Supplementary Information for "Validation-Aligned Coreset Selection for Budgeted Few-Shot Classification"

Supplementary Note 1: claim boundary

The main article reports a bounded claim. VACS-R is the strongest evaluated training-only deployable protocol in the five-dataset low-budget 3-nearest-neighbour benchmark, but the experiments do not establish broad state-of-the-art performance. The supplementary tables below document the paired fast-protocol audit, ablations, non-deployable hindsight-static reference, validation-size diagnostics, selection costs, and frozen embedding boundary checks.

Supplementary Table 1: paired low-budget fast-protocol stability

This table reports VACS-F minus herding differences over matched low-budget units.

Slice	n	Δ Acc.	95% CI	W/T/L	Sign p
Pooled	80	**+0.02**	**[-1.38,+1.20]**	28/41/11	0.009
20NG	16	**+4.54**	**[+2.43,+7.07]**	16/0/0	$< .001$
Breast cancer	16	**+0.84**	**[+0.37,+1.32]**	7/9/0	0.016
Digits	16	**+0.19**	**[-0.12,+0.53]**	3/12/1	0.625
Iris	16	**-4.31**	**[-9.86,-0.14]**	2/9/5	0.453
Wine	16	**-1.16**	**[-2.20,-0.35]**	0/11/5	0.062

Supplementary Table 2: ablation diagnostics

The ablation table separates the validation wrapper from individual selector candidates and from the final rebuild convention.

Variant	Acc.	Δ	Macro-F1	Selected
Full VACS	**70.6**	**+0.00**	**70.2**	**11.4**
No final rebuild	72.0	+1.34	71.7	11.4
No K-means candidate	70.4	-0.21	69.8	11.4
No MARC candidate	70.7	+0.08	70.3	11.4
No random candidate	70.6	-0.05	70.3	11.4

Supplementary Table 3: non-deployable dataset-hindsight static audit

The per-dataset static reference is selected with hindsight from the low-budget result table, and is therefore diagnostic rather than deployable.

Slice	Static ref.	VACS	Static	Δ [95% CI]	W/T/L
Pooled per-dataset-best	Per-dataset	**70.6**	72.3	-1.64 [-2.98,-0.61]	8/53/19
20NG	Random	**9.7**	12.1	-2.49 [-5.33,-0.03]	4/7/5
Breast cancer	K-means	**93.4**	93.5	-0.15 [-0.37,+0.00]	0/14/2
Digits	K-means	**82.3**	82.4	-0.12 [-0.50,+0.19]	2/12/2
iris	Herding	**76.4**	80.7	-4.31 [-9.86,-0.14]	2/9/5
Wine	Herding	**91.3**	92.5	-1.16 [-2.20,-0.35]	0/11/5

Supplementary Table 4: validation-size reliability

This diagnostic repeats the internal validation-ranking analysis at 10%, 20%, 25%, and 33% validation fractions.

Val. frac.	Units	VACS	Gain	Top-1	ρ	Regret	Severe
0.10	64	**69.1**	+0.94	60.9	0.81	1.58	12.5
0.20	64	**68.3**	+0.10	46.9	0.77	2.42	18.8
0.25	64	**69.3**	+1.13	60.9	0.81	1.38	9.4
0.33	64	**69.0**	+0.86	59.4	0.80	1.66	14.1

Supplementary Table 5: selection cost

The cost diagnostic reports selection-only time for herding and validation-plus-rebuild time for VACS-F over the matched low-budget decisions.

Slice	n	Herding	VACS val.	Rebuild	Total
Pooled	80	10.475	**79.820**	**7.390**	**87.210**
20NG	16	10.250	**75.500**	**6.859**	**82.359**
Breast cancer	16	0.038	**0.969**	**0.107**	**1.076**
Digits	16	0.137	**2.011**	**0.336**	**2.347**
Iris	16	0.018	**0.667**	**0.038**	**0.705**
Wine	16	0.032	**0.673**	**0.050**	**0.723**

Supplementary Table 6: frozen transformer probes

The frozen transformer probes are boundary evidence. They do not support a broad modern text-embedding superiority claim.

Probe	Dim.	VACS	Best static	Gain
DeBERTa-v3-small	768	**9.8**	Herding 9.9	-0.17
SmolLM2-360M, maxlen 64	960	**31.5**	MARC 31.4	+0.03
SmolLM2-360M, maxlen 128	960	**37.4**	Herding 37.5	-0.12
Qwen2.5-0.5B	896	**40.1**	Herding 40.2	-0.11

Supplementary Tables 7 and 8: frozen image-embedding probes

The image probes widen benchmark coverage but tie the strongest static selector in the tested one-seed, one-example-per-class settings.

Dataset	VACS	Best static	Gain	Backbone
Fashion-MNIST	**75.6**	Herding 75.6	+0.00	convnext_tiny
Pooled	**75.6**	Herding 75.6	+0.00	convnext_tiny

Dataset	VACS	Best static	Gain	Backbone
CIFAR-10	**64.7**	Herding 64.7	+0.00	resnet18
Pooled	**64.7**	Herding 64.7	+0.00	resnet18

Supplementary Table 9: full 20-class sparse-text stress boundary

The compact full-20-class TF-IDF stress check is negative for VACS relative to static random and is kept as a claim boundary.

Task	Classes	Units	VACS	Best static	Gain	Top VACS choice
20 Newsgroups (20-class)	20	16	**9.7**	Random 12.1	-2.49	Boundary 9/16

Supplementary Note 2: reproducibility files

The project package includes the Python scripts used to regenerate the primary result manifest, all tables included in the manuscript and supplementary information, the workflow and diagnostic figures, and the consistency checks that compare manuscript claims with generated result files. The public code repository is available at `https://github.com/HaotongLuan/validation-aligned-coreset-selection`. The main experiment

driver is `experiments/run_vacs_experiments.py`. Covertype audits, repeated-validation audits, validation-size diagnostics, and figure generation are implemented as separate companion scripts for auditability.