IDIOFUSE: IDENTITY-CONDITIONED SCORE FUSION FOR RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust person re-identification often combines complementary cues such as face, gait, and body shape. While adaptive fusion typically targets query quality, model strength also varies across identities. We introduce IdioFuse, an identity-conditioned score fusion framework that tailors weights to each gallery identity without training. By contrasting intra-identity consistency against cross-identity impostors, it extracts identity-specific profiles that couple with query-conditioned adaptation via a parameter-free rule. This widens the separation between true and false matches while preserving score calibration. Evaluations on three clothes-changing person re-identification benchmarks show that IdioFuse consistently outperforms statistical, rank-based, and learned baselines, achieving up to an 8.8% absolute reduction in false non-identification rate and demonstrating the value of identity-conditioned fusion in open-set person re-identification.

1 INTRODUCTION

Person re-identification (ReID) benefits from models that capture complementary biometric cues, including face, gait, and body shape (Liu et al., 2024a; 2025a). Their usefulness varies with the observations: face cues become less reliable at low resolution or oblique angles, while gait requires a suitable video sequence. Combining these models therefore calls for a fusion rule that determines how much each should contribute to a match.

Recent fusion methods address variation in cue availability by adapting to the current inputs (Gupta et al., 2020; Zhu et al., 2025; Su et al., 2025; Zhu et al., 2026). Quality estimates, learned gates, and model-selection policies adjust which models contribute and how their evidence is combined. Such adaptation can reduce reliance on a face model when the face is poorly visible or on another model whose required cues are degraded. It accounts for changes in the observations, but a usable cue need not be equally discriminative for every enrolled person.

A model's contribution also depends on how well it distinguishes the identity being matched from other people in the gallery. Similar embeddings across that identity's samples are useful only if the model also separates them from samples of other identities. A model that clearly separates one person from the rest of the gallery may provide little separation for another. Assigning the same relative model weights to every gallery identity cannot express this difference. Classical user-dependent biometric fusion has investigated such variation across individuals (Snelick et al., 2005; Fierrez-Aguilar et al., 2005; Poh et al., 2013). The labeled gallery provides evidence for estimating these differences through exemplars of each person and comparisons with other enrolled identities.

We introduce IdioFuse, an identity-conditioned score-fusion method that uses these gallery comparisons to estimate a model-weight profile for each enrolled identity. For each model, similarities within an identity measure consistency across its reference samples, while similarities to all other gallery identities indicate how readily other people could be confused with it. The separation between these two groups determines the model's weight. At query time, matching scores computed from the models' embeddings are combined using the profile of the gallery identity being considered. The same query can therefore be compared with different identities using different model weights, as illustrated in Figure 1. The profiles are reused across queries without updating the recognition models or training an additional fusion network. For open-set decisions, calibration accounts for differences in non-match scores across identities. Composition with query-dependent fusion allows the weights to be used alongside adaptation to the current observation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

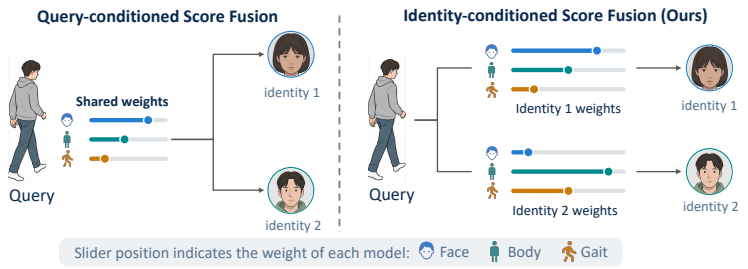


Figure 1: **Query-conditioned vs. Identity-conditioned score fusion.** Existing multi-modal gait recognition often applies a global, shared weight to all identities (left). In contrast, our proposed Identity-conditioned approach adaptively tailors weights for each identity profile (right), allowing the system to leverage the most reliable modality on an identity-specific basis. Sliders conceptually illustrate the weight assignment.

Experiments on three person re-identification benchmarks demonstrate gains over competing fusion methods using the same underlying models, both with identity-conditioned fusion alone and in composition with query-dependent methods. Further analysis shows that different identities benefit from different models, and that gallery comparisons help identify these preferences. We also study how these gains depend on the exemplar counts used to estimate the weights. Our contributions are:

- We propose an identity-conditioned score fusion method for open-set person ReID that dynamically tailors model weights to enrolled gallery profiles without training.
- We show that IdioFuse seamlessly unifies with query-conditioned fusion via a parameter-free combination rule, enabling them to jointly leverage complementary cues.
- Extensive evaluations on three benchmarks demonstrate that IdioFuse consistently outperforms statistical, rank-based, and learned fusion baselines, effectively widening the genuine–impostor separation margin.

2 Related Work

Open-set person re-identification. Open-set person ReID requires matching queries to enrolled identities and rejecting queries whose identities are absent from the gallery (Liao et al., 2014; Zheng et al., 2016; Zhu et al., 2025). Robust matching across appearance changes has motivated representations that reduce clothing dependence (Yang et al., 2021; Gu et al., 2022; Yang et al., 2023; Liang & Rawat, 2025; Pang et al., 2025; Pathak & Rawat, 2025; Kim et al., 2025) and exploit body shape or gait (Liu et al., 2023; 2024b; Li et al., 2023b; Ye et al., 2024; Zhu et al., 2024; Huang et al., 2026). Reliable rejection also requires separating true matches from high-scoring non-matches (Su et al., 2024; 2026). Open-set biometric studies show that good closed-set performance does not necessarily yield reliable rejection and introduce training losses that explicitly account for impostor scores as well as correct identity matches (Su et al., 2024; Zhu et al., 2025).

Adaptive multimodal fusion. Combining face, body shape, and gait can compensate for unreliable observations in individual modalities (Liu et al., 2024a; 2025a). Score-level fusion combines outputs from independently trained models (Ross & Jain, 2003; Nandakumar et al., 2008; Zhu et al., 2025), with normalization accounting for differences in score scale (Snelick et al., 2003; Jain et al., 2005; He et al., 2010; Zhu et al., 2026). Quality-dependent fusion further adjusts the contribution of each modality according to the reliability of the observed input (Nandakumar et al., 2006; Poh et al., 2009; Poh & Kittler, 2012; Gupta et al., 2020; Cao et al., 2024; Cho et al., 2025; Liu et al., 2025b). QME (Zhu et al., 2025) combines learned fusion strategies according to estimated query quality, while FusionAgent (Zhu et al., 2026) selects models for each query before confidence-based score fusion. HAMoBE (Su et al., 2025) conditions feature fusion on both videos in a query–gallery pair, allowing the feature weights for a query to vary with the gallery video being compared.

Identity-specific biometric fusion. Biometric matchers exhibit identity-dependent performance (Doddington et al., 1998; Yager & Dunstone, 2010), which inspired early verification studies

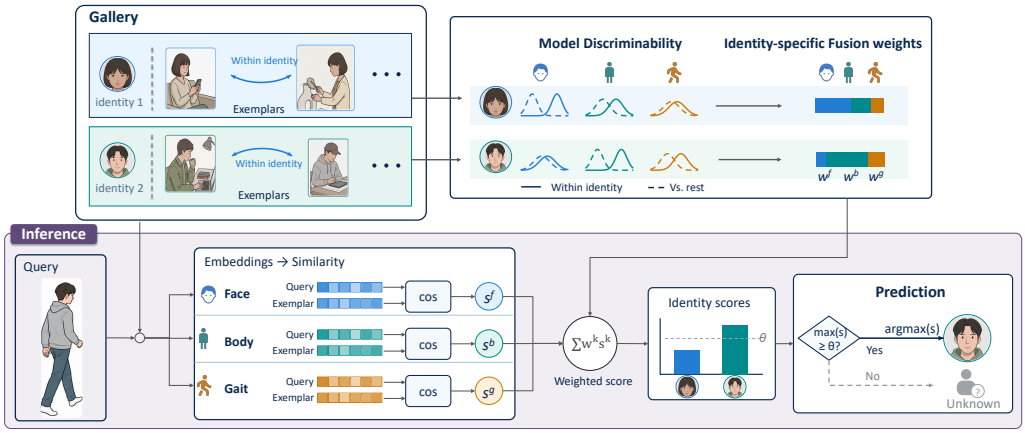


Figure 2: **Overview of IdioFuse framework.** *Top:* Our method leverages gallery intra- and inter-identity statistics to derive tailored model reliability weights w_k^i for each identity i without auxiliary training. *Bottom:* When an incoming query is compared against the gallery, its similarity scores with candidate identity i across modalities are combined using that candidate's specific profile, prioritizing reliable modalities for each individual for final open-set verification.

to weight classifiers using client-specific separation statistics (Snelick et al., 2005; Poh et al., 2013). However, these proportional weighting rules rely on an assumption that separation measured at enrollment remains valid at test time. In person Re-ID, camera and viewpoint changes frequently break this assumption, making rigid proportional weights overconfident. IdioFuse instead formulates a regularized weighting strategy that mitigates these transfer errors while integrating model consensus directly into the identity scoring pipeline.

3 Proposed Method

Problem Setup Open-set biometric identification matches a query q (an image or video) against a gallery $\mathcal{G}$ of N enrolled identities, assigning it to one of them or rejecting it as unenrolled. Each identity $i \in \mathcal{I} = \{1, \ldots, N\}$ is represented by a set of gallery exemplars $\mathcal{E}_i$, forming the gallery $\mathcal{G} = \bigcup_{i \in \mathcal{I}} \mathcal{E}_i$. Consider K fixed, pretrained models $\mathcal{M} = \{M_1, \ldots, M_K\}$, each mapping visual inputs to embeddings that capture biometric cues such as face or gait. A matching function, such as cosine similarity, compares the embeddings produced by model $M_k \in \mathcal{M}$ for query q and gallery exemplar $e \in \mathcal{G}$, yielding a matching score $s_k(M_k(q), M_k(e))$. For each candidate identity i, a fusion operator Φ_i aggregates these cross-model, multi-exemplar scores into an overall confidence score $S_{q,i} \in \mathbb{R}$ as

$$S_{q,i} = \Phi_i\Big(\big\{s_k(M_k(q), M_k(e)) \;\big|\; M_k \in \mathcal{M},\ e \in \mathcal{E}_i\big\}\Big). \tag{1}$$

Open-set identification assigns an identity only when an acceptance criterion is met; otherwise, the query is rejected as unenrolled. A common rule accepts the highest-scoring identity $\arg\max_{i \in \mathcal{I}} S_{q,i}$ if its score exceed a threshold. Closed-set one selects the highest-scoring identity without rejection.

As illustrated in Figure 2, we first estimate a mixture profile for each gallery identity from its gallery comparisons in Section 3.1). These profiles determine the score-fusion rule and its open-set decisions Section 3.2. We then describe their composition with query-dependent fusion in Section 3.3.

3.1 Gallery-derived Mixture Profiles

A biometric model may distinguish some enrolled identities more reliably than others. To determine a model's contribution for a candidate identity i, we quantify how effectively it separates identity i from the rest of the gallery. Specifically, we evaluate two factors using the reference observations in the gallery: intra-identity similarity, which reflects consistency across observations of the same identity, and inter-identity similarity, which measures distinctiveness against others. An informative

model should exhibit high intra-identity similarity and low inter-identity similarity, whereas a model assigning high scores to both groups offers little discriminative power. Intuitively, this separation margin reflects a model's reliability and governs its contribution.

Estimating identity-specific separation. To estimate the separation for identity i under model k, we compute two sets of cosine similarities: the *genuine* scores between distinct samples in its exemplars set $\mathcal{E}_i$ (capturing consistency), and the *impostor* scores comparing samples in $\mathcal{E}_i$ with $\mathcal{G}\backslash\mathcal{E}_i$ (capturing distinctiveness). Both sets are conditioned on the target identity. The difference between the genuine and impostor means measures their average separation, but the variation within each group determines how informative that difference is. Following separation-based user weighting (Snelick et al., 2005), we quantify this margin using the signal-detection sensitivity index d' (Green & Swets, 1966), which scales the difference in means by their pooled standard deviation. Letting $(\mu_{i,k}^{\text{gen}}, (\sigma_{i,k}^{\text{gen}})^2)$ and $(\mu_{i,k}^{\text{imp}}, (\sigma_{i,k}^{\text{imp}})^2)$ denote the empirical mean and variance of the genuine and impostor scores, respectively, the separation achieved by model k is defined as

$$d'_{i,k} = \frac{\mu_{i,k}^{\text{gen}} - \mu_{i,k}^{\text{imp}}}{\sqrt{\frac{1}{2}\left((\sigma_{i,k}^{\text{gen}})^2 + (\sigma_{i,k}^{\text{imp}})^2\right)}}. \tag{2}$$

Higher values of $d'_{i,k}$ reflect greater discriminability for identity i within the gallery. By evaluating models independently, this metric avoids the instability of estimating cross-model covariance matrices under scarce gallery samples, while preserving modularity when scaling to larger ensembles.

Given the resulting separation scores across all K models, $\{d'_{i,k}\}_{k=1}^{K}$, a straightforward approach would be to map them into normalized weights via softmax, favoring models with higher discriminability. However, when gallery observations are limited, empirical separation estimates carry noticeable variance, and a standard softmax can yield an overly peaked distribution that heavily overcommits to a single dominant model. To hedge against this uncertainty, we regularize the allocation by interpolating the softmax distribution with a uniform prior across all K models, forming identity i's mixture profile $\mathbf{w}^i = [w_1^i, \ldots, w_K^i]^\top$ with each component defined as

$$w_k^i = \frac{1-\alpha}{K} + \alpha \frac{\exp(d'_{i,k}/\tau)}{\sum_{k'=1}^{K} \exp(d'_{i,k'}/\tau)}\ , \tag{3}$$

where $\tau > 0$ controls the sharpness of the separation preference, and $\alpha \in [0, 1]$ balances data-driven weighting against uniform model averaging. This formulation prevents any single model from being prematurely discounted. Importantly, this mixture profile is computed directly from static gallery comparisons in closed form, requiring no additional training of a fusion network.

3.2 Identity-conditioned Score Fusion

With the identity-specific mixture profile $\mathbf{w}^i$ established, we fuse multi-model predictions for an incoming query q against candidate identity i. Directly taking a weighted sum of raw cosine scores, however, is prone to distortion because individual models inherently operate across disparate score distributions and dynamic ranges; a model with numerically dominant scores would override the calibrated mixture weights. To place all models on a shared scale, we map each raw similarity $s_k(M_k(q), M_k(e))$ into a standardized score as

$$z_k(q, e) = \frac{s_k(M_k(q), M_k(e)) - \bar{s}_k}{\rho_k}, \tag{4}$$

where $\bar{s}_k$ and ρ_k represent the empirical score mean and stabilized standard deviation for model k across the evaluation set. These normalization parameters are identity-agnostic and computed once per model without query labels.

Next, we combine these standardized scores to evaluate each candidate identity. Following standard practice in exemplar-based person Re-ID, we compute an ensemble consensus for each reference sample before aggregating across the identity's exemplar set $\mathcal{E}_i$, defining the uncalibrated fused score $F(q, i)$ between query q and candidate identity i as

$$F(q, i) = \max_{e \in \mathcal{E}_i} \sum_{k=1}^{K} w_k^i, z_k(q, e)\ . \tag{5}$$

Under this formulation, multi-model predictions are fused per exemplar prior to max-pooling over $\mathcal{E}_i$, ensuring that the aggregate score reflects the strongest reference match without being diluted by less informative observations.

While this fused score $F(q, i)$ provides an effective ranking criterion for closed-set retrieval, open-set person Re-ID requires deciding whether to accept the top candidate or reject the query as an unseen identity. This decision relies on a shared threshold; however, uncalibrated scores are susceptible to identity-specific baseline shifts, where certain candidate identities systematically elicit higher similarity from non-matching inputs. Using a uniform threshold directly on $F(q, i)$ would disproportionately produce false acceptances for these intrinsically high-scoring identities.

Open-set calibration. To align candidate scores onto a commensurable scale, we adapt score normalization principles from biometric verification (Auckenthaler et al., 2000) to calibrate each identity against its own non-match distribution. Treating each non-matching gallery sample as a pseudo-query yields an empirical impostor score set $\{F(e, i) \mid e \in \mathcal{G} \backslash \mathcal{E}_i\}$. As established in extreme value analysis for open-set recognition (Scheirer et al., 2013; 2014; Bendale & Boult, 2015), false acceptances are governed by extreme non-match outliers rather than central impostor tendencies. We therefore isolate the most competitive impostors by estimating candidate-specific tail statistics (m_i, s_i), defined as the empirical mean and standard deviation of impostor scores exceeding the 90th percentile. Because tail parameters estimated from finite gallery observations can exhibit high sample variance, we regularize them via empirical Bayes shrinkage, pulling the local estimates toward the pooled impostor statistics (m_0, s_0) across the full gallery to obtain the calibrated mean $\hat{m}_i$ and standard deviation $\hat{s}_i$ as

$$\hat{m}_i = \frac{n_i m_i + n_0 m_0}{n_i + n_0} \qquad \text{and} \qquad \hat{s}_i = \frac{n_i s_i + n_0 s_0}{n_i + n_0} \,, \tag{6}$$

where n_i denotes the number of valid impostor comparisons for identity i, and the reference count n_0 is the median of positive comparison counts across all enrolled identities. Using these regularized statistics, the calibrated score is computed via standard z-normalization as

$$\widetilde{F}(q, i) = \frac{F(q, i) - \hat{m}_i}{\hat{s}_i} \,. \tag{7}$$

During open-set person re-identification, the calibrated score serves as the final identification score, *i.e.*, $S_{q,i} = \widetilde{F}(q, i)$. A query q is assigned to the top candidate identity $\hat{i} = \arg\max_i S_{q,i}$ whenever its score $S_{q,\hat{i}}$ exceeds a shared decision threshold θ, and is rejected as an unseen identity otherwise.

3.3 Composition with Query-conditioned Fusion

The identity-conditioned weights $\mathbf{w}^i$ are derived entirely from the given gallery $\mathcal{G}$, characterizing which models best isolate candidate i from the remaining enrolled identities. While this profile captures intrinsic identity discriminability, it remains agnostic to the specific observation conditions of an incoming query q at inference time. Even if a model generally provides high discriminability for an identity, its reliability drops if that specific modality is degraded in the incoming query. Query-conditioned fusion addresses this blind spot by assessing how trustworthy each model is on the observed input. Thus, identity-conditioned fusion and query-dependent fusion rely on fundamentally different signals: gallery-side separability identifies which features make a person distinctive, while query-side gating assesses which features are clean enough to trust. Fusing these predictions lets the final decision rule account for both candidate uniqueness and query condition.

Let w_k^q denote the dynamic weight assigned to modality k by an arbitrary query-conditioned fusion rule, such as QME (Zhu et al., 2025) or FusionAgent (Zhu et al., 2026). Rather than interpolating scores, we integrate query-level visual quality with our identity-level distinctiveness directly at the weight level via element-wise multiplication. For query q and candidate identity i, the composite weight for modality k is defined as

$$\tilde{w}_i^q k = \frac{w_k^i \cdot w_k^q}{\sum_{k'} w_{k'}^i \cdot w_{k'}^q} \,, \tag{8}$$

which acts as a joint gating mechanism: a modality contributes strongly only when it is both inherently discriminative for identity i and sufficiently reliable in query q. The composite candidate

Table 1: **Statistics of person re-identification benchmarks.** Query and gallery counts refer to video sequences for CCVID (Gu et al., 2022) and MEVID (Davila et al., 2023), and images for LTCC (Qian et al., 2020). Identity counts are reported as train/test. The mean and range of gallery exemplar counts are computed across gallery identities in the test split.

Dataset	Type	# Identities	# Queries	# Gallery exemplars		
		Train / Test		Total	Mean / identity	Range / identity
CCVID	Video	75 / 151	834	1,074	7.1	3–12
MEVID	Video	104 / 54	316	1,438	26.6	1–101
LTCC	Image	77 / 75	493	7,050	94.0	4–589

score is then computed as $F_{\text{comp}}(q, i) = \sum_k \tilde{w}_k^i s_k(q, i)$, which directly replaces $F(q, i)$ in the calibration pipeline in Equation 7. We evaluate identification performance both with and without this composition in Section 4.

4 EXPERIMENTS

Datasets and Evaluation Protocol. To rigorously validate the effectiveness and robustness of our proposed method, we evaluate it on three widely used clothes-changing person re-identification benchmarks: CCVID (Gu et al., 2022) and MEVID (Davila et al., 2023) contain video sequences, while LTCC (Qian et al., 2020) contains images; detailed statistics are summarized in Table 1. Following standard evaluation practices (Zhu et al., 2025), performance is measured using Rank-1 accuracy and mean Average Precision (mAP) for closed-set ranking, alongside two standard whole-body biometric metrics: TAR@1%FAR (true accept rate at a 1% false accept rate) for verification, and FNIR@1%FPIR ($\downarrow$) for open-set identification (Grother et al., 2019). Following Zhu et al. (2025; 2026), while Rank-1, mAP, and TAR are computed over the full gallery, FNIR is averaged across 10 runs where random 80% of gallery identities serve as mates and the remaining 20% as non-mates to reflect unconstrained open-set scenarios.

Recognition models. To leverage the complementary strengths of global appearance and localized biometric traits, we evaluate heterogeneous combinations of face, body, and gait recognition backbones (AdaFace (Kim et al., 2022), CAL (Gu et al., 2022), BigGait (Ye et al., 2024), AGRL (Wu et al., 2020), and InsightFace/ArcFace (Deng et al., 2019); specified in Tables 2 and 3). For body representation, CLIP-ReID backbones (Li et al., 2023a) are evaluated either zero-shot to test generalization or with LoRA adaptation (Hu et al., 2022) on target training splits. Each model yields an independent similarity score channel. All upstream recognition models remain fixed during fusion to isolate the contribution of our fusion strategy from backbone fine-tuning. Full model details are provided in Appendix Section A.1.

Baseline setup. To fairly assess performance across different levels of supervision, we benchmark against two broad paradigms using matched recognition backbones: (1) *training-free baselines* that require no fusion training set, including fixed rules (Min/Max (Jain et al., 2005), Z-score/Min-Max normalization (Snelick et al., 2003), FarSight (Liu et al., 2025a)) and score-distribution alignment (RHE (He et al., 2010)); and (2) *training-reliant methods* optimized on the training split, including parameter-fitted rules (Asym-AO (Herbadji et al., 2020), learned Weighted-sum (Park et al., 2021)), learning-based classifiers (BSSF (Teng et al., 2022)), and multi-stage frameworks (QME (Zhu et al., 2025), FusionAgent (Zhu et al., 2026)). See more details in the Appendix.

Implementation details. Notably, our method requires no separate training split for fusion. Specifically, Ours (vanilla) performs purely identity-conditioned fusion, calibrating weights directly from the target gallery without prior training. Furthermore, following Section 3.3, we compose our framework with query-conditioned mechanisms from QME (Zhu et al., 2025) and FusionAgent (Zhu et al., 2026), reported as Ours (QME) and Ours (FusionAgent) respectively in the results, to demonstrate its plug-and-play compatibility. Across all datasets and model pools, we consistently fix hyperparameters to $\alpha = 0.5$, $\tau = 0.5$, and $\lambda = 0.5$ in Equation 3 without dataset-specific tuning.

Table 2: **Performance on CCVID and MEVID.** *Comb.* specifies the model pool shared by each block (♦ AdaFace, ♠ CAL, ▲ BigGait on CCVID and AGRL on MEVID, ♣ InsightFace/ArcFace, and ⊛ LoRA-adapted CLIP-ReID); the top block reports individual-model results and the two below it the fusion methods. **Best** and second-best fusion results are marked within each pool; highlighted rows are ours. Values are in %; FNIR is mean ± standard deviation over ten enrollment trials. R-1: Rank-1 accuracy, TAR: TAR@1%FAR, FNIR: FNIR@1%FPIR.

(a) CCVID

Comb.	Method	R-1	mAP	TAR	FNIR (↓)
♦	AdaFace	94.0	87.9	75.7	13.0 ± 3.5
♠	CAL	81.4	74.7	66.3	52.8 ± 13.3
▲	BigGait	76.7	61.0	49.7	71.1 ± 6.1
♣	InsightFace	94.1	88.2	78.9	18.1 ± 3.1
♦♠▲	QME	94.1	90.8	76.2	12.3 ± 1.4
♦♠▲	FusionAgent	93.4	92.6	85.9	10.1 ± 1.5
♦♠▲	**Ours (vanilla)**	**95.3**	92.1	81.0	11.0 ± 1.6
♦♠▲	**Ours (QME)**	95.2	91.9	79.3	10.3 ± 1.5
♦♠▲	**Ours (FusionAgent)**	94.6	**93.7**	**87.0**	**9.4 ± 1.2**
♦♠▲♣	Min-Fusion	94.4	88.6	79.6	15.1 ± 2.6
♦♠▲♣	Max-Fusion	84.2	69.2	57.5	51.2 ± 8.3
♦♠▲♣	Z-score	93.5	90.8	80.9	13.2 ± 2.5
♦♠▲♣	Min-max	93.4	90.9	78.8	13.5 ± 2.5
♦♠▲♣	RHE	93.4	90.9	79.0	13.5 ± 2.3
♦♠▲♣	Asym-AO	93.3	90.7	78.3	13.5 ± 2.5
♦♠▲♣	Farsight	93.6	90.7	79.2	13.3 ± 2.0
♦♠▲♣	Weighted-sum	94.2	89.8	81.5	13.5 ± 2.1
♦♠▲♣	BSSF	93.5	90.6	82.8	15.0 ± 3.9
♦♠▲♣	QME	93.0	91.5	78.9	11.2 ± 1.2
♦♠▲♣	**Ours (vanilla)**	**95.4**	92.2	**86.0**	**9.2 ± 1.4**
♦♠▲♣	**Ours (QME)**	94.7	**92.7**	84.2	9.6 ± 1.3

(b) MEVID

Comb.	Method	R-1	mAP	TAR	FNIR (↓)
♦	AdaFace	25.0	8.1	5.4	98.8 ± 1.2
♠	CAL	52.5	27.1	34.7	67.8 ± 7.3
▲	AGRL	51.9	25.5	30.7	69.4 ± 8.9
⊛	CLIP-ReID	69.9	45.1	47.7	49.7 ± 4.3
♦♠▲	BSSF	54.4	28.7	33.4	63.8 ± 9.4
♦♠▲	QME	55.7	28.2	32.9	64.6 ± 8.2
♦♠▲	FusionAgent	54.7	28.7	34.9	58.6 ± 7.4
♦♠▲	**Ours (vanilla)**	55.4	31.0	36.2	66.4 ± 13.6
♦♠▲	**Ours (QME)**	54.7	31.0	35.5	59.0 ± 12.6
♦♠▲	**Ours (FusionAgent)**	**57.0**	**31.2**	**37.3**	**52.0 ± 15.5**
♦♠▲⊛	Min-Fusion	47.8	21.7	28.0	70.4 ± 8.0
♦♠▲⊛	Max-Fusion	68.4	43.9	45.9	55.8 ± 8.1
♦♠▲⊛	Z-score	66.8	41.1	44.9	52.1 ± 7.2
♦♠▲⊛	Min-max	66.1	40.3	42.6	54.8 ± 8.1
♦♠▲⊛	RHE	65.2	39.0	41.2	57.0 ± 8.7
♦♠▲⊛	Asym-AO	62.3	35.0	31.3	57.8 ± 8.0
♦♠▲⊛	Farsight	61.7	36.6	38.7	58.9 ± 9.1
♦♠▲⊛	Weighted-sum	65.5	40.4	45.2	52.6 ± 5.6
♦♠▲⊛	QME	65.5	42.1	46.3	51.2 ± 6.1
♦♠▲⊛	**Ours (vanilla)**	**70.9**	**45.6**	**48.2**	49.0 ± 18.5
♦♠▲⊛	**Ours (QME)**	67.7	44.2	47.6	**48.2 ± 16.8**

4.1 Results

Tables 2a, 2b, and 3 report recognition performance on CCVID (Gu et al., 2022), MEVID (Davila et al., 2023), and LTCC (Qian et al., 2020) across unimodal models, existing fusion baselines, and our framework. Overall, Ours (vanilla) consistently outperforms both strong single-modality baselines and supervised fusion schemes without using any training data. Specifically, on CCVID, our identity-conditioned-only fusion (vanilla) improves already strong backbones without additional training, pushing Rank-1 accuracy from 94.0% (AdaFace) to 95.3%. Furthermore, composing with query-conditioned mechanisms consistently boosts performance across metrics: on MEVID, Ours (FusionAgent) improves upon standalone FusionAgent across all metrics, lifting mAP from 28.7% to 31.2% and TAR from 34.9% to 37.3%. Also, our approach is particularly effective for open-set identification, reducing FNIR@1%FPIR on LTCC from 60.7% to 29.1% via combined identity weighting and calibration.

4.2 Ablation Studies

Effect of candidate score calibration. To isolate the benefit of candidate calibration from identity-specific weighting, we evaluate open-set identification with and without this step. As shown in Table 4, calibration consistently lowers FNIR@1%FPIR across all configurations, cutting the error rate by up to 8.1%. On the CCVID dataset, for instance, it reduces FNIR from 19.1% to 11.0% while sharply compressing the cross-trial variance from ±8.1 to ±1.6. This confirms that while identity weighting provides the primary discriminative power, score calibration further improves non-mate rejection by stabilizing global thresholding across gallery identities.

4.3 Analysis

Effectiveness of identity-specific weighting. To verify that the improvement of our method comes from tailoring the mixture profile to specific identities rather than just using non-uniform values, we

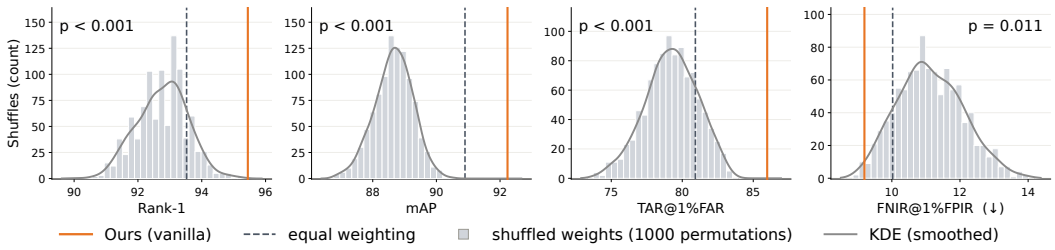


Figure 3: **Identity profile permutation on CCVID.** Expert pool: AdaFace, CAL, BigGait, and InsightFace. Gray histograms show performance distributions over 1,000 random identity-to-profile permutations while keeping models, gallery, and weight vectors identical. The genuine assignment (solid orange line) strictly outperforms all permutations across all three metrics, as well as equal weighting (dashed line), confirming that gains stem from identity-specific weights rather than arbitrary non-uniform weighting.

Table 3: **Performance on LTCC.** *Comb.* specifies the model pool shared by each block (♦ AdaFace (Kim et al., 2022), ♠ CAL (Gu et al., 2022), ■ AIM (Yang et al., 2023), and ⊛ LoRA-adapted CLIP-ReID (Li et al., 2023a; Hu et al., 2022)); the top section reports individual-model results and the section below it the fusion methods. **Best** and second-best fusion results are marked; highlighted rows are ours. Values are in %; FNIR is mean ± standard deviation over ten enrollment trials. TAR: TAR@1%FAR, FNIR: FNIR@1%FPIR.

Comb.	Method	Rank-1	mAP	TAR	FNIR (↓)
♦	AdaFace	18.5	5.9	2.4	99.8 ± 0.2
♠	CAL	74.4	40.6	36.7	59.7 ± 7.3
■	AIM	74.8	40.9	37.0	66.2 ± 7.5
⊛	CLIP-ReID (LoRA)	76.7	40.7	31.9	42.0 ± 10.3
♦♠■⊛	Min-Fusion	42.0	15.0	13.1	81.9 ± 6.0
	Max-Fusion	76.1	41.6	32.1	32.6 ± 11.0
	Z-score	79.1	47.8	40.3	48.3 ± 13.3
	Min-max	79.9	48.7	40.6	53.6 ± 15.3
	RHE	78.3	46.3	38.4	64.9 ± 14.7
	Asym-AO	76.3	38.8	32.2	63.0 ± 9.8
	Farsight	80.3	48.6	40.4	49.9 ± 14.6
	Weighted-sum	79.3	48.4	**41.9**	51.8 ± 9.1
	BSSF	79.9	49.3	**41.9**	50.9 ± 10.0
	QME	**80.7**	48.5	41.3	60.7 ± 8.5
	Ours (vanilla)	80.5	**49.6**	41.4	31.1 ± 19.0
	Ours (QME)	80.5	48.7	41.3	**29.1 ± 21.2**

Table 4: **Effectiveness of per-identity score calibration.** FNIR@1%FPIR (↓) of Ours (vanilla) before and after candidate calibration. Model symbols follow Tables 2 and 3.

Comb.	Before	After
	Dataset: CCVID	
♦♠▲	19.1 ± 8.1	11.0 ± 1.6
♦♠▲♣	15.3 ± 7.3	9.2 ± 1.4
	Dataset: MEVID	
♦♠▲	68.3 ± 14.3	66.4 ± 13.6
♦♠▲⊛	52.6 ± 20.6	49.0 ± 18.5
	Dataset: LTCC	
♦♠■	52.8 ± 12.5	52.0 ± 12.7
♦♠■⊛	37.3 ± 19.0	31.1 ± 19.0

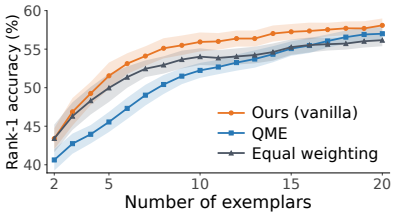
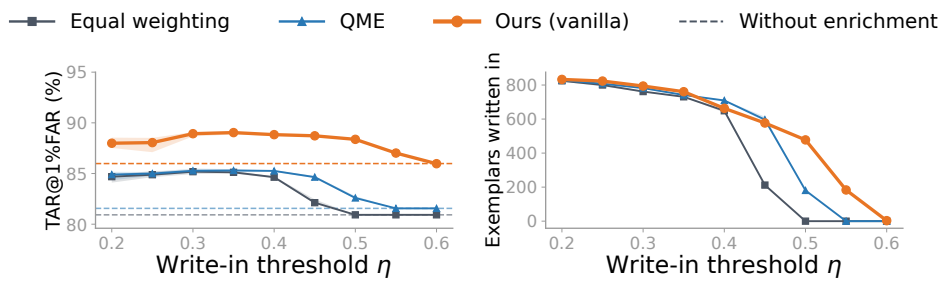


Figure 4: **Sensitivity to the number of exemplars.** Mean Rank-1 on MEVID (AdaFace, CAL, AGRL, InsightFace) over 30 random exemplar draws; shading shows ±1 SD. Ours (vanilla) leads at every exemplar count.

perform a permutation test on CCVID. Specifically, we randomly shuffle the weight-to-identity assignments over 1,000 independent runs so that each identity i receives the weight vector estimated for a different identity i', while keeping all models and gallery data identical. As shown in Figure 3, the matched assignment consistently outperforms both equal weighting and every random permutation across all three metrics. This confirms that the estimated weights encode meaningful, identity-specific priors rather than arbitrary non-uniform scaling.

Sensitivity to the number of exemplars per identity. Because identity-specific weights are estimated directly from gallery comparisons, we evaluate whether our method relies on having densely populated galleries. On a subset of MEVID (36 identities with sufficient samples, evaluated across 251 queries), we vary the number of enrolled exemplars per identity from 2 to 20 over 30 paired random draws. As shown in Figure 4, Ours (vanilla) consistently outperforms both equal weighting



(a) (b)

Figure 5: **Robustness to gallery enrichment.** Queries are added into the gallery under their predicted identity $\hat{i}$ when the calibrated match score satisfies $S_{q,\hat{i}} > \eta$. Solid curves and shaded regions denote the mean and spread across 9 arrival sequences, while dashed lines mark the static-gallery baselines. We report: (a) TAR@1%FAR and (b) number of admitted queries across enrollment thresholds η. Ours (vanilla) consistently outperforms both equal weighting and QME across all thresholds, remaining robust regardless of enrichment aggressiveness.

and QME across all exemplar counts. While overall performance scales with more exemplars for all methods, our advantage remains steady across all everage exemplar counts. This confirms that identity-specific weighting does not require extensive reference sets to be reliable, exhibiting strong robustness even in data-scarce enrollment regimes.

Robustness to gallery enrichment. Standard evaluation protocol assumes a static gallery where queries are discarded after scoring. To test whether identity-specific weighting remains robust when the gallery evolves dynamically, we evaluate recognition under gallery enrichment on CCVID using AdaFace, CAL, BigGait, and InsightFace. In each evaluation stream, a query is admitted as a new gallery exemplar if its calibrated top score exceeds an enrollment threshold η, *i.e.*, when $S_{q,\hat{i}} > \eta$ after evaluatin. For Ours (vanilla), newly admitted exemplars immediately update the identity's fusion weights without retraining. Because incoming queries can introduce pseudo-label noise and dynamic updates depend on arrival sequence, we evaluate across 9 random arrival orders and report both the mean and range. As shown in Figure 5, all methods benefit from gallery enrichment at low-to-moderate thresholds ($\eta \leq 4$) before converging to their static-gallery baselines as η increases and fewer queries qualify for admission. Notably, Ours (vanilla) consistently maintains a wide margin over equal weighting and QME across the entire threshold range, remaining resilient even when low thresholds introduce potential misidentifications. Experiments under an alternative write-in criterion show consistent gains and are reported in the Appendix.

5 Conclusion

In this work, we demonstrated that multi-model fusion for open-set identification substantially benefits from identity-specific reliability priors estimated directly from the given gallery exemplars. By exploiting intra- and inter-identity separation statistics within the gallery, our approach tailors model weights to individual identities without requiring additional training. Across multiple model pools and challenging clothes-changing benchmarks, our identity-specific weighting consistently outperforms uniform and quality-based baselines, seamlessly integrates with query-dependent fusion, and exhibits strong robustness under label noise, data scarcity, and dynamic gallery evolution.

Limitations. Our method estimates each model's reliability profile individually based on gallery comparisons, without explicitly modeling cross-model feature correlations or statistical redundancies among the candidate backbones. Extending this lightweight estimation to capture joint multi-model dependencies presents a promising direction for further refinement. Additionally, identity-conditioned score estimation requires multiple gallery exemplars per identity to gauge intra-class consistency. While this limits single-exemplar setups, standard person Re-ID benchmarks and video tracklets readily supply multiple reference frames.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure (such as generating synthetic datasets, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing experiments, implementing methods, or interpreting results), and tasks such as assisting with translation, data cleaning, or qualitative analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, specifically for editing manuscript text to improve readability. We have reviewed all AI-assisted work: all generated text, table structures, and descriptions were manually checked and verified by the authors against our empirical logs and experimental code to ensure accuracy, eliminate hallucinations, and prevent misrepresentation of findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. We use only the official research releases of CCVID, MEVID, and LTCC (Gu et al., 2022; Davila et al., 2023; Qian et al., 2020) under their license terms, and publicly released pretrained recognition models. We collect no new data, involve no human subjects, and do not attempt to identify any individual beyond the benchmark protocols.

REFERENCES

Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1–3):42–54, 2000. doi: 10.1006/dspr.1999.0360.

Abhijit Bendale and Terrance Boult. Towards open set deep networks, 2015. URL https://arxiv.org/abs/1511.06233.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628. PMLR, 21–27 Jul 2024.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust Multi-Modal Perception with Uncertainty-Aware Sensor Fusion. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 40998–41021, 2025.

Gordon Cormack, Charles Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. pp. 758–759, 07 2009. doi: 10.1145/1571941.1572114.

Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matthew Brown, Scott McCloskey, Anthony Hoogs, and Brian Clipp. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1634–1643, 2023. doi: 10.1109/WACV56688.2023.00168.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019. doi: 10.1109/CVPR.2019.00482.

George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds. SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *5th International Conference on Spoken Language Processing (ICSLP)*, 1998. doi: 10.21437/ICSLP.1998-244.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Julian Fierrez-Aguilar, Daniel Garcia-Romero, Javier Ortega-Garcia, and Joaquin Gonzalez-Rodriguez. Bayesian adaptation for user-dependent multimodal biometric authentication. *Pattern Recognition*, 38(8):1317–1319, 2005. doi: 10.1016/j.patcog.2005.01.013.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J. Radke. DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York, 1966.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 2: Identification. Technical Report NISTIR 8271, National Institute of Standards and Technology, 2019.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1060–1069, 2022. doi: 10.1109/CVPR52688.2022.00113.

Keshav Gupta, Gurjit Singh Walia, and Kapil Sharma. Quality based adaptive score fusion approach for multimodal biometric system. *Applied Intelligence*, 50(4):1086–1099, 2020. doi: 10.1007/s10489-019-01579-1.

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. *Pattern Recognition*, 43(5):1789–1800, 2010.

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*, 12(3):444, 2020.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 3rd edition, 2013.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhanbo Huang, Dingqiang Ye, Xiaoming Liu, and Yu Kong. Unlocking Motion from Large Vision Models with a Semantic and Kinematic Duality for Gait Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28379–28390, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_Unlocking_Motion_from_Large_Vision_Models_with_a_Semantic_and_CVPR_2026_paper.html.

InsightFace. InsightFace Model Zoo. https://github.com/deepinsight/insightface/tree/master/model_zoo.

Anil K. Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005. doi: 10.1016/j.patcog.2005.01.012.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18750–18759, 2022. doi: 10.1109/CVPR52688.2022.01819.

Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. SapiensID: Foundation for Human Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13937–13947, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kim_SapiensID_Foundation_for_Human_Recognition_CVPR_2025_paper.html.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023a. doi: 10.1609/aaai.v37i1.25225.

Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An In-Depth Exploration of Person Re-Identification and Gait Recognition in Cloth-Changing Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13824–13833, June 2023b.

Xin Liang and Yogesh S Rawat. DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13980–13989, June 2025.

Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z. Li. Open-set person re-identification, 2014. URL https://arxiv.org/abs/1408.0872.

Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning Clothing and Pose Invariant 3D Shape Representation for Long-Term Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19617–19626, October 2023.

Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. FarSight: A Physics-Driven Whole-Body Biometric System at Large Distance and Altitude. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6227–6236, January 2024a.

Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 256–266, June 2024b.

Feng Liu, Nicholas Chimitt, Lanqing Guo, Jitesh Jain, Aditya Kane, Minchul Kim, Wes Robbins, Yiyang Su, Dingqiang Ye, Xingguang Zhang, Jie Zhu, Siddharth Satyakam, Christopher Perry, Stanley H. Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. Person recognition at altitude and range: Fusion of face, body shape and gait, 2025a. URL https://arxiv.org/abs/2505.04616.

Wei Liu, Yufei Chen, and Xiaodong Yue. Enhancing Testing-Time Robustness for Trusted Multi-View Classification in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15508–15517, June 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Enhancing_Testing-Time_Robustness_for_Trusted_Multi-View_Classification_in_the_Wild_CVPR_2025_paper.html.

K. Nandakumar, Yi Chen, A.K. Jain, and S.C. Dass. Quality-based score level fusion in multibiometric systems. In *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006. doi: 10.1109/icpr.2006.951. URL http://dx.doi.org/10.1109/ICPR.2006.951.

K. Nandakumar, Yi Chen, S.C. Dass, and A.K. Jain. Likelihood ratio-based biometric score fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):342–347, February 2008. ISSN 0162-8828. doi: 10.1109/tpami.2007.70796. URL http://dx.doi.org/10.1109/TPAMI.2007.70796.

Zhiqi Pang, Junjie Wang, Lingling Zhao, and Chunyu Wang. Identity-Clothing Similarity Modeling for Unsupervised Clothing Change Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19251–19260, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Pang_Identity-Clothing_Similarity_Modeling_for_Unsupervised_Clothing_Change_Person_Re-Identification_CVPR_2025_paper.html.

Tae Jin Park, Manoj Kumar, and Shrikanth Narayanan. Multi-scale speaker diarization with neural affinity score fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Priyank Pathak and Yogesh S. Rawat. Colors See Colors Ignore: Clothes Changing ReID with Color Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16797–16807, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Pathak_Colors_See_Colors_Ignore_Clothes_Changing_ReID_with_Color_Disentanglement_ICCV_2025_paper.html.

N. Poh and J. Kittler. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):3–18, January 2012. ISSN 2160-9292. doi: 10.1109/tpami.2011.102. URL http://dx.doi.org/10.1109/TPAMI.2011.102.

Norman Poh, Thirimachos Bourlai, Josef Kittler, Lorene Allano, Fernando Alonso-Fernandez, Onkar Ambekar, John Baker, Bernadette Dorizzi, Omolara Fatukasi, Julian Fierrez, Harald Ganster, Javier Ortega-Garcia, Donald Maurer, Albert Ali Salah, Tobias Scheidat, and Claus Vielhauer. Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms. *IEEE Transactions on Information Forensics and Security*, 4(4):849–866, 2009. doi: 10.1109/TIFS.2009.2034885.

Norman Poh, Arun Ross, Weifeng Lee, and Josef Kittler. A user-specific and selective multimodal biometric fusion strategy by ranking subjects. *Pattern Recognition*, 46(12):3341–3357, December 2013. ISSN 0031-3203. doi: 10.1016/j.patcog.2013.03.018. URL http://dx.doi.org/10.1016/j.patcog.2013.03.018.

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 *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 71–88, 2020.

Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13): 2115–2125, September 2003. ISSN 0167-8655. doi: 10.1016/s0167-8655(03)00079-5. URL http://dx.doi.org/10.1016/S0167-8655(03)00079-5.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014. doi: 10.1109/TPAMI.2014.2321392.

R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):450–455, March 2005. ISSN 0162-8828. doi: 10.1109/tpami.2005.57. URL http://dx.doi.org/10.1109/TPAMI.2005.57.

Robert Snelick, Mike Indovina, James Yen, and Alan Mink. Multimodal biometrics: issues in design and testing. In *International Conference on Multimodal Interfaces (ICMI)*, 2003.

Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. *Open-Set Biometrics: Beyond Good Closed-Set Models*, pp. 243–261. Springer Nature Switzerland, October 2024. ISBN 9783031730337. doi: 10.1007/978-3-031-73033-7_14. URL http://dx.doi.org/10.1007/978-3-031-73033-7_14.

Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. HAMoBE: Hierarchical and adaptive mixture of biometric experts for video-based person ReID. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11525–11536, 2025.

Yiyang Su, Minchul Kim, Jie Zhu, Christopher Perry, Feng Liu, Anil Jain, and Xiaoming Liu. LocalScore: Local Density-Aware Similarity Scoring for Biometrics, 2026. URL https://arxiv.org/abs/2602.01012.

J. H. Teng, T. S. Ong, T. Connie, K. Sonai Muthu Anbananthen, and P. P. Min. Optimized score level fusion for multi-instance finger vein recognition. *Algorithms*, 15(5):161, 2022. doi: 10.3390/a15050161.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 79–88, 2018.

Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:8821–8830, 2020. doi: 10.1109/TIP.2020.3001693.

Neil Yager and Ted Dunstone. The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):220–230, 2010. doi: 10.1109/TPAMI.2008.291.

Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (6):2029–2046, 2021. doi: 10.1109/TPAMI.2019.2960509.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1472–1481, 2023.

Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210, 2024. doi: 10.1109/CVPR52733.2024.00027.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2016. doi: 10.1109/TPAMI.2015.2453984.

Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, and Ram Nevatia. SEAS: ShapE-Aligned Supervision for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 164–174, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhu_SEAS_ShapE-Aligned_Supervision_for_Person_Re-Identification_CVPR_2024_paper.html.

Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. A quality-guided mixture of score-fusion experts framework for human recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13076–13086, October 2025. doi: 10.1109/ICCV51701.2025.01215.

Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, and Xiaoming Liu. Fusionagent: A multimodal agent with dynamic model selection for human recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32756–32766, June 2026.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10492–10502, 2021. URL https://arxiv.org/abs/2103.04098.

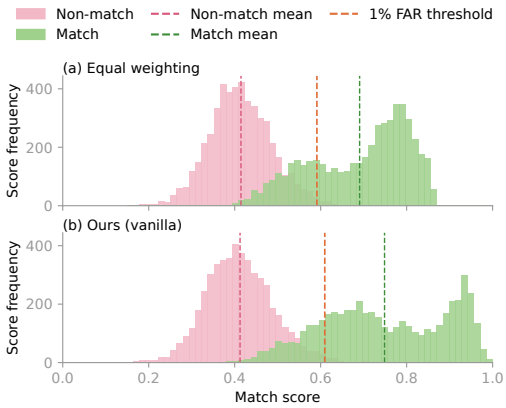


Figure A1: **Separation of match and non-match score distributions.** Dataset: CCVID; model pool: AdaFace, BigGait, and CAL. Dashed lines indicate distribution means and the 1% FAR threshold. Our identity-conditioned fusion markedly enlarges the margin between genuine matches and impostors by driving match scores rightward, effectively suppressing distribution overlap under strict verification thresholds.

A Additional Implementation Details

A.1 Recognition models and score processing

Model configurations. The evaluated model combinations are specified in Tables A4–A6. For the video benchmarks, AdaFace (Kim et al., 2022) uses ViT-Base (Dosovitskiy et al., 2021) trained on WebFace4M (Zhu et al., 2021), and CAL (Gu et al., 2022) uses a dataset-specific ResNet-50 checkpoint. We evaluate BigGait (Ye et al., 2024) on CCVID (Gu et al., 2022), AGRL (Wu et al., 2020) (ResNet-50 trained on MEVID (Davila et al., 2023)) on MEVID, and AIM (Yang et al., 2023) on LTCC (Qian et al., 2020). On CCVID, InsightFace uses the released `buffalo_l/w600k_r50` ArcFace (Deng et al., 2019) model trained on WebFace600K (InsightFace). For CLIP-ReID (Li et al., 2023a), we use ViT-B/16 (Dosovitskiy et al., 2021) checkpoints trained on MSMT17 (Wei et al., 2018), Market-1501 (Zheng et al., 2015), and DukeMTMC-reID (Gou et al., 2017), evaluated both with and without LoRA (Hu et al., 2022) adaptation on the target benchmark's training split. Each contributes as a separate model in the combinations that include it.

Similarity scores of each model. Frame embeddings are mean-pooled within each tracklet; BigGait (Ye et al., 2024) and AGRL (Wu et al., 2020) provide sequence embeddings. Face embeddings use detected and aligned frames. On CCVID and MEVID (Gu et al., 2022; Davila et al., 2023), the embedding-based fusion experiments compute cosine similarity between unit-normalized embeddings for every model, including BigGait (Ye et al., 2024). This applies to both query–gallery scoring and gallery–gallery profile estimation. On LTCC (Qian et al., 2020), query–gallery scores for AdaFace (Kim et al., 2022), CAL (Gu et al., 2022), and AIM (Yang et al., 2023) come from the released QME (Zhu et al., 2025) score matrices; CLIP-ReID (Li et al., 2023a) scores and all gallery–gallery scores are computed from embeddings using cosine similarity. Larger scores indicate stronger matches.

A.2 Additional Results

Score distribution analysis. As demonstrated in Figure A1, our identity-conditioned fusion improves distributional separability over equal weighting on CCVID (Gu et al., 2022). By prioritizing each identity's most discriminative representations, our method elevates genuine match confidence. Because the two distributions overlap far less, fewer true matches are rejected when operating under a strict 1% FAR threshold.

Table A1: **Ablation on weight combination rules for joint fusion.** Dataset: CCVID; Models pool: AdaFace, BigGait, CAL, and zero-shot CLIP-ReID. We compare different strategies for composing query-conditioned weights w_k^q and identity-specific weights w_k^i, keeping each branch's weights fixed. λ denotes the interpolation weight for the identity branch. **Bold** marks the best result among all combination strategies.

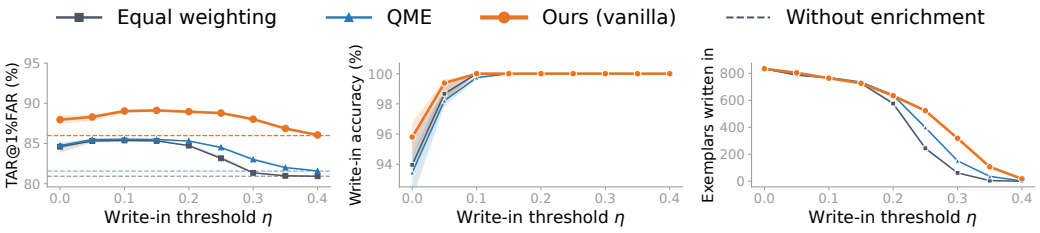
Combination Rule	Rank-1	mAP	TAR@1%FAR	FNIR@1%FPIR (↓)
Query only	93.77	91.96	82.40	10.38
Identity only	93.53	93.10	85.36	11.23
Product (Default)	**94.60**	**93.99**	85.60	**9.96**
Weighted sum ($\lambda = 0.25$)	94.00	92.94	84.45	10.27
Weighted sum ($\lambda = 0.50$)	93.89	93.38	85.94	10.51
Weighted sum ($\lambda = 0.75$)	93.64	93.31	86.08	10.72
Geometric mean	93.64	93.40	85.50	10.57
RRF ($k = 60$)	93.53	92.61	**86.40**	68.09

Table A2: **Ablation on per-identity separability metrics.** Comparison of different gallery-derived separation scores and learned per-identity weighting baselines (‡), evaluated across four CCVID and three MEVID expert pools. FNIR and TAR are measured at 1% FPIR and 1% FAR, respectively. Bold indicates the best result per column.

	CCVID				MEVID			
Approach	R-1 ↑	mAP ↑	TAR ↑	FNIR ↓	R-1 ↑	mAP ↑	TAR ↑	FNIR ↓
Equal weights (no identity score)	92.03	90.62	80.72	15.96	62.13	36.62	40.37	58.92
d' (Default)	**94.21**	**92.62**	**84.36**	10.60	**64.03**	38.10	41.75	57.57
Mean genuine similarity	93.65	91.76	83.09	**9.67**	61.50	36.10	40.12	59.97
Genuine − impostor mean	93.94	91.95	83.34	9.74	61.71	36.44	40.23	59.82
AUC	92.18	90.76	82.66	16.46	63.40	37.89	41.57	57.84
1−EER	92.24	90.87	83.07	16.41	63.08	37.82	41.44	57.79
LDA weights‡	89.24	87.95	81.99	24.56	62.34	38.47	**42.74**	**57.52**
Logistic-regression weights‡	91.28	89.56	80.39	20.09	62.24	**38.60**	42.01	58.09

Weight combination strategies. We ablate alternative rules for composing the query-conditioned weights w_k^q and identity-conditioned weights w_k^i, keeping each branch's weights fixed. Specifically, the additive baselines linearly interpolate weights via $\lambda w_k^i + (1-\lambda) w_k^q$, geometric mean computes $(w_k^i w_k^q)^{1/2}$, and RRF (Cormack et al., 2009) combines retrieval ranks instead of scores. Our product rule adopts element-wise multiplication $w_k^i \cdot w_k^q$ followed by normalization over available experts. As shown in Table A1, integrating both branches yields performance gains. Our multiplicative product outperforms additive interpolation across varying mixing weights λ, as well as geometric mean averaging, achieving the best Rank-1 (94.60%), mAP (93.99%), and open-set FNIR (9.96%). Crucially, multiplicative gating acts as a parameter-free filter that naturally activates an expert only when query quality and identity reliability align. Although RRF achieves competitive TAR, its rank-transformed scores discard continuous similarity distributions, severely disrupting per-identity score calibration and causing open-set FNIR to degrade catastrophically to 68.09%. We therefore adopt the product rule as our default composition strategy in Section 3.3.

Choice of per-identity separability metric. We ablate the gallery-derived metric used to measure model distinctiveness per identity in Table A2. Starting from equal weights, we evaluate several training-free alternatives to d'. Specifically, *Mean genuine similarity* averages similarity scores within the same identity, while *Genuine − impostor mean* measures the gap between genuine and impostor averages. We also evaluate threshold-independent separation using receiver operating characteristics, namely *AUC* and *1−EER* (Fawcett, 2006). Each score is normalized to match the global mean and variance of d' before mapping to weights. Finally, we test two learned per-identity base-



(a) (b) (c)

Figure A2: **Robustness to gallery enrichment with a margin-based admission rule.** Queries are enrolled when the gap between the top two candidate scores satisfies $S_{q,\hat{i}} - \max_{i \neq \hat{i}} S_{q,i} > \eta$. Across all subplots, curves and shaded areas indicate the mean and spread over 9 arrival orders, and dashed lines show static-gallery baselines: (a) TAR@1%FAR, (b) pseudo-label precision of admitted queries, and (c) total admitted queries ($\eta = 0$ admits all queries). Ours (vanilla) consistently maintains its performance advantage over equal weighting and QME at every threshold.

Table A3: **Sensitivity to hyperparameters.** Analysis of interpolation weight α and softmax temperature τ on CCVID using a model pool of AdaFace, BigGait, CAL, and InsightFace. FNIR and TAR are reported at 1% FPIR and 1% FAR, respectively. Bold indicates the best result per column.

(a) Interpolation weight α ($\tau = 0.5$)

α	R-1 ↑	mAP ↑	TAR ↑	FNIR ↓
0 (equal)	93.53	90.90	80.92	10.02
0.25	95.20	92.08	84.51	9.29
0.5 (default)	**95.56**	**92.33**	85.27	**8.97**
0.75	94.84	91.21	**86.32**	9.02
1	93.05	89.29	85.33	14.24

(b) Softmax temperature τ ($\alpha = 0.5$)

τ	R-1 ↑	mAP ↑	TAR ↑	FNIR ↓
0.1	94.84	91.54	86.18	9.85
0.25	95.32	92.01	**86.34**	9.45
0.5 (default)	**95.56**	**92.33**	85.27	**8.97**
1	95.44	92.23	85.98	9.20
2	95.44	92.22	84.73	9.18
4	95.32	91.98	84.21	9.35

lines, linear discriminant analysis (LDA) (Fisher, 1936) and logistic regression (Hosmer et al., 2013), which fit model weights directly on gallery pairs with sample-size shrinkage toward the global solution. Across both benchmarks, d' yields the most reliable overall performance, demonstrating that accounting for score spread provides a stronger basis for weight estimation than other methods.

Alternative admission rule for gallery enrichment. While Section 4.3 employs an absolute top-score threshold, we further verify whether our conclusions hold under an alternative, margin-based admission criterion. Under this rule, a query is admitted only when its top prediction leads the runner-up candidate by a confidence margin η, *i.e.*, $S_{q,\hat{i}} - \max_{i \neq \hat{i}} S_{q,i} > \eta$. Following the identical protocol on CCVID across nine arrival sequences, the resulting trends closely mirror those in the main text, as shown in Figure A2. All methods benefit from moderate enrichment before converging to their static-gallery baselines. Ours (vanilla) consistently outperforms equal weighting and QME across all thresholds; remarkably, its lowest mean TAR (86.1%) still surpasses the peak performance of either baseline (85.5%). The underlying enrollment dynamics in Figure A2(b) and (c) account for this advantage: even at $\eta = 0$ where all queries enter the gallery, Ours (vanilla) incurs lower label noise (95.8% accuracy vs. $\leq$ 94.0% for baselines). As η increases, it consistently admits substantially more clean exemplars than the baselines under the same threshold. This confirms that identity-specific weighting sharpens margin separation between true matches and competing candidates, safely scaling up gallery enrichment without introducing detrimental noise.

Sensitivity to hyperparameters. Table A3 analyzes the effect of the interpolation weight α and softmax temperature τ using an expert pool comprising AdaFace, BigGait, CAL, and InsightFace on CCVID. As shown in Table A3a, performance improves markedly from equal weighting ($\alpha = 0$) across intermediate values of α, peaking around $\alpha = 0.5$. Discarding the prior completely ($\alpha = 1$) degrades verification, confirming that anchoring per-identity weights to a uniform baseline prevents overfitting to sparse gallery samples. For the temperature parameter (Table A3b), recognition re-

mains stable across a wide range of values ($0.25 \leq \tau \leq 2$), with $\tau = 0.5$ providing a solid trade-off between closed-set retrieval and open-set verification without requiring aggressive tuning.

Extended evaluations across model pools. Tables A4, A5, and A6 report full results across all model pools on CCVID, MEVID, and LTCC (Gu et al., 2022; Davila et al., 2023; Qian et al., 2020). Across all benchmarks, **Ours (vanilla)** consistently outperforms single-model baselines and classical score fusion rules without requiring any parameter training. When integrated with query-adaptive methods (**Ours (QME)** and **Ours (FusionAgent)**), it achieves top-tier performance across diverse pool configurations. Notably, identity-specific weighting yields the most pronounced gains in open-set identification, substantially cutting FNIR@1%FPIR over static fusion baselines and demonstrating strong robustness against severe appearance changes.

Table A4: **Extended multi-model performance on CCVID.** *Combination* specifies the model pool shared by each block (♦ AdaFace, ♠ CAL, ▲ BigGait, ♣ InsightFace/ArcFace (Deng et al., 2019), and ● zero-shot and ⊛ LoRA-adapted CLIP-ReID (Li et al., 2023a; Hu et al., 2022)). **Best** and second-best results are marked within each model pool.

Combination	Method	Rank-1	mAP	TAR@1%FAR	FNIR@1%FPIR (↓)
♦	AdaFace (Kim et al., 2022)	94.0	87.9	75.7	13.0 ± 3.5
♠	CAL (Gu et al., 2022)	81.4	74.7	66.3	52.8 ± 13.3
▲	BigGait (Ye et al., 2024)	76.7	61.0	49.7	71.1 ± 6.1
♣	InsightFace/ArcFace (Deng et al., 2019)	94.1	88.2	78.9	18.1 ± 3.1
●	CLIP-ReID (Li et al., 2023a)	64.1	63.0	67.9	63.1 ± 2.5
⊛	LoRA-adapted CLIP-ReID (Hu et al., 2022)	85.4	84.3	80.3	42.8 ± 7.5
♦♠▲	Min-Fusion (Jain et al., 2005)	87.1	79.2	62.4	48.5 ± 8.7
	Max-Fusion (Jain et al., 2005)	89.9	89.3	73.4	23.0 ± 10.1
	Z-score (Snelick et al., 2003)	92.2	90.6	73.9	15.1 ± 1.5
	Min-max (Snelick et al., 2003)	91.8	90.9	73.9	15.4 ± 2.0
	RHE (He et al., 2010)	91.7	90.2	73.1	16.6 ± 2.5
	Asym-AO (Herbadji et al., 2020)	92.0	91.0	73.8	15.5 ± 1.8
	Farsight (Liu et al., 2025a)	92.0	91.2	73.9	13.9 ± 1.1
	Weighted-sum (Park et al., 2021)	93.5	90.8	75.8	12.6 ± 1.1
	BSSF (Teng et al., 2022)	93.6	90.7	75.5	12.9 ± 1.6
	QME (Zhu et al., 2025)	94.1	90.8	76.2	12.3 ± 1.4
	FusionAgent (Zhu et al., 2026)	93.4	92.6	85.9	10.1 ± 1.5
	Ours (vanilla)	**95.3**	92.1	81.0	11.0 ± 1.6
	Ours (QME)	95.2	91.9	79.3	10.3 ± 1.5
	Ours (FusionAgent)	94.6	**93.7**	**87.0**	**9.4 ± 1.2**
♦♠▲♣	Min-Fusion (Jain et al., 2005)	94.4	88.6	79.6	15.1 ± 2.6
	Max-Fusion (Jain et al., 2005)	84.2	69.2	57.5	51.2 ± 8.3
	Z-score (Snelick et al., 2003)	93.5	90.8	80.9	13.2 ± 2.5
	Min-max (Snelick et al., 2003)	93.4	90.9	78.8	13.5 ± 2.5
	RHE (He et al., 2010)	93.4	90.9	79.0	13.5 ± 2.3
	Asym-AO (Herbadji et al., 2020)	93.3	90.7	78.3	13.5 ± 2.5
	Farsight (Liu et al., 2025a)	93.6	90.7	79.2	13.3 ± 2.0
	Weighted-sum (Park et al., 2021)	94.2	89.8	81.5	13.5 ± 2.1
	BSSF (Teng et al., 2022)	93.5	90.6	82.8	15.0 ± 3.9
	QME (Zhu et al., 2025)	93.0	91.5	78.9	11.2 ± 1.2
	Ours (vanilla)	**95.4**	92.2	**86.0**	**9.2 ± 1.4**
	Ours (QME)	94.7	**92.7**	84.2	9.6 ± 1.3
♦♠▲●	Min-Fusion (Jain et al., 2005)	94.1	87.9	74.4	15.4 ± 1.1
	Max-Fusion (Jain et al., 2005)	79.9	73.5	75.3	55.7 ± 4.9
	Z-score (Snelick et al., 2003)	89.6	89.3	82.7	35.8 ± 7.0
	Min-max (Snelick et al., 2003)	88.6	88.2	82.5	37.8 ± 6.0
	RHE (He et al., 2010)	88.7	88.2	82.5	38.9 ± 6.6
	Asym-AO (Herbadji et al., 2020)	90.9	90.8	81.3	31.6 ± 6.2
	Farsight (Liu et al., 2025a)	91.7	91.6	80.9	15.3 ± 3.7
	Weighted-sum (Park et al., 2021)	94.0	92.4	84.4	14.0 ± 2.9
	BSSF (Teng et al., 2022)	93.9	92.3	85.1	17.0 ± 4.8
	QME (Zhu et al., 2025)	92.3	92.3	81.6	11.2 ± 1.5
	Ours (vanilla)	**94.6**	**94.0**	**85.6**	**10.0 ± 1.4**
	Ours (QME)	93.2	93.0	83.3	11.0 ± 1.4
♦♠▲⊛	Min-Fusion (Jain et al., 2005)	91.6	87.6	75.7	20.6 ± 3.9
	Max-Fusion (Jain et al., 2005)	84.3	74.4	65.5	49.2 ± 7.1
	Z-score (Snelick et al., 2003)	91.8	92.0	83.6	18.3 ± 4.7
	Min-max (Snelick et al., 2003)	91.8	91.9	83.0	18.8 ± 4.6
	RHE (He et al., 2010)	91.8	91.9	83.0	18.9 ± 4.5
	Asym-AO (Herbadji et al., 2020)	91.8	92.4	82.4	16.5 ± 4.6
	Farsight (Liu et al., 2025a)	91.8	91.8	83.2	19.5 ± 4.2
	Weighted-sum (Park et al., 2021)	**93.2**	92.7	**85.8**	12.1 ± 1.4
	BSSF (Teng et al., 2022)	92.8	92.5	**85.8**	14.5 ± 4.1
	QME (Zhu et al., 2025)	91.7	91.7	82.9	14.7 ± 2.7
	Ours (vanilla)	92.6	**93.0**	85.1	**10.9 ± 1.2**
	Ours (QME)	91.7	92.3	83.2	11.0 ± 1.2

Table A5: **Extended multi-model evaluation on MEVID.** *Combination* specifies the model pool shared by each block (♦ AdaFace, ♠ CAL, ▲ AGRL, and ● zero-shot and ⊛ LoRA-adapted CLIP-ReID (Li et al., 2023a; Hu et al., 2022)). **Best** and second-best results are marked within each block.

Combination	Method	Rank-1	mAP	TAR@1%FAR	FNIR@1%FPIR (↓)
♦	AdaFace (Kim et al., 2022)	25.0	8.1	5.4	98.8 ± 1.2
♠	CAL (Gu et al., 2022)	52.5	27.1	34.7	67.8 ± 7.3
▲	AGRL (Wu et al., 2020)	51.9	25.5	30.7	69.4 ± 8.9
●	CLIP-ReID (Li et al., 2023a)	63.0	33.0	33.6	68.3 ± 3.3
⊛	LoRA-adapted CLIP-ReID (Hu et al., 2022)	69.9	45.1	47.7	49.7 ± 4.3
♦♠▲	Min-Fusion (Jain et al., 2005)	46.8	21.2	28.0	70.4 ± 8.0
	Max-Fusion (Jain et al., 2005)	33.2	14.9	8.3	97.4 ± 1.6
	Z-score (Snelick et al., 2003)	54.1	27.4	30.7	66.5 ± 7.0
	Min-max (Snelick et al., 2003)	52.8	24.7	25.0	71.3 ± 6.1
	RHE (He et al., 2010)	52.8	24.8	25.3	71.2 ± 6.2
	Asym-AO (Herbadji et al., 2020)	52.5	22.9	23.6	71.7 ± 6.4
	Farsight (Liu et al., 2025a)	53.8	25.4	26.6	69.8 ± 7.1
	Weighted-sum (Park et al., 2021)	53.5	29.0	34.8	64.8 ± 8.9
	BSSF (Teng et al., 2022)	54.4	28.7	33.4	63.8 ± 9.4
	QME (Zhu et al., 2025)	55.7	28.2	32.9	64.6 ± 8.2
	FusionAgent (Zhu et al., 2026)	54.7	28.7	34.9	58.6 ± 7.4
	Ours (vanilla)	55.4	31.0	36.2	66.4 ± 13.6
	Ours (QME)	54.7	31.0	35.5	59.0 ± 12.6
	Ours (FusionAgent)	**57.0**	**31.2**	**37.3**	**52.0 ± 15.5**
♦♠▲●	Min-Fusion (Jain et al., 2005)	47.8	21.7	28.0	70.4 ± 8.0
	Max-Fusion (Jain et al., 2005)	62.0	35.1	35.9	63.7 ± 3.2
	Z-score (Snelick et al., 2003)	62.7	34.5	37.9	56.4 ± 8.8
	Min-max (Snelick et al., 2003)	62.0	33.2	34.9	59.8 ± 8.6
	RHE (He et al., 2010)	61.1	32.6	34.2	61.3 ± 8.5
	Asym-AO (Herbadji et al., 2020)	58.2	28.8	28.6	62.0 ± 7.7
	Farsight (Liu et al., 2025a)	56.0	29.3	30.7	65.6 ± 6.8
	Weighted-sum (Park et al., 2021)	61.1	34.0	39.4	58.4 ± 6.7
	BSSF (Teng et al., 2022)	62.0	36.1	39.6	57.5 ± 6.2
	QME (Zhu et al., 2025)	63.6	36.0	40.1	55.6 ± 8.4
	Ours (vanilla)	**65.8**	**37.7**	**40.9**	57.3 ± 16.8
	Ours (QME)	64.2	37.0	40.7	**50.1 ± 16.7**
♦♠▲⊛	Min-Fusion (Jain et al., 2005)	47.8	21.7	28.0	70.4 ± 8.0
	Max-Fusion (Jain et al., 2005)	68.4	43.9	45.9	55.8 ± 8.1
	Z-score (Snelick et al., 2003)	66.8	41.1	44.9	52.1 ± 7.2
	Min-max (Snelick et al., 2003)	66.1	40.3	42.6	54.8 ± 8.1
	RHE (He et al., 2010)	65.2	39.0	41.2	57.0 ± 8.7
	Asym-AO (Herbadji et al., 2020)	62.3	35.0	31.3	57.8 ± 8.0
	Farsight (Liu et al., 2025a)	61.7	36.6	38.7	58.9 ± 9.1
	Weighted-sum (Park et al., 2021)	65.5	40.4	45.2	52.6 ± 5.6
	QME (Zhu et al., 2025)	65.5	42.1	46.3	51.2 ± 6.1
	Ours (vanilla)	**70.9**	**45.6**	**48.2**	49.0 ± 18.5
	Ours (QME)	67.7	44.2	47.6	**48.2 ± 16.8**

Table A6: **Extended multi-model performance on LTCC.** *Combination* specifies the model pool shared by each block (♦ AdaFace, ♠ CAL, ■ AIM, and ● zero-shot and ⊛ LoRA-adapted CLIP-ReID (Li et al., 2023a; Hu et al., 2022)). **Best** and second-best results are marked within each model pool.

Combination	Method	Rank-1	mAP	TAR@1%FAR	FNIR@1%FPIR (↓)
♦	AdaFace (Kim et al., 2022)	18.5	5.9	2.4	99.8 ± 0.2
♠	CAL (Gu et al., 2022)	74.4	40.6	36.7	59.7 ± 7.3
■	AIM (Yang et al., 2023)	74.8	40.9	37.0	66.2 ± 7.5
●	CLIP-ReID (Li et al., 2023a)	70.7	33.7	26.5	47.8 ± 13.0
⊛	LoRA-adapted CLIP-ReID (Hu et al., 2022)	76.7	40.7	31.9	42.0 ± 10.3
♦♠■	Min-Fusion (Jain et al., 2005)	37.1	12.9	12.4	81.9 ± 6.0
	Max-Fusion (Jain et al., 2005)	62.5	33.3	16.8	94.8 ± 4.7
	Z-score (Snelick et al., 2003)	73.0	37.5	30.4	68.7 ± 9.2
	Min-max (Snelick et al., 2003)	73.2	38.1	31.9	75.1 ± 9.2
	RHE (He et al., 2010)	70.4	34.2	21.5	78.0 ± 10.0
	Asym-AO (Herbadji et al., 2020)	71.2	32.9	19.1	76.3 ± 8.9
	Farsight (Liu et al., 2025a)	73.2	37.8	31.3	72.4 ± 8.6
	Weighted-sum (Park et al., 2021)	75.1	**42.5**	**38.1**	59.5 ± 9.4
	BSSF (Teng et al., 2022)	75.1	**42.5**	38.0	60.0 ± 9.6
	QME (Zhu et al., 2025)	73.8	39.6	35.0	64.3 ± 8.0
	FusionAgent (Zhu et al., 2026)	75.5	41.0	37.0	50.0 ± 8.5
	Ours (vanilla)	70.8	40.3	35.1	52.0 ± 12.7
	Ours (QME)	**75.7**	42.4	37.9	**26.4 ± 10.2**
	Ours (FusionAgent)	75.5	41.0	35.2	33.6 ± 14.0
♦♠■●	Min-Fusion (Jain et al., 2005)	37.1	12.9	12.4	81.9 ± 6.0
	Max-Fusion (Jain et al., 2005)	72.4	35.6	26.5	49.4 ± 21.2
	Z-score (Snelick et al., 2003)	76.7	42.8	36.3	44.4 ± 9.2
	Min-max (Snelick et al., 2003)	77.5	43.2	36.5	51.0 ± 9.9
	RHE (He et al., 2010)	75.9	41.0	33.8	64.5 ± 9.4
	Asym-AO (Herbadji et al., 2020)	74.4	31.3	27.6	60.6 ± 7.2
	Farsight (Liu et al., 2025a)	76.3	42.1	35.9	64.2 ± 9.6
	Weighted-sum (Park et al., 2021)	77.7	**44.9**	**39.8**	54.5 ± 11.6
	BSSF (Teng et al., 2022)	77.5	44.5	39.4	56.8 ± 11.1
	QME (Zhu et al., 2025)	**77.9**	43.6	37.6	59.8 ± 7.0
	Ours (vanilla)	76.9	**44.9**	38.3	35.6 ± 16.6
	Ours (QME)	77.5	43.9	37.7	**28.5 ± 20.0**
♦♠■⊛	Min-Fusion (Jain et al., 2005)	42.0	15.0	13.1	81.9 ± 6.0
	Max-Fusion (Jain et al., 2005)	76.1	41.6	32.1	32.6 ± 11.0
	Z-score (Snelick et al., 2003)	79.1	47.8	40.3	48.3 ± 13.3
	Min-max (Snelick et al., 2003)	79.9	48.7	40.6	53.6 ± 15.3
	RHE (He et al., 2010)	78.3	46.3	38.4	64.9 ± 14.7
	Asym-AO (Herbadji et al., 2020)	76.3	38.8	32.2	63.0 ± 9.8
	Farsight (Liu et al., 2025a)	80.3	48.6	40.4	49.9 ± 14.6
	Weighted-sum (Park et al., 2021)	79.3	48.4	**41.9**	51.8 ± 9.1
	BSSF (Teng et al., 2022)	79.9	49.3	**41.9**	50.9 ± 10.0
	QME (Zhu et al., 2025)	**80.7**	48.5	41.3	60.7 ± 8.5
	Ours (vanilla)	80.5	**49.6**	41.4	31.1 ± 19.0
	Ours (QME)	80.5	48.7	41.3	**29.1 ± 21.2**