Optimization of Face-Gait Person Identification Pipelines for Edge-Deployed Cognitive Agents

Optimization of Face-Gait Person Identification Pipelines for Edge-Deployed Cognitive Agents

Federico Boscolo
Dept. of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
federico.boscolo@polito.it

Grazia De Mola
Dept. of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
s326640@studenti.polito.it

Fabrizio Lamberti
Dept. of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
fabrizio.lamberti@polito.it

***Abstract*—Cognitive robotic systems require low-latency, reliable perception to support safe and context-aware interaction with humans. Multimodal biometrics (e.g., face and gait) improve robustness under real-world variations but the resulting pipelines are often too compute-intensive and memory-heavy for edge deployment. This paper explores the possibility of exploiting multimodal redundancy to enable aggressive optimization of person recognition stacks without excessive system-level degradation. We analyze a modular face+gait pipeline and apply architectural exploration together with post-training compression, namely quantization and pruning, targeting the dominant bottleneck (face recognition). Experiments on CASIA-B under a Rank-1 protocol over near-frontal views show that while face-only accuracy degrades significantly under compression (0.971 to 0.825), score-level fusion remains comparatively stable (0.986 to 0.966) while improving throughput from 49.8 to 75.2 FPS and reducing the deployed footprint from 295.2 MB to 111.9 MB. These results indicate that multimodal fusion can serve as a resilience mechanism, enabling lightweight and optimized perception modules for real-time cognitive agents.**

***Index Terms*—Cognitive robotics, biometrics, person identification, face recognition, edge AI, model compression**

I. Introduction

Cognitive robotic systems operating in shared human environments require reliable, low-latency perception to enable adaptive behavior. In many human-centered deployments, from assistive robots and companion agents to intelligent kiosks and vehicles, context-aware interaction depends on recognizing the person interacting with the system, so the agent can personalize responses or enforce access control. Achieving such user identification robustly in realistic conditions is challenging: lighting changes, motion blur, partial occlusions, clothing variation, and viewpoint shifts can cause substantial performance degradation for unimodal biometric modules.

Multimodal biometrics provide a way to improve robustness by combining complementary sources (e.g., face as a highly discriminative cue at close range, gait as a longer-range cue useful when facial details are degraded). Classical surveys and overviews emphasize that multimodal fusion reduces sensitivity to noisy measurements and increases reliability compared to unimodal systems, at the cost of added system complexity [1]. The growing availability of embedded accelerators has made on-device multimodal inference increasingly feasible, though the co-design of accuracy and efficiency remains an open challenge [2]. In practical deployments, fusion is commonly performed at feature-, score-, or decision-level. Among these, score-level fusion is often preferred for plug-and-play integration because each modality can be processed independently and combined using normalized similarity scores; however, this requires careful score normalization due to heterogeneous score ranges and distributions [3].

Despite the robustness benefits of multimodality, deploying multimodal person recognition on embedded hardware typical of mobile agents remains challenging. Recent deep learning advances have improved face and gait recognition performance, yet these gains often come with larger models, increased compute, and higher inference latency. For cognitive agents, perception must be both accurate and computationally efficient to preserve energy, maintain responsiveness, and allow co-execution with other time-critical components (e.g. navigation, planning, control). Many human-facing scenarios favor on-device processing for privacy and latency reasons, making it essential to understand how system-level accuracy behaves under aggressive efficiency constraints.

A typical real-time recognition pipeline also relies on fast front-end localization and tracking to ensure stable spatio-temporal association of observations across frames. Contemporary one-stage detection paradigms are widely adopted in real-time applications due to their speed–accuracy trade-off [4], [5], and modern multi-object trackers such as ByteTrack provide efficient association mechanisms that are robust to intermittent low-confidence detections [6]. However, even with efficient detection and tracking, the face recognition branch frequently dominates runtime, especially when high-quality embeddings are required for identification.

In this work, we investigate whether multimodal redundancy can be leveraged to enable aggressive optimization of person recognition pipelines without excessive degradation of system-level performance. Specifically, we analyze a face-and-gait pipeline and apply (i) architectural exploration of front-end components and (ii) quantization and pruning of the most

This work is part of the project PNRR-NGEU which has received funding from the MUR – DM 117/2023.

computationally demanding module (face recognition). Our central hypothesis is that multimodal fusion provides inherent resilience: compression may degrade unimodal performance, but the fused system can preserve high identification accuracy by compensating with the complementary modality.

We validate this hypothesis on the CASIA-B dataset, a standard multi-view gait benchmark containing 124 subjects under multiple viewing angles and covariate conditions (normal walking, carrying, clothing) [7]. To reflect an embodied-agent scenario where the subject approaches the camera and facial information is available, we focus on near-frontal viewpoints (0°, 18°, 36°) and evaluate performance under a Rank-1 identification protocol. Experimental results show that aggressive compression substantially reduces face-only accuracy, yet the fused multimodal system retains high robustness (above 95% Rank-1) while achieving approximately $2\times$ throughput improvement. These findings suggest that cognitive robotic systems can exploit multimodal complementarity as a system-level safety net, enabling lightweight, efficient perception modules suitable for embedded deployment without sacrificing reliability.

This paper makes three contributions: (1) an empirical bottleneck analysis of a modular face+gait identification pipeline for cognitive agents; (2) a systematic evaluation of architectural exploration and compression strategies targeting the face branch; and (3) an experimental characterization of fusion resilience, demonstrating that score-level fusion preserves system-level performance under unimodal degradation.

The remainder of this paper is organized as follows: Section II reviews related work on multimodal biometrics, efficient perception pipelines, and model compression. Section III describes the proposed pipeline and optimization methodology, including fusion strategies. Section IV presents the experimental protocol and results on CASIA-B. Sections V and VI offer an analysis of latency–accuracy trade-offs and fusion resilience, as well as runtime and system-level considerations. Section VII concludes the paper and outlines directions for future work.

II. Related Work

A. Multimodal Person Recognition and Fusion

Unimodal biometric systems (e.g., face-only or gait-only) are sensitive to covariates such as illumination, viewpoint, occlusion, motion blur, clothing changes, and acquisition distance. Multimodal biometric systems mitigate these issues by combining complementary cues, improving robustness compared to single-modality pipelines [1]. In multimodal recognition, fusion is commonly performed at the feature, score, or decision level. Feature-level fusion can exploit cross-modal correlations but may suffer from heterogeneous embedding distributions and dimensional redundancy, often requiring careful normalization and calibration. Score-level fusion is attractive in practice because it preserves modularity and enables the use of pretrained components; however, it requires score normalization due to heterogeneous score ranges and distributions across modalities [3]. Feature-level fusion has also been explored in IoMT settings where multiple biometric streams are combined to improve robustness of user identification under resource constraints [8].

In the specific context of face and gait, fusion has been explored as a strategy for improving identification robustness under challenging conditions and variable acquisition quality. Recent work emphasizes adaptive weighting based on quality cues, where modality importance is modulated by factors such as distance, occlusion, or image quality [9]. These findings motivate the central hypothesis of this paper: fusion can act as a robustness mechanism that tolerates unimodal degradation, enabling more aggressive optimization of individual components.

Beyond quality-aware strategies, several works have explored joint representation learning across face and gait, either through cross-modal attention or shared latent spaces. These approaches aim to leverage semantic correspondence between biometric cues during training, rather than at inference time. However, such tightly coupled architectures sacrifice the modularity that makes score-level fusion attractive for practical pipelines: each branch must be retrained jointly, preventing independent optimization of individual components. Our work deliberately adopts a loosely coupled design to preserve this modularity, enabling plug-and-play replacement and compression of individual modules without retraining the entire system. This property is particularly valuable in the resource-constrained deployment scenario considered in our work.

B. Efficient Detection and Tracking for Real-Time Perception

Real-time cognitive agents typically rely on a tracking-by-detection paradigm, where a detector provides frame-wise observations and a tracker performs temporal association. One-stage detectors are widely adopted due to their favorable latency–accuracy trade-off, and recent YOLO implementations are commonly used in practical deployments [4], [5]. For multi-object tracking, simple motion models combined with efficient association strategies achieve strong performance in practice. In particular, ByteTrack associates both high- and low-confidence detections to reduce missed targets and identity switches while maintaining efficiency, making it well suited to real-time settings [6]. In our pipeline, these components serve as lightweight front-end modules that constrain downstream biometrics to a consistent spatio-temporal track.

C. Lightweight Face Recognition Models

Modern face recognition systems typically learn discriminative embeddings using margin-based classification losses that improve inter-class separability, such as SphereFace, CosFace, and ArcFace [10], [11], [12]. While high-capacity backbones yield strong accuracy, edge deployment has driven the development of compact architectures that preserve discriminative embeddings under strict compute constraints. Recent lightweight models, such as GhostFaceNet, reduce feature redundancy and optimize architectural blocks for efficient inference, making them suitable for embedded and mobile scenarios [13]. Earlier

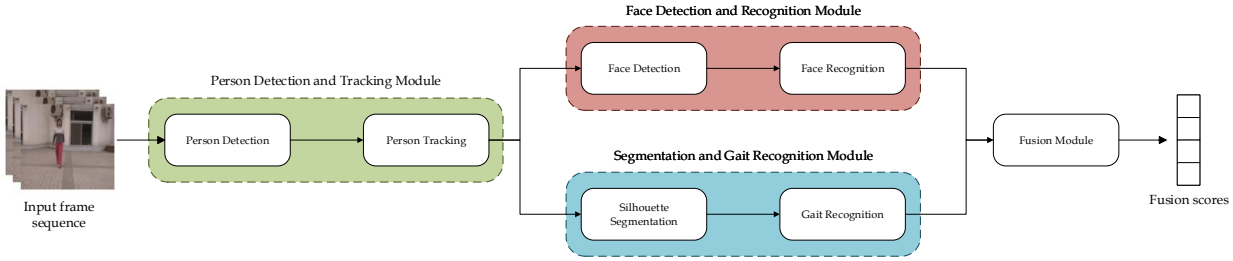

Fig. 1. Overview of the multimodal perception pipeline. RGB frames are processed by a tracking-by-detection front-end. Face and gait branches compute modality-specific templates, which are fused via feature-level fusion (FLF) or score-level fusion (SLF) for Rank-1 identification.

lightweight designs such as MobileFaceNets [14] and ShuffleFaceNet [15] established the viability of mobile-scale face embeddings, and recent surveys confirm that the accuracy gap relative to full-capacity backbones continues to narrow [16]. These trends support our choice of GhostFaceNet as a compact yet competitive base model for the face branch. Since face recognition is often the dominant computational bottleneck in multimodal pipelines, it is a primary target for compression and acceleration.

D. Gait Recognition and Silhouette-Based Pipelines

Gait recognition enables identification at longer ranges and under conditions where facial details may be unreliable. Silhouette-based methods remain widely used due to their compact representation and robustness in low-resolution settings [17]. Sequence-set modeling (e.g., GaitSet) introduced set pooling mechanisms to aggregate temporal evidence without requiring strict sequence alignment [18]. Subsequent work extended set-based modeling to part-level temporal representations (GaitPart [19]), improving discrimination under viewpoint and clothing variation. Recent surveys provide a comprehensive account of deep gait recognition architectures, datasets, and covariates [20], [21], highlighting that silhouette quality and viewing angle are the dominant sources of performance degradation in practical deployments. OpenGait provides a unified implementation and evaluation framework for a wide range of gait models, facilitating standardized benchmarking across protocols and datasets [22]. In this work, the gait branch complements face recognition by providing discriminative cues when face quality degrades, supporting system-level robustness under unimodal compression.

E. Model Compression for Edge Deployment

Quantization and pruning are widely used to reduce memory footprint and accelerate inference. Post-training quantization (PTQ) is attractive due to its low engineering cost; however, for embedding-based recognition, quantization can distort the geometry of the feature space and degrade matching performance. AdaRound addresses this issue by optimizing rounding decisions during PTQ to reduce reconstruction error relative to the full-precision model [23]. Pruning methods remove redundant parameters either in an unstructured (weight-level) or structured (channel/block-level) manner; while unstructured pruning can yield high sparsity, structured pruning more directly translates to wall-clock speedups on standard hardware [24]. Broader reviews of neural network quantization confirm that the interaction between quantization scheme, model architecture, and downstream task (e.g., embedding matching vs. classification) significantly affects the achievable accuracy–efficiency trade-off [25].

Distinct from prior work focusing on unimodal accuracy under compression, this paper examines compression through a system perspective: we empirically evaluate whether multimodal fusion compensates for unimodal degradation, enabling more aggressive optimization while preserving end-to-end identification performance.

III. Methodology

To enable robust human identification on resource-constrained cognitive agents, we developed a modular multimodal perception pipeline that extracts face and gait cues from RGB video streams and performs closed-set person identification. The pipeline follows a plug-and-play design (Fig. 1), allowing each module to be replaced or optimized independently depending on deployment constraints.

A. Problem Formulation

We address closed-set person identification in a multimodal setting. Given a gallery $\mathcal{G}$ consisting of one Normal Walking (NM) sequence per subject and a probe sequence $\mathcal{P}$, the goal is to predict the identity of $\mathcal{P}$ among the enrolled identities via Rank-1 matching.

Let $\mathbf{v}_{\text{face}}^{(t)} \in \mathbb{R}^{d_f}$ and $\mathbf{v}_{\text{gait}}^{(t)} \in \mathbb{R}^{d_g}$ denote per-frame embeddings at time t. The system produces sequence-level templates $\bar{\mathbf{v}}_{\text{face}}$ and $\bar{\mathbf{v}}_{\text{gait}}$ by temporal aggregation (Sec. III-C). Our objective is to maximize Rank-1 accuracy while minimizing average per-frame latency:

$$\mathcal{L} = \lambda_1(1 - \text{Rank-1}) + \lambda_2 T_{\text{frame}}, \tag{1}$$

where T_{frame} is the measured per-frame inference time.

B. Pipeline Architecture

The pipeline comprises three main operational blocks: (i) spatio-temporal localization, (ii) face branch, and (iii) gait branch.

1) Person Tracking (Tracking-by-Detection): We adopt a tracking-by-detection strategy. A lightweight person detector (YOLO-family) produces person bounding boxes for each frame, and ByteTrack performs temporal association to maintain track consistency under intermittent low-confidence detections and partial occlusions. This design avoids heavy re-identification models in the tracking loop while providing stable crops for downstream biometric modules.

2) Face Branch: Given a tracked person crop, a face detector localizes the face region. A face recognition backbone (GhostFaceNet family) extracts a d_f=512 embedding per frame. Since this branch is typically the dominant computational cost (Sec. III-D), it is the primary optimization target in our study.

3) Gait Branch: In parallel, a human segmenter produces binary silhouette masks from the tracked person crop. Silhouettes are then encoded using OpenGait models (GaitSet/GaitBase family) to extract gait embeddings that remain informative at longer ranges and under degraded facial detail. In our experiments, gait inference and fusion contributed a small fraction of total runtime compared to the face branch.

C. Sequence-to-Template Construction

For identification, each sequence is represented by a single template per modality. Given per-frame embeddings, we apply L_2 normalization and temporal mean pooling:

$$\bar{\mathbf{v}} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbf{v}^{(t)}}{\|\mathbf{v}^{(t)}\|_2}. \tag{2}$$

This yields one template for face and one for gait per sequence, which are then compared to gallery templates using modality-specific similarity measures (cosine similarity for face, Euclidean distance for gait).

D. Optimization Framework

A bottleneck analysis of the baseline pipeline indicated that the face branch accounts for the majority of the per-frame latency. We therefore concentrate optimization on this subsystem through a two-stage strategy.

1) Architectural Exploration: We evaluate alternative lightweight model families and variants (e.g., YOLOv5/v8/YOLO11 for detection/segmentation; lightweight face backbones) to identify efficient operating points prior to compression. This step reduces compute by selecting topologies that mitigate the latency–accuracy trade-off before applying quantization/pruning.

2) Quantization and Pruning: We apply post-training compression to reduce memory and improve throughput:

- **Quantization:** Detection models are converted from FP32 to INT8 using post-training calibration. For the face recognition backbone, we employ AdaRound-based post-training quantization, which optimizes rounding decisions to reduce deviation from the full-precision layer outputs during calibration.
- **Pruning:** We apply magnitude-based unstructured pruning (L1 criterion) to the face recognition backbone, exploring sparsity levels in the 30–60% range. Pruning is followed by model export to the deployment runtime used for timing measurements.

We emphasize that our goal is not to maximize unimodal face accuracy under compression, but to quantify system-level robustness when multimodal fusion is available.

E. Multimodal Fusion (FLF and SLF)

We implement fusion at feature-level (FLF) and score-level (SLF).

1) Feature-Level Fusion (FLF): For FLF, we construct a fused embedding by concatenating the modality templates and applying L_2 normalization:

$$\mathbf{v}_{\text{FLF}} = \left[\bar{\mathbf{v}}_{\text{face}},\ \bar{\mathbf{v}}_{\text{gait}}\right], \qquad \hat{\mathbf{v}}_{\text{FLF}} = \frac{\mathbf{v}_{\text{FLF}}}{\|\mathbf{v}_{\text{FLF}}\|_2}. \tag{3}$$

Identification is performed by cosine similarity between probe and gallery fused embeddings:

$$s_{\text{FLF}}(\mathcal{P}, \mathcal{G}_i) = \hat{\mathbf{v}}_{\text{FLF}}(\mathcal{P})^{\top}\hat{\mathbf{v}}_{\text{FLF}}(\mathcal{G}_i), \tag{4}$$

and the Rank-1 prediction corresponds to the gallery identity with maximum similarity.

2) Score-Level Fusion (SLF): For SLF, we compute modality scores independently and combine them via weighted averaging. Let s_{face} be cosine similarity and d_{gait} be Euclidean distance. We map both to a comparable similarity domain using min–max normalization computed over the gallery:

$$s' = \frac{s - \min(s)}{\max(s) - \min(s)}. \tag{5}$$

The final SLF score is:

$$S_{\text{SLF}} = \alpha\, s'_{\text{face}} + (1-\alpha)\, s'_{\text{gait}}, \tag{6}$$

where $\alpha \in [0,1]$ is tested with multiple different values during evaluation. While adaptive α is plausible in embodied settings (e.g., distance-dependent weighting), this work focuses on characterizing the resilience of SLF under compression using a constant α for controlled comparison.

F. Latency Measurement

Per-frame latency is computed by summing stage-wise times for tracking, face branch, and gait branch:

$$T_{\text{tot}} = T_{\text{trk}} + T_{\text{face}} + T_{\text{gait}}. \tag{7}$$

We report T_{tot} in ms/frame and FPS as $1000/T_{\text{tot}}$, consistent with Table I. Fusion operations (concatenation and weighted averaging) add negligible overhead and are excluded from the stage-wise breakdown.

TABLE I

SELECTED CASIA-B CONFIGURATIONS ILLUSTRATING FUSION RESILIENCE UNDER COMPRESSION. LATENCY IS AVERAGE PER-FRAME TIME (MS/FRAME); FPS $= 1000/T_{\text{TOT}}$. MEM IS THE COMBINED DEPLOYED MODEL FOOTPRINT (MB). SLF AND FLF DENOTE *score-level fusion* AND *feature-level fusion*, RESPECTIVELY; BOTH ARE REPORTED AS RANK-1 IDENTIFICATION ACCURACY.

Cfg	Tracker	Face Det.	Face Rec.	Segm.	Face R1	Gait R1	SLF R1	FLF R1	T_{trk}	T_{face}	T_{gait}	T_{tot}	FPS	Mem (MB)
A	YOLOv8n	YOLOv5-face	GhostFaceNet	Paddle	0.971	0.869	0.986	0.980	3.4	15.3	1.4	20.1	49.8	295.2
B	YOLO11n	YOLO11m-face	GhostFaceNet	YOLO11s-seg	0.949	0.901	0.987	0.971	4.0	10.8	3.9	18.7	53.5	124.6
C	YOLOv8n	YOLO11m-face	pruned_ada14	YOLO11s-seg	0.825	0.903	0.966	0.919	3.4	5.9	4.0	13.3	75.2	111.9
D	YOLOv8n	YOLO11m-face	pruned_ada14	Paddle	0.825	0.864	0.952	0.898	3.5	6.3	1.5	11.3	88.5	122.8

IV. EXPERIMENTAL RESULTS

A. Experimental Protocol

Experiments were conducted on a workstation equipped with an Intel i7 CPU and NVIDIA RTX 3090 GPU. While not an embedded platform, this setup enables controlled comparison of relative computational cost across configurations prior to edge deployment.

Latency is measured as the cumulative processing time over all gallery samples and normalized to obtain average per-frame inference time:

$$T_{frame} = \frac{T_{total}}{N_{frames}} \tag{8}$$

Gait encoding and fusion operations contributed negligibly ($<3\%$) to total latency.

Evaluation follows a Rank-1 identification protocol using CASIA-B sequences at 0°, 18°, and 36°, ensuring visibility of both facial and gait features. We report Rank-1 identification accuracy for each unimodal branch and for fusion: SLF uses normalized similarity scores combined by weighted mean, whereas FLF uses concatenated embeddings with the same matching protocol.

B. Implementation Details and Reproducibility

All configurations share the same high-level processing flow (Fig. 1). For each frame, we apply tracking-by-detection to obtain a person crop, then compute face and gait representations in parallel. For sequence-to-template construction, per-frame embeddings are L_2-normalized and temporally mean-pooled to form one template per modality. Face similarity is computed using cosine similarity, while gait similarity is computed using Euclidean distance; SLF combines normalized modality scores via weighted averaging, and FLF concatenates flattened modality templates followed by cosine similarity matching (Sec. III-E).

Latency is measured using stage-wise per-frame timing for tracking, face (face detection + recognition), and gait (segmentation + encoding), and summed as $T_{\text{tot}} = T_{\text{trk}} + T_{\text{face}} + T_{\text{gait}}$. FPS is computed as $1000/T_{\text{tot}}$. Fusion overhead is negligible relative to the network inference stages (simple concatenation or weighted averaging), and is therefore excluded from the breakdown. Model footprints are reported as the combined deployed model sizes for each configuration.

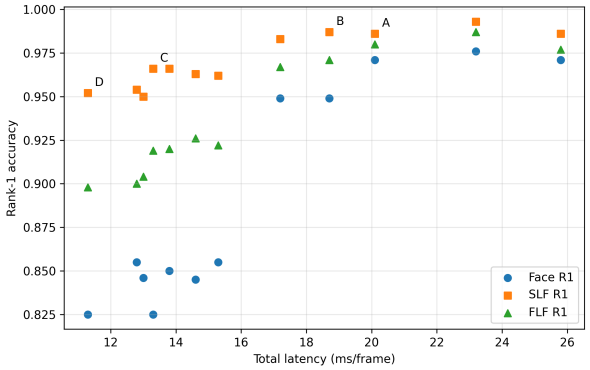

Fig. 2. Rank-1 accuracy versus total latency across CASIA-B configurations. Face-only accuracy degrades sharply under compression, while fusion remains comparatively stable. SLF (score-level fusion) preserves higher accuracy than FLF (feature-level fusion) across most operating points. Letters indicate the representative configurations in Table I.

C. Latency–Accuracy Trade-offs and Bottleneck Analysis

Table I reports four representative configurations spanning the accuracy–efficiency trade-off. In the baseline pipeline (Cfg A), the face branch dominates the computation: $T_{\text{face}} = 15.3$ ms/frame accounts for $\approx$76% of the total $T_{\text{tot}} = 20.1$ ms/frame (49.8 FPS), whereas gait processing contributes $T_{\text{gait}} = 1.4$ ms/frame. This motivates targeting the face subsystem (face detection and, in particular, face recognition) for optimization. This bottleneck is also visible in Fig. 3, where the face stage accounts for most of the baseline latency.

Cfg B illustrates the effect of architectural exploration without aggressive compression. Replacing the baseline detection/segmentation components yields a substantial reduction in memory (295.2→124.6 MB) while slightly improving end-to-end performance (SLF R1: 0.986→0.987) and maintaining a similar throughput (49.8→53.5 FPS). Notably, Cfg B reduces T_{face} from 15.3 to 10.8 ms/frame, confirming that model-family selection alone can meaningfully reduce compute load.

D. Impact of Compression on Unimodal Performance

Cfg C applies pruning and quantization to the face recognition backbone (pruned_ada14), yielding the largest reduction in face latency. Compared to the baseline (Cfg A), T_{face} decreases from 15.3 to 5.9 ms/frame (2.6× speed-up), producing a total throughput increase from 49.8 to 75.2 FPS. This acceleration comes with a marked degradation in unimodal face performance: Face R1 drops from 0.971 to

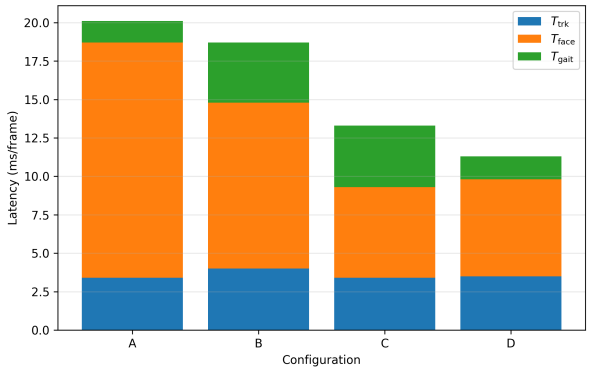

Fig. 3. Per-frame latency breakdown for configurations A–D. The face branch dominates the baseline runtime and is the primary contributor to throughput gains under optimization and compression.

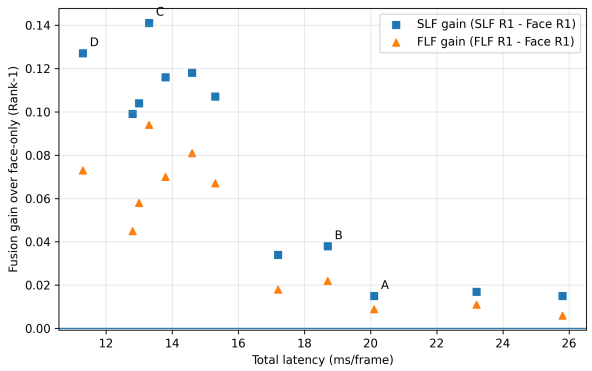

Fig. 4. Fusion compensation versus total latency across CASIA-B configurations. We report fusion gain as the difference between fused Rank-1 accuracy and face-only Rank-1 accuracy. SLF provides consistently higher compensation than FLF, and the gain increases as face-only performance degrades under stronger compression. Letters indicate configurations in Table I.

0.825. In contrast, the gait branch remains stable (Gait R1: 0.869→0.903), consistent with the fact that gait processing was not aggressively compressed in these experiments. Across all tested configurations, Fig. 2 shows that face-only Rank-1 decreases rapidly as latency is reduced, highlighting the sensitivity of facial embeddings to compression.

Cfg D pushes the system toward maximal throughput by pairing the compressed face recognizer with a lighter segmentation stage, reaching 88.5 FPS. However, this setting exhibits a larger end-to-end accuracy reduction, indicating a regime where efficiency gains begin to outweigh the benefits of multimodal redundancy.

E. Multimodal Fusion Resilience (SLF vs. FLF)

As summarized in Fig. 2, fusion accuracy degrades substantially more slowly than face-only accuracy as the pipeline is accelerated. Despite the unimodal degradation observed under compression, multimodal fusion remains comparatively robust. Under Cfg C, score-level fusion (SLF) retains high Rank-1 accuracy (0.966), corresponding to only a 2.0 percentage-point decrease relative to the baseline SLF (0.986), while providing a 1.51× FPS gain and a 2.64× reduction in memory footprint (295.2→111.9 MB). This supports the central hypothesis of this work: multimodal redundancy can compensate for unimodal performance loss induced by aggressive model compression, preserving system-level identification reliability.

Comparing fusion schemes, SLF consistently outperforms feature-level fusion (FLF) across the selected operating points (e.g., Cfg C: 0.966 vs. 0.919; Cfg D: 0.952 vs. 0.898). This aligns with the practical advantage of SLF in plug-and-play pipelines, where each modality can be optimized independently and combined through normalized similarity scores with negligible computational overhead.

Overall, the results indicate that compressed face models that would be suboptimal in isolation can remain effective within a multimodal cognitive perception stack, enabling substantial throughput and memory improvements without catastrophic degradation in end-to-end Rank-1 accuracy.

F. Limitations

Experiments were conducted on a desktop GPU environment. Although the reported trends reflect relative computational load across configurations, validation on embedded platforms (e.g., NVIDIA Jetson-class devices) remains future work. Evaluation is restricted to CASIA-B under near-frontal viewpoints $(0°, 18°, 36°)$; performance under non-frontal angles or real-world acquisition conditions (e.g., crowd scenarios, motion blur, strong illumination variation) may differ, as both face detection reliability and silhouette segmentation quality are sensitive to these factors. The closed-set Rank-1 protocol also does not capture open-set rejection behavior, which is critical for access control deployments where unknown subjects must be rejected rather than misidentified. The pruning strategy (magnitude-based unstructured pruning) does not guarantee wall-clock speedups without hardware-aware sparse execution backends; structured pruning or NAS-based approaches may yield more predictable embedded speedups.

V. Analysis and Design Guidelines

This section analyzes the overall system and provides design guidelines for possible applications.

A. Robustness of SLF and FLF under Compression

Across all tested configurations, score-level fusion (SLF) consistently outperforms feature-level fusion (FLF) (Table I and Fig. 2). A key reason is that SLF combines decision scores after mapping each modality to a comparable similarity range, whereas FLF concatenates embeddings whose statistical properties may drift differently under compression. In particular, quantization and pruning can introduce non-uniform distortions in the face embedding space, affecting angular margins and similarity distributions more severely than gait embeddings. Under FLF, such distortions propagate directly into the fused representation; under SLF, the impact is partially mitigated by (i) per-modality scoring and (ii) normalization prior to weighted averaging.

TABLE II
RECOMMENDED OPERATING POINTS DERIVED FROM TABLE I.

Goal	Config	SLF R1	FPS / Mem (MB)
Accuracy-first	A	0.986	49.8 / 295.2
Balanced	B	0.987	53.5 / 124.6
Memory-first	C	0.966	75.2 / 111.9
Throughput-first	D	0.952	88.5 / 122.8

This interpretation is supported by Fig. 4, which shows that the fusion gain over face-only accuracy increases as the pipeline is pushed toward lower latency (stronger compression), and that SLF provides consistently higher compensation than FLF. In practice, this makes SLF the preferable choice for plug-and-play cognitive perception stacks where individual modalities may be optimized independently and where the computational overhead of fusion must remain negligible.

B. Architecture Selection vs. Compression

The results suggest two complementary efficiency regimes. First, architectural exploration can yield large reductions in model footprint while preserving end-to-end performance. For instance, moving from the baseline (Cfg A) to an architecture-optimized configuration (Cfg B) reduces memory by more than $2\times$ (295.2→124.6 MB) while maintaining throughput and SLF accuracy. Second, targeted compression of the face recognizer provides the largest throughput gains. Cfg C demonstrates that heavy face compression can improve throughput by $1.51\times$ (49.8→75.2 FPS) while maintaining high SLF accuracy (0.966), despite a substantial face-only accuracy drop.

However, the extreme-throughput regime (Cfg D) indicates that beyond a certain point, additional efficiency gains lead to a sharper decrease in end-to-end accuracy. This highlights a practical trade-off boundary: multimodal redundancy can compensate for unimodal degradation, but only while at least one modality remains sufficiently informative and upstream preprocessing (detection/segmentation) remains stable.

C. Deployment-Oriented Operating Points

Based on the observed trade-offs, Table II summarizes recommended operating points for cognitive agents under different constraints. Notably, Cfg B provides a strong memory improvement while keeping SLF Rank-1 effectively equal to the baseline, by just changing tracker and face detector to more memory-efficient ones.

VI. RUNTIME BUDGET AND SYSTEM CONSIDERATIONS

A. Per-Frame Compute Decomposition

For a cognitive agent, perception modules must operate under strict real-time constraints while sharing compute with other subsystems (e.g., navigation, planning, and control). Let the per-frame budget be expressed as

$$T_{\text{tot}} = T_{\text{trk}} + T_{\text{face}} + T_{\text{gait}}, \tag{9}$$

where T_{trk} includes person detection and tracking-by-association, T_{face} includes face detection and face embedding extraction, and T_{gait} includes silhouette segmentation and gait encoding. As shown in Fig. 3, the baseline configuration is dominated by the face branch, motivating targeted optimization of T_{face}.

To quantify how optimization redistributes compute, we report the relative contribution of each stage:

$$\rho_k = \frac{T_k}{T_{\text{tot}}}, \quad k \in \{\text{trk}, \text{face}, \text{gait}\}. \tag{10}$$

In Cfg A, $\rho_{\text{face}} \approx 0.76$, indicating that improvements to tracking or gait would have limited impact on end-to-end throughput. In contrast, under compressed configurations (Cfg C/D), T_{face} is substantially reduced and the runtime becomes more balanced, which is desirable in multi-task robotic stacks where resources must be shared across modules.

TABLE III
DEPLOYMENT GUIDELINES DERIVED FROM BOTTLENECK ANALYSIS AND FUSION BEHAVIOR.

Scenario	Recommended strategy
Tight compute / real-time constraint	Prioritize optimizing T_{face} (recognizer first); use SLF to preserve accuracy under compression.
Memory-limited deployment	Prefer architecture exploration (Cfg B-like) to cut footprint before applying heavy compression.
Accuracy-critical operation	Use conservative compression (Cfg A/B-like) and SLF; avoid extreme settings where upstream preprocessing changes degrade end-to-end accuracy.
Multi-task perception stack	Aim for balanced stage contributions (Cfg C-like) to reduce contention and enable co-execution with other modules.

B. Implications for Embedded Deployment

Although our measurements were obtained on a desktop GPU, the observed trends are informative for embedded deployment because they characterize relative computational load and identify the dominant bottleneck. In practice, embedded accelerators often provide the largest benefits for quantized inference, so reductions in T_{face} and in overall model footprint can translate into improved throughput, lower energy consumption, or freed compute budget for additional perception tasks (e.g., gesture or emotion recognition).

Moreover, multimodal fusion offers a useful systems-level property: it allows individual modules to be optimized aggressively while preserving end-to-end performance. This is particularly relevant when operating under a hard real-time constraint, where the system must prefer stable, bounded-latency operations. Score-level fusion (SLF) is especially attractive in this context because it introduces negligible overhead and remains robust across the explored configurations (Fig. 4).

Table III summarizes pragmatic guidance derived from our results. These recommendations prioritize (i) maximizing the impact of optimization by targeting the dominant stage and (ii) maintaining system-level robustness via fusion.

VII. Conclusion

This paper investigated whether multimodal redundancy can be exploited to enable aggressive optimization of person recognition pipelines intended for cognitive agents. We benchmarked a face-and-gait identification stack on CASIA-B (near-frontal views $0°/18°/36°$) under a Rank-1 protocol, and evaluated architectural exploration together with face-model compression (quantization and pruning).

Results show that the face branch is the main bottleneck in the baseline configuration, and that both model-family selection and compression substantially improve throughput and reduce memory footprint. Crucially, while unimodal face performance degrades markedly under compression (e.g., Face R1: 0.971→0.825), the multimodal system is more robust: score-level fusion (SLF) retains high accuracy (0.986→0.966) while increasing throughput from 49.8 to 75.2 FPS and reducing the model footprint from 295.2 MB to 111.9 MB. Across the explored configurations, SLF consistently outperforms feature-level fusion (FLF), supporting its suitability for plug-and-play, low-overhead deployment.

These findings support the central hypothesis that multimodal fusion can act as a resilience mechanism that tolerates unimodal degradation, enabling lightweight perception without end-to-end performance loss. Future work will pursue several directions. First, validation on embedded accelerators (e.g., Jetson Orin or Hailo-class devices) will quantify absolute energy and latency benefits of quantized inference in realistic deployment conditions. Second, we plan to extend evaluation beyond near-frontal views to the full CASIA-B viewpoint range and to datasets with more challenging real-world covariates (e.g., GREW [26], Gait3D [27]). Third, we will explore adaptive fusion policies that modulate α based on per-frame quality signals such as face detection confidence, subject distance, or embedding norm, which are naturally available at inference time. Finally, structured pruning and neural architecture search for the face branch represent complementary directions for achieving hardware-friendly compression with guaranteed wall-clock speedups.

Acknowledgment

Federico Boscolo's work is part of the project PNRR-NGEU which has received funding from the MUR – DM 117/2023.

References

[1] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," *EUSIPCO*, 2004.

[2] X. Huang, H. Wang, S. Qin, and S.-K. Tang, "Embedded artificial intelligence: A comprehensive literature review," *Electronics*, vol. 14, no. 17, p. 3468, 2025.

[3] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[4] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[5] Ultralytics, "Ultralytics YOLO (yolov8/yolo11) repository," https://github.com/ultralytics/ultralytics, 2025, accessed: 2026-02-20.

[6] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *European Conference on Computer Vision (ECCV)*, 2022.

[7] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, 2006, pp. 441–444.

[8] Y. Xin, L. Kong, Z. Liu, C. Wang, H. Zhu, M. Gao, C. Zhao, and X. Xu, "Multimodal feature-level fusion for biometrics identification system on IoMT platform," *IEEE Access*, vol. 6, pp. 21 418–21 426, 2018.

[9] A. Prakash, T. S., A. Nambiar, and A. Bernardino, "Multimodal adaptive fusion of face and gait features using keyless attention based deep neural networks for human identification," arXiv:2303.13814, 2023.

[10] W. Liu, Y. Wen, Z. Yu, and M. Yang, "SphereFace: Deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[12] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "GhostFaceNets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35 429–35 446, 2023.

[14] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," arXiv:1804.07573, 2018.

[15] Y. Martinez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza, "ShuffleFaceNet: A lightweight face architecture for efficient and highly-accurate face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[16] Y. Zhang and J. Yang, "A comprehensive survey of lightweight neural networks for face recognition," *Journal of Korean Society of Industrial and Systems Engineering*, vol. 46, no. 1, pp. 55–67, 2023.

[17] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges," *arXiv preprint arXiv:2206.13732*, 2022.

[18] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[19] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 213–14 221.

[20] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges," *arXiv preprint arXiv:2206.13732*, 2023, v2, 5 Aug 2023.

[21] A. B. Mughal, R. U. Khan, A. Bermak, and A. ur Rehman, "Person recognition via gait: A review of covariate impact and challenges," *Sensors*, vol. 25, no. 11, p. 3471, 2025.

[22] C. Fan, Y. Wang, C. Li *et al.*, "OpenGait: Revisiting gait recognition toward better practicality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[23] M. Nagel, R. A. Amjad, M. van Baalen, C. Louizos, and T. Blankevoort, "Up or down? adaptive rounding for post-training quantization," *arXiv preprint arXiv:2004.10568*, 2020.

[24] H. Cheng, M. Zhang *et al.*, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[25] L. Wei, Z. Ma, C. Yang, and Q. Yao, "Advances in the neural network quantization: A comprehensive review," *Applied Sciences*, vol. 14, no. 17, p. 7445, 2024.

[26] X. Guo, Z. Zhu, T. Yang, B. Lin, J. Huang, J. Deng, G. Huang, J. Zhou, and J. Lu, "Gait recognition in the wild: A large-scale benchmark and NAS-based baseline," arXiv:2205.02692, 2024.

[27] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.