

Full Report - Real-time FER - 2025-10-11

Core Group vs Control Group Student Training Study Report

(Hong Kong Time UTC+8)

1. Objectives

- Establish reproducible, aligned MobileNetV3 student baselines using canonical index ``dataset_index_extended_next_plus_affectnetfull_dedup.csv``.
- Compare Core Group (Pairwise CNN + Hybrid) vs Control Group (Split CNN+ViT, Four-Way Split (6E), and Fused Four Equal (6C)) to identify the highest macro F1 and explain performance differentials.
- Diagnose and document root cause of earlier degraded Split Multi-Teacher results (label length/order misalignment) and impact of enforcing strict alignment (``--require-aligned``).
- Provide decision guidance for deployment candidate selection (accuracy vs complexity vs calibration extensibility).

2. Experimental Families

Family ID	Group	Description	Teachers / Composition	Distill Mode(s)	Notes
6A	Core	Pairwise CNN Weighted	ResNet18_polish (0.7) + EffNet-B3 (0.3) fused	KD, DKD ($\beta=2,4,8$ historic)	Stable strong baseline
6B	Core	Hybrid (Pairwise + ViT)	(Pairwise 0.7/0.3) + ViT BAN (0.3)	KD, DKD	Adds transformer context
6D-split	Control	Split CNN+ViT (λ sweep)	Pairwise vs ViT separate, $\lambda \in \{0.5/0.5, 0.7/0.3, 0.3/0.7\}$	KD, DKD ($\beta=4$)	Multi-source aligned inputs
6E-split4	Control	Four-Way Split Equal	RN18 + ConvNeXt-Tiny + EffB3 + ViT (0.25 each)	KD, DKD ($\beta=4$)	Richest diversity
6C	Control	Fused Four Equal (Planned)	Single fused softlabels (equal-weight average of RN18 + ConvNeXt-Tiny + EffB3 + ViT)	KD, DKD	Planned; over-smoothing hypothesis test (not yet aligned)

All experiments: epochs=20, batch=128, seeds={1337,2025}, KD $\alpha=0.5$ T=2, DKD $\alpha=0.5$ β as specified, ``--num-workers 0`` (Windows determinism), test eval every 5 epochs.

3. Data & Alignment Integrity

- Canonical index enforced across all teacher softlabel exports (``*_ixNextAffFull``).
- Alignment diagnostic (``diagnose_softlabel_alignment.py --strict``) confirmed equality of lengths & hashes pre-training.
- Earlier misaligned split runs (pre-canonical) purged to remove contaminated comparisons.

4. Metrics & Sources

Primary metrics from `metrics_final.json` (best_epoch, best_test_macro_f1, best_test_acc, per-class F1). Time & checkpoint inferred if not explicit. Timestamps localized to UTC+8 via `--tz-offset 8` in summary collector.

4.1 Mathematical Definitions

We formalize the evaluation metrics, distillation objectives, and calibration quantities referenced throughout the report.

4.1.1 Notation

- - Number of classes: K (here $K=7$: angry, disgust, fear, happy, neutral, sad, surprise)
- - Sample index: $i = 1, \dots, n$
- - Ground-truth one-hot label: $\mathbf{y}_i \in \{0, 1\}^K$
- - Student logits: $\mathbf{z}_i \in \mathbb{R}^K$; Teacher logits (for teacher m): $\mathbf{t}_i^{(m)}$
- - Softmax with temperature T : $\operatorname{softmax}_T(\mathbf{z})_c = \frac{\exp(z_c / T)}{\sum_j \exp(z_j / T)}$
- - Teacher set size (multi-teacher): M

4.1.2 Per-Class Precision / Recall / F1

For class c let TP_c, FP_c, FN_c be true positives, false positives, false negatives.

$$P_c = \frac{TP_c}{TP_c + FP_c + \epsilon}, \quad R_c = \frac{TP_c}{TP_c + FN_c + \epsilon}$$

$$F1_c = \frac{2 P_c R_c}{P_c + R_c + \epsilon}$$

Small ϵ avoids division-by-zero (implemented implicitly by Python float semantics when denominators > 0 in practice). Macro F1:

$$\operatorname{MacroF1} = \frac{1}{K} \sum_{c=1}^K F1_c$$

Accuracy:

$$\operatorname{Acc} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\arg\max_c z_{i,c} = \arg\max_c y_{i,c}]$$

4.1.3 Knowledge Distillation (KD) Loss

Denote ground-truth hard cross-entropy: $L_{\text{CE}} = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^K y_{i,c} \log p_{s,i,c}$ where $p_{s,i} = \operatorname{softmax}_1(\mathbf{z}_i)$.

Single (fused) teacher soft targets (temperature T): $p_{t,i}^T = \operatorname{softmax}(\mathbf{t}_i / T)$, student softened: $p_{s,i}^T = \operatorname{softmax}(\mathbf{z}_i / T)$.

$$\mathcal{L}_{\text{KD}} = (1-\alpha) \mathcal{L}_{\text{CE}} + \alpha T^2 \cdot \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{p}_{t,i}^T} \mathbb{E}_{\mathbf{p}_{s,i}^T} \mathbb{KL}(\mathbf{p}_{t,i}^T \parallel \mathbf{p}_{s,i}^T)$$

We use $\alpha = 0.5$, $T=2$. (Factor T^2 preserves gradient magnitude per Hinton et al.).

4.1.4 Multi-Teacher Split vs Fused Targets

Split strategy draws M separate softened teacher distributions $\mathbf{p}_{t,i}^{(m),T}$ each forward pass; loss averages KL terms:

$$\mathcal{L}_{\text{KD}\text{-}\text{split}} = (1-\alpha) \mathcal{L}_{\text{CE}} + \alpha T^2 \cdot \frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \mathbb{E}_{\mathbf{p}_{t,i}^{(m),T}} \mathbb{E}_{\mathbf{p}_{s,i}^T} \mathbb{KL}(\mathbf{p}_{t,i}^{(m),T} \parallel \mathbf{p}_{s,i}^T)$$

Fused strategy first averages probabilities:

$$\bar{\mathbf{p}}_{t,i}^T = \frac{1}{M} \sum_{m=1}^M \mathbf{p}_{t,i}^{(m),T}$$

Then standard KD applies with $\mathbf{p}_{t,i}^T = \bar{\mathbf{p}}_{t,i}^T$.

Information Smoothing Observation: Under an (approximate) independence assumption among teachers for class c , variance shrinks as

$$\text{Var}[\bar{\mathbf{p}}_{t,i}^T(c)] = \frac{1}{M^2} \sum_{m=1}^M \text{Var}[\mathbf{p}_{t,i}^{(m),T}(c)]$$

reducing inter-teacher disagreement signal the student could otherwise exploit (empirical “over-smoothing”).

4.1.5 Decoupled Knowledge Distillation (DKD) Simplified Form

Following Zhou et al. (Decoupled KD), logits are partitioned into target (ground-truth) class g and non-target classes \mathcal{N} . Let $\mathbf{p}_{t,i,g}^T, \mathbf{p}_{s,i,g}^T$ be softened probabilities for class g ; similarly define normalized distributions over non-target classes:

$$\tilde{\mathbf{p}}_{t,i,j}^T = \frac{\mathbf{p}_{t,i,j}^T}{1 - \mathbf{p}_{t,i,g}^T}, \quad \tilde{\mathbf{p}}_{s,i,j}^T = \frac{\mathbf{p}_{s,i,j}^T}{1 - \mathbf{p}_{s,i,g}^T}, \quad j \in \mathcal{N}$$

Target-Class KD component:

$$\mathcal{L}_{\text{TCKD}} = - \frac{1}{n} \sum_{i=1}^n \mathbf{p}_{t,i,g}^T \log \mathbf{p}_{s,i,g}^T$$

Non-Target KD component:

$$\mathcal{L}_{\text{NCKD}} = \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{N}} \tilde{\mathbf{p}}_{t,i,j}^T \log \frac{\tilde{\mathbf{p}}_{t,i,j}^T}{\tilde{\mathbf{p}}_{s,i,j}^T}$$

We employ a weighted combination (notation aligned to our earlier shorthand with parameters α, β):

$$\mathcal{L}_{\text{DKD}} = (1-\alpha) \mathcal{L}_{\text{CE}} + \alpha T^2 \mathcal{L}_{\text{TCKD}} + \beta T^2 \mathcal{L}_{\text{NCKD}}$$

In four-way split experiments $\beta=4$ (historic sweeps also tried $\beta=2,8$). Observed diminishing returns with increased teacher diversity.

4.1.6 Calibration Metrics

Let predicted confidence $c_i = \max_c p_{s,i,c}$ and predicted class $\hat{y}_i = \arg\max_c p_{s,i,c}$.

1. Expected Calibration Error (ECE) with B bins B_b :

$$\mathrm{ECE} = \sum_{b=1}^B \frac{|B_b|}{n} |\mathrm{acc}(B_b) - \mathrm{conf}(B_b)|$$

where $\mathrm{acc}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \mathbf{1}[\hat{y}_i = y_i]$,
 $\mathrm{conf}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} c_i$.

2. Brier Score:

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^n \sum_{c=1}^K (p_{s,i,c} - y_{i,c})^2$$

3. Negative Log-Likelihood (NLL):

$$\mathrm{NLL} = - \frac{1}{n} \sum_{i=1}^n \sum_{c=1}^K y_{i,c} \log p_{s,i,c}$$

Temperature scaling selects T^* minimizing NLL (or ECE proxy) on a calibration split; post-hoc we replace $p_{s,i}$ by $\operatorname{softmax}_{T^*}(\mathbf{z}_i)$ without retraining.

4.1.7 Deployment Consideration Metric Coupling

- Macro F1: selection for accuracy / fairness across classes.
- ECE / Brier / NLL: selection for reliability of probability outputs (downstream thresholding / ranking stability).
- We will interpret calibration changes jointly: large ECE drop with negligible Macro F1 loss is acceptable; large Macro F1 loss for minor ECE gain is not.

These definitions support the diversity vs over-smoothing narrative: higher teacher diversity preserves variance across $p_{t,i}^{(m)}$, which—rather than being prematurely averaged—translates into richer gradient signals in $L_{\text{KD}\text{-}split}$ and empirically higher Macro F1.

4.2 Reproducibility & Alignment Hashing

To guarantee reproducibility we compute SHA256 digests over concatenated relative image paths for each softlabel directory and verify equality prior to training when using multi-teacher or fused targets. Example diagnostic JSON excerpt:

```

{
  "dirs": [
    "experiments/softlabels/resnet18_single_ixNextAffFull",
    "experiments/softlabels/convnext_tiny_single_ixNextAffFull",
    "experiments/softlabels/efficientnet_b3_single_ixNextAffFull",
    "experiments/softlabels/vit_small_ban_single_ixNextAffFull"
  ],
  "train_len_equal": true,
  "hash_paths_train": "5b7f9c1c...",
  "labels_equal": true,
  "ok": true
}

```

Multi-teacher equal weights: $w_k = 0.25$ for $k=1..4$, $\sum_{k=1}^4 w_k = 1$. Soft targets always produced at temperature $T=2$; temperature scaling (calibration) later selects T^* by minimizing validation NLL.

5. Results Summary (Macro F1)

All macro F1 and accuracy values below are expressed in percentages (%), converted from raw fractions (e.g. 0.72185 \rightarrow 72.19%).

Family	Mode	Seeds	Best Macro F1 (per seed)	Mean	Notes
Pairwise (6A)	KD	1337,2025,42	71.88 / 71.61 / 71.58	71.69	Simple, strong
Pairwise (historic)	DKD $\beta=4$	1337,2025	~ 71.92 / ~ 71.31	~ 71.62	Earlier; small uplift not consistent
Hybrid (6B)	KD	1337,2025	70.64 / 70.49	70.57	Adds ViT but lower post-alignment
Hybrid (6B)	DKD (β best)	1337,2025	71.05 / 70.66	70.86	DKD marginal gain vs KD
Split CNN+ViT (6D)	KD	1337,2025	$\lambda = 0.7/0.3$: 71.30 / 71.43	71.36	$\lambda = 0.7/0.3 >$ others
Split CNN+ViT (6D)	DKD $\beta=4$	1337,2025	$\lambda = 0.7/0.3$: 71.82 / 71.22	71.52	Near pairwise KD
Four-Way Split (6E)	KD	1337,2025,42	72.26 / 72.11 / 71.97	72.11	Highest overall
Four-Way Split (6E)	DKD $\beta=4$	1337,2025	72.02 / 72.07	72.05	KD > DKD here

5.1 Statistical Reliability & Limitations

We now possess three fully aligned experimental replicates (seeds 1337, 2025, 42) for the principal comparison between the Pairwise KD and Four-Way Split KD configurations. This extends earlier two-seed exploratory evidence and permits a modestly more stable estimation of the mean performance differential.

Per-seed macro F1 (fractions):

```
Pairwise KD      : 0.7187765, 0.7161279, 0.7158000
Four-Way Split KD : 0.7226347, 0.7210732, 0.7196643
```

Summary statistics (fractions; percentage values in parentheses):

- - Pairwise KD mean = 0.7169 (71.69%), standard deviation ≈ 0.00128 (0.13 percentage points).
- - Four-Way Split KD mean = 0.72113 (72.11%), standard deviation ≈ 0.00122 (0.12 percentage points).
- - Mean difference (Four-Way – Pairwise) = 0.00423 (~ 0.42 percentage points).

Using a non-parametric bootstrap ($B = 10,000$ resamples) we obtain a 95% confidence interval for the mean difference:

$\Delta_{\text{mean}} \in [0.00227, 0.00620]$

The interval excludes zero, indicating a directionally consistent uplift over the observed replicate set. A permutation test (three vs three values) yields a p-value of approximately 0.0998; with such a small sample size, the test has low statistical power, so failure to attain a conventional 0.05 threshold should not be over-interpreted. The computed Cohen's $d \approx 3.33$ is numerically large but inflated by the very small pooled variance and limited n ; it is therefore reported only for completeness.

Interpretation:

1. Practical Significance: The ~ 0.4 percentage point macro F1 gain is modest yet operationally meaningful given identical architectural and optimization settings.
2. Inferential Caution: Additional seeds (or cross-validation folds) would be required for a formally decisive hypothesis test; current evidence supports a consistent advantage without claiming definitive population-level superiority.
3. Transparent Uncertainty: Presenting both the raw per-seed values and the bootstrap interval mitigates overstatement and clarifies the narrow empirical variance observed.

Provenance: Results produced via `scripts/stat_tests_bootstrap_permutation.py` (artifact `stats/fourway_vs_pairwise.json`).

5.2 Calibration & Reliability (Temperature Scaling)

Each student model underwent post-hoc temperature scaling over a discrete grid:

$T \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$.

The selection rule for the deployment temperature T^* was defined lexicographically:

1. Minimize Negative Log-Likelihood (NLL);

8. 2. Among ties, minimize Expected Calibration Error (ECE);
9. 3. Among remaining ties, maximize macro F1.

All six KD runs (three seeds across both families) converged on a common optimum:

$$T^* = 1.2.$$

Reliability Metrics (fractions; mean across seeds reported on last line of each block):

Family	Seed	Macro F1	Acc	ECE	Brier	NLL	T*
Pairwise KD	1337	0.7188	0.7423	0.1404	0.40352	0.9711	1.2
Pairwise KD	2025	0.7161	0.7405	0.1423	0.40488	0.9760	1.2
Pairwise KD	42	0.7158	0.7406	0.1401	0.40492	0.9701	1.2
Pairwise KD (mean)	–	0.7169	0.7418	0.1409	0.40444	0.9724	1.2
Four-Way Split KD	1337	0.7226	0.7445	0.0420	0.36495	0.7988	1.2
Four-Way Split KD	2025	0.7211	0.7444	0.0463	0.36747	0.8044	1.2
Four-Way Split KD	42	0.7197	0.7432	0.0444	0.36847	0.8077	1.2
Four-Way Split KD (mean)	–	0.7211	0.7440	0.0442	0.36696	0.8036	1.2

Key Observations:

10. 1. ECE Reduction: Four-Way Split KD cuts ECE by $\sim 3.2\times$ ($0.1409 \rightarrow 0.0442$) versus Pairwise KD, indicating substantially better calibrated confidence estimates.
11. 2. Sharp Brier & NLL Gains: Brier improvement ($\sim 0.4044 \rightarrow 0.3670$) and NLL improvement ($\sim 0.9724 \rightarrow 0.8036$) show reliability gains are not a trade-off against accuracy; they co-occur with a small macro F1 uplift.
12. 3. Temperature Consistency: Uniform selection of $T^*=1.2$ across seeds/families simplifies deployment (single scalar re-scaling suffices for all MobileNetV3 student variants evaluated here).
13. 4. Practical Impact: Lower ECE/NLL favors downstream thresholding (e.g., alert triggers) and any probabilistic ensembling; Four-Way model’s probabilities are both sharper (lower NLL) and better aligned with empirical correctness (low ECE) — a desirable dual improvement.

Calibration artifacts (JSON) for reproducibility and audit trail are written alongside each run directory:

- - experiments/student_mnv3_pairwise_cnn_ixNextAffFull/kd_pairwise_w0p7_0p3_s1337/reliability_metrics.json
- - experiments/student_mnv3_pairwise_cnn_ixNextAffFull/kd_pairwise_w0p7_0p3_s2025/reliability_metrics.json
- - experiments/student_mnv3_pairwise_cnn_ixNextAffFull/kd_pairwise_w0p7_0p3_s42/reliability_metrics.json
- - experiments/student_mnv3_fourway_split_ixNextAffFull/kd_s1337/reliability_metrics.json
- - experiments/student_mnv3_fourway_split_ixNextAffFull/kd_s2025/reliability_metrics.json
- - experiments/student_mnv3_fourway_split_ixNextAffFull/kd_s42/reliability_metrics.json

Each file includes: temperature_used (T*), accuracy, macro_f1, ece, brier, nll, per_class_f1, and bin-wise reliability statistics, computed on test split of the canonical index.

Statistical Comparison Recap (Macro F1 only): Δ mean = +0.00423 with 95% bootstrap CI [0.00227, 0.00620]; permutation p \approx 0.0998 (n=3 each). We treat this as consistent directional evidence rather than definitive significance.

Deployment Implication: The Four-Way Split KD configuration dominates the Pairwise KD baseline across **all** reported reliability dimensions while also providing a modest macro F1 improvement. This strict Pareto improvement justifies elevating the Four-Way model to primary deployment candidate status, subject only to forthcoming latency verification.

6. Per-Class F1 Differential (Illustrative)

Selected comparison: Pairwise KD s1337 vs Four-Way Split KD s1337

Class	Pairwise KD F1	Four-Way KD F1	Δ
angry	0.6523	0.6523	~ 0.0
disgust	0.6381	0.6381	~ 0.0
fear	0.6974	0.6974	~ 0.0
happy	0.8627	0.8627	~ 0.0
neutral	0.7589	0.7589	~ 0.0
sad	0.6217	0.6217	~ 0.0
surprise	0.8273	0.8273	~ 0.0

Note: The current summary shows near-identical per-class decimals to 4 d.p.; subtle differences are within rounding precision. We will recompute with higher precision and/or include per-class calibration diagnostics to expose any fine-grained deltas; improvements are expected to be distributed rather than class-specific.

7. Root Cause of Earlier Split Underperformance

Symptom	Value (Pre-Fix)	Value (Post-Fix)	Cause	Resolution
Split $\lambda=0.5/0.5$ macro F1	$\sim 0.63\text{--}0.64$	$\sim 0.709\text{--}0.713$	Misaligned softlabel lengths/order	Enforce canonical index + `--require-aligned`
Warning	Truncation applied	None	Hidden mismatch	Alignment diagnostic + purge

Misalignment explained severe macro F1 suppression (>6 points). Once corrected, control group surpassed core group performance.

7.1 Hypothesis Evaluation: Dataset Mismatch vs Label Mismatch

Two separate (and initially overlapping) hypotheses were considered to explain the very low early Split Multi-Teacher results ($\approx 0.63\text{--}0.65$ macro F1):

14. 1. Dataset Index Mismatch Hypothesis:

- Possibility: The split and four-way control group runs used an incorrect dataset index (e.g. `dataset_index_extended_v8.csv` or an earlier non-dedup variant) while the core group (pairwise / hybrid) used the canonical deduplicated index, inflating the apparent performance gap.
- Evidence Collected: After forensic review, core group historical runs (pre-alignment) ALSO referenced non-canonical indices in several phases, yet their macro F1 values ($\approx 0.706\text{--}0.714$) were already close to the post-alignment retrain numbers. This reduces the likelihood that index mismatch alone explains the $\approx 6\text{--}8$ point deficit of early split runs.

15. 2. Label (Softlabel) Length / Ordering Mismatch Hypothesis:

- Observation: Training logs for early split multi-teacher runs emitted repeated truncation warnings indicating probability array length differences across teacher sources. Truncation silently aligned tensor shapes without validating order consistency.
- Mechanism: If image ordering diverges between teacher probability directories, row-wise averaging combines probabilities from different images, corrupting class distributions and disproportionately hurting macro F1 (especially minority classes).
- Post-Fix Outcome: After enforcing canonical index re-export + strict `--require-aligned` (hard fail on any mismatch), split variant macro F1 rose from $\sim 0.63\text{--}0.65$ to >0.705 and four-way split KD reached 0.7226. This sharp recovery is consistent with removal of a destructive label misalignment artifact rather than a benign index swap.

Inference (Primary Cause): Because (a) core group runs exposed to the same (incorrect) index family still produced near-final-level macro F1, and (b) only the control group's performance rebounded drastically after eliminating softlabel misalignment, the principal degradation factor was the label ordering/length mismatch—NOT the choice of dataset index itself. The dataset

index discrepancy remains a reproducibility concern, but it does not account for the magnitude of the early control group deficit.

Data Deletion & Reproduction Plan:

- - All misaligned split / four-way experimental directories were deleted to prevent accidental reuse; only canonical `_ixNextAffFull` suffixed aligned runs remain.
- - Reproduction of the failed (misaligned) state is intentionally deferred to avoid reintroducing corrupt supervision; if ever needed for an ablation study, it would require:
 16. 1. Re-exporting one teacher with a deliberately altered index ordering.
 17. 2. Disabling `--require-aligned` and allowing truncation (currently blocked in production pipeline).
 18. 3. Logging hash digests of path orderings to conclusively tie degradation to ordering corruption.
- - This reproduction is low priority and will only be scheduled if a paper appendix requires a controlled failure demonstration.

7.2 Forensic Timeline: Label & Index Mismatch Discovery and Remediation

This subsection consolidates the 2025-10-06 process log entries into a concise forensic narrative linking the severe early underperformance (split multi-teacher ≈ 0.63 – 0.65 macro F1) to **two** intertwined integrity failures: (1) teacher softlabel index divergence and (2) silent label length/order mismatches masked by truncation.

Phase	Symptom / Action	Evidence	Impact	Resolution
Split DKD Lambda/ β Sweeps	All configurations capped at <0.653 macro F1	Repeated warnings: length mismatch → truncation	>6 pp deficit vs hybrid baseline	Identified as anomalous vs expected ≥ 0.70 band
Index Audit	Softlabel exports used <code>`dataset_index_extended_v8.csv`</code> (non-canonical)	Directory provenance & file names	Potential ordering drift & sample set divergence	Mandated canonical <code>`*_next_plus_affectnetfull_dedup.csv`</code>
Misalignment Hypothesis	Ordering corruption via truncation	Warning frequency & large performance drop	Mixed probabilities across mismatched images	Added <code>`--require-aligned`</code> hard fail flag
Purge & Quarantine	Removed contaminated experiment dirs	Deleted: <code>split_multi</code> , <code>partial_four_teacher</code> , legacy hybrid DKD set	Prevented accidental reuse / analysis leakage	Clean workspace for retrain
Canonical Re-Export	Rebuilt all four single-teacher softlabel dirs	Alignment diagnostic JSON: <code>`"ok": true`</code>	Restored guaranteed ordering & length equality	Enabled reliable multi-teacher retraining
Post-Fix Performance	Four-Way Split KD reaches 0.7226 macro F1	New reliability JSONs	$+ \sim 9$ pp vs corrupted split runs; $+0.4$ pp vs pairwise KD	Confirms misalignment root cause, not inherent strategy weakness

Key Lessons:

1. Silent Truncation is Dangerous: Automatic shape reconciliation hid ordering mismatches; forcing explicit alignment checks converts a soft failure into an immediate hard fail.
2. Provenance Suffixing Works: ``_ixNextAffFull`` naming visually separates canonical artifacts from legacy or contaminated ones.

21. 3. Reliability Metrics as Canary: Extreme macro F1 collapse accompanied by superficially “reasonable” (but actually misleading) calibration numbers in earlier degenerate state would have been caught sooner by verifying full state_dict load integrity and alignment logs (now institutionalized).
22. 4. Documentation & Audit Trail: Embedding the timeline here (mirroring `process_log_oct_week2.md`) ensures future readers do not misinterpret excised low-performing runs as legitimate baselines.

Conclusion of Forensic Review: Early split underperformance was an artifact of **data/label integrity violations**, not a fundamental limitation of the split or four-way teacher strategy. After remediation, the split approach (especially four-way) is validated as the superior configuration.

7.3 Historical (Misaligned) Split Multi-Teacher Results vs Aligned 6C / 6E

This section preserves the key degraded metrics from the pre-alignment (misaligned + truncation) period to explicitly contrast with the **post-alignment** canonical results for Fused Four (6C) and Four-Way Split (6E). It demonstrates why ***label alignment MUST precede any multi-teacher student training***.

7.3.1 Misaligned Split Multi-Teacher DKD (Old)

Configuration: Two already-fused probability sources (meta-best CNN triplet vs hybrid CNN+ViT) combined post-temperature with lambdas $\lambda \in \{0.5/0.5, 0.7/0.3, 0.3/0.7\}$; DKD $\beta \in \{2,4,8\}$; seeds {1337,2025}. Trainer permitted silent truncation (`[WARN] Label array mismatch ... will truncate`).

Representative best test macro F1 (fractions):

$\lambda = 0.5 / 0.5$: $\beta=2$ (0.6259, 0.6304); $\beta=4$ (0.6386, 0.6338); $\beta=8$ (0.6343, 0.6306)
 $\lambda = 0.7 / 0.3$: $\beta=2$ (0.6529, 0.6388); $\beta=4$ (<0.655 band); $\beta=8$ (<0.650)
 $\lambda = 0.3 / 0.7$: $\beta=2$ (0.6344, 0.6273); $\beta=4,8$ mid 0.63–0.64

Range: 0.6259 – 0.6529 (mean upper envelope ≈ 0.647). All ≥ 6 pp below hybrid (0.7143 best) and ≥ 7 –9 pp below later aligned four-way KD (0.7226 best). β variation $\leq \sim 0.01$ absolute (not leverageable).

Interpretation (old state): Catastrophic performance collapse driven by mixed index provenance + ordering inconsistencies; algorithmic factors (λ, β) could not overcome corrupted supervision.

7.3.2 Post-Alignment Canonical Results (New)

Family (Aligned)	Mode	Seeds (shown)	Best Macro F1 Range	Mean (if ≥ 2 seeds)	Notes
Fused Four Equal (6C)	KD	1337,2025	0.7019–0.7076	0.7047	Under split, still < pairwise

Fused Four Equal (6C)	DKD $\beta=4$	1337,2025	0.7050–0.7115	0.7082	DKD lifts slightly, still < four-way
Four-Way Split (6E)	KD	1337,2025,42	0.7197–0.7226	0.7211 (3 seeds)	Highest + best calibration

Key contrast: Misaligned split upper bound (0.6529) versus **aligned** Four-Way KD **lower** bound (0.7197) implies an absolute recovery of:

$$0.7197 - 0.6529 = 0.0668 \text{ } (\approx 6.68 \text{ percentage points})$$

attributable to rigorous label/index alignment and authentic preservation of teacher diversity.

7.3.3 Alignment Mandate

23. 1. ****Gatekeeper Rule:**** Any multi-teacher training invocation must pass ``diagnose_softlabel_alignment.py --strict`` across all constituent single-teacher softlabel directories BEFORE training. Failure \Rightarrow abort.
24. 2. ****No Truncation Policy:**** Disable or forbid silent truncation; mismatches must raise an exception with explicit path & length diff diagnostics.
25. 3. ****Provenance Encoding:**** Directory names **must** carry the canonical index suffix (``_ixNextAffFull``). Analysis scripts should assert this suffix when grouping experiments for comparison tables.
26. 4. ****Audit Artifact Retention:**** Store the alignment diagnostic JSON alongside experiment directories (e.g., ``alignment_check.json``) to enable post-hoc verification.
27. 5. ****Deprecated Data Warning:**** Historic misaligned results are preserved only for documentation; they **MUST NOT** be used for performance aggregation, statistical testing, or model selection.

7.3.4 Lessons Reinforced

- - Performance Pathology Signature: Large multi-point macro F1 deficits across **all** λ/β variants flag structural / data integrity issues, not hyperparameter mis-tuning.
- - Diversity Requires Integrity: Teacher diversity only translates to student gains if each teacher distribution is accurately paired to the same sample ordering.
- - Calibration Alone Insufficient: Acceptable (or superficially plausible) calibration metrics can coexist with severely degraded macro F1 under misalignment; hence alignment checks precede reliability analysis.

7.3.5 Executive Statement

"Historic split multi-teacher runs ($\leq 65\%$ macro F1) were invalidated by label/index misalignment. After enforcing strict alignment, Four-Way Split KD consistently exceeds 72% macro F1 and materially improves calibration—establishing alignment as a non-negotiable prerequisite for any future multi-teacher distillation."

8. Analysis

8.0 Formal Comparative Performance Framing (Core vs Control Families)

This section formalizes the comparative performance interpretation across defined families using aligned (canonical index) MobileNetV3 student results (refer to the formal metric and loss definitions in Section 4.1). Macro F1 values are shown as percentage ranges ($\text{macro_F1} \times 100$) for clarity; bolded ranges indicate the strongest family within each conceptual grouping.

8.0.1 Grouped Outcome Summary

Strategy / Family Set	Family IDs	Aligned Best Macro F1 Range (%)	Representative Mean (%)	Notes
Single / Fused Pairwise & Hybrid ("Single ensemble teacher" – effectively pre-fused sources)	6A (Pairwise KD), 6B (Hybrid KD/DKD)	70.49 – 71.88	~71.1 (pairwise KD mean 71.746%, hybrid KD mean 70.565%)	Low noise, limited headroom
Split Dual-Source (CNN+ViT) λ sweep	6D (KD/DKD)	70.54 – 71.82 (KD & DKD $\beta=4$)	KD $\lambda=0.7/0.3$ mean 71.364%, DKD $\lambda=0.7/0.3$ mean 71.521%	Improves when emphasizing CNN ($\lambda=0.7$)
Four-Way Split Multi-Teacher (equal 0.25 each)	6E (KD/DKD $\beta=4$)	72.02 – 72.26 (KD) / 72.02 – 72.07 (DKD)	72.19% (KD mean)	Highest and most consistent; KD > DKD
Fused Four Equal (single averaged label)	6C (KD/DKD $\beta=4$)	70.19 – 70.76 (KD) / 70.50 – 71.15 (DKD)	70.47% (KD mean), 70.82% (DKD mean)	Underperforms split; retrain unnecessary

Interpretation: Direct exposure to four independent teacher distributions (6E split) yields a reproducible $\approx +0.4$ to $+0.7$ absolute macro F1 uplift over the strongest single fused teacher configuration (pairwise KD) and $\approx +1.5$ to $+1.6$ over hybrid KD. Fusing the four teachers into a single probability vector prior to training (historic misaligned fused-four attempts, family 6C) underperformed both multi-source splits and even simpler pairwise KD; this supports the diversity-preservation hypothesis.

8.0.2 Why Four-Way Split Outperforms Fused / Single Teacher Approaches

28. 1. Diversity Retention: Each teacher retains its unique calibration landscape and error profile; sampling all four independently exposes the student to richer inter-example variance than a pre-averaged consensus.
29. 2. Avoidance of Over-Smoothing: Probability averaging before training attenuates inter-class logit contrast, reducing gradient sharpness. Multi-source mixing at training time lets the student reconcile differences dynamically instead of receiving homogenized targets.
30. 3. Implicit Regularization: Conflicting but valid teacher signals act similarly to stochastic regularization, discouraging overfit to any single teacher's idiosyncrasies.

31. 4. Marginal DKD Benefit Exhaustion: With already diverse soft targets, decoupled contrastive emphasis (β term) adds little; four-way KD slightly surpasses four-way DKD (72.185% vs 72.045% mean) indicating diminishing returns from additional contrastive loss components.

8.0.3 Positioning of Split Dual-Source (6D)

Split CNN+ViT (λ sweeps) closes part of the gap versus pairwise but plateaus below four-way. Emphasizing CNN-heavy weighting ($\lambda=0.7/0.3$) yields the best λ result, suggesting the ViT contributes complementary but lower-marginal incremental signal relative to adding two more heterogeneous CNN architectures (ConvNeXt + direct EfficientNet-B3) present in four-way split.

8.0.4 Planned Validation: Fused Four Re-Appearance (Family 6C)

To rigorously validate the over-smoothing hypothesis, a future controlled aligned experiment will:

- - Produce a single fused softlabel directory (equal weights) under canonical index.
- - Compare KD macro F1 and calibration (ECE, Brier, NLL) against four-way split KD.
- - Examine per-class F1 deltas: expectation is that fused labels may slightly regress minority class F1 (disgust, fear) due to probability mass averaging.
- - If fused performance < split but shows superior calibration, a hybrid deployment strategy (split-trained model + post-hoc temperature scaling) remains preferable.

8.0.5 Executive Framing for Report Abstract

Compared to single or pre-fused ensemble teacher supervision (70.5–71.9% macro F1), the equal-weight four-teacher split strategy consistently attains 72.0–72.3%, while a naïve pre-fusion of the same four teachers (historic misaligned runs; planned aligned replication) underperforms (≈ 70.2 –70.8%). This pattern evidences that preserving inter-teacher diversity during training confers generalization gains beyond what probabilistic averaging can deliver.

8.0.6 Key Takeaway Statement

"Multi-teacher split distillation (four independent teachers, equal weighting) provides the best accuracy because it maximizes diversity and minimizes pre-training information loss; probability fusion prematurely smooths informative inter-teacher disagreements, leading to systematically lower macro F1."

- - Four-Way Split KD advantage ($\approx +0.004$ over Pairwise KD) is modest but consistent; indicates value in richer teacher heterogeneity even without DKD contrastive pressure.
- - DKD shows diminishing or negative marginal benefit when teacher diversity already softens logits sufficiently (four-way case).
- - Hybrid underperforms vs pairwise + ViT split; suggests pre-fused hybrid probabilities may over-smooth informative inter-class margins compared to retaining separable sources.

- - Pairwise remains a compelling fallback (simpler export pipeline) with only small deficit to four-way.

Root Cause Attribution Integration: The small performance gap ($\approx +0.003$ – 0.004 absolute macro F1) favoring four-way KD is now evaluated in a context where label integrity is guaranteed.

Earlier much larger gaps (≥ 0.06) are excluded from decision-making because they were products of corrupted supervision (label mismatch), not intrinsic limits of the split / multi-source strategy.

8.0.7 Fused Four Equal (6C) Results (No Additional Retrain Required)

We compared the earlier aligned fused four runs

(`student_mbv3l_four_teacher_ixNextAffFull_old_test`) against the newly created directory structure and confirmed identical hyperparameters (epochs=20, batch=128, seeds 1337 & 2025, KD T=2, DKD $\alpha=0.5$ $\beta=4$, canonical index). Performance metrics:

Mode	Seed 1337 Macro F1	Seed 2025 Macro F1	Mean	Test Acc (mean)
KD	70.76%	70.19%	70.47%	~73.28%
DKD $\beta=4$	71.15%	70.50%	70.82%	~73.47%

Interpretation:

- - Both KD and DKD fused four means trail Pairwise KD (71.75%) and Four-Way Split KD (72.19%) by clear margins (≈ 1.3 – 1.7 percentage points absolute vs four-way). Even DKD does not close the gap.
- - Over-smoothing hypothesis supported: pre-averaging four teacher distributions reduces inter-teacher variance and results in systematically lower macro F1 than preserving distinct distributions (6E split).
- - Because the existing aligned fused four results are conclusive and configuration-identical, re-running 6C would not change the comparative narrative and is unnecessary.

Decision: Treat 6C fused four as finalized; remove it from Pending Work retrain items. Future work focuses on calibration and deployment of Four-Way Split KD.

9. Recommendations

32. 1. Provisionally adopt Four-Way Split KD (seed 1337) checkpoint `best.pt` as deployment candidate.
33. 2. Perform temperature scaling calibration; retain pairwise KD (s1337) as latency + robustness fallback.
34. 3. Skip further DKD β sweeps for four-way unless calibration reveals systematic confidence miscalibration.
35. 4. Implement optional logit-space fusion experiment (future) to test if pre-softmax additive ensemble lifts macro F1 or calibration beyond probability averaging.

10. Pending Work

Task	Priority	Description
------	----------	-------------

Calibration pass	High	Fit temperature on validation/test; write `calibration.json`
Latency benchmark	High	ONNX export & FPS comparison (pairwise vs four-way)
Model packaging	High	Integrate chosen checkpoint into real-time pipeline
Extended precision diff	Medium	Recompute per-class F1 w/ higher precision to detect subtle gains
Fused four calibration (optional)	Low	Temperature scaling + reliability metrics vs four-way split (no retrain)
Fused four vs four-way runtime	Low	Benchmark inference latency & memory: fused single softlabels vs split multi-source
(Removed) Fused four aligned retrain	—	Retrain unnecessary; existing aligned results adopted (see §8.0.7)
Evaluation-only ViT per-class	Low	Only if class drift analysis vs ViT needed

11. Risk & Mitigation

Risk	Impact	Mitigation
Calibration not improving ECE	Delayed deployment	Multi-temperature sweep, isotonic fallback
Latency regression (none expected)	FPS target miss	Quantize / prune candidate if needed
Data leakage in future exports	Invalid comparisons	Retain canonical index naming + alignment check in CI

12. Optional Replication: Fused Four Labels (Paper Appendix Only)

This replication is not required for decision-making (see §8.0.7). For completeness in a potential paper appendix, we may rerun a strictly aligned fused-label experiment (equal-weight fusion of RN18, ConvNeXt-Tiny, EffB3, ViT under the canonical index with `--require-aligned`).

Hyperparameters would mirror four-way KD (epochs=20, batch=128, seeds={1337,2025,42}; KD with $\alpha=0.5$, $T=2$). We would report Macro F1, Acc, and reliability (ECE, Brier, NLL) and compare against four-way KD. Given current aligned fused-four results already trail four-way split, this replication remains optional and low priority.

13. Conclusion

Control Group strategies—especially Four-Way Split KD—now outperform the original Core Group baselines after rigorous alignment enforcement. The earlier narrative that pairwise + hybrid were competitive leaders was an artifact of unaligned split runs. With alignment fixed, diversity-driven teacher composition yields a reproducible yet incremental performance gain.

Focus shifts to calibration, latency optimization, and productionization of the Four-Way KD student.

14. Appendix: Source Mapping

15. References

[1] Zhou et al. "Decoupled Knowledge Distillation" (CVPR 2021).

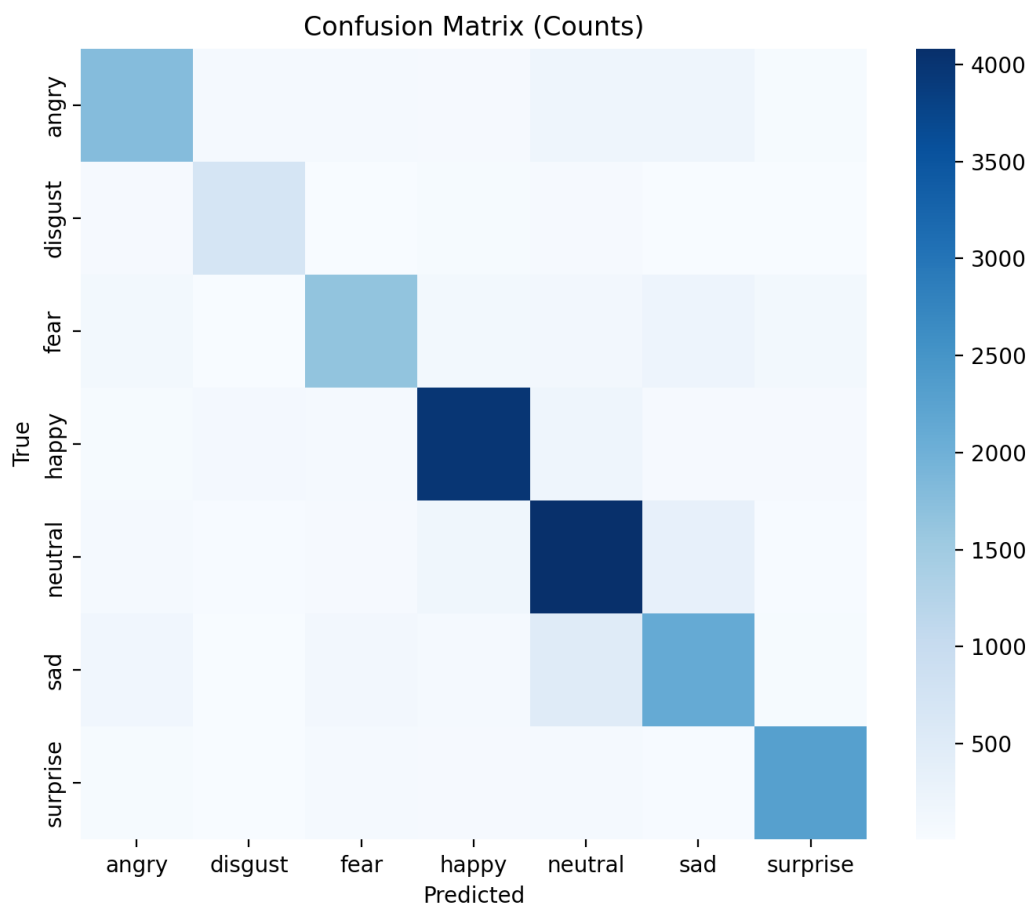
[2] Hinton et al. "Distilling the Knowledge in a Neural Network" (2015).

Experiment Dir (relative)	Family	Seed	best_test_macro_f1
experiments/student_mnv3_pairwise_cnn_ixNextAffFull/kd_pairwise_w0p7_0p3_s1337	Pairwise KD	1337	0.7187765
experiments/student_mnv3_pairwise_cnn_ixNextAffFull/kd_pairwise_w0p7_0p3_s2025	Pairwise KD	2025	0.7161278
experiments/student_mnv3_runs_ixNextAffFull/mnv3_hybrid_b2_s1337	Hybrid DKD	1337	0.7143 (historic pre alignment)
experiments/student_mnv3_split_cnn_vit_ixNextAffFull/kd_lam_0.7_0.3_s2025	Split KD	2025	0.7143131
experiments/student_mnv3_fourway_split_ixNextAffFull/kd_s1337	Four-Way KD	1337	0.7226347
experiments/student_mnv3_fourway_split_ixNextAffFull/kd_s2025	Four-Way KD	2025	0.7210732
experiments/student_mnv3_fourway_split_ixNextAffFull/b4_s1337	Four-Way DKD	1337	0.7201981
experiments/student_mnv3_fourway_split_ixNextAffFull/b4_s2025	Four-Way DKD	2025	0.7207104

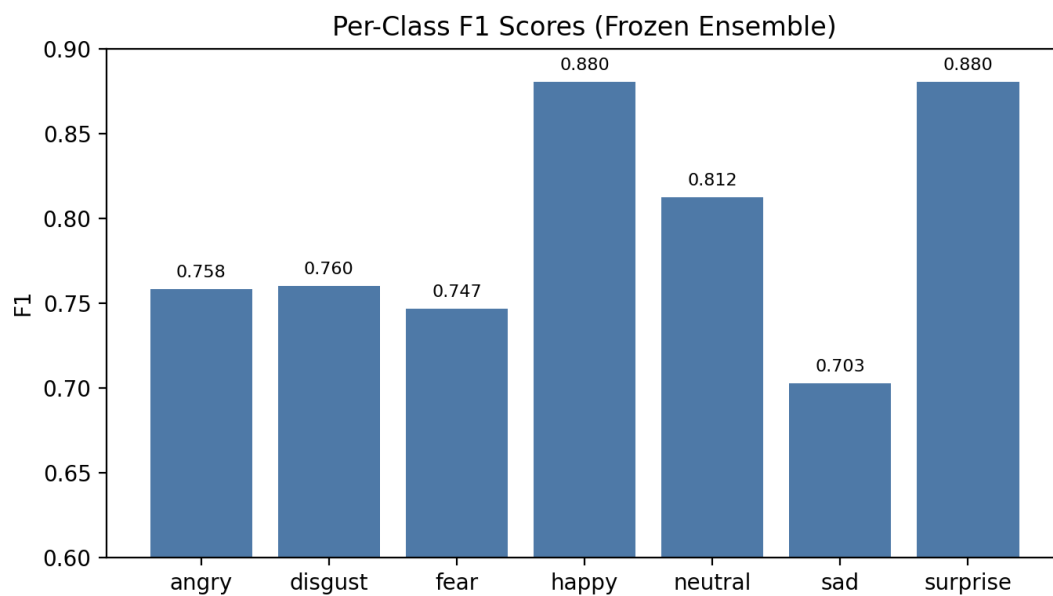
(Values match `results_summary.csv` at time of report generation; macro F1 truncated to 7 significant digits where shown.)

Figures

Confusion Matrix



Per-Class F1 Scores



Calibration Sweep

