 # Facial Expression Recognition Research Plan (November 2025)

Last Updated: 2025-11-24

Research Question

Can Nested Learning (NL) and Negative Learning (NegL) improve student model training beyond classical Knowledge Distillation (KD) and Decoupled Knowledge Distillation (DKD)?

 # Specific Goals

1. Preserve long-term knowledge through learnable memory modules

2. Adaptively optimize based on student capacity and class imbalance

3. Improve calibration (ECE, Brier, NLL) beyond static KD/DKD

4. Enhance minority class performance (angry, disgust, fear)

Background: Why KD/DKD Need Improvement

 # Classical KD Weaknesses

1. Coupled loss structure — Target CE and non-target KL entangled, suppressing informative high-confidence teacher signals

2. Limited flexibility — Single α parameter cannot independently tune target vs non-target knowledge transfer

3. Logit-only guidance — Missing spatial/feature-level signals for richer generalization

4. Capacity mismatch brittleness — Larger teachers can over-smooth logits, causing student underfitting

 # DKD Weaknesses

1. Still logit-centric — No spatial/positional cues (lacks attention/feature alignment)

2. Hyperparameter sensitivity — α/β interaction + temperature scaling sharply affects gradient magnitudes

3. Teacher reliability dependence — Miscalibrated teacher probabilities amplify non-target KL emphasis

4. Diminishing returns with diversity — When teacher ensemble already rich (e.g., four-way split), additional contrastive weighting (β) adds noise

 # How NL and NegL Address These Issues

Nested Learning (NL) Addresses:

- Catastrophic forgetting → Memory retains cross-class structure across epochs

- Hyperparameter sensitivity → Difficulty-aware gates dynamically adjust weighting

- Real-time jitter → Temporal memory smooths output trajectories

- Teacher miscalibration → Consistency checks flag and dampen overconfident guidance

Negative Learning (NegL) Addresses:

- Poor calibration → Penalizes overconfident wrong predictions

- Blurry boundaries → Adds "repulsion" around incorrect classes

- Domain shift sensitivity → Synthetic noise training improves robustness

Combined Synergy:

- Memory-informed rejection (NL consistency triggers selective NegL)

- Adaptive thresholds (NL smoothing + NegL calibration)

- Minority class protection (NL memory + NegL boundary sharpening)

Research Framework & Progress Tracking

Step	Hypothesis	Status	Key Results	Next Action
1. Verify NL	NL prevents catastrophic forgetting	✅ DONE	RN18 smoke: F1 0.53 (3 epochs), stable training, no gradient explosions	-
2. Verify NegL	NegL improves calibration	⏳ PLANNED	-	Run NegL Phase 1 (see below)
3. Expand Data	More minority samples reduce domain gap	✅ DONE	Added 11,203 images; angry 9.12%, disgust 8.84% (vs 4.95%/low before)	-
4. Choose Teacher	EfficientNet-B3 provides best soft labels	✅ DONE	EffB3 ArcFace: Macro-F1 0.79, minority-F1 0.74, well-calibrated	-
5. Choose Student	MobileNetV3-Large best for calibration	✅ DONE	MBV3-L baseline: Macro-F1 0.72, ECE 0.042 (best calibration)	-
6. Run NL+KD Test	NL+KD improves F1 or calibration	🚫 BLOCKED (OOM)	Phase 1 failed: OOM at batch 128/64/32/16/8 before first checkpoint	Apply Tier 1 mitigation (see below)
6b. Run NegL+KD Test	NegL+KD improves calibration independently	⏳ PLANNED	-	Run after NegL Phase 1
7. Run NL+NegL+KD	Combined NL+NegL synergy improves both F1 and calibration	⏸️ WAITING	-	Run after steps 6 and 6b unblocked
8. Observe Results	Analyze comparative benefits (NL-only vs NegL-only vs NL+NegL)	⏸️ WAITING	-	Run after step 7
9. Decide Next	Hard-sample mining or per-class calibration	⏸️ WAITING	-	Based on step 8 results

 # Literature Weaknesses Addressed

Our Step	Literature Weakness	Our Improvement
1	Prior KD/continual learning suffers catastrophic forgetting	Introduce NL memory layers to preserve old knowledge
2	Existing NegL methods slow convergence, treat all samples equally	Control negative-label ratio, monitor calibration metrics
3	FER datasets lack real-time diversity	Add webcam-style augmentations + minority samples
4	Fused multi-teacher KD causes over-smoothing	Use weighted ensemble split to preserve variance
5	Prior students optimized only for accuracy	Select MBV3 for accuracy-calibration balance
6	KD pipelines rarely combine NL + NegL	Integrate both for memory + boundary improvements
7	Prior work focuses on accuracy, neglects calibration	Explicitly evaluate ECE/NLL alongside F1
8	Studies stop at baseline without iterative refinement	Add adaptive steps (hard-sample mining, per-class calibration)

Current Status: Step 6 Blocked by OOM

 # Phase 1 OOM Incident (2025-11-24)

Event: First full 20-epoch MBV3-L NL+KD attempt (AUGMENTED index, 228,615 samples) failed with repeated CUDA OOM despite batch size reduction 128→64→32→16→8.

Observed Patterns:

- OOM during early epoch 1 before first checkpoint
- Loss guard (>10) triggers frequently
- Gradient norm spikes up to ~150
- Meta-optimizer memory modules (hidden_dim=64, layers=2) + create_graph=True inflate activation retention

Root Causes (Ranked):

1. Absence of AMP — FP32 activations ~1.6–1.8× memory of mixed precision
2. Large memory module — hidden_dim=64, layers=2 too big for initial student NL
3. Meta-graph retention — create_graph=True every 50 batches without warmup
4. Early high KD loss scale — α=0.5, T=2.0 producing spikes & loss guard skips
5. No gradient accumulation — High instantaneous activation footprint

Mitigation Roadmap (Prioritized)

Tier 1: Critical Fixes (Target: 2025-11-24)

Action	Owner	Success Metric
Enable AMP (autocast + GradScaler)	Code	Peak allocated <10.5 GiB
Reduce memory-hidden-dim 64→32; layers 2→1	Code	OOM resolved at batch ≥16
Add gradient accumulation (phys 16 ×4 = eff 64)	Code	Stable epoch 1 completion

Tier 2: Stabilization (Target: 2025-11-25)

Action	Owner	Success Metric
KD α ramp (0.3→0.4 →0.5 epochs 1–4)	Config	<5% batches skip loss guard
Increase loss guard threshold to 15 (disable after epoch 2)	Config	Reduced skipped graph builds
Meta warmup (activate memory at epoch 3)	Code	Lower early graph size

Tier 3: Conditional (Target: 2025-11-26)

Action	Owner	Success Metric
Image size 224→192 (conditional)	Config	Activation memory ↓ ~25%
Selective gradient checkpointing (large blocks)	Code	Additional memory recovery if needed

Unblock Criteria:

- Two consecutive epochs without OOM (physical batch ≥16, effective ≥64 via accumulation)

- Peak allocated <10.5 GiB after step 500

- First `best.pt` checkpoint written

Fallback Plan: If still blocked after Tier 1+2 → run baseline KD (with AMP) on AUGMENTED index to secure updated metrics; resume NL optimization separately.

Alternative Methods (Parallel Exploration)

While fixing OOM, these simpler methods can run in parallel:

 # Tier 1 Alternatives (Immediate)

1. Focal KD (γ=1.5, λ=0.25) — Emphasizes hard examples, addresses minority class directly

2. Class-Balanced α (angry α=0.6, others α=0.5) — Simple per-class weighting

3. Temperature Scheduling (T: 3.0→2.0 over epochs 1-5) — Stabilizes early loss spikes

 # Tier 2 Alternatives (If Tier 1 insufficient)

4. Attention Transfer — Adds spatial supervision (lightweight AT loss, λ_AT=0.1)

5. DKD + Focal NCKD — Applies focal weight to non-target KL

6. Similarity-Preserving KD — Preserves intra-batch structure

Decision Rule:

- If Focal KD alone achieves angry F1 ≥ 0.58: Document simpler method success, compare with NL if/when it works

- If DKD+Focal synergizes (≥ +0.03 F1 over Focal alone): Prioritize DKD path

- Continue NL exploration in parallel for research contribution

Negative Learning (NegL) Standalone Experiments

 # Phase 1: NegL+KD Baseline (Independent Validation)

Purpose: Validate NegL effectiveness independently of NL before attempting combined integration. This establishes whether NegL alone improves calibration and boundary sharpness.

Tier 1: Conservative NegL (Target: 2025-11-25)

Configuration:

- Backbone: MobileNetV3-Large (timm `mobilenetv3_large_100`)
- Base Loss: KD (α=0.5, T=2.0)
- NegL Method: Complementary labels (uniform distribution over non-target classes)
- Negative Ratio: 0.05 (5% of batches receive complementary labels)
- Application: Random sampling (no NL-gating yet, establish baseline)
- Epochs: 20
- Batch Size: 128 (with AMP enabled)
- Seed: 1337

Expected Outcomes:

- ECE ↓ 10-15% vs baseline KD (teacher miscalibration dampened)
- Precision ↑ on confusing pairs (fear/surprise, neutral/sad)
- Macro-F1 ≥ baseline KD (no significant F1 drop)
- Reduced confidence peaks >0.95 (overconfidence mitigation)

Success Criteria:

- ECE ↓ ≥10% over baseline KD (0.042 → ≤0.038)

- Macro-F1 within ±0.3pp of baseline (0.72 ±0.003)

- Angry F1 maintained or improved (≥baseline)

- Training stability: no NaN/Inf, grad norms <80

Failure Indicators:

- Macro-F1 drops >0.5pp → NegL ratio too high (retry with 0.03)

- ECE increases → complementary label distribution suboptimal (try targeted complementary)

- Minority F1 drops >2pp → NegL hurting recall (reduce ratio, add class-specific gates)

Tier 2: Adaptive NegL (Target: 2025-11-26)

Trigger: Tier 1 shows ECE improvement ≥10% but Macro-F1 drops 0.3-0.5pp

Configuration:

- Base: Tier 1 config

- Negative Ratio: 0.08 → 0.05 ramped over epochs 1-5

- Application: Uncertainty-gated (apply NegL only to samples with prediction entropy >threshold)

- Entropy Threshold: 1.5 (empirically tuned; ~uncertain samples get NegL)

- Class-Specific Ratio: Minority classes (angry/disgust) get 0.03 ratio, majority 0.05

Expected Outcomes:

- ECE ↓ 15-20% (adaptive gating improves calibration without harming F1)

- Minority F1 protected (lower NegL exposure for sparse classes)

- Precision/recall balance maintained

Success Criteria:

- ECE ↓ ≥15% over baseline (0.042 → ≤0.036)

- Macro-F1 within ±0.2pp of baseline

- Angry F1 ≥ baseline (no degradation)

- Confusion matrix: off-diagonal reduction for fear/surprise, neutral/sad

Tier 3: Hard Negative Mining (Target: 2025-11-27, Conditional)

Trigger: Tier 2 ECE improves but precision on specific pairs (fear/surprise) still low

Configuration:

- Base: Tier 2 config

- Complementary Label Strategy: Targeted (for fear → emphasize "not surprise"; for neutral → emphasize "not sad")

- Negative Ratio: 0.1 for hard pairs, 0.05 for others

- Mining: Identify hard pairs from Tier 2 confusion matrix; apply focused NegL

Expected Outcomes:

- Precision ↑ 3-5pp on targeted pairs

- ECE ↓ 20-25% (sharper boundaries reduce calibration error)

- Macro-F1 stable or slight improvement (+0.2-0.5pp)

Success Criteria:

- Fear precision ≥0.70 (vs baseline ~0.65)

- Neutral precision ≥0.72 (vs baseline ~0.68)

- ECE ≤0.035

- Macro-F1 ≥0.72

 # Phase 2: NegL+DKD Composition (Optional Extension)

Trigger: NegL Tier 1 or 2 successful (ECE ↓ ≥10% without F1 loss)

Purpose: Test whether NegL synergizes better with DKD's decoupled structure than KD.

Configuration:

- Base Loss: DKD (α=0.5, β=2.0 conservative, T=2.0)
- NegL Application: Apply to NCKD term only (non-target component receives complementary labels)
- Negative Ratio: 0.05 (start conservative)
- Rationale: DKD's explicit TCKD/NCKD split allows targeted NegL to non-target knowledge

Expected Outcomes:

- ECE ↓ 15-20% (DKD + NegL may synergize better than KD + NegL)
- Angry F1 ↑ (DKD emphasis on non-target + NegL boundary sharpening)
- Macro-F1 ↑ 0.3-0.5pp over baseline KD

Success Criteria:

- ECE ≤0.035
- Macro-F1 ≥0.725 (beat baseline KD)
- Angry F1 ≥0.56
- Grad norms stable (<60)

Decision Gate:

- If NegL+DKD > NegL+KD (≥+0.2pp Macro-F1 or ≥5pp ECE reduction): Prioritize DKD path for NL+NegL integration
- If NegL+KD ≈ NegL+DKD: Keep KD path (simpler, less hyperparameter sensitivity)

NL + NegL Combined Integration Plan (Post-Mitigation)

 # Prerequisites (Run First)

1. NL OOM Mitigation: Tier 1 fixes applied (AMP, memory downsizing, accumulation)

2. NegL Phase 1: Tier 1 completed successfully (ECE ↓ ≥10%, F1 stable)

3. Baseline Established: Both NL-only and NegL-only results available for comparison

 # Phase A: Baseline NL+KD (NL-Only, No NegL)

- Config: hidden_dim=32, layers=1, meta-interval=100, NL only (no NegL)

- Goal: Verify NL stability and minority retention (establishes NL-only baseline)

- Success: Macro-F1 ≥ baseline KD; angry F1 stable epoch 1→20; grad norms <50; ECE ↓ ≥10%

- Duration: ~6-8 hours (seed=1337)

 # Phase B: NL+NegL+KD (Memory-Informed Rejection)

- Prerequisite: Phase A successful AND NegL Phase 1 Tier 1 successful

- Config:

  - NL: hidden_dim=32, layers=1, meta-interval=100

  - NegL: negative ratio=0.05 → 0.08 ramped epochs 1-5

  - Synergy Mechanism: NL consistency score triggers NegL application

    - If NL consistency <0.6 (uncertain/drift-prone) → apply NegL complementary loss

    - If NL consistency ≥0.6 (stable) → pure KD loss

  - Class-Aware Gating: Minority classes (angry/disgust) get lower NegL ratio (0.03 vs 0.05)

- Goal: Test synergy—does memory-informed NegL improve both F1 and calibration?

- Success:

  - Macro-F1 ≥ max(NL-only, NegL-only) (synergy improves over individual methods)

  - ECE ↓ ≥20% over baseline KD (0.042 → ≤0.034)

  - Angry F1 ≥ max(NL-only, NegL-only) (minority protection works)

  - Grad norms <60 (NL+NegL stable)

- Duration: ~20 hours

 # Phase C: NL+NegL+DKD (Full Decoupled Integration)

- Prerequisite: Phase B shows synergy (improvement over both NL-only and NegL-only)

- Config:

  - Base: DKD (α=0.5, β=2.0 conservative, T=2.0)

  - NL: hidden_dim=32, layers=1, meta-interval=100

  - NegL: Applied to NCKD term only, negative ratio=0.05, NL-gated

  - Synergy Enhancement:

    - NL consistency <0.6 AND sample in top-50% NCKD loss → apply NegL (double-gated)

    - Minority class override: angry/disgust always use lower NegL ratio (0.03)

- Goal: Maximize synergy—NL memory + DKD decoupling + NegL calibration

- Success:

  - Macro-F1 ≥0.725 (beat baseline KD 0.7226)

  - Angry F1 ≥0.56 (minority improvement)

  - ECE ≤0.035 (strong calibration)

  - Grad norms <50 (stable full integration)

- Duration: ~20 hours; expand to multi-seed (seeds 2025, 42) if successful

 # Phase C+: NL+NegL+DKD+Focal (Maximum Composition)

- Prerequisite: Phase C successful (Macro-F1 ≥0.725, Angry F1 ≥0.56)

- Config:

  - Base: Phase C (NL+NegL+DKD)

  - Add: Focal term (γ=1.5) to NCKD component

  - Add: Class-balanced α (angry α=0.6, others α=0.5)

  - Rationale: Focal emphasizes hard examples; NL memory + NegL protect minorities

- Goal: Push minority F1 to research target (angry F1 ≥0.58)

- Success:

  - Angry F1 ≥0.58 (primary target achieved)

  - ECE ≤0.033 (calibration maintained with focal)

  - Macro-F1 ≥0.73 (overall improvement)

  - Grad norms <50

- Duration: ~20 hours; if successful, this becomes production candidate

 # Phase D: Deployment Tuning (Real-Time Validation)

- Prerequisite: Phase C or C+ successful

- Config:

  - Best model from Phase C/C+

  - Hysteresis (window=5 frames, threshold_delta=0.1)

  - Per-class thresholds tuned to NL+NegL calibration metrics

  - Adaptive Threshold Logic:

    - Use NL smoothed probabilities as base

    - Apply NegL calibration offset (learned from validation set)

    - Add hysteresis for temporal stability

- Goal: Validate real-time stability on webcam

- Success: Flip rate <10%; user feedback positive; latency <50ms

- Duration: Real-time eval on 10 test videos (varied lighting/pose)

 # Decision Gates

After Phase A (NL-Only):

- If Macro-F1 < baseline KD → reduce meta-interval to 200 or downsize hidden_dim to 16; retry

- If Macro-F1 ≥ baseline but ECE no improvement → NL helps F1 but not calibration; still proceed to Phase B (NegL may help ECE)

- If both Macro-F1 ↑ and ECE ↓ → NL alone sufficient; Phase B tests if NegL adds further gains

After NegL Phase 1 (NegL-Only):

- If ECE ↓ ≥10% and F1 stable → NegL validated independently; proceed to Phase B (combined NL+NegL)

- If ECE ↓ <10% → NegL weak; try Tier 2 (adaptive gating) before combining with NL

- If Macro-F1 drops >0.5pp → NegL too aggressive; reduce ratio to 0.03 and retry

After Phase B (NL+NegL+KD):

- Synergy Check: Compare Phase B results vs Phase A (NL-only) and NegL Phase 1 (NegL-only)

  - If Phase B Macro-F1 ≥ max(A, NegL) AND ECE ≤ min(A, NegL) → Synergy confirmed, proceed to Phase C

  - If Phase B ≈ Phase A → NegL adds no value to NL; document null result, skip to deployment with NL-only

  - If Phase B ≈ NegL Phase 1 → NL adds no value to NegL; consider NegL-only path

  - If Phase B < both A and NegL → Negative interaction; analyze (possibly NL consistency gating conflicts with NegL application); fallback to better individual method

- If ECE ↓ ≥20% but Macro-F1 ↓ 0.3-0.5pp → reduce NegL ratio to 0.03, retry

- If precision ↑ without F1 drop and synergy confirmed → proceed to Phase C (DKD integration)

After Phase C (NL+NegL+DKD):

- If Macro-F1 ≥0.725 and Angry F1 ≥0.56 → Success, proceed to Phase C+ (add Focal) or directly to Phase D

- If Macro-F1 <0.725 but ≥0.72 → Marginal gain; compare with Tier 1 alternatives (Focal KD, Class-Balanced α); choose simpler if comparable

- If Angry F1 <0.56 → Add Phase C+ (Focal term + class-balanced α)

- If ECE >0.035 → Add per-class temperature scaling post-hoc; re-evaluate

After Phase C+ (NL+NegL+DKD+Focal):

- If Angry F1 ≥0.58 and ECE ≤0.033 → Production candidate, proceed to Phase D

- If Angry F1 <0.58 → Add class-specific NegL (higher ratio for angry, 0.08 vs 0.03); retry

- If gradient instability (norms >80) → Reduce Focal γ to 1.2, reduce NegL ratio to 0.03; retry

After Phase D (Deployment):

- If flip rate <10% and latency <50ms → Deployment ready

- If flip rate >10% → Increase hysteresis window to 7 frames or raise threshold_delta to 0.15

- If latency >50ms → Optimize NL memory module (quantization/pruning); re-test

- If user feedback negative → Analyze failure cases; add confidence-based UI cues

 # Monitoring Signals

Training-Time:

Signal	Meaning	Remediation
Grad norm spikes >100	NL memory unstable or NegL too aggressive	Reduce NegL ratio; increase grad clip; lower meta-LR
Loss guard triggers >10%	Early instability (KD loss too high)	Apply temperature scheduling (T: 3.0→2.0); raise loss guard to 15
Minority F1 drops after epoch 10	Catastrophic forgetting despite NL	Increase NL memory retention; add minority-class weighting
ECE plateaus despite NegL	Teacher calibration poor or NegL ratio too low	Increase NegL negative ratio to 0.15; add label smoothing (ε=0.05)

Deployment-Time:

Signal	Meaning	Remediation
Flip rate >15%	Temporal smoothing insufficient	Increase hysteresis window; lower per-class threshold sensitivity
Confidence peaks >0.95	Overconfidence persists	Increase NegL during training; apply stricter temperature scaling (T*<1.0)
Accuracy drops in low-light	Domain shift not covered	Add augmentation (brightness/contrast jitter); retrain with harder NegL
User reports "uncertain" too often	Threshold tuning too conservative	Lower per-class thresholds; tighten hysteresis delta

---

Dataset & Baseline Metrics

 # Current Dataset (AUGMENTED)

File: `dataset_index_extended_next_plus_affectnetfull_dedup_new_augmented_angry_disgust.csv`

Class Distribution:

Class	Count	Percent
angry	20,861	9.12%
disgust	20,200	8.84%
fear	28,819	12.61%
happy	48,416	21.18%
neutral	44,576	19.50%
sad	34,701	15.18%
surprise	31,042	13.58%
Total	228,615	100%

Key Improvement:

- Angry: 10,644 → 20,861 (+96% increase to 9.12% from 4.95%)

- Disgust: Significantly boosted for better minority representation

- ✅ Path validation complete: 0 missing files

 # Baseline Metrics (Four-Way Split KD)

Metric	Value	Target
Macro-F1	0.7226	≥ 0.7226 (maintain or improve)
Accuracy	0.7445	≥ 0.7445
ECE (calibrated)	0.0420	≤ 0.0420
Brier	0.36495	≤ 0.36495
NLL	0.7988	≤ 0.7988
Calibrated T*	1.2	(report observed)

Teacher: EfficientNet-B3 ArcFace (Macro-F1 0.79, Minority-F1 0.74)

Student: MobileNetV3-Large (timm `mobilenetv3_large_100`)

Phase 0 Smoke Tests (COMPLETE ✅)

MBV3-L NL+KD Smoke (3 epochs)

- Status: ✅ PASSED
- Val Loss: 1.54–1.95 (stable, well below 3.5 threshold)
- Best Val Macro-F1: 0.4685
- Final Test Macro-F1: 0.4572
- Stability: No loss-guard triggers, gate warmup completed

MBV3-L NL+DKD Safe Smoke (3 epochs)

- Status: ✅ PASSED
- Config: kd_alpha=0.3, dkd_beta=1.0 (conservative)
- Val Loss: ~1.95 (stable)
- Best Val Macro-F1: 0.4595
- Final Test Macro-F1: 0.4546
- Fix Applied: T^2 scaling only to softened KL term (not target CE)

Immediate Next Steps (Priority Order)

 # 1. Implement Tier 1 OOM Mitigation (TODAY - 2025-11-24)

- [ ] Enable AMP (autocast + GradScaler) in `scripts/train_student_nested_learning.py`
- [ ] Add flags: `--amp`, `--memory-hidden-dim` (default 32), `--memory-layers` (default 1), `--accum-steps` (default 4)
- [ ] Update `tools/run_phase1_augmented_dataset.ps1` to pass new flags
- [ ] Success: Peak allocated <10.5 GiB after 500 steps; 2 consecutive epochs without OOM

 # 2. Implement NegL Training Script (TODAY - 2025-11-24)

- [ ] Create `scripts/train_student_negative_learning.py` or add `--enable-negl` flag to existing script

- [ ] Add NegL hyperparameters: `--negl-ratio`, `--negl-method` (complementary/uniform/targeted), `--negl-gate-threshold`

- [ ] Implement complementary label generation (uniform over non-target classes)

- [ ] Add uncertainty-based gating (entropy threshold)

- [ ] Add logging: NegL application rate, per-class NegL exposure, complementary loss magnitude

- [ ] Success: Script runs 3-epoch smoke test without errors

 # 3. Run NegL Phase 1 Tier 1 (TOMORROW - 2025-11-25)

- [ ] Execute NegL+KD (negative ratio=0.05, 20 epochs, seed=1337)

- [ ] Monitor: ECE, Brier, NLL, confusion matrix, confidence distribution

- [ ] Success: ECE ↓ ≥10% vs baseline KD; Macro-F1 within ±0.3pp

- [ ] Decision: If successful, proceed to Tier 2 (adaptive gating); if not, adjust ratio and retry

 # 4. Re-run NL Phase A After OOM Mitigation (PARALLEL - 2025-11-25)

- [ ] Execute NL+KD (hidden_dim=32, layers=1, 20 epochs, seed=1337)

- [ ] Monitor memory usage every 100 steps

- [ ] Success: First `best.pt` checkpoint written; val Macro-F1 ≥ 0.45; ECE ↓ ≥10%

 # 5. Run Tier 1 Alternative Methods (PARALLEL - 2025-11-25)

- [ ] Focal KD (γ=1.5, λ=0.25)

- [ ] Class-Balanced α (angry α=0.6, others α=0.5)

- [ ] Temperature Scheduling (T: 3.0→2.0 over epochs 1-5)

- [ ] Decision: If any achieves angry F1 ≥ 0.58, document as viable simpler alternative

 # 6. Execute NL+NegL Phase B (AFTER 3 & 4 - 2025-11-26)

- [ ] Prerequisite Check: Both NL Phase A and NegL Phase 1 Tier 1 successful

- [ ] Execute NL+NegL+KD with memory-informed rejection (20 epochs, seed=1337)

- [ ] Synergy Analysis: Compare vs NL-only (Phase A) and NegL-only (Phase 1 Tier 1)

- [ ] Apply decision gate: if synergy confirmed, proceed to Phase C; else document interaction

- [ ] Success: Macro-F1 ≥ max(NL-only, NegL-only); ECE ≤ min(NL-only, NegL-only)

 # 7. Execute NL+NegL Phase C (CONDITIONAL - 2025-11-27)

- [ ] Prerequisite: Phase B shows synergy (improvement over individual methods)

- [ ] Execute NL+NegL+DKD (20 epochs, seed=1337)

- [ ] Monitor: Angry F1, ECE, grad norms, NCKD loss magnitude

- [ ] Apply decision gate: if Macro-F1 ≥0.725 and Angry F1 ≥0.56, consider Phase C+ (Focal)

- [ ] Success: Macro-F1 ≥0.725; Angry F1 ≥0.56; ECE ≤0.035

 # 8. Multi-Seed Expansion (CONDITIONAL - 2025-11-28)

- [ ] Prerequisite: Phase C or C+ successful (single-seed targets met)

- [ ] Run winning configuration with seeds 2025 and 42

- [ ] Compute aggregate statistics (mean ± std for Macro-F1, Angry F1, ECE)

- [ ] Success: Mean Macro-F1 ≥0.725; Mean Angry F1 ≥0.56; ECE std <0.005

 # 9. Calibration & Reporting (FINAL - 2025-11-29)

- [ ] Temperature scaling grid search (T ∈ {0.8, 0.9, 1.0, 1.1, 1.2, 1.3})

- [ ] Generate comparative analysis: NL-only vs NegL-only vs NL+NegL vs baselines

- [ ] Generate `Nested_Learning_Student_Study_Report.md`

- [ ] Update process log with final outcomes and decision rationale

- [ ] Archive artifacts and update this plan with final status

Research Contributions (Expected)

1. Methodology: First documented integration of NL (learnable optimizer memory) + NegL (complementary labels) in FER distillation with systematic ablation study

2. Empirical — Comparative Analysis:

- NL-only effects: Minority class retention across epochs, catastrophic forgetting mitigation

- NegL-only effects: Calibration improvement (ECE/NLL reduction), boundary sharpening (precision gains)

- NL+NegL synergy: Combined benefits exceeding individual methods, memory-informed rejection mechanism

- Quantitative synergy metrics: ΔMacro-F1(NL+NegL) vs max(ΔNL, ΔNegL), ΔECE(NL+NegL) vs min(ECE_NL, ECE_NegL)

3. Interaction Analysis:

- Positive synergy: Memory-informed NegL gating improves sample-wise adaptation

- Null interaction: NL and NegL effects independent (additive gains)

- Negative interaction: Consistency-gating conflicts with NegL application (document when/why)

4. Negative Results (if applicable):

- NL memory overhead prohibitive → contributes to "when not to use meta-learning" literature

- NegL harms minority recall → refines understanding of complementary label limitations in imbalanced tasks

- No synergy observed → documents that memory retention and boundary sharpening may be orthogonal in some contexts

5. Deployment Insights: Practical guidelines for hysteresis + per-class thresholds tuned to NL+NegL calibration metrics; transferable to other real-time classification tasks

---

References & Documentation

- Process Log: `research/process_log/process_log_nov_week4.md`

- Path D Results: `research/report/Path_D_Results_Summary_report.md`

- NL Study Report: `research/report/Nested_Learning_Student_Study_Report.md` (to be created)

- Core Baseline Study: Core Group Student Training Study Report §5.1-5.2

- Literature: Zhou et al. CVPR 2021 (DKD), Hinton et al. 2015 (KD), FER surveys

---

Document Version: 2.0 (Restructured 2025-11-24)

Status: Step 6 blocked (OOM); Tier 1 mitigation in progress

Next Review: After Phase 1 unblock (estimated 2025-11-25)