FusionAgent: A Multimodal Agent with Dynamic Model Selection for Human Recognition

Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, Xiaoming Liu
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824
{zhujie4, guoxia11, suyiyan1, jain, liuxm}@msu.edu

Abstract

Systematic human recognition requires integrating multiple biometric traits such as face, gait, and body shape, through specialized models to achieve robustness in unconstrained scenarios. However, existing score-fusion strategies typically adopt a static design, combining all models for every test sample regardless of sample quality. This not only increases unnecessary computation but can degrade performance by incorporating noisy or unreliable modalities. To overcome these limitations, we propose ***FusionAgent****, a novel agentic framework that leverages a Multimodal Large Language Model (MLLM) to perform dynamic, sample-specific model selection. Each model is treated as a tool, and through Reinforcement Fine-Tuning (RFT) with a metric-based reward, the agent learns to adaptively determine the optimal model combination for each test input. To address the model score misalignment and embedding heterogeneity, we introduce Anchor-based Confidence Top-k (ACT) score-fusion, which anchors on the most confident model and integrates complementary predictions in a confidence-aware manner. Extensive experiments on multiple whole-body biometric benchmarks demonstrate that FusionAgent significantly outperforms SoTA methods, underscoring the critical role of dynamic, explainable, and robust model fusion in real-world recognition systems. The proposed framework is scalable and adaptable to a wide range of multi-modal and multi-model tasks, such as vision-language retrieval, indicating its potential relevance to broader application scenarios. The code and model will be publicly released upon publication.*

1. Introduction

Human (whole-body) recognition is a challenging task that leverages multimodal traits (*e.g.*, face, gait, and body shape) and multiple specialized models to identify individuals accurately [44]. In contrast to traditional unimodal biometric systems that rely on a single modality, such as face recognition (FR) [6, 13, 14], gait recognition (GR) [40, 43], or person re-identification (ReID) [7, 17, 34], systematic whole-body recognition integrates complementary information from diverse modalities to improve the recognition performance. This fusion enhances robustness but introduces challenges in effectively and efficiently combining multi-model outputs. Such integration is critical to overcoming the limitations of unimodal systems; for instance, FR models struggle with occluded faces, where GR or ReID models can provide reliable alternatives. These capabilities are vital in real-world surveillance applications, which demand high accuracy and robustness despite low-quality imagery [16].

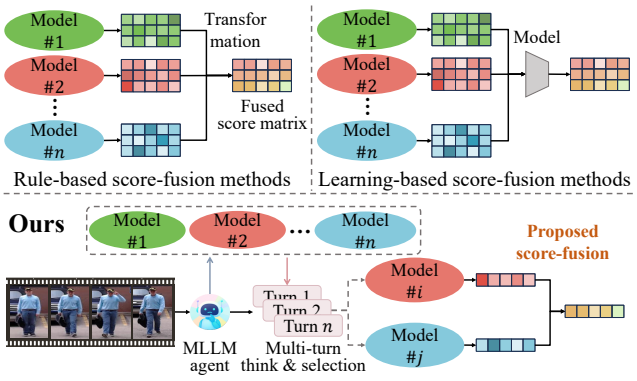

Figure 1. **Comparison of score-fusion methods. Top:** Rule-based methods apply predefined transformations to fuse all model scores, while learning-based methods infer a fusion model from data but still assume that every model contributes to all test samples. **Bottom:** our framework leverages an MLLM agent to dynamically select a subset of models, followed by the proposed score-fusion strategy, enabling adaptive and robust integration.

Score-level fusion is essential in human recognition systems. As shown in Fig. 1, existing methods fall into two main categories: rule-based and learning-based. Empirical approaches use predefined rules to combine scores [11, 32], making them computationally efficient and training-free.

In contrast, learning-based methods learn fusion strategies from data [22, 35, 44], capturing complex relationships at the cost of requiring training data and model design. A key limitation of both approaches is their inherent assumption that all models provide complementary information, leading them to treat the model combination as fixed for every input, even those from low-quality or unreliable modalities. For instance, in cases where a video frame only captures the back view of a person, incorporating FR models would be inappropriate. Although some methods, such as QME [44], attempt to mitigate poor input quality within the fusion process, they still allow low content inputs to influence the output. This suggests that the conventional fusion designs may be suboptimal. An ideal fusion framework should instead: 1) dynamically select the most suitable model combination for each input, and 2) effectively integrate the outputs from the selected models. To systematically explore the potential of such an approach, we focus on the following question:

*How can we adaptively choose the **optimal model combination for each test sample** while rigorously testing whether such selection improves performance even with an **ad hoc** score-fusion algorithm?*

To address the question, we introduce a unified framework centered around two core components, each designed to systematically respond to model selections and score-fusion methods. In response to model selections, we propose **FusionAgent**, an MLLM-based agent that performs dynamic model selection per sample. Each biometric model is wrapped as a tool, providing the agent with a score vector and predicted label. Through Reinforcement Fine-Tuning (RFT) guided by the **proposed metric-based reward**, the agent learns to identify sample-specific model combinations by analyzing query patterns, effectively answering how optimal ensembles can be found for each input. To explore score-fusion methods, we design a lightweight **Anchor-based Confidence Top-k score-fusion method (ACT)**. This method leverages the most confident model selected by the agent, then integrates only the top-matching scores scaled by confidence weights. By doing so, ACT mitigates misalignment from sample-wise model selection and embedding heterogeneity, demonstrating that even a simple fusion rule can yield strong performance when guided by adaptive model selection. The proposed method can be extended to various retrieval-oriented scenarios. For instance, it could facilitate cross-modal retrieval between visual and textual data or enhance domain-specific search systems. Our main contributions are as follows:

- We propose FusionAgent, an agentic framework that leverages an MLLM to perform explainable, sample-wise dynamic model selection.
- We introduce an Anchor-based Confidence Top-k score-fusion method (ACT) to mitigate score misalignment arising from dynamic model selection, ensuring robust integration of heterogeneous score outputs.
- We design a metric-based reward function, central to our optimization, which directly aligns the agent's selection strategy with final performance metrics.
- Extensive experiments on multiple whole-body biometric benchmarks demonstrate the superiority of our approach over state-of-the-art methods, even using conventional score-fusion methods.

2. Related Work

2.1. Whole-body Biometric Recognition

Whole-body biometric recognition systems combine detectors (*e.g.*, for whole body and face), embedding models, and fusion modules to leverage multi-modal cues such as face and body features [5, 16, 18]. The key challenge is to effectively integrate the complementary strengths of different modalities and their dedicated models to maximize overall performance. For example, FR models excel with high-quality frontal faces but struggle under adverse conditions like oblique angles [13, 14]. In contrast, GR models focus on clothing-invariant dynamic body attributes [40, 43], while ReID models take a holistic approach to extract comprehensive appearance features [7, 17, 37, 38]. Prior fusion methods typically use all available models for every query [3, 16, 18, 22, 35, 44], ignoring the sample-dependent nature of optimal model combinations. A clear case is a low-resolution image with a side-view of a person: FR models may be unreliable, making GR or ReID models more critical for that specific sample. Therefore, we propose FusionAgent to dynamically select an optimal model combination for each test sample through explainable analysis, thereby tailoring the fusion process to individual inputs.

2.2. Score-fusion

Score-level fusion integrates similarity scores from multiple modalities to improve recognition performance [31]. This approach is broadly categorized into two paradigms: rule-based and learning-based methods. Rule-based methods employ fixed rules—such as Z-score, Min-max normalization, max/min fusion [11, 12, 26, 41], and likelihood ratio-based fusion [8, 21, 23, 24, 36], offering simplicity and efficiency. In contrast, learning-based methods optimize fusion during training [3, 22, 35]. These include recent quality-aware approaches like QME [44], which performs weighted score-fusion based on input quality. In contrast, we introduce a simple score-fusion algorithm with an anchor. This design, combined with selective model input, leads to superior performance compared with existing fusion strategies.

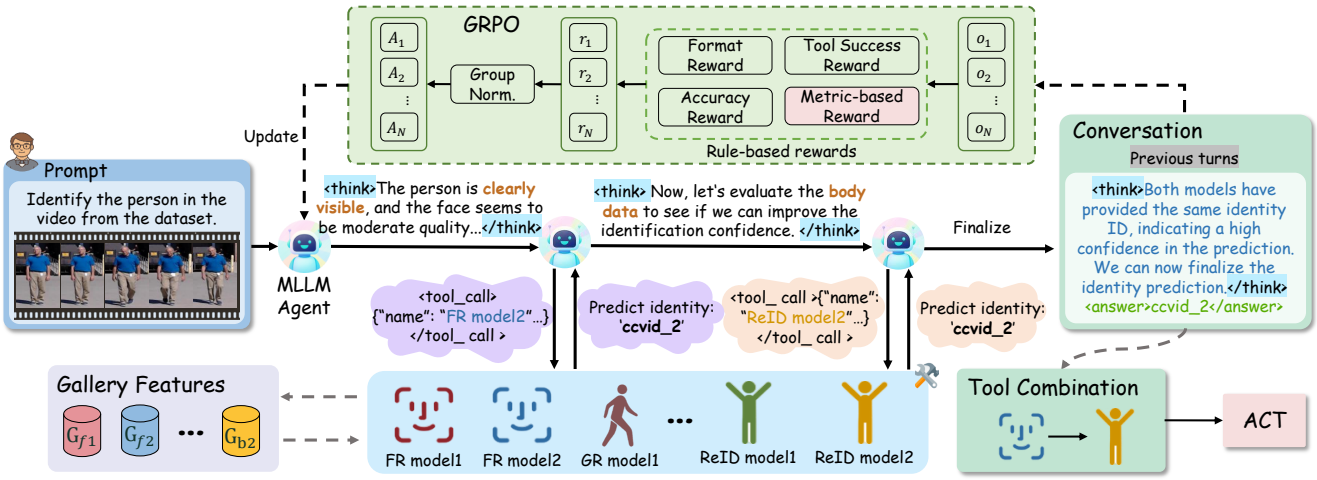

Figure 2. **Overview of the FusionAgent framework.** Recognition models are wrapped as tools to generate score vectors and predicted identities based on gallery features. The MLLM agent receives multimodal biometric inputs and performs a reasoning-action step through multi-turn, selectively invokes tools, and integrates predictions into a final identity decision and fused score vector. The agent is optimized with reinforcement fine-tuning using rule-based rewards, including the proposed metric-based reward.

2.3. Agents and Reinforcement Learning

The advent of tool-augmented MLLMs has spurred the development of agentic systems: models capable of planning, reasoning, and interacting with external tools to solve complex tasks. Recent studies [2, 20, 39, 45] have equipped MLLMs with tools to address diverse challenges such as visual grounding, image super-resolution, and advanced visual understanding. Furthermore, subsequent research has employed RFT with rule-based rewards, such as Group Relative Policy Optimization (GRPO) [28], which offers improved generalizability with lower data dependency [19, 29, 42]. Building upon these advancements, we further propose a metric-based reward to optimize model selections of FusionAgent, effectively capturing data-dependent patterns.

3. Proposed Method

Problem Formulation. In biometric recognition, a query (or probe) is a sample sequence that is to be identified (1:N comparisons) or verified by comparison (1:1 comparison) with a gallery of previously enrolled subjects in the system. Let Q denote the set of query videos and G be the gallery. M is a predefined model set $\{m_1, m_2, \ldots, m_Z\}$ that is used for feature extraction. For a query $q \in Q$ and a model $m \in M$ (from a predefined model set M), m extracts features from q and produces a similarity score vector $\mathbf{s}_{m,q} \in \mathbb{R}^{|G|}$ with respect to each gallery item $g \in G$. The human recognition pipeline comprises two stages: (i) *sample-level model subset selection* and (ii) *query-based score fusion of selected models*. For each q, we seek an optimal subset of models $\mathbf{M}_q \subseteq M$ that maximizes performance. Given $\mathbf{M}_q$, we stack the per-model scores into a score matrix $\mathbf{S}_q \in \mathbb{R}^{|M_q| \times |G|}$ whose i-th row is $\mathbf{s}_{m_i,q}$. A score-fusion function f then maps $\mathbf{S}_q$ to a fused score vector $\mathbf{s}'_q = f(\mathbf{S}_q) \in \mathbb{R}^{|G|}$. We detail the model-selection procedure using an agent in Sec. 3.1, the supervision signals (reward functions) in Sec. 3.2, and the proposed score-fusion method in Sec. 3.3.

3.1. Agentic Training Framework

Training the agent presents two challenges: (i) the lack of ground-truth dialogue data for Supervised Fine-tuning (SFT), since grid searching optimal model combinations per sample is infeasible; and (ii) the intractability of dataset-level grid search for the best model combination per sample due to exponential growth in search space with models and samples. We thus employ GRPO, which enables trial-based learning without pre-labeled data. Additionally, we design a metric-based reward function that promotes exploration of diverse model combinations and facilitates implicit learning of the relationship between input characteristics and model performance (details in Sec. 3.2).

The overall framework is depicted in Fig. 2. For a given query q, the agent first analyzes its content and selects an initial model $a_q \in M_q$. For example, FR model 2 is selected as a clearly visible face is detected. Upon a successful function call, the tool returns $s_{a_q}(q)$ and the predicted label. Since score vectors are not directly interpretable by the agent, the predicted identity label is returned explicitly, while $s_{a_q}(q)$ is retained internally. The agent then decides whether to invoke another model or output its decision. This sequential decision-making process enables the flexible maximization of a long-term objective. [30, 39]. If the agent decides to output its decision, it also provides a

reasoning summary that justifies the model selections. Reasoning at each step is necessary to enhance transparency and decision traceability. The whole conversation o_i, including the successful model selections, is then used to measure the overall rewards. In GRPO, we sample N rollouts (*i.e.*, responses) for each query and jointly compute their advantages $\{A_1, A_2, \ldots, A_N\}$ from the corresponding rewards $\{r_1, r_2, \ldots, r_N\}$ to update the agent. Details of GRPO are provided in the supplementary.

Multi-turn Design. We adopt a ReAct-style (reason-before-act) multi-turn controller for tool use, rather than a single-shot plan generator [39]. This design is motivated by the need to decompose the complex task of multi-tool usage into a sequence of simpler, more manageable decisions. A single-turn approach would require the agent to generate a complete and static execution plan at once, a task plagued by a combinatorial action space and an inability to handle unexpected outcomes. This interleaving of reasoning with actions decomposes multi-tool execution into atomic steps and yields the following benefits:

i) Simplified Learning: Reduces the vast action space of generating a full plan to a single decision at each step, making the policy significantly more tractable to learn.

ii) Dynamic Adaptation: Allows the agent to observe tool outputs and adjust its strategy in real time, enabling error correction and flexible reasoning. These sequential actions also support effective credit assignment during inference.

3.2. Reward Functions

Reward functions play a central role in RFT-based agent training. Unlike metric learning or SFT, which minimize loss functions, RFT seeks to maximize reward. We design four reward functions: format reward, tool success reward, answer accuracy reward, and metric-based reward.

Format Reward. The agent is trained to produce structured responses that separate reasoning, tool calls, and final answers [19, 20, 28]. We adapt this to a multi-turn setting where each turn must be a structured format, such as *<think>...</think> <answer>...</answer>*. The reward is computed per turn, and the overall reward is averaged across all turns.

Tool Success Reward. This reward assesses whether the agent's tool calls are executable and yield valid results. Each tool call receives a binary success/failure score, and the reward is defined as the success rate across all tool calls in a trajectory. This incentivizes the agent to produce syntactically correct tool inputs.

Answer Accuracy Reward. This reward measures the correctness of the agent's final prediction a_i extracted from the answer tag *<answer>...</answer>* of the response o_i against the ground truth label y. Its primary objective is to prioritize factual accuracy over procedural correctness, which enables the agent to implicitly assess model reliability, especially when predictions conflict:

$$R_{acc}(a_i, y) = \begin{cases} 1, & \text{if } a_i = y, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Metric-based Reward. The metric-based reward guides the agent toward effective and dynamic model selection. A key challenge is to encourage sufficient exploration of diverse model combinations without prior knowledge of their optimality. To address this, the reward is computed based on the performance of the agent-selected model combination on the training set, using the score-fusion method from Sec. 3.3. The combination is derived from successfully executed conversations per sample.

However, exhaustively searching for the ground-truth optimal model combination for each sample is computationally infeasible. To promote exploration, we construct an augmented model selection $\mathbf{M_Q} \in \mathbb{R}^{|Q|\times|M|}$ based on the model combination M_{o_i} of the response o_i. Parameter $\gamma \in [0, 1]$ is the ratio of samples in $\mathbf{M_Q}$ whose model combination is the same as M_{o_i}. For the remaining $(1 - \gamma)$, the mask entries are sampled from a Bernoulli distribution to diversify the model combinations explored. The column corresponding to the anchor model is always set to true, ensuring its consistent inclusion.

We apply the ACT (details in Sec. 3.3) based on $\mathbf{M_Q}$. The performance of the resulting fused scores matrix $\mathbf{S}'_{o_i}$ is then evaluated across the entire training set using four key metrics: True Accept Rate at a False Acceptance Rate (TAR@FAR), mean Average Precision (mAP), Rank-1 accuracy, and False Non-Identity Rate at a specified False Positive Identification Rate (FNIR@FPIR). The final metric-based reward R_{mat} is a composite score formulated to provide a holistic performance signal:

$$\mathbf{S}'_{o_i} = \text{ACT}(\mathbf{M_Q}, k), \tag{2}$$

$$\begin{aligned} R_{mat}(o_i) = \text{Rank}(\mathbf{S}'_{o_i}) + \text{mAP}(\mathbf{S}'_{o_i}) \\ +\text{TAR}(\mathbf{S}'_{o_i}) - \text{FNIR}(\mathbf{S}'_{o_i}). \end{aligned} \tag{3}$$

This formulation naturally aligns with both our optimization objective and real-world deployment needs. It rewards the agent for achieving higher accuracy and retrieval performance (TAR, mAP, Rank) while penalizing missed identifications (FNIR), thereby capturing operational performance requirements in a comprehensive manner. By consolidating these diverse metrics into a single scalar, the agent receives a clear and reliable training signal, effectively guiding its policy toward discovering superior model combinations.

3.3. Anchor-based Confidence Top-k Score-fusion

The proposed ACT (Anchor-based Confidence Top-k) score-fusion approach aims to dynamically and effectively

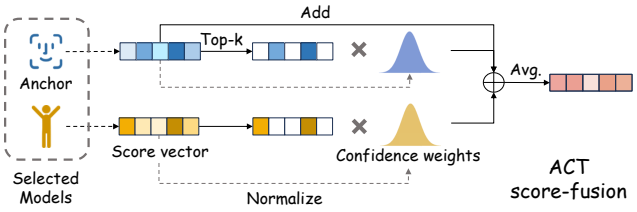

Figure 3. **Overview of the ACT score-fusion.** Based on tool execution results (*i.e.*, score vectors) and selected model combination, the first selected model serves as the anchor, and ACT produces the final score vector via confidence weighting and top-k filtering.

combine scores from multiple models for a given query. The overview is shown in the Fig. 3. Our approach is built on the principle of leveraging a stable and powerful "anchor model" $m_a \in \mathbf{M}_q$ to provide a robust score vector, while selectively incorporating normalized, high-confidence scores from the set of models. Let m_a be the first selected model in $\mathbf{M}_q$, and $s_{m,q,g}$ denote the similarity score between q and g for model m.

We begin by computing a contribution score $c_{m,q,g}$. This score is designed to leverage the most confident predictions while filtering out potential noise from low-scoring candidates, which is achieved through a Top-k selection process. Formally, for a score $c_{m,q,g}$ in $\mathbf{c}_{m,q} \in \mathbb{R}^{|G|}$, the contribution score is defined as:

$$c_{m,q,g} = \begin{cases} z_{m,q,g} \cdot s_{m,q,g} & \text{if } g \in \mathcal{T}_{m,q}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $\mathcal{T}_{m,q}$ is the set of indices of the k highest-scoring gallery items for query q from model m. The term $z_{m,q,g}$ represents the Z-score normalized score from m between q and g. This standardization makes the scores from different models with potentially different scales comparable. It is calculated as:

$$z_{m,q,g} = \frac{s_{m,q,g} - \mu_{m,q}}{\sigma_{m,q}}, \tag{5}$$

where $\mu_{m,q}$ and $\sigma_{m,q}$ are the mean and standard deviation of $\mathbf{s}_{m,q}$. The final fused score vector $\mathbf{s}'_q$ is computed via:

$$\mathbf{s}'_q = \frac{1}{1 + |\mathbf{M}_q|} \left(\mathbf{s}_{m_a,q} + \sum_{m \in \mathbf{M}_q} \mathbf{c}_{m,q} \right). \tag{6}$$

The term $(1 + |M_q|)$ serves as the normalization factor, ensuring a balanced contribution from all models. This strategy balances robust, general performance with specialized, high-confidence insights. The anchor model provides a more substantial contribution by establishing a global ranking structure, as its scores are applied unconditionally to all candidates. In contrast, the selected models provide sparse, localized refinements only for their top-k predictions. Fig. 4 provides a toy example of ACT score-fusion.

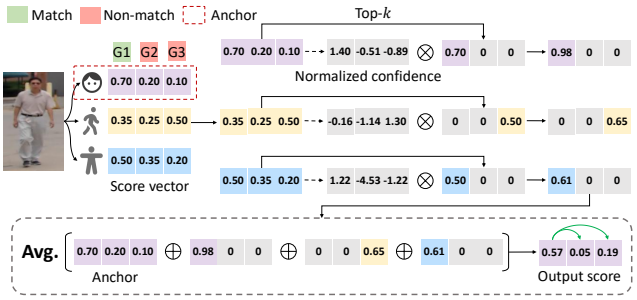

Figure 4. **A toy example of the proposed ACT score-fusion.** Three models are used with the FR model as the anchor and $k = 1$. ACT amplifies the gap between match and non-match scores through confidence-based top-k and anchor weighting, improving verification and open-set search performance.

4. Experiments

Datasets and Evaluation Protocol. We evaluate on popular human recognition datasets: CCVID [7], MEVID [4], and LTCC [25]. They comprise multi-view captures and cross-modal biometric data with diverse resolutions and temporal dynamics. Our evaluation protocol follows [44] for fair comparison. Performance is measured via standard metrics in general settings, *i.e.*, Rank-1 accuracy, mean Average Precision (mAP), verification rate (TAR@1.0%FAR), and open-set search metric (FNIR@1.0%FPIR), which collectively reflect real-world deployment requirements. For open-set search evaluation, we follow the protocol of [33, 44] to construct 10 random subsets of gallery subjects. Each subset, containing $\sim$20% of the subjects in the test set, serves as a non-mated list. We report the mean and standard deviation for the 10 trials.

Baselines. Our baselines include the following methods: Min/Max score-fusion [11], Z-score and min-max normalization [32], RHE [8], Farsight [16], and learning-based methods like BSSF [35], Weighted-sum [22], AsymA-O1 [9], and QME [44] (ICCV'25). We follow [44] to construct the biometric model pool (details provided in the Supp.). We also compare with the SoTA multimodal human recognition model [15] (CVPR'25).

Implementation Details. Following [44], We extract the center embeddings of subjects for training and gallery embeddings for testing by applying average pooling. Considering the demand for faster responses in practical applications, FusionAgent is based on Qwen2.5-VL-3B [1] and fine-tuned via GRPO [28]. For efficiency, we set a turn limit of 4 and train for 200 steps. All biometric models are frozen in training. For GRPO, we set $N = 6$ and a KL coefficient $\beta = 0.04$. LoRA [10] is applied with rank $r = 64$ and scaling factor $\alpha = 128$. The learning rate follows a linear decay schedule starting from 2×10^{-5}. The metric-based reward uses $\gamma = 0.8$. Based on training-set performance, we set

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [13]	♦	94.0	87.9	75.7	13.0 ± 3.5
CAL [7]	♠	81.4	74.7	66.3	52.8 ± 13.3
*BigGait** [40]	♣	76.7	61.0	49.7	71.1 ± 6.1
SapiensID [15]	●	92.6	77.8	-	-
Min-Fusion [11]	♦♠♣	87.1	79.2	62.4	48.5 ± 8.7
Max-Fusion [11]	♦♠♣	89.9	89.3	73.4	23.0 ± 10.1
Z-score [32]	♦♠♣	92.2	90.6	73.9	15.1 ± 1.5
Min-max [32]	♦♠♣	91.8	90.9	73.9	15.4 ± 2.5
Weigthed-sum [22]	♦♠♣	91.7	90.6	73.6	15.4 ± 1.8
Asym-AO1 [9]	♦♠♣	92.3	90.0	74.0	15.9 ± 1.7
BSSF [35]	♦♠♣	91.8	91.1	73.9	14.1 ± 1.3
Farsight [16]	♦♠♣	92.0	91.2	73.9	13.9 ± 1.1
QME [44]	♦♠♣	**94.1**	90.8	76.2	12.3 ± 1.4
FusionAgent (DA)	♦♠♣	92.8	92.2	85.8	10.5 ± 1.5
FusionAgent (CoT)	♦♠♣	93.4	**92.6**	**85.9**	**10.1 ± 1.5**

Table 1. **Performance on CCVID.** [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; ♦: AdaFace for face modality; ♣: BigGait for gait modality; ♠: CAL of body modality; ●: SapiensID for face and body modalities; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [13]	♦	18.5	5.9	2.4	99.8 ± 0.2
CAL [7]	♠	74.4	40.6	36.7	59.7 ± 7.3
AIM [38]	■	74.8	40.9	37.0	66.2 ± 7.5
SapiensID [15]	●	72.0	34.6	-	-
Min-Fusion [11]	♦♠■	38.1	13.5	12.4	81.9 ± 6.0
Max-Fusion [11]	♦♠■	62.5	33.3	16.8	94.8 ± 4.7
Z-score [32]	♦♠■	73.0	37.5	30.4	68.7 ± 9.2
Min-max [32]	♦♠■	73.2	38.1	31.9	75.1 ± 9.2
Weigthed-sum [22]	♦♠■	73.2	37.8	31.3	72.4 ± 8.6
Asym-AO1 [9]	♦♠■	71.2	32.9	19.1	76.3 ± 8.9
BSSF [35]	♦♠■	73.5	39.1	34.2	68.9 ± 8.5
Farsight [16]	♦♠■	73.2	37.8	31.3	72.4 ± 8.6
QME [44]	♦♠■	73.8	39.6	35.0	64.3 ± 8.0
FusionAgent (DA)	♦♠■	**75.5**	**41.0**	36.5	50.3 ± 9.0
FusionAgent (CoT)	♦♠■	**75.5**	**41.0**	**37.0**	**50.0 ± 8.5**

Table 2. **Performance on LTCC.** [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; ♦: AdaFace for face modality; ♠: CAL of body modality; ■: AIM for body modality; ●: SapiensID for face and body modalities; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

$k = 10$ for CCVID and $k = 40$ for MEVID and LTCC. During training, we sample a continuous 4-frame clip from each video for CCVID and MEVID, and 1 frame for LTCC as it is an image-based dataset. Training is conducted on 4 H100 GPUs with an effective batch size of 4, and takes nearly 4 hours. During evaluation, we disable sampling in FusionAgent to ensure results are reproducible.

4.1. Experimental Results

In-domain Evaluation. As shown in Tab. 1, 2, and 3, statistical baselines (*e.g.*, Z-score, FarSight) often fail to surpass the strongest single model across all metrics, while learning-based methods (*e.g.*, QME) achieve better gains and occasionally outperform the best unimodal model, but remain limited when fused models are not complementary (*e.g.*, LTCC). By contrast, FusionAgent consistently achieves superior performance on all three datasets, outperforming both unimodal and fusion baselines on most metrics. A consistent trend across the tables is that FNIR benefits the most from fusion while Rank-1 improves the least: this can be explained because FNIR is particularly sensitive to outliers, whereas Rank-1 is largely determined by the strongest modality and only marginally affected by fusion. Through top-k selection and confidence weighting, our method effectively bounds the increase in non-match scores. On CCVID, the largest TAR arise as FusionAgent identifies the high quality of facial inputs and primarily anchors on the FR model, which excels at distinguishing matches from non-matches. On LTCC, FNIR reduction

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [13]	♦	25.0	8.1	5.4	98.8 ± 1.2
CAL [7]	♠	52.5	27.1	34.7	67.8 ± 7.3
AGRL [37]	■	51.9	25.5	30.7	69.4 ± 8.9
Min-Fusion [11]	♦♠■	46.8	21.2	28.0	70.4 ± 8.0
Max-Fusion [11]	♦♠■	33.2	14.9	8.3	97.4 ± 1.6
Z-score [32]	♦♠■	54.1	27.4	30.7	66.5 ± 7.0
Min-max [32]	♦♠■	52.8	24.7	25.0	71.3 ± 6.1
Weigthed-sum [22]	♦♠■	54.1	27.3	30.3	66.3 ± 7.0
Asym-AO1 [9]	♦♠■	52.5	22.9	23.6	71.7 ± 5.8
BSSF [35]	♦♠■	53.5	27.4	30.5	65.9 ± 7.2
Farsight [17]	♦♠■	53.8	25.4	26.6	69.8 ± 6.4
QME [44]	♦♠■	**55.7**	28.2	32.9	64.6 ± 8.2
FusionAgent (DA)	♦♠■	52.5	**28.7**	34.8	60.8 ± 7.3
FusionAgent (CoT)	♦♠■	54.7	**28.7**	**34.9**	**58.6 ± 7.4**

Table 3. **Performance on MEVID.** [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; ♦: AdaFace for face modality; ♠: CAL of body modality; ■: AGRL for body modality; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

Method	Rank1↑	mAP↑	TAR↑	FNIR↓
CCVID → LTCC (Zero-shot)				
FusionAgent (ACT)	60.4	11.9	7.7	60.3 ± 8.6
FusionAgent (Farsight)	68.2	31.7	17.0	81.8 ± 9.6
CCVID → LTCC (10-shot)				
FusionAgent (ACT)	73.6	39.8	34.8	53.5 ± 8.5
MEVID → LTCC (Zero-shot)				
FusionAgent (Farsight)	75.3	42.3	37.7	59.3 ± 9.5
FusionAgent (ACT)	75.3	41.1	36.1	50.1 ± 8.4

Table 4. **Cross-domain Performance on LTCC.** FusionAgent is trained on CCVID with its model combination (Tab. 1) and evaluated on LTCC using LTCC's model combination (Tab. 2). [Keys: TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

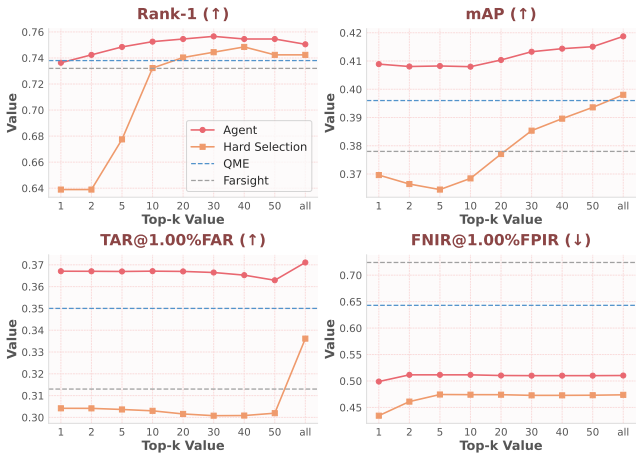

Figure 5. **Performance comparison on LTCC in 4 metrics.** FusionAgent consistently outperforms baselines, including the hard selection (i.e., using all models), which highlights the effectiveness of dynamic model selection.

is most pronounced due to the proposed ACT score-fusion strategy (see ablation in Tab. 5).

Efficiency in Real-world Scenarios. We introduce two inference modes: Direct Answering (DA) and Chain-of-Thought (CoT). DA omits explicit reasoning, reducing average inference time from 2.81s to 1.03s (QME takes 0.67s) while maintaining competitive performance on an H100 device. CoT offers interpretable results, but at the cost of higher latency. Users can flexibly choose between the two to balance efficiency, interpretability, and accuracy.

Cross-domain Evaluation. In real-world deployments, agents might be deployed in unseen environments with different tool combinations, making cross-domain evaluation critical for assessing transferability. In Tab. 4), we evaluate FusionAgent by training on CCVID and testing on LTCC. In the zero-shot setting, although the tool pool is changed, FusionAgent generalizes well and successfully executes tool calls without any additional training. Nevertheless, in CCVID $\rightarrow$ LTCC, where face quality degrades, FusionAgent experiences a performance drop due to (1) the unpredictable behavior of newly introduced models or weights on the target dataset, and (2) the domain gap in data distribution. With only 10 samples per subject (*i.e.*, 8% of the training data) and 50 training steps, FusionAgent rapidly adapts, achieving performance comparable to training on the full in-domain dataset. In MEVID $\rightarrow$ LTCC, the zero-shot setting attains similar performance, further demonstrating the robustness of FusionAgent.

4.2. Ablation Studies

Agent Selection *vs*. Hard Selection. We compare dynamic model selection by the agent (red line) with hard selection, which uses all available models (orange line) in Fig. 5. We use the averaged scores as the anchor for the hard selection. In the agent case, the first selected model serves as the anchor, whereas for hard selection such anchor information is unavailable. Therefore, we use the averaged scores as a surrogate anchor for hard selection. Our method yields notable improvements across all metrics compared to previous SoTA approaches, underscoring the efficacy of the proposed agent and the ACT algorithm. Furthermore, it consistently outperforms hard selection in terms of Rank-1, mAP, and TAR, highlighting the critical role of adaptive model selection. The significantly lower performance of hard selection relative to prior SoTA confirms that dynamic model selection is the key driver of performance gains.

Top-K Values. We investigate the impact of top-k values in ACT. Fig. 5 indicates that the choice of k affects different metrics distinctively: larger k values generally improve Rank-1, mAP, and TAR@FAR, while FNIR remains relatively stable. Compared with hard selection, which exhibits large fluctuations, especially at small k, the proposed agent maintains both higher performance and greater stability across a wide range of k values. This shows that the agent not only adapts to sample-level variations by selectively exploiting complementary models but also avoids the noise introduced by averaging redundant scores. The robustness of ACT under varying k highlights its effectiveness in supporting adaptive model selection, ensuring consistent improvements across diverse evaluation settings.

Alternative Score-fusion Methods. Tab. 5 compares the proposed ACT against alternative statistical score-fusion methods using the same agent-selected model combinations. Even with standard fusion techniques such as Z-score and Farsight, our approach already surpasses the QME, highlighting the decisive role of dynamic model selection. Interestingly, Z-score and Farsight achieve nearly identical results, suggesting that the choice of statistical fusion has a limited impact once adaptive selection is applied. In contrast, the proposed ACT yields further consistent gains, with a substantial reduction in FNIR (down to 51.0). This demonstrates that beyond dynamic selection, robust score integration is crucial for handling challenging open-set search scenarios, making ACT more reliable and generalizable across diverse evaluation metrics.

Statistics of Dynamic Model Selection. Fig. 6 presents the frequency of model and anchor selections per dataset. On CCVID, where faces are often clearly visible, the agent frequently selects AdaFace as the anchor and consistently combines it with CAL, while BigGait is selected less often, suggesting limited complementary value for combination. In contrast, LTCC and MEVID—collected under surveillance conditions with multi-view and low-quality faces—lead the agent to rely on ReID models.

Dataset-level Model Selection. Since exhaustive sample-

Method	Rank1	mAP	TAR@1%FAR	FNIR@1%FPIR
QME [44]	73.8	39.6	35.0	64.3 ± 8.0
Z-score	74.8	**41.7**	37.1	63.7 ± 9.5
Farsight	74.8	**41.7**	**37.2**	62.5 ± 9.7
ACT (Ours)	**75.5**	41.4	36.5	$\mathbf{51.0 \pm 9.4}$

Table 5. **Comparison of score-fusion methods with agent selection on LTCC.** All fusion methods combined with agent-based selection outperform QME, confirming the value of dynamic model selection. The proposed ACT yields the best overall performance, particularly in open-set search.

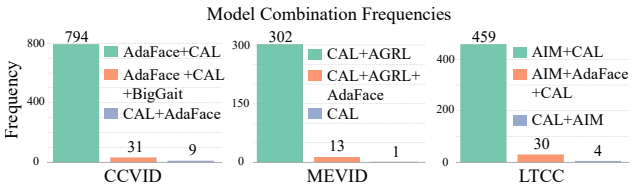

Figure 6. **Statistics of model selection.** The first model is the anchor model for each category. The combination distribution reveals the characteristics of each dataset and which model combination provides robust performance.

Comb.	Rank1	mAP	TAR@1%FAR	FNIR@1%FPIR
AIM+CAL	74.8	41.0	36.1	53.4 ± 9.2
FusionAgent	75.5	41.4	36.5	51.0 ± 9.4

Table 6. **Ablation on model combination and anchor model selection on LTCC.** The results show that FusionAgent closely aligns with the grid-searched results but achieves slightly better performance due to sample-level diversity.

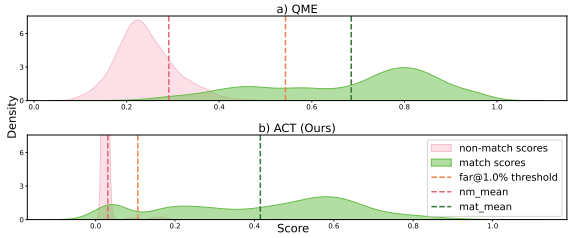

Figure 7. **Comparison of score distribution on CCVID.** [Keys: nm mean = mean value of non-match scores; mat mean = mean; value of match scores.]

level grid search is computationally prohibitive, we conduct grid search at the dataset level to determine the optimal model combination and anchor model. As shown in Tab. 6, AIM+CAL is the best model combination using ACT for the grid search, which is highly consistent with the combination chosen by our agent. Nevertheless, the agent achieves superior performance, attributed to its ability to exploit sample-level diversity during model selection.

4.3. Qualitative Results

Score Distribution. Fig. 7 plots the distribution of non-match (imposter) scores, match (genuine) scores, and the

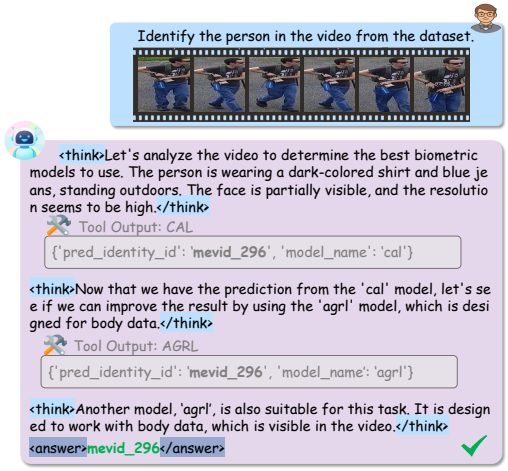

Figure 8. **CoT of FusionAgent on MEVID.** FusionAgent performs interpretable reasoning and dynamically selects the suitable tool combination for each test sample.

threshold for FAR=1% for QME and FusionAgent. For a better visualization, we align the y-axis and normalize the scores. Our method demonstrates a markedly larger margin between the match scores and the FAR threshold, while effectively suppressing non-match scores near zero. This improvement stems from two key mechanisms: (1) by summing only the top-k confident scores, we amplify the contribution of true matches and widen the separation from non-matches, which are predominantly excluded; and (2) our dynamic model selection provides a more robust set of models per query, leading to higher-quality scores for fusion.

CoT for Model Selection. Fig. 8 illustrates the CoT mode of FusionAgent, where the agent interprets the input and dynamically selects suitable models for each query sample. By leveraging the learned reliability and complementarity of different models, FusionAgent reasons over the query's characteristics to determine the optimal model combination. This CoT process enhances interpretability by explicitly revealing the agent's decision path and selection rationale.

5. Conclusion

We propose FusionAgent, a novel agentic framework for human recognition with dynamic model selection. We introduce an Anchor-based Confidence Top-k score-fusion method (ACT) for sample-dependent and adaptive model integration. Multiple reward functions are designed to guide the agent in tool-use and exploration across diverse model combinations. Extensive experiments and analyses validate the effectiveness of FusionAgent and ACT, highlighting the benefit of query-wise model selection and fusion. Our approach provides a scalable and extensible solution for multi-modal and multi-model tasks, with potential applicability to broader scenarios such as vision-language retrieval.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint arXiv:2503.10200*, 2025. 3

[3] Mohamed Cheniti, Zahid Akhtar, Chandranath Adak, and Kamran Siddique. An approach for full reinforcement-based biometric score fusion. *IEEE Access*, 2024. 2

[4] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *WACV*, 2023. 5

[5] Maria De Marsico, Michele Nappi, and Daniel Riccio. Cabala—collaborative architectures based on biometric adaptable layers and activities. *PR*, 2012. 2

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 1, 2, 5, 6

[8] Mingxing He, Shi-Jinn Horng, Pingzhi Fan, Ray-Shine Run, Rong-Jian Chen, Jui-Lin Lai, Muhammad Khurram Khan, and Kevin Octavius Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *PR*, 2010. 2, 5

[9] Abderrahmane Herbadji, Zahid Akhtar, Kamran Siddique, Noubeil Guermat, Lahcene Ziet, Mohamed Cheniti, and Khan Muhammad. Combining multiple biometric traits using asymmetric aggregation operators for improved person recognition. *Symmetry*. 5, 6

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 5

[11] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *PR*, 2005. 1, 2, 5, 6

[12] Anil Jain, Ruud Bolle, and Sharath Pankanti. Introduction to biometrics. In *Biometrics: personal identification in networked society*. Springer, 2011. 2

[13] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022. 1, 2, 6

[14] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. KeyPoint Relative Position Encoding for Face Recognition. In *CVPR*, 2024. 1, 2

[15] Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *CVPR*, 2025. 5, 6

[16] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, et al. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *WACV*, 2024. 1, 2, 5, 6

[17] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation. In *CVPR*, 2024. 1, 2, 6

[18] Feng Liu, Nicholas Chimitt, Lanqing Guo, Jitesh Jain, Aditya Kane, Minchul Kim, Wes Robbins, Yiyang Su, Dingqiang Ye, Xingguang Zhang, et al. Person recognition at altitude and range: Fusion of face, body shape and gait. *arXiv preprint arXiv:2505.04616*, 2025. 2

[19] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3, 4

[20] Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025. 3, 4

[21] Karthik Nandakumar, Yi Chen, Sarat C Dass, and Anil Jain. Likelihood ratio-based biometric score fusion. *TPAMI*, 2007. 2

[22] Tae Jin Park, Manoj Kumar, and Shrikanth Narayanan. Multi-scale speaker diarization with neural affinity score fusion. In *ICASSP*, 2021. 2, 5, 6

[23] Norman Poh and Josef Kittler. A unified framework for biometric expert fusion incorporating quality measures. *TPAMI*, 2011. 2

[24] Norman Poh, Josef Kittler, and Thirimachos Bourlai. Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization. In *BTAS*, 2007. 2

[25] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 5

[26] Arun Ross and Anil Jain. Information fusion in biometrics. *PR letters*, 2003. 2

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3, 4, 5

[29] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 3

[30] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*, 2023. 3

[31] Maneet Singh, Richa Singh, and Arun Ross. A comprehensive overview of biometric fusion. *Information Fusion*, 52, 2019. 2

[32] Robert Snelick, Mike Indovina, James Yen, and Alan Mink. Multimodal biometrics: issues in design and testing. In *ICMI*, 2003. 1, 5, 6

[33] Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Open-set biometrics: Beyond good closed-set models. In *ECCV*, 2024. 5

[34] Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid. In *ICCV*, 2025. 1

[35] Jackson Horlick Teng, Thian Song Ong, Tee Connie, Kalaiarasi Sonai Muthu Anbananthen, and Pa Pa Min. Optimized score level fusion for multi-instance finger vein recognition. *Algorithms*, 2022. 2, 5, 6

[36] Mayank Vatsa, Richa Singh, and Afzel Noore. Integrating image quality in 2ν-svm biometric match score fusion. *International Journal of Neural Systems*, 2007. 2

[37] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *TIP*, 2020. 2, 6

[38] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, 2023. 2, 6

[39] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023. 3, 4

[40] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. BigGait: Learning Gait Representation You Want by Large Vision Models. In *CVPR*, 2024. 1, 2, 6

[41] Mustafa Berkay Yılmaz and Berrin Yanıkoğlu. Score level fusion of classifiers in off-line signature verification. *Information Fusion*, 2016. 2

[42] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 3

[43] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 1, 2

[44] Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. A quality-guided mixture of score-fusion experts framework for human recognition. In *ICCV*, 2025. 1, 2, 5, 6, 8, 3

[45] Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V Wang, James Zou, et al. 4kagent: agentic any image to 4k super-resolution. *arXiv preprint arXiv:2507.07105*, 2025. 3

FusionAgent: A Multimodal Agent with Dynamic Model Selection for Human Recognition

Supplementary Material

6. Additional Methodology

6.1. Group Relative Policy Optimization (GRPO)

In contrast to RL algorithms such as Proximal Policy Optimization (PPO) [27]— which rely on a critic model to assess policy performance, GRPO eliminates the need for the critic model by directly comparing groups of candidate responses. As shown in Fig. 2, for a given input x, GRPO requires the model to sample N diverse responses $\{o_1, o_2, \ldots, o_N\}$ from the curren t model π_θ and obtains overall rewards $\{r_1, r_2, \ldots, r_N\}$ for o_i based on the reward function $R(x, o_i)$. In our case, it can be formatted as:

$$\begin{aligned} R(x, o_i) = w_f R_f(o_i) + w_{tool} R_{tool}(o_i) \\ + w_{acc} R_{acc}(a_i, y) + w_{mat} R_{mat}(o_i), \end{aligned} \tag{7}$$

where w_f, w_{tool}, w_{acc}, and w_{mat} are the reward weights for format reward R_f, tool success reward R_{tool}, answer accuracy reward R_{acc}, and metric-based reward R_{mat}, respectively. GRPO assesses the relative quality by normalizing r_i using the mean and standard deviation of the group reward:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}, \tag{8}$$

where A_i denotes the advantage of the i-th response. With the group normalization, GRPO encourages the model to sample preferred answers with a higher reward. The model is updated via:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)} \\ \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \right. \right. \\ \left. \text{clip} \left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \right) A_i \right) \\ \left. - \beta D_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right) \right], \end{aligned} \tag{9}$$

where ε and β are the GRPO clipping hyperparameters and the coefficient weight for controlling the Kullback–Leibler (KL) penalty [27], respectively. π_{ref} is the reference model.

$$\frac{1}{1 + Euc(q, g)}. \tag{10}$$

Dataset	Type	#Subjects (Train/Test/Non-mated)	#Query	#Gallery
CCVID	Video	75 / 151 / 31	834	1074
MEVID	Video	104 / 54 / 11	316	1438
LTCC	Image	77 / 75 / 15	493	7050

Table 7. Statistics of the evaluation set of human recognition benchmarks. The number of query and gallery indicate the number of images/sequences for image/video datasets.

Prompts. We provide the system prompt used in the agent training and inference in Fig. 9. The tool schema is the function documentation with input and output formats and meaning. Model type dict is the modality type of each biometric model.

6.2. Tool Design.

Tool design should be in an efficient way for agent learning. If the number of tools or the number of parameters of tools is large, it will become more difficult for the agent to execute the tools. In our scenario, the goal of the agent is to call different biometric recognition models to extract features based on the input. Therefore, we design a universal tool that is suitable for every biometric model. The tool takes sequential images, the biometric model name as the input, and returns the similarity matrix and the predicted label of the images.

Tool Results. FusionAgent only receives the predicted identity label from the tool execution, while the score vectors and selected models are stored within the system. We do not return the similarity scores or confidence weights of the predicted identity, as doing so was found to prematurely halt the exploration of model combinations during training. This design choice is justified because a high confidence or similarity score from one model does not preclude further performance gains through fusion with other models, given that each model contributes independently.

Error Handling. When the tool calling fails, for example, calling a wrong model name or an invalid JSON format. It is important to let the agent know the reason for the failure during training. We design error handling for unexpected behaviors and enable the agent to identify the reason for the failures.

7. Additional Implementation Details

Datasets. The dataset statistics are summarized in Tab. 7. They comprise multi-view captures and cross-modal bio-

Judge Prompt

Role & Objective
You are an expert-level biometric analysis agent. Your primary mission is to achieve the highest possible identification performance by strategically analyzing an input images/videos and selecting the optimal combination of biometric models. Prioritize the model you think is the most suitable. Do not select the same model more than once. Your final answer should be a fused identity prediction based on the evidence from your chosen models.
Loop
Work step-by-step. Each turn you must output exactly TWO blocks—first <think>, then ONE action: <tool_call> or <answer>. Wait for <tool_result> before the next turn.
Strict Output Format (no extra text, no markdown)
1) <think>...</think><tool_call>{JSON}</tool_call>
2) <think>...</think><answer>...</answer>
Tag Rules
- <think> (required, first): Briefly describe what you get, and explain the current decision.
  • If calling a tool: you MUST first analyze the input video's characteristics. Consider factors like: Is the face clearly visible? Is the subject close to the camera with high resolution, or far away and low-resolution? etc.
  • If answering: summarize tools results, key evidence, and your final prediction.
- <tool_call>: JSON with exactly two keys, "name" and "parameters". You can call ONLY ONE tool per turn.
- <answer>: Identity: The ID of the recognized person.
Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags: <tools> {TOOL_SCHEMA} </tools>
For each function call, return a json format object with function name and arguments within <tool_call></tool_call> XML tags: <tool_call> {{"name": <function-name>, "parameters": <args-json-object>}} </tool_call>. Only call declared tools.
Model Type
You have access to a suite of specialized models. Your key challenge is to understand when to use them for maximum impact: {MODEL_TYPE_DICT}
Stopping Condition
End with <answer> when evidence is sufficient. Never invent tool outputs or identities.

Figure 9. System prompt for FusionAgent.

metric data, enabling rigorous evaluation of generalization across diverse resolutions, viewpoints, and temporal dynamics. This comprehensive benchmarking setup ensures robustness against real-world challenges such as occlusion, motion blur, and sensor heterogeneity, thereby validating the practicality of the proposed approach in unconstrained environments. During training, we only use $2,000$ samples (or medias) maximum for each dataset to perform training set score matrices.

Model Pools. We follow [44] to construct the same model pools: AdaFace [13] (ViT-Base, WebFace4M), CAL [7] (ResNet50, CCVID/MEVID/LTCC), BigGait [40] (DINOv2-Small, CCPG), AIM [38] (ResNet50, LTCC), and AGRL [37] (ResNet50, MEVID), where the former denotes the model architecture and the latter indicates the training dataset. We use $\{model\}\_\{dataset\}$ to denote the difference of checkpoints (*i.e.*, CAL_CCVID and CAL_LTCC) during training and inference.

Center Features of Training Set. We follow QME [44] to extract center features as the gallery features for the training set. We compute the center features based on the subject ID, camera ID, and clothing ID. Therefore, each subject may have multiple center features.

Similarity Distances of Each Model. We follow QME [44] to measure the distances between features. AdaFace [13], CAL [7], AIM [38], AGRL [37], and CLIP3DReID [17] use cosine-similarity to measure the distance, while BigGait [40] uses Euclidean distance. We use Eq. 10 to transform Euclidean distance into similarity scores.

Cross-domain Training. We conduct cross-domain evaluation through zero-shot testing and few-shot training. For the few-shot setting, we adopt a 10-shot protocol (*i.e.*, each training subject provides 10 images/videos). The same procedure is applied to extract center features and score matrices from the training set, and only the 10-shot dataset is used for FusionAgent training. Few-shot data size for each

Dataset	#Media (Full)	#Media (10-shot)	Percentage (%)
LTCC	9576	768	8.0

Table 8. Comparison of the full dataset size and the few-shot (10-shot) subset used for training. The percentage indicates the proportion of data used in the few-shot setting relative to the full dataset.

dataset is shown in Tab. 8

Additional In-Domain Evaluation.

Accuracy From Agent Answer. As shown in Table 9, the *Agent* predictions are derived by selecting identity labels from the outputs of different tools, rather than directly relying on the score fusion results. Despite this discrete decision process, the agent achieves performance comparable to the score-based fusion method (ACT). We attribute this performance improvement primarily to the use of the answer accuracy reward. However, ACT consistently yields higher accuracy across all datasets, demonstrating that score-level fusion effectively integrates complementary information from multiple models and provides more reliable identity estimation than selection based solely on predicted labels. However, since the agent's answer can only reflect label prediction accuracy—and not other metrics such as ranking quality or calibration—different evaluation criteria may be adopted depending on application requirements. This result also provides a future direction on whether MLLMs or agents can even have a better performance w.r.t. metric results like verification (TAR) and open-set search (FNIR).

However, we do not observe the performance gain on CCVID. We hypothesize this is due to the performance gap between training set and test set. CAL is trained on CCVID, but AdaFace is not. CAL has a better Rank1 result on the training set, which make the agent more reliant on the decision on CAL.

Time-consuming. Since the agent only decides the model combination, the task itself is relatively simple. Considering the demand for faster responses in practical applications, we adopt a lightweight 3B model for efficiency. As shown in Tab. 10, the CoT inference of FusionAgent, including tool executions and score-fusion, takes 2.81s per sample on a H100 device, while Direct Answering takes 1.03s. For comparison, QME [44] takes 0.67s per sample on average, including tool executions, quality estimating, and score-fusion.

7.1. Additional Ablation Experiments

Confidence Weights. Tab. 11 ablates the effect of confidence weights in ACT. Compared to Z-score, Min–max normalization underperforms by 0.9% points in mAP and 0.4% in Rank-1, with an even larger drop observed in FNIR (9.8%). Omitting confidence weighting leads to the largest FNIR degradation (62.4), confirming that confidence-aware scaling is essential in open-set search. Among the strategies, Z-score consistently yields the best overall results, reducing FNIR by more than 10% over other variants. This advantage likely stems from its robustness to outliers, which allows for more stable calibration across heterogeneous models. These findings indicate that while confidence weighting has a limited effect on closed-set metrics (Rank-1, TAR), it plays a critical role in improving reliability under stricter false-positive constraints.

Method	CCVID	MEVID	LTCC
ACT	93.4	79.4	98.9
Agent	81.8	76.9	96.8

Table 9. Answer accuracy (Rank 1) performance on CCVID, MEVID, and LTCC. ACT is the Rank 1 result evaluated from the score matrix. Agent is the accuracy evaluated from the agent responses.

Method	*Time/sample (s)*			
	CCVID	**MEVID**	**LTCC**	**Avg**
QME [44]	0.72	0.66	0.64	0.67
FusionAgent (DA)	1.20	0.98	0.91	1.03
FusionAgent (CoT)	2.80	2.43	3.19	2.81

Table 10. Time-consuming Comparison of FusionAgent on Different Datasets. [Keys: DA=Direct answering.]

Norm.	Rank1	mAP	TAR@1%FAR	FNIR@1%FPIR
None	75.0	40.8	36.5	62.4 ± 9.2
Min-max	75.1	40.5	36.6	60.8 ± 10.7
Z-score	75.5	41.4	36.5	51.0 ± 9.4

Table 11. **Ablation on confidence weighting strategies in ACT on LTCC.** Both Min-max and Z-score normalization improve over no weighting, with Z-score achieving the best overall performance and substantially reducing FNIR. [Keys: Norm.= the method for confidence weights.]

Effects of Top-k Selection. As shown in Table 12, the effect of Top-k selection is consistent across all three datasets. The overall score is computed as the sum of Rank-1, mAP, and TAR, minus FNIR. On CCVID, applying Top-k selection brings clear improvements in Rank-1 accuracy (92.3→93.4), TAR (83.3→85.8), and the overall score (64.6→65.5), while maintaining comparable FNIR. For MEVID, the performance remains relatively stable, with a slight increase in Rank-1 (54.1→54.7) but minor fluctuations in other metrics. Similarly, on LTCC, Top-k selection provides marginal gains in Rank-1 (75.1→75.5) and overall score (25.5→25.6), with negligible changes in TAR and FNIR. Overall, Top-k selection consistently achieves a

Top-k	Rank1	mAP	TAR	FNIR	Overall
CCVID					
✗	92.3	92.5	83.3	9.7 ± 1.3	64.6
✓	93.4	92.7	85.8	9.9 ± 1.5	65.5
MEVID					
✗	54.1	29.1	35.6	58.2 ± 8.5	15.2
✓	54.7	28.7	34.9	58.6 ± 7.4	14.9
LTCC					
✗	75.1	41.8	37.2	51.9 ± 9.4	25.5
✓	75.5	41.4	36.5	51.0 ± 9.4	25.6

Table 12. **Effects of Top-k selection.** Top-k selection gains a better overall performance on three datasets.

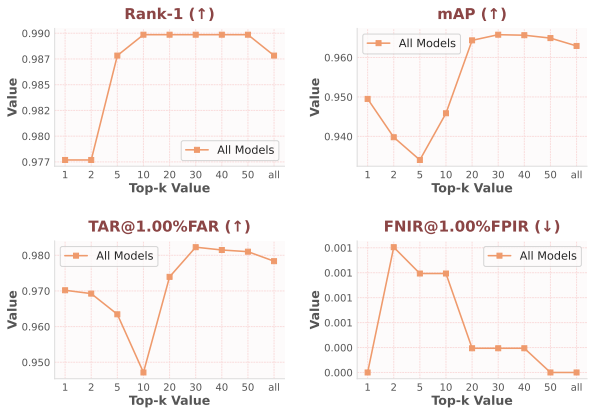

Figure 10. **Effects of Top-k Values on LTCC Training Set.** The overall performance reaches the peak when $k = 40$.

more favorable trade-off and leads to a better aggregated performance across datasets, confirming its robustness and general effectiveness.

Effects of Top-k Values on Training Set. We visualize the performance comparison on the LTCC training set in Fig. 10. The results show that the overall performance gradually improves as k increases and reaches its peak at $k = 40$, after which further growth of k does not yield additional benefits. Therefore, we adopt $k = 40$ for the testing stage to achieve a good balance between effectiveness and stability. We follow the same strategy when selecting Top-k values for the other datasets as well.

7.2. Additional Qualitative Results

Additional Conversations. Figures 11 and 12 present additional CoT examples from FusionAgent. Depending on the input query pattern, FusionAgent dynamically selects the most suitable models. Figure 13 illustrates the DA (direct answering) mode of FusionAgent, which bypasses reasoning generation to accelerate inference while maintaining comparable performance.

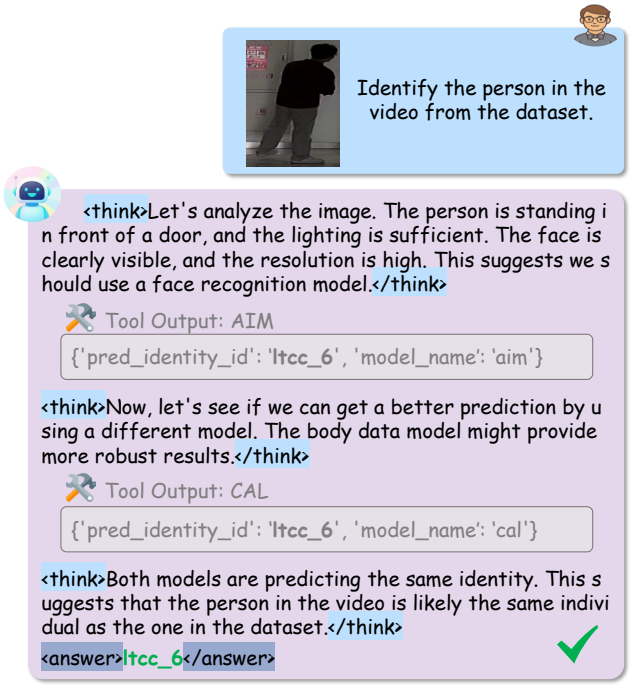

Figure 11. **CoT example of FusionAgent on LTCC.**

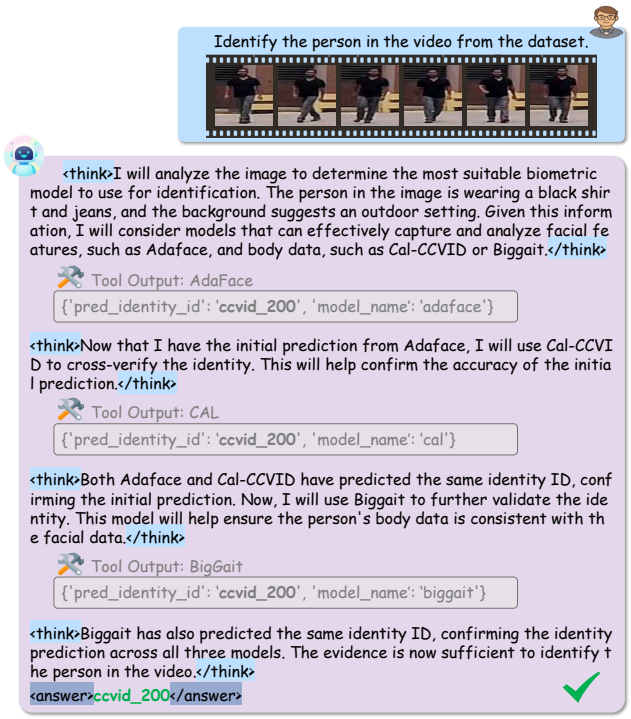

Figure 12. **CoT example of FusionAgent on CCVID.**

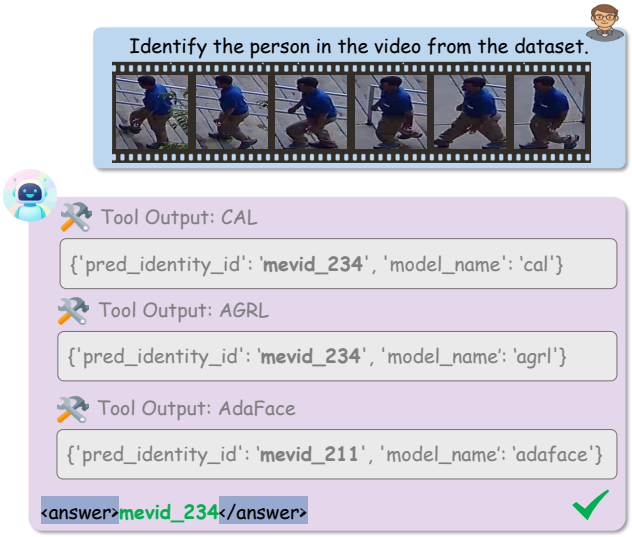

Figure 13. **DA example of FusionAgent on LTCC. FusionAgent directly outputs the tool-use code and answer without thinking.**

8. Limitations

Reasoning Collapse. RFT may lead to unstable or degenerate reasoning behaviors during training. For instance, the agent's reasoning content can become repetitive, disregarding the actual differences in model-predicted identities as training progresses. This phenomenon likely arises from *reward hacking*, since no explicit supervision is provided on the quality of reasoning. Stabilizing the reasoning process and ensuring consistent answer quality remains an important direction for future research.

Model Combination Estimation. In our setting, each sample can be associated with multiple possible model combinations. As the number of samples increases, the search space grows exponentially, making grid search for the ground-truth optimal combination computationally infeasible. Consequently, we adopt the proposed metric-based reward to estimate the overall performance without exhaustively enumerating all combinations. In the future, exploring more efficient or learning-based strategies for model combination estimation could further enhance scalability and accuracy.

9. Potential Societal Impacts

Our paper leverages multiple public biometric datasets for research purposes, with a focus on the similarity score domain, which is less directly tied to sensitive biometric data. As biometric recognition tasks grow increasingly complex, integrating multiple models has become a key trend to enhance system performance. It is essential to ensure that the use of biometric datasets and recognition systems adheres to ethical standards and complies with privacy regulations.