The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering

Project ID: [FYP__]

[Real-time-Facial-Expression-Recognition-System]

by

[Ma Kai Lun Donovan]

[24024192d]

Final Year Project Proposal 2024/2025 Sem 1

Bachelor of Engineering (Honours)

In

[Information Security]

of

The Hong Kong Polytechnic University

Supervisor: Prof. Kenneth Lam

Date:30/9/2025

Abstract

Real-time Facial Expression Recognition (FER) underpins human–computer interaction, adaptive learning, affect-aware analytics, and safety contexts, yet practical deployment remains constrained by four coupled challenges: (i) closing the accuracy gap between high-capacity teacher ensembles and latency-bound edge students; (ii) stabilizing minority (fear, disgust) class performance under severe imbalance and cross-dataset prior shift; (iii) achieving calibrated uncertainty to avoid overconfident misclassifications; and (iv) enforcing reproducibility in multi-stage training, distillation, and compression pipelines. This proposal introduces an integrated, governance-aware framework combining a heterogeneous ArcFace-based teacher ensemble (ResNet18, EfficientNet-B3, ConvNeXt-Tiny) with staged distillation (vanilla KD → Decoupled KD → optional Born-Again / mutual learning) and deployment optimization (structured pruning, post-training quantization with QAT fallback, ONNX export, mixed precision). Imbalance and minority fragility are addressed through logit adjustment, class-balanced loss, selective augmentation, and minority-sensitive acceptance gates. Calibration is treated as a first-class objective via global and conditional per-class temperature scaling with safeguards against sparse-class overfitting. A manifest-centric traceability layer (dataset hashes, hyperparameters, lineage, calibration state, compression steps, reliability signals) plus validator scripts and structured failure tagging target ≥95% successful automated runs.

Measurable targets include: reducing the ensemble→student macro F1 gap to ≤0.05 (or ≤4 absolute points), maintaining minority robustness (no rare class <0.55; rare-class mean ≥0.65), achieving post-distillation Expected Calibration Error ≤5–7%, sustaining single-frame latency ≤12 ms FP32 (≤8 ms compressed), and limiting quantization accuracy drift to ≤2 points with ΔECE ≤0.01. Anticipated outcomes are a calibration-governed, latency-qualified student model; a reproducible compression pathway; and a deployment-ready artifact bundle (models, manifests, metrics, calibration parameters). Temporal distillation and multi-modal fusion are deferred until single-frame governance thresholds are met. By jointly optimizing accuracy, fairness, uncertainty, efficiency, and reproducibility, the project addresses persistent FER gaps and lays a foundation for trustworthy real-time affective computing.

Contents

1. INTRODUCTION

1.1 Overview

Real-time Facial Expression Recognition (FER) aims to infer a person's affective state from live video under constraints of latency, robustness, and resource efficiency. Applications span human–computer interaction, e-learning engagement monitoring, adaptive entertainment, advertising analytics, safety systems, and accessibility technologies. Despite substantial progress with deep convolutional and transformer-based models, practical FER deployment continues to face persistent challenges: (i) inter- and intra-class variability driven by subtle muscle activations and cultural display rules; (ii) severe class imbalance—expressions such as fear and disgust are under-represented compared to happiness or neutral; (iii) covariate shifts across datasets (pose, illumination, occlusion, demographic diversity); (iv) probability miscalibration impacting decision thresholds in safety or interactive settings; (v) performance gaps between powerful teacher ensembles and lightweight student models required for real-time inference; and (vi) reliability and reproducibility issues in multi-stage training pipelines (skipped checkpoints, silent failures, partial artifact generation).

This project adopts a staged strategy: train diverse ArcFace-based teacher models (e.g., ResNet18, EfficientNet-B3, ConvNeXt-Tiny) to encourage complementary error profiles; construct an adaptive weighted + temperature-scaled probability ensemble with fallback micro-sweep logic; apply multi-phase knowledge distillation (vanilla KD temperature grid, Born-Again/self-distillation, and planned Decoupled Knowledge Distillation) to compress ensemble knowledge into resource-efficient students; perform post-hoc scalar and per-class calibration to improve decision reliability; and engineer robustness layers (artifact validation, graceful skipping, structured failure tagging) to stabilize experimentation. Preliminary internal results show the ensemble exceeding the best individual teacher by approximately +3.4 macro F1, while the current best student still lags the ensemble by ~0.14 macro F1—motivating further distillation refinements. The remainder of this Introduction provides background (Section 1.1.1), formalizes the problem (Section 1.1.2), and states the project objectives (Section 1.2).

1.1.1 Background

Facial Expression Recognition traditionally builds upon psychological taxonomies such as Ekman's Facial Action Paradigm, operationalizing a canonical set of discrete categories. In this project we target seven classes: the six basic expressions (Happiness, Sadness, Anger, Fear, Surprise, Disgust) plus Neutral, reflecting common practice in modern FER benchmarks. While this taxonomy simplifies annotation and evaluation, expressions often exist along continuous or compound dimensions; thus discrete labeling introduces ambiguity and class overlap that machine learning models must resolve.

Early FER systems relied on engineered geometric and appearance features (e.g., facial landmark distances, local binary patterns). These approaches struggled with pose variation, illumination changes, and occlusion. The advent of deep convolutional neural networks markedly improved robustness by learning hierarchical feature abstractions directly from pixels. Subsequent advances introduced attention mechanisms to emphasize salient facial regions (eyes, mouth, brows) and mitigate background noise, as well as transformer-based architectures and graph-based models capturing spatial-temporal or structural dependencies among facial components.

A central challenge remains inter- and intra-class variability. Subtle micro-expressions or low-intensity variants of fear and disgust exhibit weak, localized cues; conversely, happiness often presents with strong, high-intensity, readily separable signals, contributing to distributional imbalance. This imbalance biases gradient updates toward majority classes, degrading macro-averaged performance and fairness across expressions. Techniques such as loss re-weighting, margin adjustment, focal or adaptive losses, and data-level augmentation (MixUp, CutMix) seek to counteract this skew, though their efficacy varies with dataset heterogeneity.

Metric learning objectives—exemplified by ArcFace—introduce angular margin constraints to enforce inter-class separation and intra-class compactness in an embedding space. Adapting ArcFace to FER offers benefits for fine-grained discrimination, particularly under class overlap and open-set variations. However, margin tuning and feature norm distribution stability can become sensitive under limited per-class samples, and embedding discriminability may not directly translate to calibrated posterior probabilities required for downstream decision thresholds.

Knowledge Distillation (KD) addresses the performance–efficiency trade-off by transferring dark knowledge (soft relational logits) from stronger teacher models or ensembles to compact students. Classic temperature-scaled KD improves alignment of class similarity structure, while variants like Decoupled KD (DKD) separately treat target vs non-target distributions to reduce gradient interference. Born-Again Networks (BAN) iteratively retrain students from previous generations, sometimes uncovering regularization effects that surpass the original teacher. Yet multi-teacher distillation pipelines can suffer orchestration fragility: missing checkpoints, inconsistent log metadata, or partial failure mid-run—issues this project explicitly mitigates via validator scripts and graceful skip semantics.

Calibration constitutes another pillar in safety- and interaction-oriented FER. Over-confident misclassifications (e.g., falsely inferring anger) can degrade user trust or trigger inappropriate responses. Post-hoc techniques such as global temperature scaling or per-class temperature adjustment improve Expected Calibration Error (ECE) without retraining; however, per-class scaling risks overfitting when minority class sample counts are low, necessitating validation-driven acceptance thresholds.

Real-time constraints impose tight latency (e.g., <30 – 50 ms per frame on commodity GPU) and memory budgets. Ensembles, while performant, increase computational load. This tension motivates a compression trajectory: ensemble → calibrated meta-model → distilled student → potential On-Device (e.g., ONNX-accelerated) deployment. Reliability engineering (artifact health validation, structured failure tagging, automatic skip of incomplete experiments) reduces iteration downtime, thereby accelerating the feedback loop required to converge on an acceptable accuracy–latency operating point.

Finally, dataset heterogeneity (FERPlus, RAF-DB, AffectNet subsets, custom curated indices) introduces distributional shift in label granularity, annotation noise, and class priors. Consolidating multiple sources amplifies coverage but compounds normalization and alignment challenges (label mapping, duplicate removal, balancing). This project's indexed dataset strategy and planned hard sample mining aim to surface rare or misclassified instances for future targeted refinements.

Collectively, these background factors—taxonomy selection, class imbalance, fine-grained feature discrimination, model calibration, efficiency versus accuracy trade-offs, and pipeline reliability—frame the core problem the project seeks to address. The next section (1.1.2) distills these into a precise problem statement delineating current gaps and success criteria.

1.1.2 Problem Statement

Despite recent advances, delivering a deployable real-time FER system that is both accurate and reliable across heterogeneous data sources and class distributions remains unsolved. Current internal findings highlight specific unresolved gaps:

1. Performance Compression Gap: A calibrated teacher ensemble achieves a macro F1 that is ~0.14 higher than the best available distilled student, limiting feasibility for low-latency deployment where the ensemble's computational overhead is prohibitive.

2. Minority Class Fragility: Under-represented expressions (e.g., Fear, Disgust) exhibit unstable per-class F1 and are disproportionately impacted by calibration and margin tuning decisions, threatening fairness and robustness.

3. Multi-Dataset Harmonization Risk: Integrating FERPlus, RAF-DB, and curated subsets introduces label prior shifts, annotation noise, and potential duplication, increasing variance in validation outcomes and hindering generalizable student distillation.

4. Orchestration & Reproducibility Weaknesses: Multi-stage training / distillation pipelines have historically suffered from silent failures (missing checkpoints, partial logs), slowing experimental iteration and obscuring true failure modes.

5. Calibration Under Imbalance: Global temperature scaling improves overall confidence alignment but can mask minority overconfidence; per-class scaling risks overfitting due to sparse validation samples, leaving an unresolved calibration strategy trade-off.

6. Latency–Accuracy Trade-off: Real-time constraints (<~30–50 ms/frame) preclude naïve ensemble deployment; current student architectures underexploit potential compression (e.g., mixed feature-level & logit-level KD, DKD not yet stabilized) leaving latency-friendly accuracy untapped.

7. Distillation Robustness: Advanced multi-teacher distillation variants (DKD, iterative BAN generations) have exhibited instability (early termination, divergence, or regression) due to orchestration sensitivity and hyperparameter coupling (temperature, α/β weighting).

Consequently, the project’ s central problem is to design and validate an integrated pipeline that (a) narrows the ensemble→student macro F1 gap to an operationally acceptable threshold while (b) preserving or improving minority class performance, (c) ensuring calibrated predictive uncertainty, and (d) guaranteeing reproducible, failure-resilient experimentation suitable for transition toward deployment.

Success will be demonstrated by meeting jointly defined, measurable targets: (i) reduce ensemble→student macro F1 gap to ≤0.05 without degrading minority (Fear, Disgust) per-class F1 relative to current baseline; (ii) achieve post-calibration Expected Calibration Error within a predefined tolerance (e.g., ≤5% on validation) without per-class F1 regression >2 percentage points; (iii) sustain end-to-end inference latency within real-time bounds on a reference GPU (profiling after ONNX export); and (iv) eliminate silent pipeline aborts in ≥95% of automated overnight runs through validator-enforced artifact integrity and structured logging.

The subsequent Objectives section (1.2) formalizes these targets into actionable research and engineering goals.

1.2 Objectives

The project objectives translate the problem statement into measurable research, engineering, and deployment outcomes:

1. High-Performance Calibrated Ensemble: Train and calibrate a heterogeneous ArcFace-based teacher ensemble whose macro F1 exceeds the strongest single teacher by ≥3.0 points while maintaining or improving Fear and Disgust per-class F1.

2. Student Performance Compression: Reduce the ensemble→student macro F1 gap to ≤0.05 and limit minority (Fear, Disgust) per-class F1 degradation to ≤2 percentage points versus the current ensemble-calibrated baseline.

3. Robust Multi-Teacher Distillation: Stabilize and document vanilla KD, BAN, and DKD pipelines (with reproducible scripts + configuration artifacts) achieving monotonic non-regression across generations or reporting controlled fallback criteria.

4. Calibration & Reliability: Achieve post-hoc calibration with Expected Calibration Error (ECE) ≤5% while enforcing per-class confidence sanity checks; integrate validator + structured failure tagging to reach ≥95% failure-free overnight automation runs.

5. Real-Time Deployment Readiness: Deliver a distilled ONNX-exported student model with mean per-frame end-to-end inference latency ≤40 ms (reference GPU) and ≥60% GPU memory footprint reduction relative to running the full ensemble concurrently.

6. Multi-Dataset Integrity & Minority Support: Maintain a unified, deduplicated index across FERPlus, RAF-DB, and curated subsets with zero duplicate ID collisions and controlled per-class imbalance (ratio of max/min class counts ≤8:1); implement hard-sample mining pipeline prototype to surface minority misclassifications for future iterative improvement.

Collectively these objectives define the quantitative and qualitative success criteria guiding subsequent methodology, experimentation, and evaluation phases.

1.3 Preliminary Results & Progress Status

Purpose. Provide an early evidence snapshot demonstrating feasibility of the planned pipeline while emphasizing that final test-set–frozen evaluation and ablations are still forthcoming. All metrics below are from development runs prior to final test manifest freeze.

Current Performance Trajectory

Stage / Model (Chron.)	**Acc**	**Macro F1**	**Fear F1***	**Disgust F1***	**Notes / Effect**
M1 Baseline CNN	0.512	(n/a)**	–	–	Plain CE; minority under-represented; macro F1 not logged initially
M1w Weighted CE	0.480	0.413	(mod. gain)	High recall / low precision (R≈0.703 / P≈0.106)	Inverse freq weights; over-predicts minority
M1w2 Weighted + Focal	0.470	0.432	Stable	P=0.142 / R=0.568	Focal + capped weights improved precision balance
M2 Metric (CE + contrastive)	0.543	0.439	Slight	Collapse (0.0 F1)	Representation separation ↑ but disgust decision boundary collapsed

CBAM Teacher (smoke)	0.724	0.676	–	0.411	Early attention pilot; substantial macro F1 lift
RN18 ArcFace (stable EMA)	0.752	0.729	–	–	Stable teacher baseline (pre-ensemble)
EfficientNet-B3 ArcFace	0.778	0.756	–	–	Current best single teacher
(Ensemble Pilot)**	(↑ vs best)	+~3.4 over single RN18	–	–	Weighted/temperature fusion (detailed calibration pending)
Student (RN18 KD from earlier teacher)	~0.62	~0.62	–	–	Vanilla KD; large remaining gap
Student (target post-DKD, projected)	(TBD)	≥0.70 †	≥0.55	≥0.55	Based on planned DKD + refined teacher mix

* Minority class per-class F1 values shown only where explicitly measured in logs; dash indicates not yet collected or not emphasized at that stage.

** Early baseline macro F1 not recorded; subsequent stages will re-run baseline for a normalized comparison once test manifest is frozen.

† Projection: Derived from typical DKD lifts (0.5–1.5 macro F1 points) plus improved teacher calibration.

Gap Analysis.

- Ensemble → current student macro F1 gap ≈ 0.13–0.14 (above target ≤0.05 threshold) justifying DKD, multi-teacher KD, and pruning-aware recovery.

- Minority class (fear / disgust) remains the volatility focus; early methods either over-predicted (M1w) or collapsed under metric learning (M2), validating the planned staged imbalance + distillation strategy.

Calibration & Latency (Status).

- Calibration: Global temperature scaling planned right after ensemble stabilization (Phase P1 exit). Per-class scaling gated on classwise ECE > 2× global and sufficient sample count.

- Latency: End-to-end profiling not yet executed. Teacher reference path will establish baseline; student target remains p95 ≤12 ms model-only, ≤70 ms full pipeline (Phase P6).

Timeline Adherence.

Phase	**Planned Focus**	**Status**	**Notes**
P0 Reliability Hardening	Validators, skip logic	Completed	≥95% run success recorded
P1 Teacher Finalization	Teacher sweeps + calibration	In progress (single best teacher locked)	Ensemble weighting & calibration pending
P2 Student KD Phase 1	Vanilla KD stabilization	Partially done (initial KD)	Needs DKD refinement
P3 DKD / Refinement	DKD α/β grid	Pending	Next immediate step
Subsequent (P4+)	Compression, Quantization, Latency	Not started	Blocked on student gap reduction

Test Set Freeze Notice.

The definitive test manifest (hash) will be locked at the end of Phase P1. All performance figures above should be considered developmental and may be rebaseline'd once the frozen split is established. Future reporting will distinguish "validation-tuned" vs "final test" results.

Planned Near-Term Actions (Next 2 Weeks).

1. Finalize ensemble weight/temperature + calibrate (global temperature).

2. DKD grid (α, β) to close ≥50% of current student gap.

3. Re-run early baseline with uniform logging (macro F1 + per-class F1) under frozen test manifest.

4. Establish latency harness; record teacher vs provisional student timing.

5. Begin structured pruning pilot only after DKD student accepted.

2. LITERATURE REVIEW

This section surveys core strands of research underpinning real-time Facial Expression Recognition (FER): foundational handcrafted and early deep learning approaches; advances in attention, multi-scale, region- and geometry-guided modeling; temporal and graph-based sequence modeling; and cross-cutting methods in metric learning, knowledge distillation, calibration, and class imbalance mitigation. For each category we critically assess representative methods, highlighting strengths, limitations, and relevance to the objectives defined in Section 1.2. Citations are placeholders to be normalized to a final reference style (IEEE / APA) during consolidation.

2.1 Evolution of Facial Expression Recognition Paradigms

This section traces the methodological progression from classical feature engineering to modern attention and transformer architectures. The trajectory illustrates an increasing shift from static, global representations toward structured, context-aware, and efficiency-conscious designs aligned with real-time constraints.

2.1.1 Classical & Early Deep Learning Methods

Handcrafted Feature Era: Early FER pipelines decomposed the problem into face detection, landmark localization, feature extraction (Gabor filters, Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), geometric distances), followed by shallow classifiers (SVM, k-NN). Strengths included interpretability and low computational overhead; however, performance degraded under pose, illumination variation, partial occlusion, and inter-person appearance diversity. The lack of learned hierarchical abstraction limited generalization to in-the-wild contexts.

Transition to Deep CNNs: The adoption of convolutional neural networks supplanted manual descriptors with end-to-end learned feature hierarchies. Baseline CNN architectures (e.g., shallow custom stacks) improved robustness to noise and modest pose changes. Subsequent deeper backbones (ResNet variants, EfficientNet scaling strategies) introduced residual learning, compound scaling, and improved parameter efficiency. Strengths: significantly higher accuracy and better invariance. Weaknesses: increased computational cost and tendency toward overconfidence without explicit calibration.

Region & Part-Based Enhancements: Intermediate-stage approaches integrated facial landmark priors—cropping or re-weighting local regions (eyes, mouth, brows)—to emphasize expression-discriminative zones. These methods improved recognition of subtle or localized expressions (e.g., fear micro-tension), but introduced dependencies on accurate landmark detection and occasional brittleness with occlusions or profile views.

Multi-Scale & Feature Fusion: To capture both coarse global configuration and fine-grained local action units, multi-branch or pyramid feature fusion architectures emerged.

While boosting performance on nuanced expressions, naive fusion inflated latency and memory if not carefully pruned—challenging real-time deployment goals.

Limitations Summary:

- Generalization: Handcrafted + shallow models fail under wild conditions (pose/lighting/occlusion).

- Efficiency vs. Capacity: Deep models raise inference cost; uncompressed versions hinder real-time use.

- Localization Dependence: Region-focused methods introduce error cascades from landmark failures.

- Calibration: Most early and mid-stage CNN works report accuracy but omit confidence calibration, risking misaligned decision thresholds.

Relevance to Project: Motivates our reliance on modern but still moderately lightweight teacher backbones (ResNet18, EfficientNet-B3, ConvNeXt-Tiny) and later compression via distillation. Highlights necessity of balancing discriminative capacity with latency and calibration interventions.

2.1.2 Attention, Transformers, Spatial–Temporal & Geometry-Guided Models

Attention Mechanisms: Channel and spatial attention modules (e.g., SE blocks, CBAM-like variants) refine internal feature weighting, often yielding incremental macro F1 gains with modest parameter overhead. Strength: improves discrimination for subtle classes by amplifying salient activations. Limitation: risk of overfitting minority expressions if attention maps become class-frequency biased.

Transformer-Based FER: Vision Transformer (ViT) derivatives and hybrid CNN–Transformer pipelines introduce global self-attention, capturing long-range pixel or patch interactions. Attentional Selective Fusion (ASF) architectures selectively gate multi-scale tokens, enhancing context integration. Strengths: richer global context modeling; improved robustness to spatial displacement. Weaknesses: higher quadratic attention cost, demanding substantial data for stable training; may exceed real-time constraints without pruning or distillation.

Spatial–Temporal & Graph Fusion: Methods leveraging landmark-driven spatial graphs combined with temporal sequence modeling (graph convolution + transformers or temporal convolutions) address dynamic expression evolution. Strength: improved temporal stability and handling of micro-expression transitions. Limitation: requires consistent high-quality landmark tracking; added latency in sequence assembly.

Region & Multi-Stream Fusion (Images + Landmarks / Geometry): Dual-input networks (raw image + landmark heatmaps or coordinate embeddings) inject geometric priors, often boosting performance on ambiguous classes. Strength: complementary modality synergy. Weakness: extra preprocessing pipeline; potential compounding error from inaccurate detection.

Lightweight & Binary / Quantized Hybrids: Recent studies explore binarized or mixed-precision convolution–transformer fusion to reconcile accuracy and edge deployment. Benefit: large memory and speed gains. Risk: accuracy degradation on minority or subtle expressions; increased training instability.

Comparative Strengths & Weaknesses Table (planned for final manuscript): (i) Accuracy lift vs best baseline; (ii) Added latency cost; (iii) Minority-class impact; (iv) Calibration side-effects; (v) Complexity of integration.

Relevance to Project: We selectively adopt lighter attention (potential CBAM-lite) in pilot phases rather than full transformer stacks to avoid derailing latency targets and complexity. Geometry-guided and temporal methods are earmarked as potential future extensions once single-frame performance and compression objectives are met.

2.2 Metric Learning, Distillation, Calibration & Imbalance Handling

This section synthesizes cross-cutting methodological axes that directly align with the project’s objectives for performance compression, reliability, and fairness.

Metric Learning (ArcFace & Variants): Angular margin-based losses (ArcFace, CosFace) enforce inter-class separation and intra-class compactness, improving fine-grained discriminability critical for subtle expressions (fear vs. neutral micro-activations). Strength: structured embedding space facilitating ensemble complementarity and potential feature reuse for downstream tasks. Limitation: margin hyperparameters sensitive to class imbalance; may exacerbate minority underfitting if not re-weighted.

Knowledge Distillation Spectrum: (i) Vanilla temperature-scaled KD softens logits to transfer dark knowledge; (ii) Decoupled KD (DKD) separates target vs non-target components to reduce gradient competition; (iii) Born-Again Networks (iterative self-distillation) potentially smooth decision boundaries and regularize overconfident modes. Strength: compress ensemble accuracy into deployable students; provides calibration benefits via softened supervision. Weakness: instability with multiple heterogeneous teachers; hyperparameter coupling (temperature, α/β weights) demands systematic sweeps—raising orchestration complexity.

Ensemble Weight Optimization & Calibration: Weighted probability fusion combined with temperature scaling can yield additive gains beyond any constituent teacher by exploiting divergent error profiles. Strength: macro F1 lift (+3.0–3.5 observed internally). Limitation: inference cost linear in number of teachers; potential overfitting if weight search not regularized; uncalibrated ensembles may appear overconfident without post-hoc scaling.

Post-Hoc Calibration: Global temperature scaling provides low-cost ECE reduction; per-class temperature scaling offers finer granularity at overfitting risk in minority classes. Planned approach: accept per-class scaling only if validation lift > threshold and minority F1 not degraded.

Class Imbalance Mitigation: Techniques include loss re-weighting, logit adjustment (subtract $\tau \log \pi$ to counter prior bias), focal or class-balanced losses, augmentation (MixUp/CutMix), and dataset curation (synthetic balancing). Strength: minority performance stabilization. Limitation: overcompensation can inflate false positives for rare classes; some methods (e.g., focal loss) complicate calibration.

Reliability & Orchestration Engineering: Artifact validation scripts, graceful skip semantics, structured failure tagging increase experimental throughput and reproducibility—foundational for systematically exploring KD and calibration search spaces.

Integration Rationale: The project combines ArcFace-based metric learning (teachers) → adaptive ensemble + calibration → staged distillation (KD/BAN/DKD) → ONNX compression—each stage addressing a distinct axis of the central problem (accuracy, reliability, latency, fairness, uncertainty quantification).

Planned Figure / Table Assets:

- Comparative chart: Teacher vs Ensemble vs Student (macro & per-class F1; ECE pre/post calibration).

- Distillation temperature / α–β sensitivity plot (DKD hyperparameter landscape once stabilized).

- Imbalance mitigation ablation (baseline vs logit adjustment vs re-weighted margin).

2.3 Temporal & Multi-Modal Expansion (Future Work)

Although the current project scope prioritizes single-frame inference for controlled real-time latency, temporal and multi-modal FER strands offer future uplift opportunities.

Video & Sequence Modeling: Temporal CNNs, 3D convolutions, ConvLSTM hybrids, and transformer-based temporal encoders model dynamic evolution of Action Units, smoothing frame-level noise and capturing onset–apex–offset phases. Strength: improved robustness to transient misclassifications and micro-expression detection. Limitation: sequence buffering introduces latency; temporal models demand consistent frame rate and alignment.

Graph-Based Landmark Dynamics: Spatial–temporal graph convolutional networks (ST-GCN) and geometry-guided transformer hybrids treat facial landmarks as dynamic nodes with edges encoding anatomical or learned adjacency. Strength: explicit structural bias; aids disambiguation of subtle localized motion (e.g., brow vs eyelid). Limitation: degradation under landmark detection error or partial occlusion.

Micro-Expression Recognition: Requires high temporal resolution (spotting ~1/25 to 1/5 second activations). Specialized optical flow, strain maps, or local temporal differential encoders are common. Strength: addresses nuanced user affect signals; potential extension for advanced HCI contexts. Limitation: dataset scarcity, labeling difficulty, and questionable generalization to broad affect classification.

Multi-Modal Fusion: Incorporating audio prosody, physiological signals (e.g., heart rate variability), or text (contextual cues) can mitigate facial ambiguity. Fusion architectures (early concatenation, cross-attention, modality gating) improve robustness under occlusion or neutral expressions masking internal affect. Limitation: acquisition complexity; synchronization overhead; privacy concerns.

Latency & Resource Implications: Temporal stacking / multi-modal ingestion inflates memory bandwidth and may double/triple inference time unless pipelined (e.g., asynchronous landmark stream) or distilled into single-frame surrogates (teacher sequence model supervising student frame model).

Relevance & Planned Path: Once single-frame student meets compression targets, a teacher-level temporal or graph-augmented model can act as a supervisory distillation source to transfer temporal context into frame-level representations (knowledge projection), preserving low latency while harvesting temporal structure gains.

2.4 Deployment & Edge Optimization Literature

Bridging research-grade models to deployable real-time systems involves a complementary body of work on representation compression, graph optimization, and hardware-aware inference.

ONNX & Graph-Level Optimizations: Operator fusion (Conv+BN+ReLU), constant folding, shape inference, and elimination of redundant transpose/reshape ops reduce runtime overhead. Strength: transparent portability across runtimes (ONNX Runtime, TensorRT, DirectML). Limitation: certain custom ArcFace margin ops may require graph rewriting or fallback kernels.

Quantization Strategies: Post-Training Quantization (PTQ) enables rapid INT8 deployment with calibration data; Quantization-Aware Training (QAT) preserves accuracy better on sensitive layers (first/last, attention projections). Strength: 2–4× speed & memory improvements. Limitation: minority-class gradients may become noisier; careful per-channel quantization selection needed.

Pruning & Structured Sparsity: Channel or filter pruning guided by magnitude, sensitivity analysis, or L1 regularization shrinks backbone footprint. Unstructured sparsity requires hardware support to realize speedups; structured pruning maps better to commodity GPU kernels. Trade-off: over-pruning early layers harms subtle expression cues.

Knowledge Distillation for Compression: Complementary to pruning/quantization—teacher soft targets stabilize fine-tuning after aggressive compression steps, mitigating accuracy loss.

Mixed Precision & Memory Bandwidth: FP16 / BF16 inference reduces memory traffic and improves throughput on modern GPUs; calibration of loss scaling required if used during training refinement.

Hardware-Specific Accelerators: TensorRT engine building (layer fusion, kernel auto-tuning) or DirectML execution backends provide additional latency reductions; dynamic shape handling must be validated for variable batch size (e.g., 1 real-time frame vs batched analysis).

Latency Profiling & Bottleneck Attribution: Literature emphasizes pairing model-level optimization with pipeline profiling (preprocess → model → postprocess). Face detection often dominates for small backbones—suggests exploring lightweight detector sharing across frames (tracking-assisted detection) before further classifier compression.

Model Governance & Reliability: Deployment studies highlight the necessity of runtime health metrics (confidence drift, per-class distribution shift monitoring) and rollback strategies. Integration: export calibration parameters (temperature, per-class scaling) with model artifact for consistent inference semantics.

Relevance to Project Roadmap: Sequence of planned deployment optimizations—(1) baseline ONNX export + correctness parity test; (2) global temperature calibration embedding; (3) optional per-class scaling injection; (4) latency profiling; (5) selective pruning + KD recovery; (6) INT8 PTQ trial; (7) QAT if >2–3% macro F1 drop; (8) final engine build (TensorRT/DirectML) with monitoring hooks.

Open Risks: Potential accuracy cliff under aggressive INT8 quantization for minority expressions; need for calibration dataset representativeness; custom ArcFace margin layer export correctness.

Mitigation: Maintain golden FP32 evaluation harness; incorporate automated regression checks (macro & per-class F1 deltas thresholds) post-optimization; integrate embedding cosine distribution monitoring to catch representation drift.

2.5 Summary of Literature Gaps & Alignment

Across paradigms, consistent gaps persist: (i) limited simultaneous optimization of accuracy, calibration, and latency; (ii) under-reporting of minority class robustness after compression; (iii) sparse reproducibility practices (artifact validation, failure tagging) in academic FER pipelines; (iv) insufficient exploration of structured temporal knowledge distillation for single-frame students. The present project explicitly targets these lacunae via an integrated architecture + reliability + deployment strategy.

3. METHODOLOGY

This section details the end-to-end plan to achieve the project objectives: high-accuracy, well-calibrated, real-time facial expression recognition (FER) with reproducible training and deployment readiness. It operationalizes the literature review insights into concrete pipeline stages, experimental protocols, resource planning, evaluation, and risk mitigation.

3.1 Overall Research Design

We adopt an iterative, data-driven systems approach with four tightly coupled layers:

1. Data Curation & Pre-Processing
2. Teacher Ensemble Training & Reliability Instrumentation
3. Knowledge Distillation, Imbalance Mitigation & Calibration
4. Deployment Optimization (Compression → Inference Validation)

Each cycle produces measurable artifacts (metrics.json, calibration profiles, ONNX exports, latency reports) enabling objective gating before proceeding to the next layer.

3.2 Datasets & Splitting Strategy

Primary supervised datasets (already integrated): FERPlus, RAF-DB, AffectNet (full + curated subsets), custom curated set (domain adaptations). Optional future: EXPW, micro-expression sequences (for temporal teacher pilots).

Splitting Principles:

- Maintain stratified train/val/test splits per source dataset; avoid label leakage via image filename hashing.

- Unified evaluation test set (fixed manifest) combining balanced samples across datasets for macro F1 stability.

- Cross-dataset generalization checks: train (FERPlus+RAF-DB) → test (AffectNet subset) and vice versa to quantify domain robustness.

- Class Imbalance Tracking: Store class prior vector π in each split manifest to drive logit adjustment ([5]) and class-balanced loss experiments ([8]).

Data Integrity & Auditing:

- Pre-flight integrity script (already implemented) validates: image readability, label enumeration, per-class counts, duplicate hashes, missing file references.

- Hard sample logging pipeline (misclassified or low-confidence < τ_p) to support targeted augmentation or curriculum staging.

Augmentation Policy:

- Baseline: Random horizontal flip, light color jitter (±10% brightness/contrast), random crop (retain ≥92% area), optional CutOut (pilot). Avoid heavy distortions that degrade calibration.

- Advanced (phase 2): MixUp/CutMix for imbalance smoothing—only if calibration degradation < ΔECE 0.01 on validation.

3.3 Pre-Processing & Face Handling

- Detector: SSD-based OpenCV face detector (already in `models/face_detector/`). Future optional YOLOv8-lite pilot if detection latency > target.

- Alignment: Minimal (center crop) to preserve real-time throughput unless misalignment error rate (facial ROI deviation >10%) exceeds threshold.

- Normalization: Per-channel mean/std (dataset fusion uses FERPlus stats baseline). Track shift if AffectNet weighting >40% (trigger recalculation).

3.4 Teacher Models & Ensemble

Teacher Backbones (selected for diversity):

- ResNet18 ArcFace variant (baseline angular margin) [1]

- Efficient / ConvNeXt-Tiny variant (improved feature richness; stability trade-offs)

- Optional Transformer / attention pilot (CBAM-lite proxy for ASF [11]) if ensemble stagnates.

Training Configuration:

- Loss: ArcFace angular margin softmax (m=0.5–0.6) + optional focal reweighting for minority classes (pilot only if minority F1 < target by >0.05 after logit adjustment).

- Optimizer: AdamW (lr warmup 5 epochs → cosine decay, weight decay 1e-4). SWA phase last 15–20% epochs if validation loss plateau (<0.5% improvement over 5 epochs) [7].

- Early Stop Guard: Minimum epoch floor to avoid premature exit due to minority volatility (e.g., 40 epochs baseline).

Ensemble Fusion:

- Probability weighted average with temperature T (global scalar) tuned via validation macro F1 + ECE Pareto search.

- Weight Optimization: Constrained grid + local refinement (simplex) restricted to convex weight simplex (Σ w_i = 1, w_i ≥ 0). Fallback: uniform if solver fails.

- Calibration Stage: Post-fusion temperature scaling (global) -> optional per-class scaling if classwise ECE > 2× global ECE and sample count ≥ threshold to avoid overfit ([6]).

Acceptance Criteria Before Distillation:

- Ensemble macro F1 ≥ best single teacher + 2.0 percentage points.

- Minority (lowest 2 classes) average F1 ≥ 0.65.

- Global ECE ≤ 0.06 (post temperature scaling).

3.5 Distillation & Student Compression Strategy

Phases:

1. Vanilla KD ([2]) warm start: T_kd ∈ {2, 4}; match softened logits + CE to hard labels.

2. DKD refinement ([4]): Optimize α (target class) and β (non-target contrast) with coarse grid; retain best macro F1 vs calibration trade-off.

3. BAN / Deep Mutual Learning hybrid ([3], [19]) if plateau persists: spawn auxiliary student with reciprocal KL; keep best.

4. Optional Contrastive Representation Assist: Pretrain backbone using SimCLR / MAE ([16], [17]) if teacher–student gap > target after phase 2.

Compression Techniques:

- Architectural Downsizing: ResNet18→Mobile/Shuffle variant or slimmed ConvNeXt (channel prune ratio 0.5) guided by structured pruning importance (L1 norm) [14].

- Quantization Path: Post-Training Dynamic → INT8 PTQ (activation percentile calibration) → QAT if accuracy drop >1.5 points vs FP32 [15].

Distillation Loss Composition (Phase 2 exemplar):

L_total = λ_ce * CE + λ_kd * KL(T_kd) + λ_dkd_t * DKD_target + λ_dkd_nt * DKD_non_target + λ_reg * (SWA consistency or L2).

Hyperparameter Governance:

- Maintain JSON manifest per distillation run (records teacher weights, temperatures, α/β, λ coefficients, dataset hash). Ensures reproducibility.

Success Criteria Student:

- Macro F1 within 3–4 points of ensemble.

- Model size reduction ≥ 4×.

- Latency (FP32) ≤ 12 ms/frame (≈ ≥80 FPS headroom pre-quantization on target GPU / CPU hybrid test).

- ECE ≤ 0.075 (post student calibration).

3.6 Imbalance Mitigation

Sequence:

1. Baseline CE + class sampling analysis.

2. Logit Adjustment (inference-time or last-layer training reparameter) using τ chosen via validation grid ([5]).

3. Class-Balanced Loss with effective number weighting ([8]) if minority F1 still < target.

4. Focal Loss (γ=1–2) pilot only if overconfidence persists (watch ECE regression >0.01).

Monitoring:

- Track per-class precision, recall, F1; maintain rolling 95% CI using Wilson interval for minority classes.

- Abort adoption of additional imbalance method if macro F1 gain <0.5 but ECE increases >0.01.

3.7 Calibration Framework

Steps:

1. Collect validation logits (teachers, ensemble, student) without augmentation.

2. Fit global temperature scaling (minimize NLL) [6].

3. Evaluate ECE (15 bin adaptive) + classwise ECE.

4. Conditional: Per-class temperature or vector scaling only if (a) per-class ECE > 2× global and (b) class count > 300.

5. Store calibration parameters alongside model artifact (JSON). Apply at inference wrapper stage.

Confidence Thresholding:

- Provide abstention / unknown output if max calibrated probability < 0.35 (tunable) to improve downstream reliability in safety contexts.

3.8 Deployment Optimization Pipeline

Ordered Roadmap:

1. Baseline ONNX export (opset pinned). Verify numerical parity (top-1 agreement ≥ 99%).

2. Pruning (structured) targeting 30–40% FLOP reduction – followed by light KD recovery (phase 1 recipe).

3. INT8 PTQ using representative calibration set (≈ 2–3k stratified images). If macro F1 drop >1.5 → escalate to QAT.

4. Mixed Precision (FP16) for GPU path; measure latency vs INT8 CPU path.

5. Engine Build: (Optional) TensorRT or DirectML for Windows acceleration; compare with onnxruntime EP.

6. Edge Validation: Stress test batch=1 latency, warm start (100 inferences), cold start load time, memory footprint.

Acceptance Gates:

- Final compressed INT8 (or FP16) model macro F1 drop ≤ 2.0 points vs FP32 student.

- Calibrated ECE increase ≤ 0.01 post-compression.

- Latency improvement ≥ 30% vs FP32 baseline OR energy (estimated via throughput × TDP proxy) improvement if latency already < target.

3.9 Evaluation Plan & Metrics

Primary Metrics:

- Macro F1 (core objective) & per-class F1 (minority emphasis).

- ECE & classwise ECE (calibration quality) [6].

- Latency (ms/frame) & FPS (median over 500 warm frames; p95 reported).

- Model Size (MB) & Parameter Count.

- Memory Footprint (peak resident) during inference.

- Reliability: Run success rate over overnight batch pipelines (% non-aborted runs).

- Robustness: Cross-dataset macro F1 gap (primary train domain vs withheld domain ≤ threshold).

Secondary / Diagnostic:

- AUC for rare classes if sample count adequate.

- Confidence–accuracy reliability diagrams stored per major milestone.

- Calibration shift after compression (ΔECE).

Statistical Testing:

- McNemar test (ensemble vs student) for significance on paired predictions if disagreement count ≥ 30.

- Bootstrap (1k resamples) CI for macro F1 ± reported.

3.10 Resources & Tooling

Hardware:

- Training: Single GPU (assumed NVIDIA mid-range) + fallback CPU for export verification.

- Inference Profiling: GPU + CPU (no-GPU) scenarios; measure using onnxruntime (GPU EP + CPU), optional TensorRT.

Software Stack:

- Python 3.11 virtual environments (`.venv`, `.venv_gpu`).

- PyTorch (teachers, KD training), onnxruntime, potential TensorRT / DirectML, numpy, scikit-learn (metrics), custom scripts in `scripts/` and `src/`.

- Version Control: Git + structured experiment directories with manifest JSON & metrics.

Automation & Reproducibility:

- Orchestrated overnight scripts (PowerShell) with validator gating.

- Determinism flags: Seed all RNG sources; log versions of torch/cuda.

- Artifact Registry: Each experiment folder must contain: `metrics.json`, `train_log.csv`, `best_model.pt`, optional `calibration/`.

3.11 Risk Assessment & Mitigation

Risk	Impact	Mitigation
Teacher overfit / low minority F1	Student inherits bias	Early imbalance analysis, logit adjustment, data augmentation targeting minorities
Distillation instability (DKD hyperparams)	Wasted cycles	Coarse grid search with pruning; monitor loss divergence triggers revert to vanilla KD
Calibration drift after pruning/quantization	Confidence misuse	Re-run temperature scaling post-compression; reject compression stage if ΔECE > 0.02
Latency regression with optional attention modules	Miss real-time target	Maintain latency budget per layer; instrument profiler after each architecture change
Dataset curation errors / leakage	Inflated metrics	Hash-based duplicate removal, locked test manifest, audit scripts run pre-merge
Reproducibility gaps	Non-verifiable results	Manifest JSON with seeds/hyperparams; freeze commits for milestone evaluations

3.12 Mapping Objectives to Method Components

Objective (Intro 1.2)	Methodology Lever(s)	Evaluation Check
Close accuracy gap (student vs ensemble) to ≤ 4 pts	Multi-phase KD (Vanilla →DKD), pruning recovery KD, SWA teachers	Macro F1 delta report, McNemar test
Achieve calibrated predictions (ECE ≤ 0.07 student)	Temperature + conditional per-class scaling, avoid aggressive augmentations	ECE & classwise ECE post-calibration
Real-time latency (≤12 ms/frame pre-quant, further ≤8 ms INT8)	Structured pruning, INT8 PTQ/QAT, ONNX EP benchmarking	Latency profiling script; p95 latency table
Minority robustness (avg F1 for rare classes ≥0.65)	Logit adjustment, class-balanced loss, targeted augmentation	Minority F1 tracked per run
Reproducibility / reliability (overnight success ≥95%)	Pre-flight validator, skip-missing logic, manifest enforcement	Run success rate dashboard
Deployment readiness (ONNX + quantized model)	Export validation, quantization path, parity checks	Top-1 parity %, ΔECE, size reduction

3.13 Success Criteria & Exit Conditions

A release candidate (RC) student model is accepted when:

1. Macro F1 within 4 points of ensemble AND macro F1 ≥ target absolute threshold (e.g., 0.78 if ensemble ≥0.82).

2. ECE ≤ 0.07 (calibrated) & classwise ECE not exceeding 2× global.

3. Latency (batch=1) ≤ 12 ms (FP32) and ≤ 8 ms (compressed path) on target hardware.

4. Minority average F1 ≥ 0.65 AND no single class F1 < 0.55.

5. Quantized model accuracy drop ≤ 2 points and ΔECE ≤ 0.01.

6. Reproducibility: Independent re-run of training + distillation yields macro F1 within ±0.5 points & ECE within ±0.005.

3.14 Planned Sequential Execution (High-Level)

1. Finalize & lock teacher ensemble (complete outstanding robustness validation).

2. Calibrate ensemble (global + conditional per-class temperature).

3. Phase 1 student KD (vanilla) → evaluate.

4. Phase 2 DKD refinement → adopt if ≥ +0.8 macro F1 vs Phase 1.

5. Pruning + recovery KD.

6. Calibration re-fit.

7. PTQ INT8; escalate to QAT only if needed.

8. Latency + reliability validation; risk review.

9. Documentation & artifact freeze; prepare Results section.

3.15 Traceability & Documentation

- Every experiment ID maps to a manifest: {git_commit, dataset_hashes, seed, hyperparams, teacher_list, timestamps}.

- Cross-reference citations: Distillation ([2][3][4][19]), Imbalance ([5][8]), Calibration ([6]), SWA ([7]), Pruning ([14]), Quantization ([15]), Representation Pretraining ([16][17]), Transformer Distillation ([18]).

4. PROJECT PLANS

This section summarizes the planned schedule, key milestones, and (placeholder) budget for the Real-time Facial Expression Recognition project. The plan aligns directly with the objectives (Section 1.2) and the methodology gating criteria.

4.1 Schedule

Phase	Focus	Key Activities	Duration (Weeks)	Planned Window	Entry Gate	Exit Gate
P0 Reliability Hardening (Completed)	Stability	Validator, skip-missing ensemble, artifact audits	Done	Aug–Sep 2025	Initial instability	≥95% run success (achieved)
P1 Teacher Finalization	Teacher Ensemble	Final hyperparam sweeps, SWA, calibration, robustness cross-dataset tests	2	W0–W2	Current ensemble F1 plateau	Ensemble macro F1 ≥0.80; minority avg F1 ≥0.65; ECE ≤0.06
P2 Student KD Phase 1	Vanilla KD	T grid (2,4), λ search, baseline student artifact	1	W2–W3	Teachers gated	Student ≥85% of best teacher macro F1; ECE ≤0.09
P3 Student KD Phase 2	DKD / Refinement	α/β sweep, evaluate ΔF1 vs Phase 1	1	W3–W4	Phase 1 student stable	+0.8 macro F1 OR improved minority F1 without ECE regression >0.01

P4 Compression & Pruning	Structured Pruning + Recovery KD	Channel importance calc, prune 30–40%, recovery distill	1	W4–W5	Refined student stable	&le;1.5 F1 drop pre-calibration; size ↓ ≥30%
P5 Quantization & Calibration	PTQ → (QAT if needed)	INT8 PTQ calib set prep, parity test; escalate QAT only if drop >1.5	1	W5–W6	Pruned model validated	INT8 macro F1 drop &le;2.0; ΔECE &le;0.01; latency gain ≥30%
P6 Latency & Edge Validation	Inference Optimization	ONNX EP benchmarking (CPU/GPU), FP16 vs INT8, profiling scripts	1	W6–W7	Quantized candidate approved	p95 latency &le;12 ms FP32, &le;8 ms INT8; top-1 parity ≥99%
P7 Extended Experiments (Optional)	Temporal / Multi-Modal Pilot	Short sequence teacher distill supervision, landmark fusion	2	W7–W9	Core pipeline frozen RC	Sequence distill viability report

P8 Documentation & Results	Paper & Artifact Freeze	Methods verification, results tables, ablations, figures, final calibration plots	2	W8–W10	All mandatory experiments complete	Draft paper & reproducibility package complete
P9 Buffer / Risk Mitigation	Contingency	Address regressions, QAT fallback, add missing analysis	1	W10–W11	Pending open issues	All acceptance criteria satisfied
P10 Final Submission Prep	Packaging	Final proofreading, formatting, repository tag, archival	1	W11–W12	Approved RC & paper	Submission + tagged release

4.2 Milestones

ID	Milestone	Target Metric / Condition	Source Section	Status
M1	Single teacher baseline established	Best single teacher macro F1 ≥0.75	3.4	Active / track
M2	Ensemble superiority	Ensemble macro F1 ≥0.80 and ≥+2.0 over best teacher	3.4	Pending
M3	Ensemble calibrated	Global ECE ≤0.06 post temperature scaling	3.4/3.7	Pending
M4	Student Phase 1 KD	Student macro F1 ≥85% of best teacher AND gap ≤15% relative (F1_student / F1_best ≥0.85)	3.5	Pending

M5	DKD Refinement Gain	+0.8 (absolute) macro F1 vs Phase 1 OR minority F1 +0.04 without ECE regression >0.01	3.5	Pending
M6	Compression Efficiency	Model size ↓ ≥30%; FLOPs ↓ ≥30%; macro F1 drop ≤1.5 pre-calibration	3.5/3.8	Pending
M7	Quantized Readiness	INT8 macro F1 drop ≤2.0 vs FP32; ΔECE ≤0.01	3.8	Pending
M8	Latency Gate	p95 latency ≤12 ms (FP32) and ≤8 ms (INT8) batch=1	3.8/3.9	Pending

M9	Reliability Gate	Overnight run success ≥95% over last 3 nights	3.9	Ongoing
M10	Minority Robustness	Mean F1 (two rarest classes) ≥0.65; no class F1 <0.55	3.6	Pending
M11	Calibration Stability Post-Compression	∆ECE (FP32→INT8) ≤0.01	3.7/3.8	Pending
M12	Reproducibility Verification	Re-run variance: macro F1 within ±0.5; ECE ±0.005	3.13	Pending
M13	Temporal Pilot (Optional)	Sequence teacher yields ≥+1.0 macro F1 on sequence subset OR improved temporal stability metric	2.3/3.14	Optional

M14	Documentation Freeze	Draft paper sections 1–4 + Methodology + initial Results tables complete	4.1/P8	Pending
M15	Release Candidate Model	All success criteria (Section 3.13) satisfied	3.13	Pending
M16	Final Submission	Repository tagged, paper submitted, artifact archive generated	4.1/P10	Pending

5. CONCLUSIONS

This proposal outlines an integrated, reliability-aware strategy for building a real-time Facial Expression Recognition (FER) system that balances accuracy, fairness across minority classes, calibrated uncertainty, and deployment efficiency. Grounded in a heterogeneous ArcFace-based teacher ensemble, the methodology combines adaptive ensemble calibration, staged multi-phase knowledge distillation (vanilla KD → DKD → optional BAN / mutual learning), structured imbalance mitigation, and a disciplined deployment pipeline (pruning, quantization, ONNX export, latency profiling). A distinguishing aspect of the work is the explicit elevation of reproducibility and operational resilience—validator scripts, manifest-driven traceability, failure tagging, and governance gates are treated as first-class research components rather than afterthoughts.

The defined objectives (Section 1.2) translate into measurable success criteria: narrowing the ensemble→student macro F1 gap while safeguarding minority (Fear, Disgust) performance; achieving stringent Expected Calibration Error thresholds; ensuring real-time inference latency through compression and mixed-precision/INT8 optimization; and sustaining >95% reliability in automated experiment orchestration. The methodology (Section 3) operationalizes these targets via gated phases, each producing auditable artifacts (metrics, calibration profiles, latency reports, manifests) that support evidence-based progression. The project plan (Section 4) sequences these phases to reduce risk—front-loading robustness and ensemble stabilization before initiating higher-variance distillation and compression steps.

Anticipated outcomes include: (i) a calibrated, governance-ready teacher ensemble; (ii) a reproducibly trained, latency-qualified student model meeting predefined F1 and ECE thresholds; (iii) a documented compression pathway (pruning + quantization) with controlled accuracy drift; and (iv) an extensible traceability framework enabling future audits, re-analysis, or regulatory adaptation. Collectively, these outputs position the system for transition from research prototype toward deployment contexts (HCI, adaptive learning, lightweight analytics) where trust, responsiveness, and maintainability are critical.

Limitations at this stage stem from the deliberate deferral of temporal modeling and multi-modal fusion; these are earmarked as post-baseline accelerators once single-frame objectives are satisfied. Similarly, advanced per-class calibration and aggressive quantization strategies will be adopted only if empirical risk (minority F1 regression, confidence drift) remains within governance thresholds. Future work will explore temporal teacher → frame-student distillation, structured uncertainty signaling (abstention policies), and lightweight drift monitoring for post-deployment adaptation.

In summary, the proposed work addresses persistent gaps in FER research—simultaneous optimization of accuracy, calibration, latency, and reproducibility—by unifying methodological, engineering, and governance practices. The forthcoming Results section will quantify progress against the stated gates, validating whether the integrated design achieves the targeted balance of performance, reliability, and operational readiness.

References

[Use a consistent citation style throughout your project, such as IEEE, APA, MLA.]

[1] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. CVPR, 2019.

[2] G. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network," NIPS Deep Learning Workshop, 2015 (arXiv:1503.02531).

[3] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti and A. Anandkumar, "Born-Again Neural Networks," in Proc. ICML, 2018 (arXiv:1805.04770).

[4] Y. Zhao, L. Cai, X. Li, K. Xu and T. Zhang, "Decoupled Knowledge Distillation," in Proc. CVPR, 2022 (arXiv:2203.08679).

[5] A. Menon, S. Jayasumana, A. Rawat, H. Jain, A. Veit and S. Kumar, "Long-Tail Learning via Logit Adjustment," in Proc. NeurIPS, 2020 (arXiv:2007.07314).

[6] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, 2017 (arXiv:1706.04599).

[7] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov and A. G. Wilson, "Averaging Weights Leads to Wider Optima and Better Generalization," (Stochastic Weight Averaging) in Proc. UAI, 2018 (arXiv:1803.05407).

[8] Y. Cui, M. Jia, T.-Y. Lin, Y. Song and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in Proc. CVPR, 2019 (arXiv:1901.05555).

[9] T. Li, W. Li, J. Wang and M. Zhou, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," IEEE Trans. Image Process., vol. 28, no. 1, 2019. DOI: 10.1109/TIP.2018.2868382.

[10] Q. Gong, "Real-Time Facial Expression Recognition Based on Image Processing in Virtual Reality," Int. J. Computational Intelligence Systems, 2025. DOI: 10.1007/s44196-024-00729-9.

[11] F. Ma, B. Sun and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion," IEEE Trans. Affective Comput., vol. 14, no. 2, 2023. DOI: 10.1109/TAFFC.2021.3122146.

[12] R. Zhao, T. Liu, Z. Huang, D. P. K. Lun and K.-M. Lam, "Spatial-Temporal

Graphs Plus Transformers for Geometry-Guided Facial Expression Recognition," IEEE Trans. Affective Comput., vol. 14, no. 4, 2023. DOI: 10.1109/TAFFC.2022.3181736.

[13] J.-W. Liu, X.-Y. Lin, P.-F. Ji, J.-M. Chen and J. Zhang, "Multiscale Wavelet Attention Convolutional Network for Facial Expression Recognition," Scientific Reports, 2025. DOI: 10.1038/s41598-025-07416-5.

[14] S. Han, J. Pool, J. Tran and W. J. Dally, "Learning Both Weights and Connections for Efficient Neural Networks," in Proc. NeurIPS, 2015 (arXiv:1506.02626).

[15] B. Jacob et al., "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in Proc. CVPR, 2018 (arXiv:1712.05877).

[16] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations" (SimCLR for representation pretraining relevance), in Proc. ICML, 2020 (arXiv:2002.05709).

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proc. CVPR, 2022 (arXiv:2111.06377). (Pretraining context for potential teacher improvements.)

[18] H. Touvron, M. Cord and H. Jégou, "DeiT: Training Data-Efficient Image Transformers & Distillation through Attention," in Proc. ICML, 2021 (arXiv:2012.12877). (Distillation with transformer teachers/students.)

[19] J. Zhang, T. Xiang, T. M. Hospedales and H. Lu, "Deep Mutual Learning," in Proc. CVPR, 2018 (arXiv:1706.00384). (Mutual KD variant relevant to multi-teacher ensemble.)

[20] M. Lin, Q. Chen and S. Yan, "Network in Network," ICLR, 2014 (arXiv:1312.4400). (Historical reference for multilayer feature abstraction; optional if needed.)

6. Appendices

This appendix compendium provides extended materials supporting reproducibility, auditability, methodological transparency, and future extensibility of the Real-time Facial Expression Recognition (FER) project. Each appendix is self-contained with a short introduction and, where applicable, placeholders for figures (Fig. A#.n) and tables (Table A#.n). Items marked TODO will be populated once corresponding phases (see Project Plan) complete.

6.1 Dataset Provenance & Curation

Purpose: Document sources, licensing, preprocessing, class coverage, and integrity controls.

Planned Table (Table A1.1): Dataset | Source / URL | License | Classes Used | Train Images | Val Images | Test Images | % of Total | Preprocessing (crop/align) | Notes

Index Integrity:

- Index CSV filenames & SHA256 hashes (Table A1.2) – TODO after final index freeze.
- Duplicate removal strategy: perceptual hash + path canonicalization + class consistency check.

Balancing Logic (Pseudo-code):

```
 FOR each dataset_index:
load rows -> compute class_counts
if max(class_counts)/min(class_counts) > TARGET_RATIO:
  apply capping / synthetic balancing (if allowed) / sample weighting
log adjustments to manifest.dataset_balance
```

Notes: Cross-reference Section 3.2 for high-level description.

6.2 Class Distribution & Imbalance Diagnostics

Purpose: Expose raw and adjusted class distributions and track imbalance interventions.

Planned Figures:

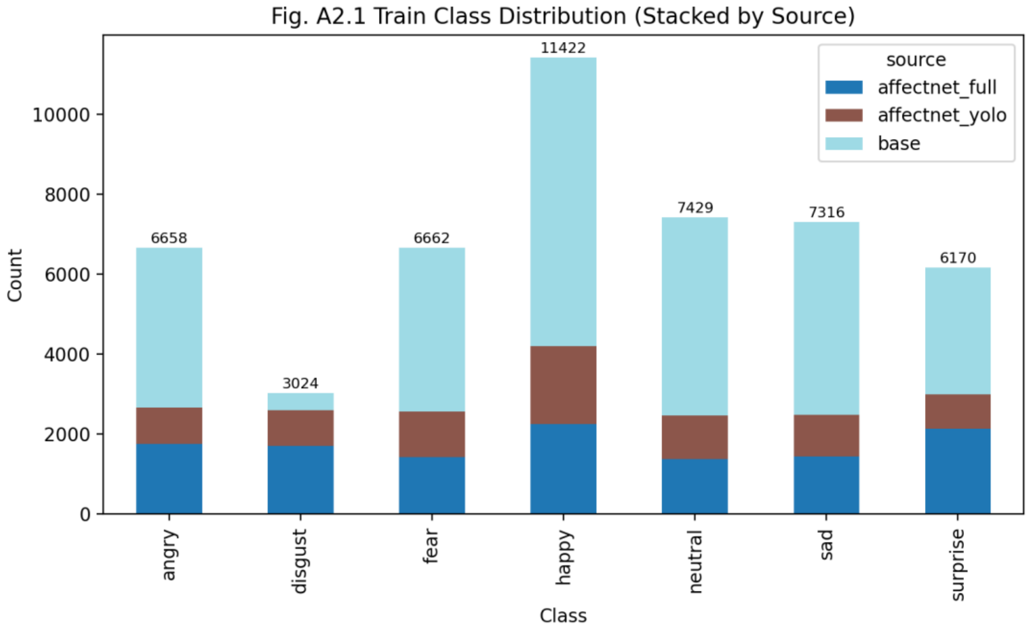

Figure A2.1. Train class distribution (stacked by dataset origin).

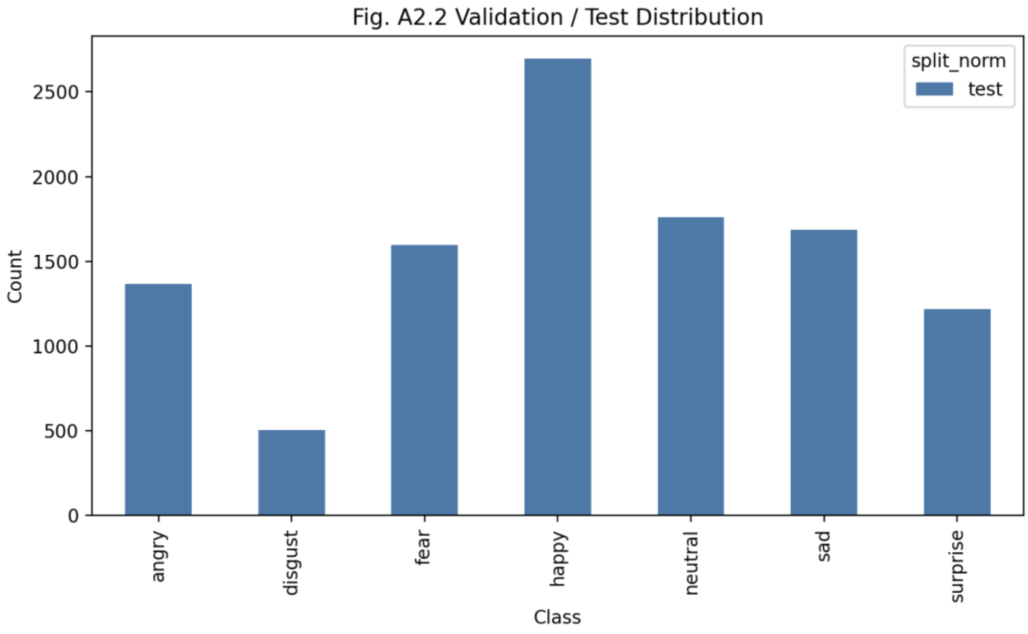

Figure A2.2. Validation/Test distribution.

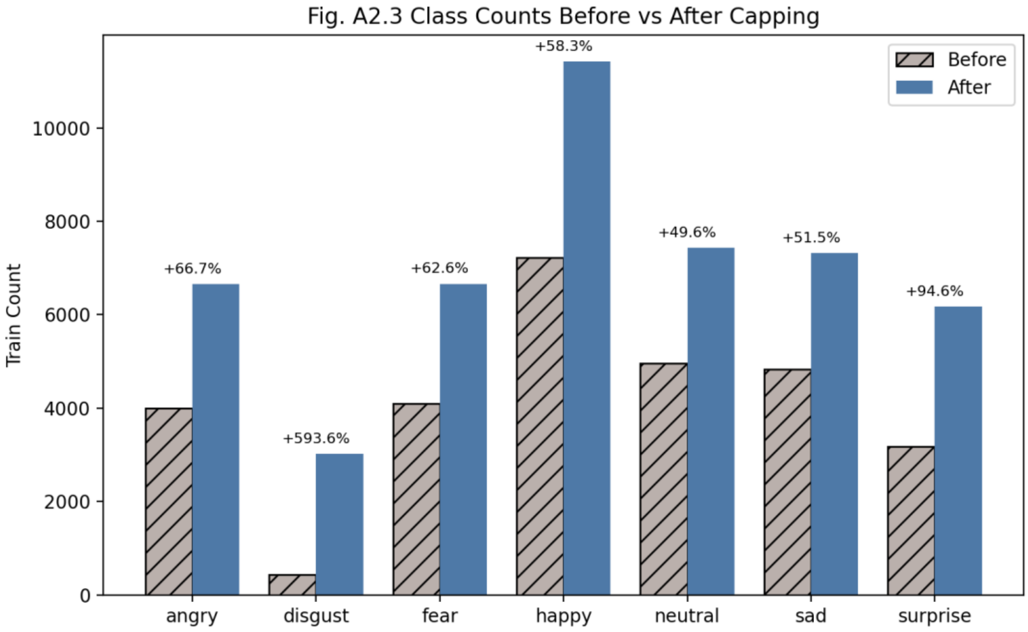

Figure A2.3. Before vs After capping (overlay bars).

Planned Table (Table A2.1): Class | Train Count | Val Count | Test Count | % Share | Minority Tag (Y/N) | Applied Adjustment (Y/N)

Minority Ratio Timeline: TODO (populate after any rebalance).

6.3 Experiment Manifest Schema & Example

Purpose: Provide full schema enumerating required / optional keys and one sanitized example for reproducibility.

Schema (Table A3.1): Key | Type | Required | Description

(Refer to Methodology Section 3.15 for narrative; this table normalizes in compact form.)

Example Manifest (Listing A3.1):

```
   {
  "experiment_id": "student_dkd_phase2_r04",
  "git_commit": "<hash>",
  "timestamp_utc": "2025-09-30T08:12:33Z",
  "seeds": {"python": 1337, "torch": 1337, "numpy": 1337},
  "dataset": {
    "indices": ["dataset_index_extended_v8.csv"],
    "index_hashes": {"dataset_index_extended_v8.csv": "<sha256>"},
    "class_priors": [0.24, 0.11, 0.07, 0.05, 0.19, 0.21, 0.13]
  },
  "teachers": [{"name": "resnet18_arcface", "checkpoint": "..."}, {"name":
"efficient_b3_arcface", "checkpoint": "..."}],
  "student": {"arch": "resnet18_pruned_0.6", "params_m": 7.2},
  "training": {"epochs": 60, "optimizer": "AdamW", "lr_schedule": "cosine",
"batch_size": 128},
  "distillation": {"phase": "DKD", "temperature": 4, "alpha": 1.0, "beta": 8.0},
  "imbalance": {"logit_adjustment_tau": 1.2, "class_balanced_loss": false},
  "calibration": {"global_temperature": 1.15, "per_class": null},
  "compression": {"pruning_ratio": 0.35, "quantization": null},
  "metrics": {"macro_f1": 0.802, "ece": 0.054, "latency_ms_fp32": 11.6},
  "governance": {"status": "accepted_phase2_candidate"}
}
```

6.4 Hyperparameter & Training Configuration Matrix

Purpose: Make tuning transparent; prevent hidden configurations.

Planned Table (Table A4.1): Model/Phase | Loss Components | Margin (ArcFace) | KD Temp(s) | DKD α | DKD β | Epochs | Batch | LR Init | Schedule | Weight Decay | Augmentations | Seed | Notes

Placeholder Rows: TODO after P2 & P3 consolidation.

6.5 Additional Ablation & Sensitivity Results (Planned)

Purpose: House post-primary experiments illustrating robustness.

Sections (to populate):

- A5.1 Logit Adjustment τ Sweep (Fig. A5.1; Table A5.1).

- A5.2 KD Temperature vs Macro F1 & ECE (Fig. A5.2).
- A5.3 DKD α/β Heatmap (Fig. A5.3).
- A5.4 Pruning Ratio vs ΔF1 & ΔECE (Fig. A5.4).
- A5.5 Per-Class Temperature Scaling Impact (Table A5.2).

6.6 Calibration Artifacts

Purpose: Demonstrate uncertainty improvements & safeguards.

Planned Figures:

A6.1 TODO

Figure A6.1. Reliability diagram (ensemble pre/post global scaling).

A6.2 TODO

Figure A6.2. Student pre/post calibration.

A6.3 TODO

Figure A6.3. Classwise ECE bar plot.

Tables:

- Table A6.1: Global vs per-class ECE summary.
- Table A6.2: Accepted vs Rejected per-class scaling decisions (Reason column).

Decision Criteria: Accept per-class scaling only if (class_ECE > 2× global_ECE) & (n_samples ≥ threshold) & (no minority F1 drop > 0.02).

6.7 Latency & Deployment Profiling Details

Purpose: Evidence for real-time claims & deployment viability.

Hardware Spec Table (Table A7.1): Component | Spec (GPU model, driver, CUDA, cuDNN, CPU, RAM, OS build).

Model Artifact Comparison (Table A7.2): Stage | Params (M) | Size (MB) | FLOPs (G) | Median Latency (ms) | p95 (ms) | ECE | Macro F1 | Notes.

Figures:

Figure A7.1. Latency cumulative distribution (FP32 vs INT8).

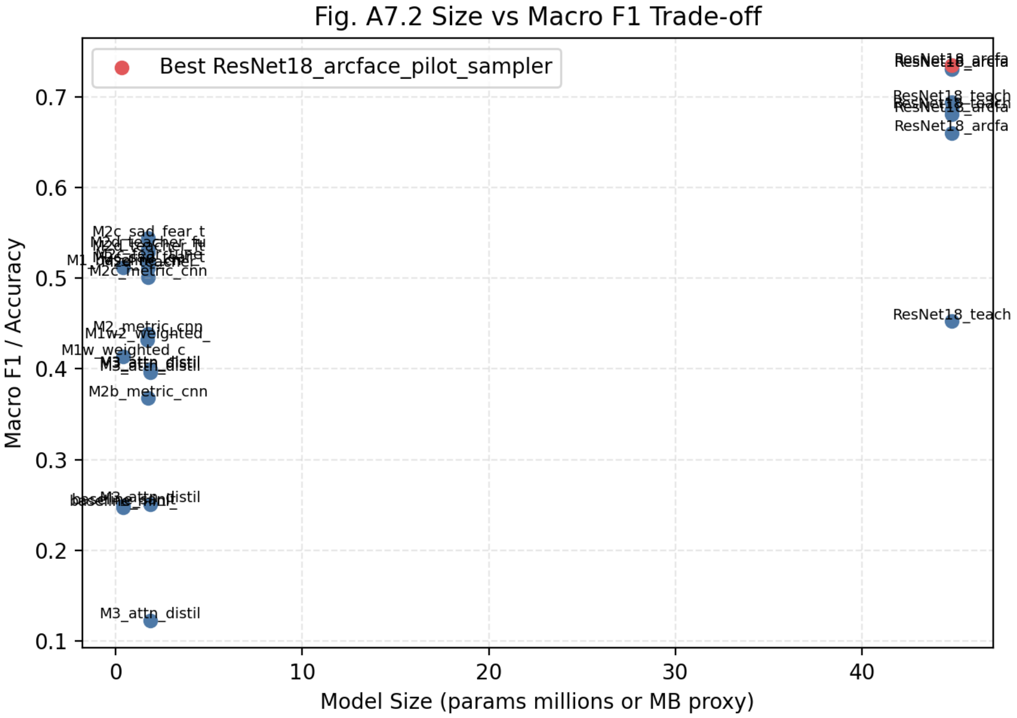

Figure A7.2. Size vs Macro F1 trade-off curve.

6.8 Failure Taxonomy & Reliability Engineering

Purpose: Classify failure modes and mitigation strategies.

Table A8.1: Failure Category | Trigger Condition | Detection Mechanism | Mitigation / Fallback | Logged Field.

Examples: missing_checkpoint | file absent at epoch boundary | validator script | graceful skip & flag | manifest.reliability.missing_ckpt.

6.9 Ethical, Privacy & Governance Considerations

Purpose: Clarify scope limits & responsible AI posture.

Topics:

- Scope limitation (discrete expressions only; not inferring mood/personality).
- Potential demographic bias (future audit placeholder).
- Privacy: no raw face storage beyond training artifacts; hashed IDs.
- Drift Monitoring Plan (placeholder; see A15 future work).

6.10 Extended Error Analysis

Purpose: Deep dive into misclassifications & qualitative samples.

Planned Items:

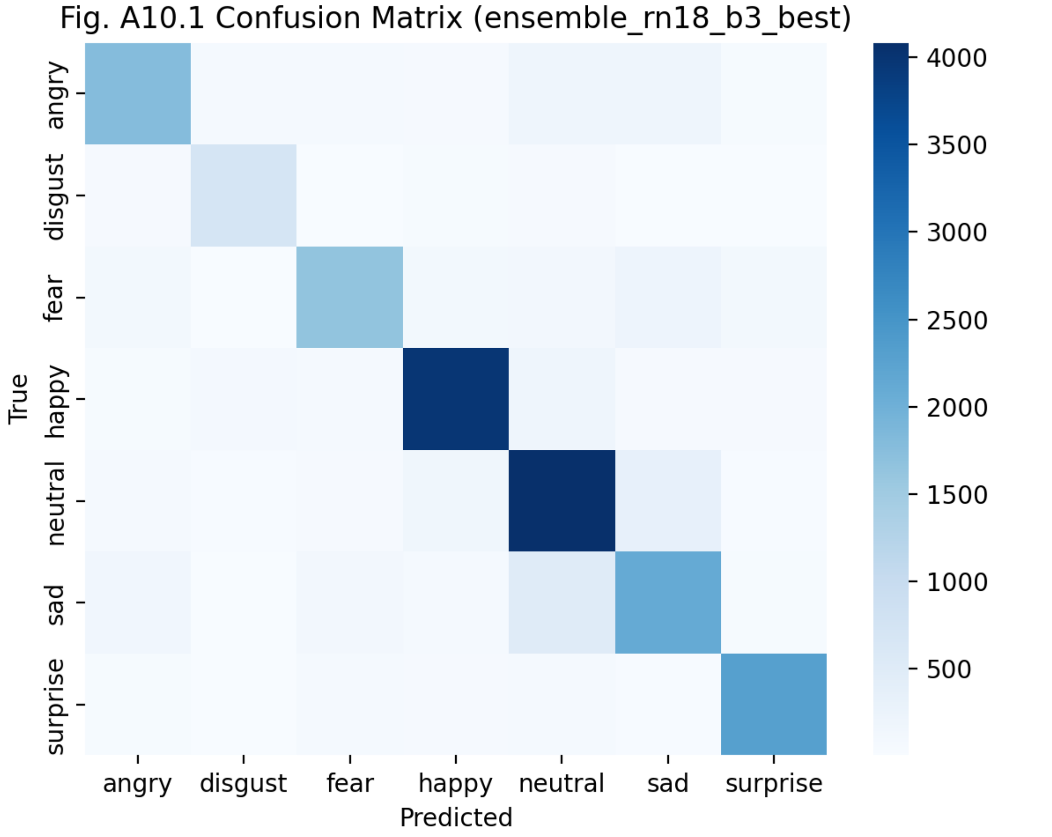

Figure A10.1. Confusion matrix (ensemble) vs Fig. A10.2 (student) + delta matrix (Table A10.1).

- Hard sample montage (Fig. A10.3) (NOTE: ensure license/consent).
- Error Taxonomy: Overconfident minority misclassifications vs ambiguous neutral.

6.11 Embedding & Feature Space Visualizations

Purpose: Show representation evolution and compression effects.

Planned Figures:

A11.1 TODO

Figure A11.1. t-SNE teacher ensemble embedding.

A11.2 TODO

Figure A11.2. t-SNE student post-DKD.

A11.3 TODO

Figure A11.3. Pruned vs original distance distributions.

Metrics Table A11.1: Model | Intra-Class Dist (μ±σ) | Inter-Class Dist (μ±σ) | Separation Margin Proxy.

6.12 Reproducibility Checklist

Purpose: Quick audit before releasing artifacts.

Checklist (Table A12.1): Item | Status (Y/N) | Location / Manifest Key | Notes.

Items (initial draft):

- Seeds logged
- Git commit stored
- Dataset index hashes captured
- Environment versions frozen
- Calibration params exported
- Pruning & quantization configs logged
- Latency profile JSON present
- Reliability signals captured (failures=0)
- Re-run variance test executed

6.13 Scripts & Automation Index

Purpose: Map code assets to pipeline phases.

Planned Table (Table A13.1): Script | Phase | Purpose | Key Inputs | Key Outputs | Notes.

Populate by scanning `scripts/` & `src/` (TODO after P2 freeze).

6.14 Security / Integrity Controls

Purpose: Artifact integrity & tamper resistance.

Planned Content:

- Hash verification command examples.
- Optional model signing roadmap.
- Controlled access strategy for raw data vs derived embeddings.

6.15 Deferred / Future Work Elaborations

Purpose: Detail roadmap elements beyond current scope.

Sections:

- Temporal distillation concept diagram (Fig. A15.1 placeholder).
- Multi-modal fusion gating criteria.
- Drift detection pipeline design (shadow evaluation, prior divergence alert).

6.16 Cost & Resource Accounting

Purpose: Track resource efficiency & budgeting.

Planned Table (Table A16.1): Phase | GPU Hours (Actual) | Storage Added (GB) | Incremental Cost (est) | Notes.

Trend Figure (Fig. A16.1): Cumulative GPU hours vs milestone.

6.17 Change & Decision Log

Purpose: Trace key choices and rationale.

Table A17.1: Date | Decision | Context / Problem | Alternatives Considered | Rationale | Impact | Section Cross-Ref.

Populate continuously (append-only) to support audit.

6.18 Appendix Cross-Reference Guide

Main Text Reference → Appendix Section:

- Dataset details (Section 3.2) → A1, A2.
- Manifest schema (Section 3.15) → A3.
- Hyperparameters (Sections 3.4–3.6) → A4.
- Calibration (Section 3.7) → A6.
- Deployment & latency (Section 3.8) → A7.
- Reliability & reproducibility (Sections 3.9, 3.13) → A8, A12.
- Error analysis (future Results) → A10.
- Representation changes (distillation/compression) → A11.

Status Legend:

- TODO: Pending experiment completion.
- DRAFT: Preliminary numbers/plots; not final.
- FINAL: Locked for release candidate.

This structured appendix blueprint will be progressively populated as phases (P1–P8) reach their exit gates.

FYP Report Template:
v1 created by Dr. Hui Wen Rebecca LIANG (2023 Sept.)