Supplementary Materials: The following supporting information is available:

Table S1, Detailed results of acquisition-stage label-predictability and input-ablation diagnostics;
Table S2, Detailed results of ordinal modeling, MixStyle, and data-composition ablations;
Table S3, Detailed configuration and performance of the 2-s temporal-exclusion experiments;
Table S4, Configuration of the fixed-test-set temporal diagnostic;
Table S5, Performance on the fixed S07 test set under the three training conditions;
Table S6, Multi-seed performance of Protocols A and C on their respective fixed data partitions;
Table S7, Multi-seed performance on the fixed S07 test set under three training conditions;
Figure S1, Evolution of source-stage weights during Stage-GroupDRO training with S07 held out;
Figure S2, Seed-level variability of Stage-GroupDRO relative to Weighted-ERM.

Table S1. Detailed results of acquisition-stage label-predictability and input-ablation diagnostics.

Diagnostic task	Stages included	Input condition	Number of stage classes	1/K reference (%)	Stage-classification Macro-F1 (%)
Medium-only stage classification	S01–S08	Original RGB	8	12.5	94.28
Source-stage input ablation	S01/S03/S04/S06	Original RGB	4	25.0	96.05
Source-stage input ablation	S01/S03/S04/S06	Grayscale	4	25.0	97.29
Source-stage input ablation	S01/S03/S04/S06	Gray-World white balance	4	25.0	97.42
Source-stage input ablation	S01/S03/S04/S06	Expanded-crop contraction, 5% per side	4	25.0	97.15
Source-stage input ablation	S01/S03/S04/S06	Expanded-crop contraction, 10% per side	4	25.0	97.19
Source-stage input ablation	S01/S03/S04/S06	Center 70% crop	4	25.0	97.30

Note: The eight-stage diagnostic used only medium-grade garlic-bulb instances to reduce confounding from residue-grade composition. The 1/K values are uniform-class guessing references rather than exact expected Macro-F1 values under the empirical stage distributions. The input ablations were used to test whether stage-associated information could be explained mainly by simple color, white-balance, or crop-boundary factors.

Table S2. Detailed results of ordinal modeling, MixStyle, and data-composition ablations.

Experiment category	Evaluation setting	Method / condition	Replicates	Macro-F1 (%)	Change vs. matched reference (pp)	Interpretation
Ordinal supervision	Fixed within-source-folder chronological split	Standard CE	3 training seeds	88.60 ± 2.20	Reference	Nominal-classification baseline
Ordinal supervision	Fixed within-source-folder chronological split	CE + Expected Rank	3 training seeds	88.41 ± 1.74	−0.19	No stable improvement over CE
Feature-style randomization	S07 validation	Baseline without MixStyle	Single configuration	78.58	Reference	Baseline for MixStyle ablation
Feature-style randomization	S07 validation	CE + MixStyle	Single configuration	74.64	−3.94	Feature-style randomization reduced validation performance
Data composition	LOSO on S05–S08	Mixed-only source data	4 folds	73.00	Reference	Only mixed stages S05–S08 used as the candidate source pool
Data composition	LOSO on S05–S08	All-source data	4 folds	73.26	+0.26	Small mean increase with inconsistent fold-wise directions

Note: The ordinal-supervision experiment used a single fixed within-source-folder chronological partition, with three independent training seeds (0, 1, and 2). The MixStyle comparison used the S07 validation set and one training seed (42), whereas the data-composition experiment used four held-out test stages (S05–S08) with seed 42. The Mixed-only and All-source conditions differed in source-stage composition, training-set size, and class distribution. Because the three experiment categories used different evaluation settings, their absolute Macro-F1 values should not be compared directly. Differences were calculated only relative to the reference condition within each category. Fold-specific test Macro-F1 values for Mixed-only versus All-source were 58.81% versus 60.21% (S05), 86.61% versus 86.82% (S06), 73.13% versus 75.56% (S07), and 73.44% versus 70.42% (S08).

Table S3. Detailed configuration and performance of the 2-s temporal-exclusion experiments.

Seed	Train/Val/Test scenes	Temporal rule	Best epoch	Accuracy (%)	Macro-F1 (%)	MAE	QWK
42	530/112/112	No same-stage cross-subset image pair within 2 s	1	87.40	87.43	0.1292	0.8726
0	530/112/112	No same-stage cross-subset image pair within 2 s	6	86.25	86.32	0.1464	0.8495
1	530/112/112	No same-stage cross-subset image pair within 2 s	3	84.85	84.39	0.1515	0.8462
2	530/112/112	No same-stage cross-subset image pair within 2 s	3	86.76	87.05	0.1381	0.8678
Mean ± SD	530/112/112	—	—	85.95 ± 0.99	85.92 ± 1.38	0.1453 ± 0.0067	0.8545 ± 0.0117

Note: The mean and sample standard deviation were calculated from seeds 0, 1, and 2 only. The seed-42 run was the primary comparison and was not included in the three-seed summary. Under the 2-s temporal-exclusion rule, no same-stage image pair assigned to different subsets could have an absolute time difference of ≤2 s.

Table S4. Configuration of the fixed-test-set temporal diagnostic.

Condition	Test stage	Test scenes	Test instances	Train scenes	Same validation set	Same class counts as Neighbor-included	S07 training available	≤2-s test-neighboring S07 training images
Neighbor-included	S07	26	405	400	Yes	Reference	Yes	Yes
2-s-excluded	S07	26	405	400	Yes	Yes	Yes	No
Stage-excluded	S07	26	405	400	Yes	Not used as matching constraint	No	No

Note: All three conditions used the same 26 S07 test scene images containing 405 garlic-bulb instances, the same validation set, and 400 training scene images. The Neighbor-included and 2-s-excluded conditions additionally had identical class-specific training-instance counts. In the 2-s-excluded condition, all candidate S07 training images with an absolute time difference of ≤2 s from any fixed test image were removed and replaced by more temporally distant images from the same stage. The Stage-excluded condition removed all S07 images from training and was included as a reference rather than as a fully class-matched single-factor comparison.

Table S5. Performance on the fixed S07 test set under the three training conditions.

Condition	Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)	MAE	QWK	Class-wise Recall/F1 (%)
Neighbor-included	99.75	99.76	99.66	0.0025	0.9964	low 100.00/100.00; medium 100.00/99.81; high 98.98/99.49
2-s-excluded	74.81	74.12	73.86	0.2519	0.6305	low 91.84/91.84; medium 81.78/80.53; high 47.96/50.00
Stage-excluded	71.85	76.92	81.47	0.2840	0.6589	low 100.00/98.00; medium 62.79/74.14; high 81.63/58.61

Note: Class-specific values are reported as Recall/F1 in the order low, medium, and high. All three conditions used the same 26 S07 test scene images and 405 test instances. Neighbor-included and 2-s-excluded also used the same validation set, the same number of training scenes, and identical class-specific training-instance counts. Stage-excluded was included only as a reference and was not constrained to match the training-class composition.

Table S6. Multi-seed performance of Protocols A and C on their respective fixed data partitions.

Protocol	Seed	Accuracy (%)	Macro-F1 (%)	MAE	QWK
A	42	97.59	97.49	0.0247	0.9738
A	0	97.91	97.77	0.0209	0.9789
A	1	97.78	97.69	0.0222	0.9776
A	2	97.72	97.61	0.0234	0.9751
A	Mean ± SD	97.80 ± 0.10	97.69 ± 0.08	0.0222 ± 0.0013	0.9772 ± 0.0020
C	42	76.76	76.32	0.2350	0.7742
C	0	78.47	77.31	0.2179	0.7772
C	1	75.68	74.93	0.2464	0.7577
C	2	78.91	78.36	0.2141	0.7914
C	Mean ± SD	77.69 ± 1.75	76.87 ± 1.76	0.2261 ± 0.0177	0.7754 ± 0.0169

Note: Seed 42 represents the original experiment and is reported separately. The mean and sample standard deviation were calculated using only the three additional training seeds (0, 1, and 2). Data partitions, model configuration, and checkpoint-selection criteria were held fixed within each protocol. Protocols A and C used different test images; therefore, their performance differences should be interpreted as sensitivity to the data-partitioning strategy.

Table S7. Multi-seed performance on the fixed S07 test set under three training conditions.

Condition	Seed	Accuracy (%)	Macro-F1 (%)	MAE	QWK
Neighbor-included	42	99.75	99.76	0.0025	0.9964
	0	99.26	99.30	0.0074	0.9895
	1	100.00	100.00	0.0000	1.0000
	2	100.00	100.00	0.0000	1.0000
Neighbor-included	Mean ± SD	99.75 ± 0.43	99.77 ± 0.40	0.0025 ± 0.0043	0.9965 ± 0.0061
2-s-excluded	42	74.81	74.12	0.2519	0.6305
	0	82.72	82.61	0.1728	0.7523
	1	65.93	64.55	0.3407	0.4949
	2	68.40	66.13	0.3210	0.5786
2-s-excluded	Mean ± SD	72.35 ± 9.07	71.10 ± 10.00	0.2782 ± 0.0918	0.6086 ± 0.1313
Stage-excluded	42	71.85	76.92	0.2840	0.6589
	0	69.38	69.83	0.3111	0.6078
	1	75.56	79.41	0.2469	0.6862
	2	63.95	66.34	0.3679	0.6105
Stage-excluded	Mean ± SD	69.63 ± 5.81	71.86 ± 6.77	0.3086 ± 0.0605	0.6349 ± 0.0445

Note: All conditions were evaluated using the same fixed S07 test set comprising 26 RGB scene images and 405 annotated garlic-bulb instances. Seed 42 denotes the original experiment, while the mean and sample standard deviation were calculated from additional training seeds 0, 1, and 2. Neighbor-included and 2-s-excluded used matched training-scene and class-specific instance counts. Stage-excluded was retained as a reference with a different training-class composition. The paired Neighbor-included minus 2-s-excluded Macro-F1 differences were 16.69, 35.45, and 33.87 percentage points for seeds 0, 1, and 2, respectively (mean ± SD: 28.67 ± 10.40 percentage points).

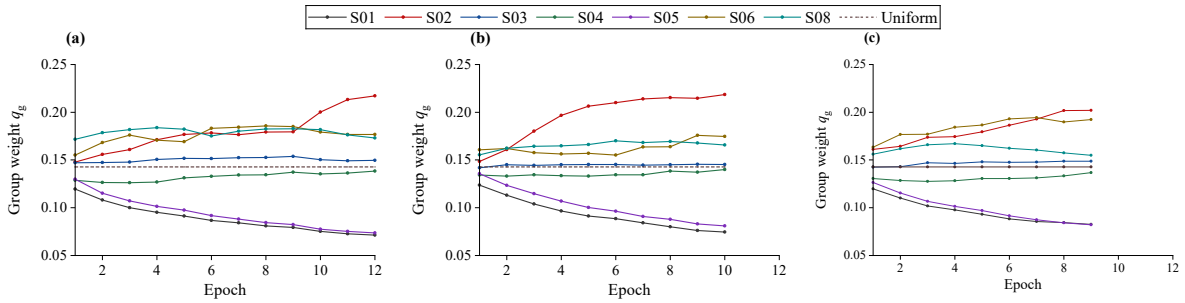


Figure S1. Evolution of source-stage weights during Stage-GroupDRO training with S07 held out. (**a–c**) Epoch-wise source-stage weights for random seeds 0, 1, and 2. The seven source stages were S01, S02, S03, S04, S05, S06, and S08. The dashed line marks the uniform initialization 1/7 (≈0.1429). The trajectories show how Stage-GroupDRO redistributed optimization emphasis among source stages during training; they do not define a fixed ranking of intrinsic stage difficulty.

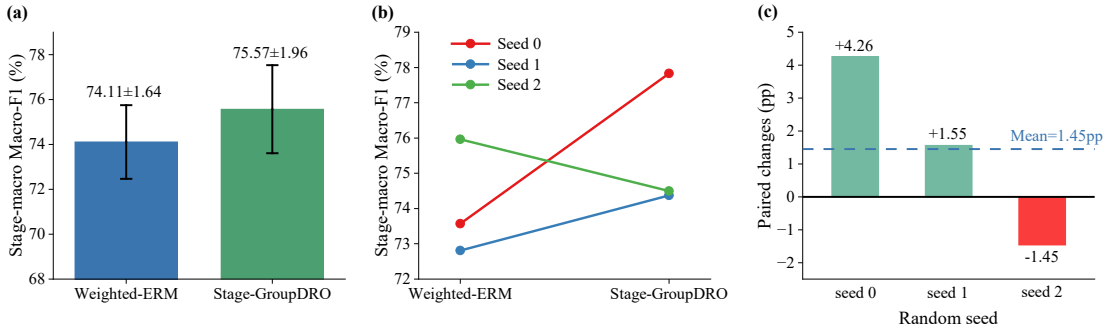


Figure S2. Seed-level variability of Stage-GroupDRO relative to Weighted-ERM. Each value represents the difference in Stage-macro Macro-F1 between Stage-GroupDRO and Weighted-ERM for the same random seed under the four-fold LOSO evaluation (Stage-GroupDRO − Weighted-ERM). Positive values indicate higher Stage-macro Macro-F1 for Stage-GroupDRO, whereas negative values indicate higher performance for Weighted-ERM. The comparison used the same data splits, model architecture, training settings, and random seed within each paired run.