects IoU under joint variation of τ and β. Columns correspond to the tested calibration fractions. The reliability-gap threshold is fixed at $\gamma = 0.02$, and each cell reports the mean over the four Lin3D leave-one-plot episodes. One fixed calibration sample is used per fraction and episode across all parameter combinations.

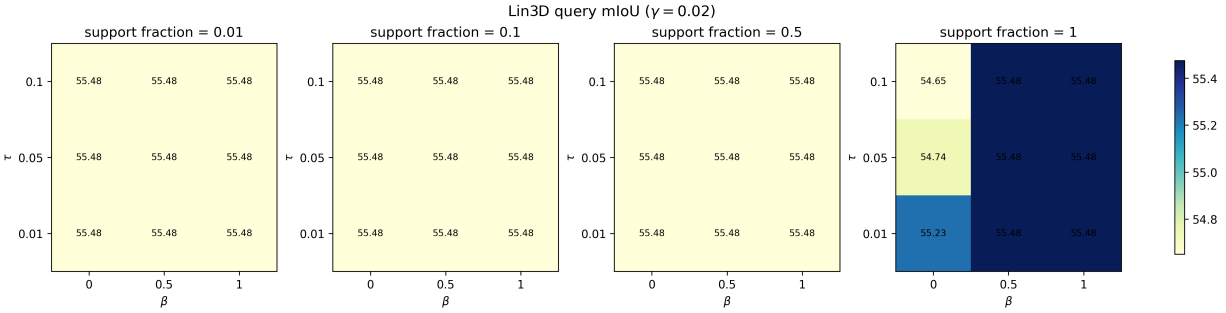


Figure S3: Limited-label sensitivity of Lin3D mIoU under joint variation of τ and β. Columns correspond to the tested calibration fractions. The reliability-gap threshold is fixed at $\gamma = 0.02$, and each cell reports the mean over the four Lin3D leave-one-plot episodes. One fixed calibration sample is used per fraction and episode across all parameter combinations.

S8.7 Primary Query Confusion Matrices

Table S22 reports the query confusion matrices for Shared UCRG on FOR-instance, EvoMLS, and the four Lin3D leave-one-plot episodes. Rows correspond to ground-truth classes and columns to predicted classes. The complete machine-readable confusion matrices for all primary methods, including Calibration-optimized class-wise fusion and KNORA-E dynamic ensemble selection, are provided in `test/reviewer2_classwise_primary_audit/confusion_matrices.json`.

Table S22: Query confusion matrices for Shared UCRG. Rows correspond to ground-truth classes and columns to predicted classes.

Dataset/fold	Ground-truth class	Predicted 1	Predicted 2	Predicted 3	Predicted 4
FOR-instance: Ground, Low vegetation, Wood, Crown					
	Ground	2,088,587	369,407	6,273	35,747
	Low vegetation	206,240	5,420,947	13,491	31,493
	Wood	7,146	24,450	1,696,529	1,002,770
	Crown	408	2,833	262,535	17,047,145
EvoMLS: Ground, Understorey, Wood, Foliage					
	Ground	1,155,640	123,621	282	52
	Understorey	820,982	4,209,610	32,634	96,704
	Wood	10,473	111,889	2,054,829	57,867
	Foliage	1,635	108,347	150,780	9,409,613
Lin3D, fold 1: Ground, Lower objects, Wood, Foliage					
	Ground	36,297	12,396	32	144
	Lower objects	2,048	7,898	318	444
	Wood	0	14	76,940	7,755
	Foliage	99	116,060	50,344	794,879
Lin3D, fold 2: Ground, Lower objects, Wood, Foliage					
	Ground	36,417	14,968	27	139
	Lower objects	2,207	9,579	675	1,038
	Wood	1	9	73,553	7,469
	Foliage	603	53,767	49,815	863,269
Lin3D, fold 3: Ground, Lower objects, Wood, Foliage					
	Ground	46,473	13,730	7	2
	Lower objects	2,010	7,465	643	809
	Wood	1	14	88,266	6,609
	Foliage	649	124,412	49,664	824,605
Lin3D, fold 4: Ground, Lower objects, Wood, Foliage					
	Ground	42,920	12,199	30	144
	Lower objects	1,855	5,643	420	1,084
	Wood	1	11	84,893	5,848
	Foliage	592	100,336	49,441	836,867

S8.8 Fold-Level Paired Comparison

Table S23: Fold-level paired comparison between Shared UCRG and calibration-based hard selection. Differences are computed as Shared UCRG minus hard selection.

Dataset/fold	Hard mIoU	Shared mIoU	Δ mIoU	Hard OA	Shared OA	Δ OA	Δ class IoU
FOR-instance	79.2855	78.8306	−0.4549	93.2079	93.0437	−0.1642	(−0.1428, −0.4311, −1.1300, −0.1157)
EvoMLS	77.1000	77.9780	+0.8780	91.3279	91.7402	+0.4123	(+2.4075, +1.3649, +0.0389, −0.2992)
Lin3D, fold 1	53.9039	53.9039	0.0000	82.8471	82.8471	0.0000	(0, 0, 0, 0)
Lin3D, fold 2	56.7165	55.7477	−0.9688	88.8737	88.2610	−0.6127	(−2.4366, −0.8568, −0.0200, −0.5620)
Lin3D, fold 3	55.4048	55.4048	0.0000	82.9623	82.9623	0.0000	(0, 0, 0, 0)
Lin3D, fold 4	55.8750	55.8750	0.0000	84.9459	84.9459	0.0000	(0, 0, 0, 0)
Episode mean	–	–	**−0.0909**	–	–	–	–

Class-wise differences are Shared UCRG minus calibration-based hard selection. The class order is Ground, Low vegetation, Wood, Crown for FOR-instance; Ground, Understorey, Wood, Foliage for EvoMLS; and Ground, Lower objects, Wood, Foliage for Lin3D. The descriptive 95% t-based interval for the six episode-level mIoU differences is $[-0.7322, 0.5503]$ percentage points. The exact paired sign-flip p-value is $p = 0.75$. The four Lin3D query sets overlap.

S8.9 Joint UCRG Parameter Sensitivity

Table S24: Full-support joint UCRG parameter sensitivity summary. Each row reports one (τ, β, γ) combination. Mean mIoU and OA are averaged across the three protocol-level results (FOR-instance, EvoMLS, and the four-episode mean for Lin3D). The negative-transfer count is computed over the six evaluation episodes.

τ	β	γ	Mean mIoU (%)	Mean OA (%)	Neg. transfer
0.01	0.00	0.00	70.6704	89.8389	4
0.01	0.25	0.00	70.7435	89.8839	2
0.01	0.50	0.00	70.7206	89.8701	1
0.01	0.75	0.00	70.6501	89.8352	1
0.01	1.00	0.00	70.5067	89.7682	1
0.03	0.00	0.00	70.3172	89.6110	5
0.03	0.25	0.00	70.5136	89.7587	5
0.03	0.50	0.00	70.5100	89.7650	5
0.03	0.75	0.00	70.4680	89.7475	5
0.03	1.00	0.00	70.4146	89.7235	4
0.05	0.00	0.00	70.0913	89.4463	5
0.05	0.25	0.00	70.3364	89.6442	5
0.05	0.50	0.00	70.4012	89.7015	5
0.05	0.75	0.00	70.4064	89.7115	5
0.05	1.00	0.00	70.3707	89.6975	6
0.10	0.00	0.00	69.6464	89.0953	6
0.10	0.25	0.00	69.8731	89.2894	6
0.10	0.50	0.00	70.0113	89.4054	6
0.10	0.75	0.00	70.0713	89.4612	6
0.10	1.00	0.00	70.0788	89.4746	6
0.15	0.00	0.00	69.1989	88.7310	6
0.15	0.25	0.00	69.3694	88.8842	6
0.15	0.50	0.00	69.4957	88.9944	6

τ	β	γ	Mean mIoU (%)	Mean OA (%)	Neg. transfer
0.15	0.75	0.00	69.5626	89.0556	6
0.15	1.00	0.00	69.5871	89.0801	6
0.01	0.00	0.02	70.6805	89.8460	2
0.01	0.25	0.02	70.7494	89.8862	1
0.01	0.50	0.02	70.7352	89.8758	1
0.01	0.75	0.02	70.6501	89.8352	1
0.01	1.00	0.02	70.5067	89.7682	1
0.03	0.00	0.02	70.4989	89.7265	2
0.03	0.25	0.02	70.6057	89.8052	1
0.03	0.50	0.02	70.5875	89.7956	1
0.03	0.75	0.02	70.4685	89.7481	1
0.03	1.00	0.02	70.4147	89.7237	1
0.05	0.00	0.02	70.4404	89.6856	2
0.05	0.25	0.02	70.5653	89.7831	1
0.05	0.50	0.02	70.5563	89.7773	1
0.05	0.75	0.02	70.4287	89.7280	1
0.05	1.00	0.02	70.3862	89.7090	2
0.10	0.00	0.02	70.3962	89.6531	2
0.10	0.25	0.02	70.5342	89.7658	2
0.10	0.50	0.02	70.5305	89.7626	2
0.10	0.75	0.02	70.3867	89.7086	2
0.10	1.00	0.02	70.3635	89.6980	2
0.15	0.00	0.02	70.3812	89.6424	2
0.15	0.25	0.02	70.5231	89.7592	2
0.15	0.50	0.02	70.5209	89.7575	2
0.15	0.75	0.02	70.3685	89.7001	2
0.15	1.00	0.02	70.3575	89.6952	2
0.01	0.00	0.05	70.6706	89.8392	2
0.01	0.25	0.05	70.7436	89.8840	1
0.01	0.50	0.05	70.7206	89.8701	1
0.01	0.75	0.05	70.6501	89.8352	1
0.01	1.00	0.05	70.5067	89.7682	1
0.03	0.00	0.05	70.3746	89.6619	2
0.03	0.25	0.05	70.5530	89.7854	1
0.03	0.50	0.05	70.5146	89.7684	1
0.03	0.75	0.05	70.4685	89.7481	1
0.03	1.00	0.05	70.4147	89.7237	1
0.05	0.00	0.05	70.2619	89.5974	2
0.05	0.25	0.05	70.4835	89.7527	1
0.05	0.50	0.05	70.4539	89.7397	1
0.05	0.75	0.05	70.4287	89.7280	1
0.05	1.00	0.05	70.3862	89.7090	2
0.10	0.00	0.05	70.1575	89.5394	3
0.10	0.25	0.05	70.4213	89.7241	2
0.10	0.50	0.05	70.4062	89.7169	2
0.10	0.75	0.05	70.3867	89.7086	2

τ	β	γ	Mean mIoU (%)	Mean OA (%)	Neg. transfer
0.10	1.00	0.05	70.3635	89.6980	2
0.15	0.00	0.05	70.1203	89.5188	3
0.15	0.25	0.05	70.3956	89.7124	2
0.15	0.50	0.05	70.3854	89.7079	2
0.15	0.75	0.05	70.3685	89.7001	2
0.15	1.00	0.05	70.3575	89.6952	2
0.01	0.00	0.10	70.6704	89.8389	4
0.01	0.25	0.10	70.7435	89.8839	2
0.01	0.50	0.10	70.7206	89.8701	1
0.01	0.75	0.10	70.6501	89.8352	1
0.01	1.00	0.10	70.5067	89.7682	1
0.03	0.00	0.10	70.3182	89.6118	4
0.03	0.25	0.10	70.5236	89.7671	2
0.03	0.50	0.10	70.5146	89.7684	1
0.03	0.75	0.10	70.4685	89.7481	1
0.03	1.00	0.10	70.4147	89.7237	1
0.05	0.00	0.10	70.1084	89.4604	4
0.05	0.25	0.10	70.4023	89.7012	2
0.05	0.50	0.10	70.4539	89.7397	1
0.05	0.75	0.10	70.4287	89.7280	1
0.05	1.00	0.10	70.3862	89.7090	2
0.10	0.00	0.10	69.8454	89.2626	5
0.10	0.25	0.10	70.2537	89.6151	3
0.10	0.50	0.10	70.4062	89.7169	2
0.10	0.75	0.10	70.3867	89.7086	2
0.10	1.00	0.10	70.3635	89.6980	2
0.15	0.00	0.10	69.7296	89.1742	5
0.15	0.25	0.10	70.1877	89.5740	3
0.15	0.50	0.10	70.3854	89.7079	2
0.15	0.75	0.10	70.3685	89.7001	2
0.15	1.00	0.10	70.3575	89.6952	2