Minimal rebuild plan checkpoints

Below is a sharp, dependency-ordered checklist to ensure nothing critical is missing before you execute the five rebuild steps. I've grouped items by area, and highlighted common failure points you've hit before (alignment, calibration, backbone mismatches, NL memory).

---

0. Global foundations and integrity gates

- Canonical index & alignment: Freeze the dataset index you'll use (e.g., the AUGMENTED ixNextAffFull) and run strict softlabel alignment checks before any KD/DKD training.

- Run alignment diagnostic once per teacher softlabel export.

- Keep per-run alignment JSON near experiments; abort on any mismatch.

- Provenance & manifests: Ensure each experiment dir includes args.json, classes.json, calibration.json, reliability_metrics.json, and seed info for reproducibility.

- Backbone consistency: Standardize on timm mobilenetv3large100 for students; refuse loading mismatched checkpoints (<60% param match).

- DKD correctness: Confirm T^2 is applied only to KL terms (not CE). Add a tiny unit test on gradient magnitudes.

- Seed replication: Aim for ≥3 seeds for student comparisons (1337, 2025, 42), and record mean/std.

- Calibration policy: Use lexicographic selection (min NLL → min ECE → max macro-F1). Store T* in calibration.json.

---

1. Teachers (single) for KD/DKD

- Minimum set: RN18, RN50 (optional), EfficientNet-B3, ConvNeXt-Tiny, ViT (only if needed for split diversity rather than performance).

- Exports & integrity:

  - Soft labels: Export per-teacher probabilities at T=2 with canonical ordering (ixNextAffFull suffix).

  - Reliability: Temperature grid and ECE/Brier/NLL JSON for each teacher.

- Ensemble-ready metadata: Per-class F1 and confusion matrices to support later NegL design (teacher-guided complementary labels).

Potential omission:

- Strict alignment artifacts saved with each teacher softlabel dir.

- Teacher calibration artifacts (T*) finalized and stored.

---

2. Ensembles (ensumable teachers) for KD/DKD

- Pairwise & four-way split readiness:

  - Pairwise: RN18+B3 (0.7/0.3).

  - Four-way split: RN18 + ConvNeXt-Tiny + EffB3 + ViT (0.25 each).

- Split vs fused targets: Prefer split (diversity retention) over pre-fused averages unless you need a calibration comparison.

- Class-aware weights (optional): Keep a simple heuristic or a config to skew weights per class if needed.

Potential omission:

- Strict gatekeeper step to verify all constituent teacher dirs pass alignment before starting split KD.

- Documented weight ratios and temperature settings per ensemble (so KD students can replicate supervision).

---

3. Students (core group only)

- Baseline KD/DKD configs:

  - KD: α =0.5, T=2.0; DKD: α =0.5, β ≈4.0, T=2.0 (with T^2 only on KL).

  - 20 epochs, batch 128 (adjust for VRAM), AdamW, cosine schedule.

- Backbone lock: timm mobilenetv3large100; embed classes metadata.

- Reliability pass: Temperature sweep; store calibration.json + reliability_metrics.json.

- Multi-seed runs: ≥3 seeds; aggregate mean/std; bootstrap CI if you're comparing pairwise vs four-way.

Potential omissions:

- Result aggregation into a single resultssummarystudents.csv for quick comparisons.

- Smoke test script (3 – 5 epochs) to catch data/config mismatches before full runs.

---

4. Real-time demo (manual label, video label, live tuning, detectors)

- Manual labeling UI: Keyboard map + clickable bar; logs per-frame CSV and events CSV.

- Transition-fair scoring: Min hold (≥600 ms), 250 ms exclusion windows, jitter/min, time-to-lock.

- Runtime tuning: EMA α, hysteresis δ, vote window/min-count hotkeys; emo-ratio overlay.

- Detectors & geometry hygiene: YuNet preferred; min-face 96; align-eyes; CLAHE; crop-square 224 with margin ~0.30.

- Calibration loading: auto-load calibration.json; global T*; optional per-class thresholds.

Potential omissions:

- Objective scoring script standardized and appended to demoresultssummary.csv.

- Latency benchmarks (CPU/GPU/DirectML) + ONNX export for the chosen student.

- Per-class thresholds and (gentle) logit bias configs saved and versioned for reproducibility.

---

5. NL and NegL on students (KD/DKD)

- NL memory safety before scale:
  - AMP enabled (autocast + GradScaler).
  - Memory module downsized (64→32 dim, 2→1 layers).
  - Gradient accumulation (e.g., 4× to reach effective batch).
  - KD ramp (α 0.3→0.5) to soften early KL.
  - Gradient clipping and memory logging.
- NegL design:
  - Complementary labels guided by teacher confusion matrices.
  - Class-aware negative ratio (lower for minorities to protect recall).
  - Uncertainty gating (apply NegL only on high-entropy/unstable samples).
- Phased integration: NL+KD smoke → NL full run → add NegL → explore NL+NegL+DKD only if stable.

Potential omissions:

- A tiny testbed experiment (few epochs, reduced data) that logs memory and grad norms before full NL runs.
- Config files for NegL ratios, gating thresholds, and class-aware rules, tracked in VCS.

---

6. Domain adaptation and validation (recommended add-ons)

- Hard-sample mining: Extract live segments where pred ≠ truth on minority classes; build webcam validation subset.
- Targeted fine-tuning: 1 – 3 epochs with webcam-style augmentations

(gamma/exposure, motion blur, sensor noise, JPEG).

- Per-class temperature scaling: Vector T per class on webcam validation; re-evaluate thresholds after calibration.

Potential omissions:

- A small, labeled webcam validation set and before/after logs to quantify gap closure.

- Fairness checks: Stratify live metrics by lighting/pose; optionally by participant demographics if feasible.

---

7. Packaging, safety, and communication

- ONNX export + providers: FP16 export; DirectML/CUDA providers; record latency.

- Privacy defaults: No frame saving by default; explicit consent for saved clips; feature-only logs if needed.

- Professor-ready artifacts: One-page executive summary, reproducibility manifest, alignment checks, calibration plots, live demo protocol, planned NL mitigations, and risk log.

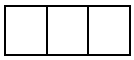
Potential omissions:

- A simple smoke-run and "go/no-go" checklist per stage (Teachers → Ensembles → Students → Demo → NL/NegL).

- Thank-you/update note mapping progress to prior advice and next steps (keeps trust high).

Quick status table and priorities

| Area | Must-have artifacts | Priority |



| Single teachers | Softlabels @ T=2, alignment JSON, calibration.json | High |

| Ensembles | Split configs, alignment verified, weights logged | High |

| Students | 3-seed KD/DKD, calibration + reliability metrics | High |

| Demo | Labeling UI, scoring script, protocol-lite defaults | High |

| NL safety | AMP, downsized memory, grad accumulation, smoke logs | High |

| NegL configs | Teacher-guided complementary labels, class-aware ratios | Medium |

| ONNX + latency | Export + provider benchmarks, config snapshot | Medium |

| Domain adaptation | Hard-sample mining + webcam validation set | Medium |

- Direct answer: You’ re largely complete, but you’ re most likely missing strict per-teacher softlabel alignment artifacts, a standardized calibration.json for each model (teachers and students), backbone consistency guards for students, a smoke test gate before full student/NL runs, ONNX + latency benchmarks, and saved configs for NegL (ratios/gating/class-aware rules). If you want, tell me which of the seven areas you want templates for and I’ ll generate the exact config files and check scripts.