TransFIRA: Transfer Learning for Face Image Recognizability Assessment

Allen Tu[1,2] Kartik Narayan[3] Joshua Gleason[1] Jennifer Xu[1]
Matthew Meyn[1] Tom Goldstein[2] Vishal M. Patel[3]
[1] Systems and Technology Research [2] University of Maryland, College Park [3] Johns Hopkins University
https://transfira.github.io/

***Abstract*— Face recognition in unconstrained environments such as surveillance, video, and web imagery must contend with extreme variation in pose, blur, illumination, and occlusion, where conventional visual quality metrics fail to predict whether inputs are truly recognizable to the deployed encoder. Existing FIQA methods typically rely on visual heuristics, curated annotations, or computationally intensive generative pipelines, leaving their predictions detached from the encoder's decision geometry. We introduce TransFIRA (Transfer Learning for Face Image Recognizability Assessment), a lightweight and annotation-free framework that grounds recognizability directly in embedding space. TransFIRA delivers three advances: (i) a definition of recognizability via class-center similarity (CCS) and class-center angular separation (CCAS), yielding the first natural, decision-boundary–aligned criterion for filtering and weighting; (ii) a recognizability-informed aggregation strategy that achieves state-of-the-art verification accuracy on BRIAR and IJB-C while nearly doubling correlation with true recognizability, all without external labels, heuristics, or backbone-specific training; and (iii) new extensions beyond faces, including encoder-grounded explainability that reveals how degradations and subject-specific factors affect recognizability, and the first recognizability-aware body recognition assessment. Experiments confirm state-of-the-art results on faces, strong performance on body recognition, and robustness under cross-dataset shifts. Together, these contributions establish TransFIRA as a unified, geometry-driven framework for recognizability assessment – encoder-specific, accurate, interpretable, and extensible across modalities – significantly advancing FIQA in accuracy, explainability, and scope.**

I. INTRODUCTION

Face recognition in unconstrained domains such as surveillance, long-form video, and web imagery must contend with severe variation in pose, blur, illumination, and occlusion. Under such conditions, recognition performance cannot be inferred reliably from superficial measures of image quality. What ultimately matters is whether samples are *recognizable* to the deployed encoder, since even a few unrecognizable frames can corrupt template construction and compromise downstream matching. Bridging this gap between appearance-based quality and encoder-specific recognizability defines *face image quality assessment (FIQA)* as a central challenge for robust recognition.

Despite progress in face recognition, FIQA methods remain limited in generalization and interpretability. Hand-crafted cues such as sharpness, pose, or illumination [7, 8, 16, 22, 23] often fail to reflect recognizability in the encoder's embedding space. Regression- and uncertainty-based pipelines [11, 20, 24] rely on curated labels, auxiliary supervision, or costly inference, making them indirect and resource-heavy. Model-based methods [3, 15, 19] couple FIQA to specific backbones, restricting adaptability, while generative and multimodal pipelines [1, 2, 21] improve performance but add complexity without grounding in encoder decision geometry. Crucially, none provide a self-supervised, geometrically principled definition of recognizability, leaving predictions as proxies rather than encoder-specific signals.

We introduce **TransFIRA** (Transfer Learning for Face Image Recognizability Assessment), the first framework to adapt any pretrained backbone for recognizability prediction without human quality annotations or IQA supervision. TransFIRA derives supervision directly from embeddings via *class-center similarity* (CCS) and *class-center angular separation* (CCAS), labels deterministically tied to the encoder's discrimination ability. This yields predictions that are encoder-specific, geometry-driven, and interpretable. Beyond prediction, TransFIRA enables **recognizability-informed aggregation**: a natural cutoff ($CCAS > 0$) that filters unrecognizable frames and CCS-based weighting that emphasizes compact, reliable embeddings. Together, these operations replace heuristics with transparent, geometry-grounded rules.

Extensive experiments on BRIAR Protocol 3.1 [5] and IJB-C [18] show that TransFIRA consistently surpasses prior FIQA methods such as MagFace [19], CR-FIQA [4], and DifFIQA [1], while remaining lightweight and broadly adaptable. Beyond faces, we extend the framework to body recognition via a sigmoid calibration strategy that restores discriminative power for CCS and CCAS. Finally, we show that recognizability predictions provide the first *encoder-grounded explainability*: they capture how degradations such as blur or occlusion alter recognizability, flag subjects who are inherently difficult to identify, and expose conditions where recognition is likely to fail.

In summary, we propose the following contributions:

- **Encoder-specific recognizability supervision:** CCS and CCAS derived directly from embeddings provide the first geometrically grounded basis for FIQA.
- **Flexible transfer learning:** Any pretrained encoder can be extended with a lightweight regression head, enabling recognizability prediction without retraining recognition or altering backbone design.
- **Recognizability-informed aggregation and explainability:** A natural $CCAS > 0$ cutoff and CCS-based weighting yield state-of-the-art results on BRIAR and IJB-C, generalize to body recognition, and enable encoder-grounded explainability.

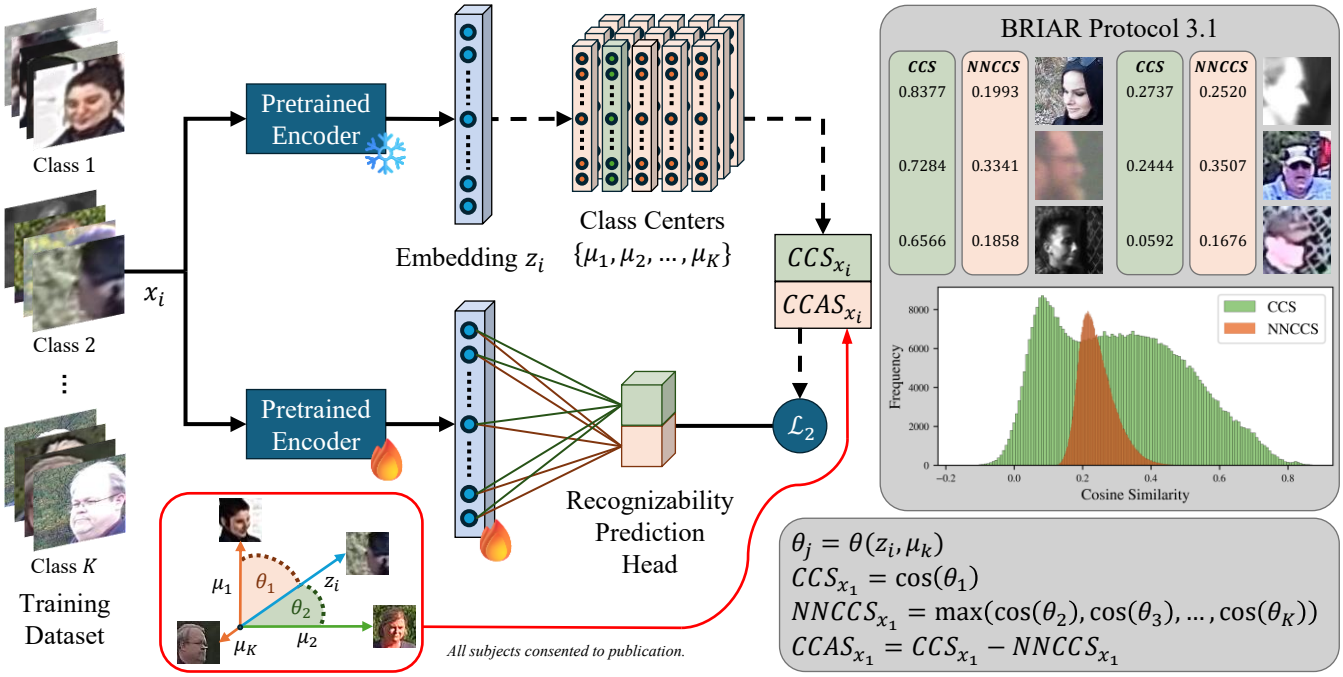

Fig. 1: **Overview of TransFIRA.** Visual quality does not reliably predict recognizability; for instance, blurred faces may still be discriminative while clear ones may fail. TransFIRA avoids such proxies by deriving recognizability labels directly from class-center similarities (CCS, $NNCCS$, $CCAS$) and fine-tuning any pretrained encoder end-to-end with a prediction head. Scores remain tied to the encoder's embedding space, enabling two key operations: *recognizability-weighted aggregation* using CCS and *principled filtering* with the natural cutoff $CCAS > 0$. This encoder-agnostic design improves robustness across challenging benchmarks (e.g., BRIAR Protocol 3.1 [5]) and generalizes naturally to other modalities (Section IV-E).

II. Related Work

Research in face image quality assessment (FIQA) has evolved along three main directions. Analytical methods estimate quality from hand-crafted image statistics such as sharpness, pose, or illumination [7, 8, 16, 22, 23]. While simple, these approaches generalize poorly because visual degradations do not consistently align with recognizability, leaving their scores detached from the decision geometry of modern recognition models.

Regression and uncertainty-based methods learn mappings from images or embeddings to continuous quality scores. FaceQNet [11] fine-tunes a ResNet-50 on automatically labeled VGGFace2 data to approximate commercial system behavior, SDD-FIQA [20] introduces an unsupervised similarity distribution distance, and SER-FIQ [24] estimates reliability via Monte Carlo dropout. While correlated with accuracy, these methods depend on curated labels, auxiliary supervision, or stochastic sampling, making them less direct and more resource-intensive.

Model-based methods integrate FIQA directly into recognition training. MagFace [19] encodes quality through embedding magnitudes, CR-FIQA [3] learns relative classifiability in angular-margin space, and GraFIQs [15] derives quality from gradient magnitudes induced by batch-normalization discrepancies. While effective, these approaches are tightly coupled to specific backbones and objectives, limiting their transferability to arbitrary encoders.

Recent work has also explored heavier alternatives. CLIB-FIQA [21] fits quality with multi-factor calibration via a vision–language framework, while DifFIQA [1] and eDifFIQA [2] perturb images with diffusion models to approximate recognizability. Though these pipelines advance the state of the art, they remain indirect proxies for encoder behavior and introduce significant computational overhead compared to lightweight embedding-based approaches.

In contrast, TransFIRA provides the first lightweight, self-supervised framework that is encoder-specific and geometrically grounded. By deriving labels directly from class-center similarity and separation, it avoids heuristic proxies, curated labels, and recognition-specific pipelines. TransFIRA adapts seamlessly to any pretrained encoder with only a regression head, produces interpretable decision-boundary–aligned scores, and uniquely extends to body recognition through calibration, establishing the first unified framework for recognizability-aware modeling across modalities.

III. Method

TransFIRA adapts a pretrained backbone to predict *recognizability*, defined by embedding-space similarity and separation that indicate whether an input will be correctly recognized. Unlike prior FIQA methods based on visual proxies, TransFIRA ties recognizability directly to the encoder's decision geometry and uses it for interpretable aggregation.

Our framework has three stages. First, we derive recognizability labels directly from embeddings by computing class-center similarities and separations (Section III-A), requiring

no human annotations or IQA supervision. Second, we fine-tune the backbone with a lightweight regression head to predict these labels from raw images (Section III-B), producing encoder-specific recognizability scores. Finally, we use these predictions for **recognizability-informed aggregation** (Section III-C), combining a natural cutoff ($CCAS > 0$) that filters unrecognizable frames with CCS-based weighting that emphasizes compact, reliable embeddings. Together, these components form a principled framework that is encoder-specific, geometrically interpretable, and lightweight.

A. Image Recognizability

To supervise recognizability prediction, we derive labels directly from encoder embeddings by comparing each image to the center of its class. This grounds recognizability in the embedding space of the chosen encoder, ensuring that the scores reflect its actual discrimination ability rather than superficial proxies such as blur or illumination. In this way, recognizability labels are obtained automatically from the encoder itself, without relying on handcrafted measures or other IQA supervision.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training dataset of N face images $x_i \in \mathbb{R}^{H\times W\times 3}$ with identity labels $y_i \in \{1, \ldots, K\}$, where K is the number of classes. A pretrained face encoder ϕ maps each image to a d-dimensional embedding:

$$z_i = \phi(x_i). \tag{1}$$

We measure the **recognizability** of an image x_i by its similarity to the **class center** of its identity. When a separate gallery set $\mathcal{G}$ is available, class centers are computed only from gallery embeddings:

$$\mu_j \;=\; \frac{1}{n_j^{\mathcal{G}}} \sum_{\substack{i\in\mathcal{G}\\ y_i=j}} z_i, \tag{2}$$

where $n_j^{\mathcal{G}}$ is the number of gallery samples for class j. If no gallery–probe split exists, class centers are computed from all embeddings of identity j in $\mathcal{D}$:

$$\mu_j \;=\; \frac{1}{n_j} \sum_{i:\, y_i=j} z_i, \tag{3}$$

where n_j is the number of samples for class j.

We define the **Class Center Angular Similarity (CCS)** as the cosine similarity between z_i and its class center μ_{y_i}:

$$CCS_{x_i} = \frac{z_i^\top \mu_{y_i}}{\|z_i\|_2 \, \|\mu_{y_i}\|_2}, \tag{4}$$

which captures how closely an embedding aligns with its own class center.

We then define the **Nearest Nonmatch Class Center Angular Similarity (NNCCS)** as the maximum similarity to any other class center:

$$NNCCS_{x_i} = \max_{j\neq y_i} \frac{z_i^\top \mu_j}{\|z_i\|_2 \, \|\mu_j\|_2}, \tag{5}$$

capturing how strongly the image resembles its most confusable impostor class. Fig. 1 illustrates how CCS and NNCCS are computed and distributed on BRIAR Protocol 3.1 [5], highlighting their role in defining recognizability.

Finally, we define the **Class Center Angular Separation (CCAS)** as their difference:

$$CCAS_{x_i} = CCS_{x_i} - NNCCS_{x_i}. \tag{6}$$

Large positive values of $CCAS_{x_i}$ indicate strong confidence in the correct class, negative values imply misclassification, and values near zero suggest ambiguity. Importantly, a natural cutoff emerges at $CCAS > 0$: the encoder assigns higher similarity to the correct class center than to any impostor, placing the sample on the correct side of the decision boundary. This provides a principled and interpretable criterion for recognizability, which we use for filtering in Section III-C.

Although CR-FIQA's **Certainty Ratio (CR)** [4] is also a function of CCS and $NNCCS$, it does not align with the encoder's decision geometry. CR is defined as a ratio of match to impostor similarity with an added constant in the denominator, which distorts the underlying margin. For example, consider two samples with identical separations $CCS - NNCCS = 0.1$: one with $(CCS, NNCCS) = (0.9, 0.8)$ and another with $(0.3, 0.2)$. Both are equally well-separated in the embedding space ($CCAS = 0.1$), yet their CR values differ sharply (0.5 versus 0.25), leading to inconsistent judgments of recognizability.

By contrast, both CCS and $CCAS$ map directly to geometric properties of the encoder's embedding space: CCS reflects alignment with the correct class center, and $CCAS$ reflects separation from impostors. The intrinsic cutoff $CCAS > 0$ therefore provides the first principled definition of recognizability: a sample is recognizable precisely when it lies closer to its own class center than to any impostor. This direct interpretability makes CCS and CCAS not only more faithful to the encoder's decision rule but also more transparent for explainability analyses.

This formulation ties recognizability labels to the encoder's embedding geometry, ensuring that both labels and thresholds reflect its actual discrimination ability. Because it depends only on embeddings and class centers, it generalizes naturally across modalities. In Section IV-E, we extend it to body recognition, where training details and model characteristics cause raw CCS and NNCCS to saturate near one; sigmoid calibration restores variation, yielding well-behaved scores and preserving CCAS as a reliable measure of separability. Appendix A further shows how CCS and CCAS provide encoder-grounded explainability, revealing how degradations like Gaussian blur affect recognizability. Together, CCS, NNCCS, and CCAS offer a geometrically grounded basis for recognizability, independent of handcrafted quality labels or supervision from other IQA methods, and always specific to the encoder in use.

B. Recognizability Prediction Network

Because recognizability labels are deterministic once embeddings are computed, they can be precomputed and reused during training. We predict recognizability directly from images by extending the pretrained encoder ϕ with a lightweight **recognizability prediction head** h_ψ, implemented as a multi-layer perceptron (MLP). Given an input x_i,

the model outputs a two-dimensional recognizability vector:

$$\hat{\mathbf{r}}_i = [\hat{CCS}_{x_i}, \hat{CCAS}_{x_i}]^\top = h_\psi(\phi(x_i)). \tag{7}$$

The network is trained end-to-end, with both the backbone and prediction head optimized jointly, using mean squared error against ground-truth recognizability labels:

$$\mathbf{r}_i \equiv [CCS_{x_i}, CCAS_{x_i}]^\top, \tag{8}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{r}}_i - \mathbf{r}_i\|_2^2. \tag{9}$$

By explicitly regressing CCS and CCAS, recognizability is tied to the discrimination ability of the *chosen encoder*. The same image may be highly recognizable to one model yet ambiguous to another, depending on their embedding spaces. This encoder-specific property is central: predicted scores reflect the behavior of the deployed system rather than generic image quality. Fine-tuning the entire backbone alongside the prediction head ensures recognizability is learned as an integrated extension of the encoder, rather than as a detached add-on.

As shown in Table I, full end-to-end training with pretrained initialization provides clear gains over both random initialization and head-only fine-tuning. The models converge after only a single epoch when initialized from a pretrained backbone, compared to roughly 20 epochs without, underscoring the efficiency benefits of transfer learning. The same table also ablates recognizability supervision with CCS, CCAS, and CR-FIQA's Certainty Ratio (CR) [4], where CR is computed from CCS and NNCCS. Among these, CCS provides the most stable supervisory signal for training, while CCAS offers a principled margin-based criterion that we later exploit for filtering. By contrast, CR's ratio form distorts the underlying margin and leads to less reliable supervision, making CCS a more effective alternative.

This embedding-space formulation distinguishes our approach from prior FIQA methods that rely on visual proxies [12, 13, 15, 20], perturbation-based uncertainty [24], or generative/multimodal pretraining [1, 2, 21]. Unlike methods that entangle recognizability with recognition [4, 19], our framework decouples the two tasks, preserving recognition accuracy while learning recognizability as a separate, embedding-grounded signal. This two-stage design makes the approach encoder-specific, geometrically interpretable, and readily extensible across modalities. Importantly, because recognizability is predicted at the image level, it can also serve as an *explainability signal*: by exposing how the encoder itself perceives reliability under perturbations, our method enables analyses that were previously infeasible. As shown in Appendix A, predicted CCS and CCAS track recognizability dynamics, providing a principled way to identify hard-to-recognize subjects, anticipate failure cases, or guide data augmentation strategies for robustness.

We refer to this framework as **TransFIRA (Transfer Learning for Face Image Recognizability Assessment)**. It offers a general-purpose pipeline in which any pretrained backbone, regardless of architecture or training set, can be extended with recognizability prediction through end-to-end fine-tuning. Training reuses the backbone's existing components (e.g., normalization layers, dropout, augmentation) while adding only a lightweight regression head. Because labels are derived automatically from CCS and CCAS, no human annotation, IQA supervision, or architectural changes are required. The resulting models integrate seamlessly into downstream systems, enabling recognizability-informed aggregation (Section III-C), cross-modality generalization (Section IV-E), and encoder-grounded explainability (Appendix A). Taken together, CCS, CCAS, and their predictions establish the first framework that is (i) encoder-specific, (ii) grounded in decision geometry, and (iii) decoupled from recognition itself – making image-level recognizability predictions a principled foundation for both template aggregation and explainable, more reliable recognition systems.

TABLE I: **Ablation by training scope and predicted label.** Evaluation on BRIAR Protocol 3.1 [5] with CosFace [25], reporting Spearman correlation (SC) with ground truth and FNMR–ERC AUC at fixed target FMRs. *P* indicates whether the backbone is pretrained, *B* and *H* whether the backbone and head are finetuned, *Score* denotes the label being regressed. The highest SC and lowest AUCs are **bolded**.

Method				BRIAR [5]: FNMR-ERC AUC			
P	B	H	Label	SC	10^{-3}	10^{-4}	10^{-6}
✓		✓	CCS	0.4858	0.2422	0.4020	0.6278
	✓	✓	CCS	0.7448	0.1791	0.2990	0.4994
✓	✓	✓	CR [4]	0.8553	0.1549	0.2739	0.4770
✓	✓	✓	CCAS	0.8230	0.1558	0.2734	**0.4752**
✓	✓	✓	CCS	**0.8566**	**0.1536**	**0.2722**	0.4753

C. Recognizability-Informed Template Aggregation

Image-level recognizability predictions are especially valuable for template aggregation, the standard protocol in BRIAR [5], IJB-C [18], and real-world surveillance, where subjects are represented by many heterogeneous frames. In this setting, even a few unrecognizable images can corrupt the representation and degrade accuracy. A template is typically reduced to a single feature vector by averaging, but uniform weighting treats all frames equally, leaving the system vulnerable to poor-quality inputs. TransFIRA instead exploits recognizability predictions to filter and weight frames, producing cleaner and more discriminative templates.

We first apply **recognizability-based filtering**, retaining only images with predicted $\hat{CCAS}_{x_i} > 0$. As established in Section III-A, this condition means the sample is predicted to lie on the correct side of the encoder's decision boundary, so it can be considered recognizable. Unlike tuned thresholds or external quality scores, this natural cutoff is parameter-free, interpretable, and intrinsic to the deployed encoder. We evaluate filtering with CR-FIQA in Appendix D.

Next, we perform **recognizability-weighted aggregation**, where each retained embedding is scaled by its predicted $\hat{CCS}_{x_i}$ before averaging:

$$\hat{\mu}^{(s)} = \frac{\sum_{k\in\mathcal{T}^{(s)}} \hat{CCS}_{x_k}\, z_k}{\sum_{k\in\mathcal{T}^{(s)}} \hat{CCS}_{x_k}}, \tag{10}$$

where $\hat{\mu}^{(s)}$ is the aggregated template representation for

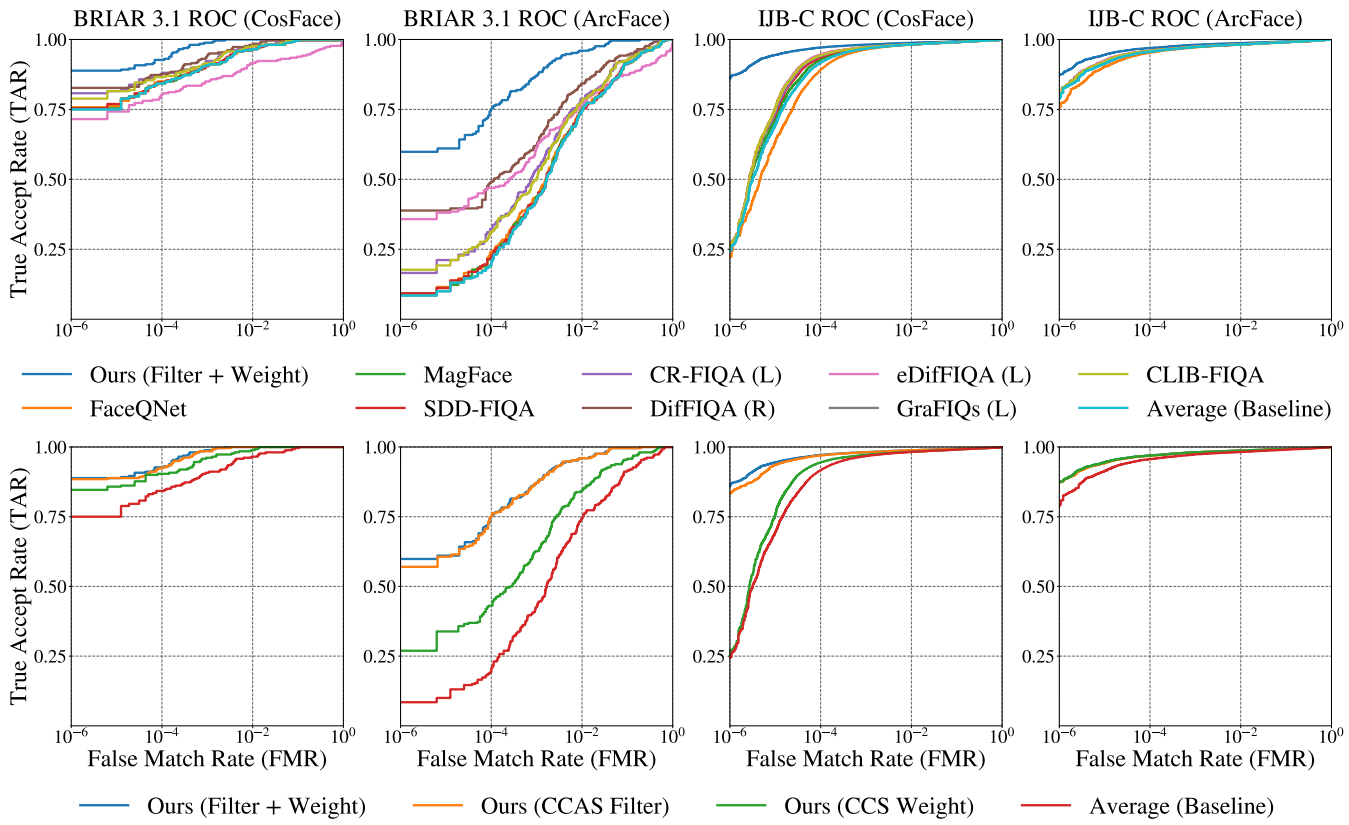

Fig. 2: **Overall ROC analysis.** Top: template-level ROC comparisons across different IQA methods; for clarity, only the strongest variant of each is shown. Bottom: ablation study illustrating the individual and combined effects of CCAS-based filtering and CCS-based weighting. Metrics correspond to Table II, and *Average* denotes uniform mean aggregation.

subject s, and $\mathcal{T}^{(s)}$ denotes the set of images retained after filtering. Because CCS measures proximity to the class center, weighting emphasizes samples that are more compactly embedded within their class, directly reflecting intra-class consistency in the encoder's embedding space. Rather than relying on visual heuristics or external IQA models, our approach derives weights directly from the deployed encoder itself, ensuring that aggregation is guided by the same geometry in which recognition is performed.

Filtering and weighting are complementary operations that together define a recognizability-informed aggregation strategy – their individual and combined effects are ablated in Fig. 2 and Table II. First, filtering introduces the *first principled, parameter-free cutoff*: $CCAS > 0$, which indicates that a frame lies closer to its own class than to any impostor. Second, weighting uses CCS to measure proximity to the class center, favoring embeddings that are more compactly aligned with their class. Unlike prior aggregation schemes that rely on appearance cues or external proxies, both operations are intrinsic to the deployed encoder and map directly onto geometric properties of its embedding space. This grounding makes the method not only effective but also interpretable, as filtering and weighting can be explained transparently in terms of encoder geometry (Appendix A) and extended to other modalities (Section IV-E).

In effect, TransFIRA introduces the first aggregation strategy that is simultaneously principled, encoder-specific, and geometry-driven. It forms a two-layer safeguard for template construction: filtering removes harmful frames, while weighting amplifies the most reliable ones. The result is both stronger recognition performance and clear decision criteria absent from prior aggregation methods.

IV. EXPERIMENTS

We evaluate TransFIRA across face and body recognition tasks to answer four key questions: (i) does recognizability-aware aggregation improve template verification accuracy, (ii) do predicted scores align with ground-truth recognizability at the image level, (iii) how well does the framework generalize across training and evaluation domains, and (iv) can it transfer beyond faces to body recognition? Additional results on explainability (Appendix A) and cross-dataset generalization (Appendix C) complement the main experiments.

A. Datasets

We evaluate on two challenging template-based benchmarks: IJB-C and BRIAR Protocol 3.1. Recognition in these settings operates over *templates*: sets of frames from the same subject aggregated into a single representation. Template quality is highly sensitive to outliers, since even one unrecognizable frame can corrupt the aggregate and degrade accuracy. **IJB-C [18]** is the standard benchmark for unconstrained face recognition, containing diverse web and video imagery with strong open-source baselines. **BRIAR Protocol**

TABLE II: **Overall ROC performance.** TAR at fixed FMRs for BRIAR Protocol 3.1 [5] and IJB-C [18] using the backbones in Section IV-B. All baseline methods perform *weighted* aggregation using FIQA scores, while our **CCAS Filter** and **Filter + Weight** apply our $CCAS > 0$ cutoff. *Average* denotes uniform aggregation without quality weighting. For each column, the best result is **bolded and underlined**, and the second-best is **bolded**. Appendix D compares filtering with CR-FIQA [4].

Method	BRIAR Protocol 3.1 [5]: TAR at Fixed FMR						IJB-C [18]: TAR at Fixed FMR					
	CosFace (BRIAR) [5, 25]			ArcFace (WebFace) [9, 27]			CosFace (BRIAR) [5, 25]			ArcFace (WebFace) [9, 27]		
	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}
Average (Baseline)	0.9077	0.8423	0.7500	0.4269	0.1923	0.0846	0.9675	0.9168	0.2440	0.9734	0.9564	0.7881
FaceQNet [10, 12]	0.9077	0.8500	0.7577	0.4385	0.2385	0.0923	0.9642	0.8912	0.2224	0.9717	0.9529	0.7567
MagFace [19]	0.9077	0.8423	0.7538	0.4385	0.2154	0.0923	0.9694	0.9280	0.2504	0.9736	0.9575	0.7970
SDD-FIQA [20]	0.9038	0.8462	0.7538	0.4269	0.2154	0.0923	0.9708	0.9361	0.2525	0.9738	0.9590	0.7977
CR-FIQA (S) [4]	0.9038	0.8462	0.7577	0.4538	0.2577	0.1077	0.9721	0.9434	0.2522	0.9742	0.9595	0.7966
CR-FIQA (L) [4]	0.9231	0.8731	0.8077	0.5346	0.3231	0.1654	0.9800	0.9617	0.3711	**0.9811**	**0.9700**	**0.8726**
DifFIQA (R) [1]	0.9423	0.8769	0.8269	0.6192	0.4885	0.3885	0.9720	0.9447	0.2499	0.9739	0.9596	0.7972
eDifFIQA (S) [2]	0.8885	0.8462	0.7962	0.5500	0.3731	0.2808	0.9720	0.9454	0.2531	0.9739	0.9597	0.7958
eDifFIQA (M) [2]	0.6769	0.5923	0.4423	0.3962	0.3000	0.2115	0.9717	0.9482	0.2518	0.9740	0.9597	0.7956
eDifFIQA (L) [2]	0.8500	0.8000	0.7154	0.6154	0.4692	0.3577	0.9721	0.9468	0.2529	0.9739	0.9599	0.7971
GraFIQs (S) [15]	0.9077	0.8423	0.7500	0.4269	0.1923	0.0846	0.9676	0.9175	0.2441	0.9735	0.9565	0.7884
GraFIQs (L) [15]	0.9077	0.8423	0.7500	0.4269	0.1923	0.0846	0.9675	0.9169	0.2440	0.9734	0.9564	0.7881
CLIB-FIQA [21]	0.9154	0.8654	0.7885	0.5038	0.3115	0.1769	0.9714	0.9428	0.2513	0.9740	0.9594	0.7979
Ours (CCAS Filter)	**0.9846**	**0.9231**	**0.8846**	**0.8715**	**0.7510**	**0.5703**	**0.9811**	**0.9697**	**0.8334**	**0.9811**	0.9685	0.8712
Ours (CCS Weight)	0.9615	0.9038	0.8462	0.6269	0.4308	0.2692	0.9718	0.9447	0.2512	0.9737	0.9578	0.7970
Ours (Filter + Weight)	**0.9885**	**0.9269**	**0.8885**	**0.8715**	**0.7510**	**0.5984**	**0.9815**	**0.9711**	**0.8601**	0.9809	**0.9693**	**0.8724**

3.1 [5] targets surveillance scenarios with extreme degradations. Because it is private, open-source encoders have never been exposed to its distribution, making it uniquely representative of deployment. In this setting, conventional IQA cues provide little signal because most frames appear degraded, yet high-capacity encoders still achieve strong recognition, underscoring the need for encoder-specific FIQA.

Traditional datasets such as LFW [14] are saturated at near-perfect accuracy, where nearly all images are trivially recognizable, and thus no longer differentiate FIQA methods. In contrast, IJB-C and BRIAR stress-test recognizability prediction in complementary ways: IJB-C through diverse unconstrained imagery, and BRIAR through severe real-world degradations. Together, they move beyond proxy benchmarks to expose the limitations of methods that rely on visual cues or generic training objectives. This dual evaluation positions FIQA not just as an academic exercise but as a foundation for robust recognition in deployment, highlighting the need for encoder-specific approaches like TransFIRA that remain faithful to the decision geometry of the deployed system.

B. Experimental Setup

We evaluate with two complementary backbones aligned with our table headings. The first, **CosFace (BRIAR)**, is a Swin-B Transformer trained with CosFace on the private BRIAR dataset, achieving state-of-the-art results on BRIAR Protocol 3.1 [5, 17, 25]. The second, **ArcFace (WebFace)**, is the open-source iResNet50 from the InsightFace repository, trained on WebFace12M, which provides the strongest IJB-C performance among publicly available InsightFace models [6, 9, 18, 27]. Each backbone is extended with a lightweight regression head for CCS and CCAS, trained end-to-end with mean squared error loss under a unified protocol: disjoint splits (a subset of BRIAR for BRIAR evaluations, 10,000 WebFace12M identities for IJB-C), 50 epochs of AdamW (1×10^{-6} learning rate, 1×10^{-4} weight decay, batch size 64), and checkpoint selection by validation Spearman correlation. This setup isolates the impact of recognizability-aware modeling while spanning both a domain-matched SOTA benchmark and a fully reproducible public baseline.

C. Recognizability-Aware Template Aggregation

Template aggregation is the primary evaluation protocol in BRIAR and IJB-C, and is critical for real-world applications where recognition must operate over sets of frames rather than single images. Fig. 2 and Table II evaluate whether recognizability-aware filtering and weighting improve verification accuracy relative to existing FIQA methods.

On BRIAR Protocol 3.1, CCS-weighting achieves the strongest unfiltered performance across operating points, already surpassing prior FIQA methods. Filtering with the intrinsic cutoff $CCAS > 0$ provides even larger improvements, and combining filtering with weighting delivers the best results overall. TAR at 10^{-6} FMR improves by +0.1385 with Swin and more than triples with ArcFace compared to uniform averaging. To our knowledge, these are the largest FIQA-driven gains yet reported in surveillance settings.

On IJB-C, CCS-weighting remains highly competitive for both backbones, while our CCAS-filtering and combined variants achieve the best results with Swin. For ArcFace, TransFIRA is either the strongest or nearly indistinguishable from the best baseline, trailing CR-FIQA by only 0.001 TAR in some cases. This narrow difference is likely due to training alignment, since both ArcFace and CR-FIQA are optimized on WebFace12M. More broadly, the results suggest that FIQA methods aligned with an encoder's training distribution

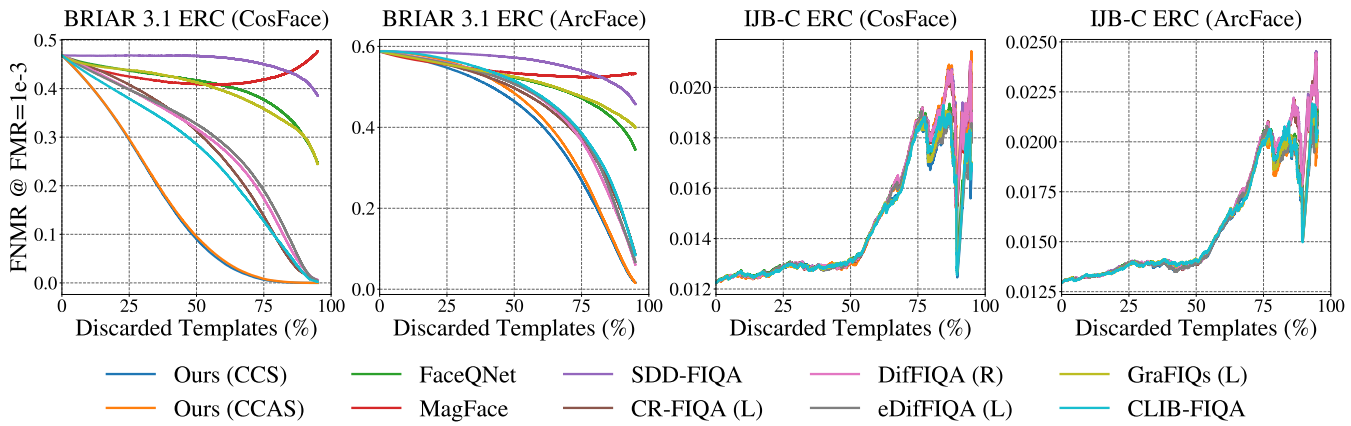

Fig. 3: **Image-level ERC curves at a target FMR of 10^{-3}.** Curves closer to the bottom-left indicate better trade-offs between discarding fraction and FNMR. AUCs and Spearman correlations are reported in Table III, with stricter operating points (10^{-4} and 10^{-6}) reported in Appendix E. For clarity, Only the strongest variant of each method is shown.

TABLE III: **Image-level recognizability evaluation.** Image-level Spearman correlation (SC) and FNMR–ERC AUC at fixed FMRs for BRIAR Protocol 3.1 [5] and IJB-C [18], using the backbones in Section IV-B. SC is computed against CCS for all methods except for **Ours (CCAS)**, where it is computed against CCAS. For each dataset and backbone, the method with the highest SC or lowest AUC is highlighted in **bold and underlined**, while the second-best is shown in **bold**.

Method	BRIAR Protocol 3.1 [5]: FNMR-ERC AUC								IJB-C [18]: FNMR-ERC AUC							
	CosFace (BRIAR) [5, 25]				ArcFace (WebFace) [9, 27]				CosFace (BRIAR) [5, 25]				ArcFace (WebFace) [9, 27]			
	SC	10^{-3}	10^{-4}	10^{-6}	SC	10^{-3}	10^{-4}	10^{-6}	SC	10^{-3}	10^{-4}	10^{-6}	SC	10^{-3}	10^{-4}	10^{-6}
FaceQNet [10, 12]	0.1589	0.3957	0.5225	0.6903	0.1194	0.5037	0.7606	0.8924	0.3607	0.0143	0.0438	0.6716	0.3677	0.0154	0.0288	0.1239
MagFace [19]	0.1367	0.4314	0.5616	0.7336	0.0467	0.5426	0.8054	0.9200	0.5684	0.0144	0.0439	0.6663	0.6030	0.0156	0.0290	0.1237
SDD-FIQA [20]	0.0293	0.4513	0.5771	0.7335	0.0112	0.5527	0.8069	0.9148	0.5374	0.0144	0.0441	0.6668	0.5682	0.0156	0.0291	0.1242
CR-FIQA (S) [4]	0.1107	0.4199	0.5463	0.7093	0.0570	0.5409	0.7874	0.8975	0.4218	0.0142	0.0436	0.6700	0.4589	0.0154	0.0287	0.1232
CR-FIQA (L) [4]	0.4143	0.2708	0.3839	0.5640	0.3654	0.4387	0.6815	0.8386	0.5571	0.0144	0.0440	0.6670	0.5902	0.0156	0.0290	0.1237
DifFIQA (R) [1]	0.4416	0.2782	0.3820	0.5490	0.4657	0.4375	0.6622	0.8194	0.4729	0.0144	0.0440	0.6688	0.5193	0.0156	0.0290	0.1240
eDifFIQA (S) [2]	0.3610	0.2978	0.4064	0.5763	0.2577	0.4644	0.7035	0.8433	0.5236	0.0144	0.0439	0.6672	0.5663	0.0156	0.0290	0.1236
eDifFIQA (M) [2]	0.2870	0.3249	0.4267	0.5836	0.2265	0.4517	0.6895	0.8316	0.5277	0.0146	0.0444	0.6683	0.5715	0.0158	0.0292	0.1248
eDifFIQA (L) [2]	0.4377	0.2840	0.3855	0.5489	0.4415	0.4407	0.6658	0.8198	0.5344	**0.0141**	**0.0433**	0.6674	0.5803	0.0153	0.0286	**0.1229**
GraFIQs (S) [15]	0.0350	0.4363	0.5646	0.7344	-0.0194	0.5325	0.8102	0.9287	0.1050	0.0143	0.0438	0.6793	0.0911	0.0154	0.0288	0.1237
GraFIQs (L) [15]	0.1831	0.3881	0.5059	0.6732	0.1434	0.5074	0.7643	0.8914	0.3307	0.0142	0.0436	0.6733	0.3440	0.0154	0.0287	0.1232
CLIB-FIQA [21]	0.5004	0.2536	0.3591	0.5340	0.4404	0.4508	0.6758	0.8281	0.5112	0.0142	0.0435	0.6677	0.5580	0.0154	0.0287	0.1231
Ours (CCAS)	**0.8230**	**0.1558**	**0.2734**	**0.4752**	**0.6568**	**0.4071**	**0.6335**	**0.8068**	**0.6245**	0.0144	0.0442	**0.6659**	**0.6306**	**0.0153**	**0.0285**	0.1230
Ours (CCS)	**0.8566**	**0.1536**	**0.2722**	**0.4753**	**0.6401**	**0.3934**	**0.6315**	**0.8076**	**0.6111**	**0.0141**	**0.0433**	**0.6659**	**0.6245**	**0.0153**	**0.0285**	**0.1228**

can gain slight performance advantages under in-domain evaluation. Appendix D compares filtering with CR-FIQA.

Overall, recognizability-aware aggregation establishes state-of-the-art TAR while introducing the first explicit geometry-derived filtering rule. On BRIAR, it sets a new frontier for FIQA under surveillance imagery. On IJB-C, it achieves new best results with Swin and nearly matches the strongest ArcFace-aligned baseline. By grounding both filtering and weighting in the encoder's decision geometry, TransFIRA improves not only accuracy but also interpretability, offering clear and transparent decision rules absent from proxy-based methods. Cross-dataset experiments in Appendix C further confirm that these benefits extend beyond the training domain, supporting recognizability as a general principle for robust template aggregation.

D. Image-Level ERC Analysis

While template aggregation measures downstream recognition accuracy, image-level evaluation tests whether predicted recognizability scores faithfully reflect the encoder's decision geometry. We assess this alignment using two complementary metrics: Spearman correlation (SC), which measures monotonic agreement with ground-truth recognizability, and the FNMR–ERC metric at target FMRs of 10^{-3}, 10^{-4}, and 10^{-6}, which quantifies the impact of recognizability-based filtering. Fig. 6 and Table III summarize these results, with stronger alignment indicated by higher SC and lower ERC AUC (area under the curve).

On BRIAR Protocol 3.1, TransFIRA substantially outperforms all baselines. With Swin, SC reaches 0.86, nearly double the best competing method (below 0.45), while ERC AUCs are cut almost in half. ArcFace shows the same trend, with both CCS and CCAS providing consistently faithful alignment. These findings confirm that grounding recognizability in encoder geometry captures discriminability far more accurately than visual proxies, directly explaining the large aggregation gains reported above.

On IJB-C, where diffusion-based and multimodal FIQA pipelines compete, TransFIRA again ranks first or second across nearly all metrics and backbones. CCS and CCAS deliver the highest SC and lowest ERC AUCs at most operating points, with a minor exception at 10^{-4} for Swin,

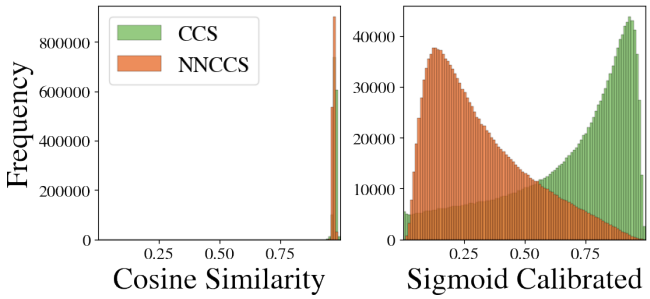

Fig. 4: **Sigmoid calibration of BRIAR [5] recognizability labels for the SemReID [26] body encoder.** Left: raw CCS/NNCCS distributions (0.97 mean). Right: sigmoid calibrated distributions (0.50 mean). Appendix B reports full metrics, including raw and calibrated CCS/CCAS variants.

where we trail eDifFIQA [2] by 0.001 AUC. Despite its simplicity, TransFIRA matches or surpasses these more complex pipelines while remaining lightweight and encoder-specific.

Overall, these results establish TransFIRA as both the most accurate and the most principled framework for image-level FIQA. By grounding recognizability in CCS and CCAS, it provides the first geometrically interpretable confidence signal without generative modeling or supervision by other IQA methods. This strong image-level fidelity naturally extends to template aggregation, confirming that TransFIRA unites accuracy, interpretability, and efficiency in a way unmatched by prior FIQA approaches. Appendix A further demonstrates how TransFIRA enables encoder-grounded explainability, revealing both failure modes and hard-to-recognize subjects.

E. Body Recognizability

To evaluate whether TransFIRA extends beyond faces, we apply it to body recognition on BRIAR Protocol 3.1 [5] using the SemReID encoder [26]. This domain is especially challenging since clothing changes, pose variation, and occlusion often overwhelm appearance-based similarity. As in our face experiments, recognizability labels are derived from cosine similarity to the correct (CCS) and nearest nonmatch (NNCCS) class centers.

Fig. 4 shows that raw CCS and NNCCS values collapse near one, providing little discriminative variation. To address this, we introduce **sigmoid calibration**, which maps CCS values toward one and NNCCS values toward zero, spreading both across the $[0, 1]$ interval. This restores meaningful variation, allowing CCAS to be computed stably and recognizability labels to reflect separability rather than saturation. While sigmoid rescaling is standard in calibration, its application here to recognizability labels is, to our knowledge, novel in the context of FIQA and body recognition.

We present the strongest calibrated CCAS results in Table IV, where weighting consistently raises TAR the most and combining filtering with weighting provides complementary gains at the strictest operating point. This indicates that *relative separability* (CCAS) is a more reliable indicator of recognizability in body recognition than absolute similarity, which tends to saturate. We therefore adopt CCAS as the primary signal for body recognizability. Appendix B reports the full results, including raw CCS and CCAS as well as calibrated CCS and CCAS variants, all of which confirm that CCAS outperforms CCS. These findings extend recognizability prediction beyond faces for the first time, establishing TransFIRA as a unified framework for modeling recognizability across modalities.

TABLE IV: **TAR at fixed FMRs for body recognition on BRIAR Protocol 3.1 [5] using the SemReID encoder [26].** The method with the best performance for each operating point is **bolded**. Full results are reported in Appendix B.

Method	TAR@1e-3	TAR@1e-4	TAR@1e-6
Average (Baseline)	0.5934	0.3278	0.0851
Calibrated CCAS Filter	0.6037	0.3485	0.0933
Calibrated CCAS Weight	**0.6452**	**0.3693**	0.1037
Calibrated Filter + Weight	0.6328	0.3568	**0.1079**

V. LIMITATIONS AND FUTURE WORK

While TransFIRA is compact in that it adds only a lightweight regression head, it still relies on transfer learning with a full backbone, which ties recognizability prediction directly to the pretrained encoder. A natural alternative is to distill recognizability into smaller student models, enabling lighter deployment at the cost of slower convergence and weaker alignment with the encoder's geometry. Beyond biometrics, encoder-grounded reliability prediction can extend to other domains where decisions depend on embedding quality like reconstruction. Our explainability experiments in Appendix A further suggest that recognizability predictions are not only interpretable but also actionable. They can flag hard-to-identify subjects, reveal operational failure modes such as blur or occlusion, and guide targeted data collection, highlighting their potential as practical tools for system monitoring and robustness.

VI. CONCLUSION

We introduced **TransFIRA**, a transfer learning framework that redefines FIQA by grounding recognizability directly in the geometry of the deployed encoder. Unlike prior methods that rely on visual proxies, heuristics, or generative pipelines, TransFIRA derives supervision from embeddings through class-center similarity and angular separation, offering the first self-supervised and geometrically interpretable foundation for recognizability.

Our approach contributes three advances: a principled filtering rule via the natural decision-boundary cutoff ($CCAS > 0$), CCS-based weighting for robust template aggregation, and encoder-grounded explainability that exposes how degradations alter recognizability. Experiments on BRIAR [5] and IJB-C [18] show clear state-of-the-art gains, while extensions to body recognition through sigmoid calibration demonstrate robustness across modalities. By closing the gap between visual quality and recognizability, TransFIRA establishes the first annotation-free, geometry-driven framework for recognizability-aware modeling, advancing recognition in both accuracy and interpretability.

ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

ETHICAL IMPACT STATEMENT

This research was conducted in accordance with the FG 2026 Ethical Impact Guidelines. We train our models on the BRIAR dataset [5] and WebFace12M [27], and evaluate on both the BRIAR Protocol 3.1 benchmark and the IJB-C benchmark [18]. The BRIAR dataset, collected by IARPA under Institutional Review Board (IRB) approval, consists of surveillance imagery acquired with informed oversight and does not pose harm to the individuals represented. All BRIAR subjects whose images appear in this paper explicitly consented to publication. WebFace12M was collected from publicly available internet sources; while individual subject consent is not feasible at this scale, the dataset is broadly used in academic research and released for non-commercial research purposes. The IJB-C dataset contains unconstrained face images and is publicly available; our use complies with its licensing and ethical use standards.

No new participants were recruited for this research, and thus considerations of participant compensation and vulnerable populations are not applicable. All datasets used reflect a broad demographic spectrum without targeted oversampling.

We recognize that face recognition research carries potential risks, particularly regarding privacy, surveillance misuse, and social bias. To mitigate these risks, we restricted our study to IRB-approved and publicly available benchmarks, maintained data security, and limited analysis to aggregate recognition performance. We acknowledge that recognition technologies can have dual-use applications and emphasize that our contributions are intended solely for academic research. Any code or models released will follow community standards for reproducibility while respecting dataset licensing restrictions. The benefits of this work—advancing recognizability-aware modeling and developing methods that can make recognition systems more transparent and explainable—outweigh minimal risks. We further advocate for ethical oversight and clear communication of limitations as key strategies to prevent misuse.

REFERENCES

[1] Ž. Babnik, P. Peer, and V. Štruc. DifFIQA: Face Image Quality Assessment Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2023.

[2] Ž. Babnik, P. Peer, and V. Štruc. eDifFIQA: Towards Efficient Face Image Quality Assessment based on Denoising Diffusion Probabilistic Models. *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 2024.

[3] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5836–5845, 2023.

[4] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. Cr-fiqa: Face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5845, June 2023.

[5] D. Cornett, J. Brogan, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[7] P. Dhar, C. Castillo, and R. Chellappa. On measuring the iconicity of a face. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2137–2145, 2019.

[8] X. Gao, S. Z. Li, R. Liu, and P. Zhang. Standardization of face image sample quality. In *International Conference on Biometrics*, pages 242–251. Springer, 2007.

[9] J. Guo and J. Deng. Insightface: 2d and 3d face analysis project. `https://github.com/deepinsight/insightface`, 2021.

[10] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay. Biometric quality: Review and application to face recognition with faceqnet, 2021.

[11] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[12] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. Faceqnet: Quality assessment for face recognition based on deep learning, 2019.

[13] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[15] J. N. Kolf, N. Damer, and F. Boutros. Grafiqs: Face image quality assessment using gradient magnitudes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 1490–1499. IEEE, 2024.

[16] Z. Lijun, S. Xiaohu, Y. Fei, D. Pingling, Z. Xiangdong, and S. Yu. Multi-branch face quality assessment for face recognition. In *2019 IEEE 19th international conference on communication technology (ICCT)*, pages 1659–1664. IEEE, 2019.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[18] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[19] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021.

[20] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[21] F.-Z. Ou, C. Li, S. Wang, and S. Kwong. Clib-fiqa: Face image quality assessment with confidence calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1694–1704, 2024.

[22] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch. Automatic face quality assessment from video using gray level co-occurrence matrix: An empirical study on automatic border control system. In *2014 22nd International Conference on Pattern Recognition*, pages 438–443. IEEE, 2014.

[23] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Serfiq: Unsupervised estimation of face image quality based on stochastic

embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020.

[24] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659. IEEE, 2020.

[25] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[26] K. Zhu, Y. Huang, M. Xu, F. Zhu, F. Zheng, and R. Ji. Identity-guided human semantic parsing for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5126–5135, 2022.

[27] Z. Zhu, G. Huang, J. Deng, J. Guo, Y. Lu, J. Zhou, X. Ji, D. Gong, J. Du, Y. Deng, Z. He, C. Li, J. Huang, H. Yang, X. Li, F. Zhou, and S. Zafeiriou. 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 (CVPR)*, pages 10492–10501, 2021.

APPENDIX

This appendix presents additional analyses, including encoder-grounded explainability under perturbations, extended body recognition experiments, cross-dataset generalization, CR-FIQA filtering, and supplementary ERC curves.

A. Encoder-Grounded Explainability

A distinctive advantage of TransFIRA is that it provides the first *encoder-specific explainability signal* for FIQA. By deriving recognizability directly from embedding geometry, it reveals how the deployed model separates identities rather than relying on visual proxies such as blur, pose, or handcrafted uncertainty. This stands in sharp contrast to prior approaches, which approximate failure modes but remain detached from the encoder's decision boundary.

We illustrate this with Gaussian blur on IJB-C [18], using an ArcFace encoder finetuned on WebFace12M [9, 27]. Blur generally reduces discriminability, but its effect on recognizability is not monotonic. Fig. 5 shows cases where predictions decline smoothly, remain stable, or even improve under mild blur – when blur suppresses distractors or mitigates misalignment. These examples highlight that visual degradation does not reliably indicate recognizability, underscoring the need for encoder-grounded measures.

Table V provides quantitative evidence. While mean CCS and CCAS decrease with stronger blur, predicted values remain strongly correlated with ground truth across all conditions. Spearman correlations reach ≈0.62 for CCS and ≈0.66 for CCAS without blur and remain high under light and moderate blur, showing that predictions preserve the *relative ordering of recognizability across samples*. This property is critical for explainability: it means predictions not only capture global degradation trends but also identify which images remain robust or degrade most sharply.

A major deployment advantage is that TransFIRA extends to *unseen subjects*, including those with only a single enrollment image where ground-truth CCS/CCAS cannot be computed. In such cases, predicted recognizability is the only practical way to assess whether a subject is reliably identifiable, enabling operators to flag inherently hard-to-recognize individuals, prompt re-enrollment, or add safeguards against likely failures.

High correlation ensures that recognizability predictions reflect the encoder's own reliability rather than detached quality proxies, making explainability both scalable and actionable. Beyond subject-level monitoring, recognizability dynamics under perturbations can guide augmentation strategies for adversarial robustness and reveal operational conditions most prone to error.

Together, these findings establish TransFIRA as the first framework to provide *encoder-grounded explainability*. By aligning predictions with decision geometry, extending them to unseen subjects, and preserving relative structure under perturbations, TransFIRA yields explanations that are faithful, interpretable, and practically useful – capabilities absent from prior FIQA methods.

TABLE V: **Effect of Gaussian blur on recognizability.** Mean ground-truth (GT) and predicted (Pred) CCS and CCAS values after applying Gaussian blur of varying intensities to all images in IJB-C, along with their Spearman correlation (SC) across images. All predictions are obtained using an ArcFace encoder finetuned on WebFace12M [9, 27].

Blur Level	CCS			CCAS		
	GT	Pred	SC	GT	Pred	SC
None	0.6335	0.7436	0.6171	0.3351	0.4677	0.6594
Light	0.6338	0.7434	0.6157	0.3344	0.4676	0.6596
Moderate	0.6040	0.7050	0.5554	0.2891	0.4277	0.6314
Heavy	0.4865	0.5316	0.3143	0.0834	0.2545	0.3962

TABLE VI: **TAR at fixed FMRs for body recognition on BRIAR Protocol 3.1 [5] with the SemReID encoder [26].** For each operating point, the best result is **bolded and underlined**, and the second-best is **bolded**.

Method	TAR@1e-3	TAR@1e-4	TAR@1e-6
Average (Baseline)	0.5934	0.3278	0.0851
Raw CCAS Filter	0.5934	0.3278	0.0851
Raw CCS Weight	0.5934	0.3257	0.0851
Filter + Weight	0.5934	0.3257	0.0851
Raw CCAS Weight	0.6000	0.3340	0.0851
Filter + Weight	0.6000	0.3340	0.0851
Calibrated CCAS Filter	0.6037	0.3485	0.0933
Calibrated CCS Weight	0.6120	0.3402	0.0934
Filter + Weight	0.6203	0.3465	0.0996
Calibrated CCAS Weight	**0.6452**	**0.3693**	**0.1037**
Filter + Weight	**0.6328**	**0.3568**	**0.1079**

B. Extended Body Recognition Results

Table IV reports the full body recognition results on BRIAR Protocol3.1 with the SemReID encoder [26], complementing the main findings in Section IV-E. As shown in Fig. 4, raw CCS and NNCCS values collapse near one (mean 0.97, variance 3.69×10^{-5}), providing little discriminative variation. Sigmoid calibration spreads the distributions across $[0, 1]$ (mean 0.50, variance 0.09, Brier score 0.1564), restoring usable variation and stabilizing CCAS computation.

The extended results confirm that raw CCS- and CCAS-based weighting do not substantially alter performance relative to unweighted aggregation. In contrast, calibration makes a clear difference: calibrated CCAS weighting consistently yields the strongest results across all operating points, outperforming both raw and calibrated CCS. For instance, at 10^{-3} and 10^{-4} FMR, calibrated CCAS achieves the highest TARs, and at the strictest 10^{-6} FMR, filtering combined with calibrated CCAS weighting delivers the best overall performance. These trends demonstrate that calibration is essential when similarity scores collapse into a narrow range and lose discriminative power, with sigmoid calibration restoring discriminative variation and enabling more effective downstream use.

Taken together, these findings establish CCAS, particularly when calibrated, as the most reliable basis for recognizability prediction under conditions of degenerate similarity distributions. In contrast, CCS remains effective for face

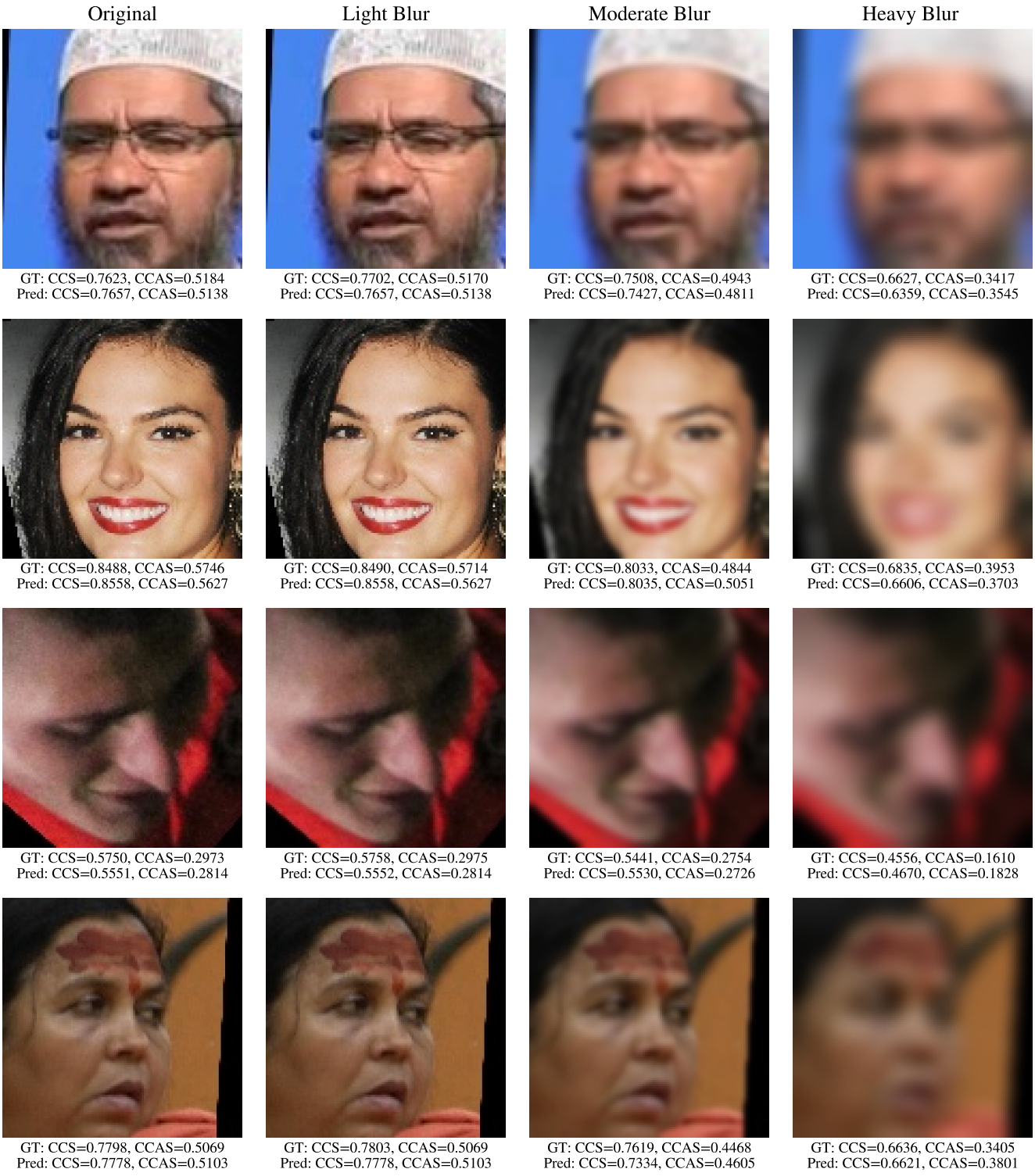

Fig. 5: **Explainability via Gaussian blur with ArcFace [9, 27] on IJB-C [18]** Across all four examples, adding *light blur* slightly **improves** recognizability, as both CCS and CCAS increase relative to the no-blur condition. In contrast, *moderate and heavy blur* produce a monotonic decline in recognizability, with the sharpest drop under heavy blur. These patterns highlight that visual degradation does not map directly to recognizability: mild blur can reduce misalignment or suppress distractors, enhancing embeddings, while stronger blur consistently erodes discriminability. This illustrates the value of encoder-grounded signals for capturing nuanced, non-monotonic effects of perturbations.

TABLE VII: **Cross-dataset generalization.** TAR at fixed FMRs when recognizability prediction is finetuned on one dataset and evaluated on another. **Backbone** indicates the base encoder used in TransFIRA, **Finetune** the dataset used for recognizability training, **Encoder** the recognition model being tested, and **Evaluate** the benchmark dataset. Rows marked with * denote domain-matched settings where finetune and evaluation datasets align. The best TAR for each evaluation configuration is **bolded**.

Backbone	Finetune	Encoder	Evaluate	TAR@1e-3	TAR@1e-4	TAR@1e-6
CosFace [25]*	BRIAR [5]	CosFace [25]	BRIAR [5]	**0.9615**	**0.9038**	**0.8462**
ArcFace [9]	BRIAR [5]	CosFace [25]	BRIAR [5]	0.9423	0.8923	0.8385
CosFace [25]	WebFace [27]	CosFace [25]	BRIAR [5]	0.9423	0.8923	0.8192
ArcFace [9]	WebFace [27]	CosFace [25]	BRIAR [5]	0.9000	0.8423	0.7500
CosFace [25]	BRIAR [5]	ArcFace [9]	BRIAR [5]	0.5846	0.3538	0.2115
ArcFace [9]*	BRIAR [5]	ArcFace [9]	BRIAR [5]	**0.6269**	**0.4308**	**0.2692**
CosFace [25]	WebFace [27]	ArcFace [9]	BRIAR [5]	0.5654	0.3423	0.1962
ArcFace [9]	WebFace [27]	ArcFace [9]	BRIAR [5]	0.4154	0.2154	0.0855
CosFace [25]	BRIAR [5]	CosFace [25]	IJB-C [18]	**0.9720**	**0.9470**	0.2493
ArcFace [9]	BRIAR [5]	CosFace [25]	IJB-C [18]	0.9716	0.9452	0.2255
CosFace [25]*	WebFace [27]	CosFace [25]	IJB-C [18]	0.9718	0.9447	**0.2512**
ArcFace [9]	WebFace [27]	CosFace [25]	IJB-C [18]	0.9697	0.9276	0.2502
CosFace [25]	BRIAR [5]	ArcFace [9]	IJB-C [18]	0.9737	0.9594	0.7970
ArcFace [9]	BRIAR [5]	ArcFace [9]	IJB-C [18]	0.9735	0.9593	0.7958
CosFace [25]	WebFace [27]	ArcFace [9]	IJB-C [18]	0.9738	0.9589	0.7978
ArcFace [9]*	WebFace [27]	ArcFace [9]	IJB-C [18]	**0.9809**	**0.9693**	**0.8724**

recognition, where similarity values are less saturated. To our knowledge, this is the first systematic study of recognizability prediction for body recognition, and the results demonstrate that encoder-grounded recognizability is modality-agnostic: while CCS dominates in face recognition, CCAS emerges as the key signal for bodies. This cross-modality flexibility highlights TransFIRA's adaptability and underscores its practical value for real-world deployments, where recognition systems must operate seamlessly across both facial and bodily inputs.

C. Cross-Dataset Generalization

While domain-matched finetuning yields the strongest results (e.g., CosFace on BRIAR [5, 25], ArcFace on WebFace/IJB-C [6, 18, 27]), Table VII shows that TransFIRA remains highly competitive even when training and evaluation domains differ. In several cases, cross-dataset models approach – or even surpass – baselines trained in-domain, underscoring that recognizability prediction generalizes more robustly than heuristic or proxy-based FIQA measures.

A particularly notable case is CosFace finetuned on BRIAR but evaluated on IJB-C, which achieves the best TAR at 10^{-3} and 10^{-4} despite the domain mismatch. This transferability arises from two factors: (i) recognizability supervision is defined by gallery–probe separation, which mirrors IJB-C's verification protocol, and (ii) CCAS emphasizes relative separability over absolute similarity, making it inherently more resilient to shifts in data distribution. These properties allow TransFIRA to retain discriminative power even when the training and evaluation domains differ substantially.

Cross-dataset generalization is especially challenging in FIQA because datasets differ not only in image quality but also in domain characteristics: BRIAR emphasizes surveillance imagery with severe degradations, while IJB-C emphasizes unconstrained but curated web and video imagery. Despite these differences, TransFIRA consistently aligns with recognition performance across domains. This robustness highlights that recognizability prediction reflects a transferable principle rooted in encoder geometry. While domain-specific finetuning provides the strongest results, the ability to generalize across datasets without retraining is crucial for deployment, where FIQA labels are rarely available for each new distribution and systems must remain reliable under shifting conditions.

D. CR-FIQA Filtering

CR-FIQA [4] introduces the Certainty Ratio (CR) as a confidence measure but does not specify a filtering rule. For completeness, we evaluate a natural candidate: $CR > 0.5$, the midpoint of the $[0, 1]$ scale covered by CR. From the definition,

$$CR = \frac{CCS}{NNCCS + 1 + \epsilon},$$

where $\epsilon = 10^{-9}$ is a numerical stabilizer, this cutoff effectively requires the genuine similarity to exceed half of the shifted nonmatch similarity. However, the addition of $1 + \epsilon$ in the denominator arbitrarily shifts the CR scale,

TABLE VIII: **CR-FIQA vs. Ours: ROC performance with filtering.** TAR at fixed FMRs for BRIAR Protocol 3.1 [5] and IJB-C [18] using the backbones in Section IV-B. We directly compare CR-FIQA [4] with a filtering threshold of $CR > 0.5$ against our encoder-grounded $CCAS > 0$ rule. The best result for each column is highlighted in **bold and underlined**, while the second-best is highlighted in **bold**. Rows labeled "Filter" (without $+$) denote filtering alone without weighting.

Method	BRIAR Protocol 3.1 [5]: TAR at Fixed FMR						IJB-C [18]: TAR at Fixed FMR					
	CosFace (BRIAR) [5, 25]			ArcFace (WebFace) [9, 27]			CosFace (BRIAR) [5, 25]			ArcFace (WebFace) [9, 27]		
	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}	10^{-3}	10^{-4}	10^{-6}
CR-FIQA (L) [4] (CR Filter)	0.9077	0.8654	0.7808	0.4731	0.2577	0.1000	0.9782	0.9496	0.3617	0.9811	0.9686	0.8711
CR-FIQA (L) [4] (CR Weight)	0.9231	0.8731	0.8077	0.5346	0.3231	0.1654	0.9800	0.9617	0.3711	0.9811	**0.9700**	**0.8726**
CR-FIQA (L) [4] (Filter + Weight)	0.9231	0.8808	0.8231	0.5577	0.3385	0.1846	0.9800	0.9620	0.3712	**0.9811**	**0.9701**	**0.8726**
Ours (CCAS Filter)	**0.9846**	**0.9231**	**0.8846**	**0.8715**	**0.7510**	**0.5703**	**0.9811**	**0.9697**	**0.8334**	**0.9811**	0.9685	0.8712
Ours (CCS Weight)	0.9615	0.9038	0.8462	0.6269	0.4308	0.2692	0.9718	0.9447	0.2512	0.9737	0.9578	0.7970
Ours (Filter + Weight)	**0.9885**	**0.9269**	**0.8885**	**0.8715**	**0.7510**	**0.5984**	**0.9815**	**0.9711**	**0.8601**	0.9809	0.9693	0.8724

meaning the 0.5 threshold has no intrinsic connection to the decision geometry. In contrast, our analytical $CCAS > 0$ rule corresponds exactly to the condition $CCS > NNCCS$ that defines correct recognition. To our knowledge, this work provides the first systematic evaluation of CR-FIQA as a filtering rule.

As shown in Table VIII, the difference is consequential. Filtering with $CR > 0.5$ adds little beyond weighting alone, whereas $CCAS > 0$ filtering improves TAR even on its own and, when combined with weighting, consistently outperforms $CR > 0.5$ across both Swin and ArcFace backbones on BRIAR and IJB-C. For instance, with Swin on BRIAR, $CCAS > 0$ filtering achieves 0.9846 TAR at 10^{-3} FMR compared to 0.9077 for $CR > 0.5$, and the combined Filter+Weight variant further boosts performance to 0.9885. These results highlight that while CR-FIQA's weighting remains strong, its implicit midpoint filtering is arbitrary and of limited value, whereas our decision-boundary–grounded $CCAS > 0$ criterion offers both interpretability and measurable gains.

E. Additional Image-Level ERC Curves

Fig. 6 extends the main analysis in Section IV-D by showing ERC curves at stricter target FMRs of 10^{-4} and 10^{-6}. The trends remain consistent: CCS- and CCAS-based predictions achieve the highest Spearman correlations and lowest AUCs across all settings. The only exception is Swin at 10^{-4}, where our method performs nearly identically to eDifFIQA (L). These results reinforce that recognizability-grounded supervision provides state-of-the-art alignment with recognition reliability, even under the strictest operating conditions.

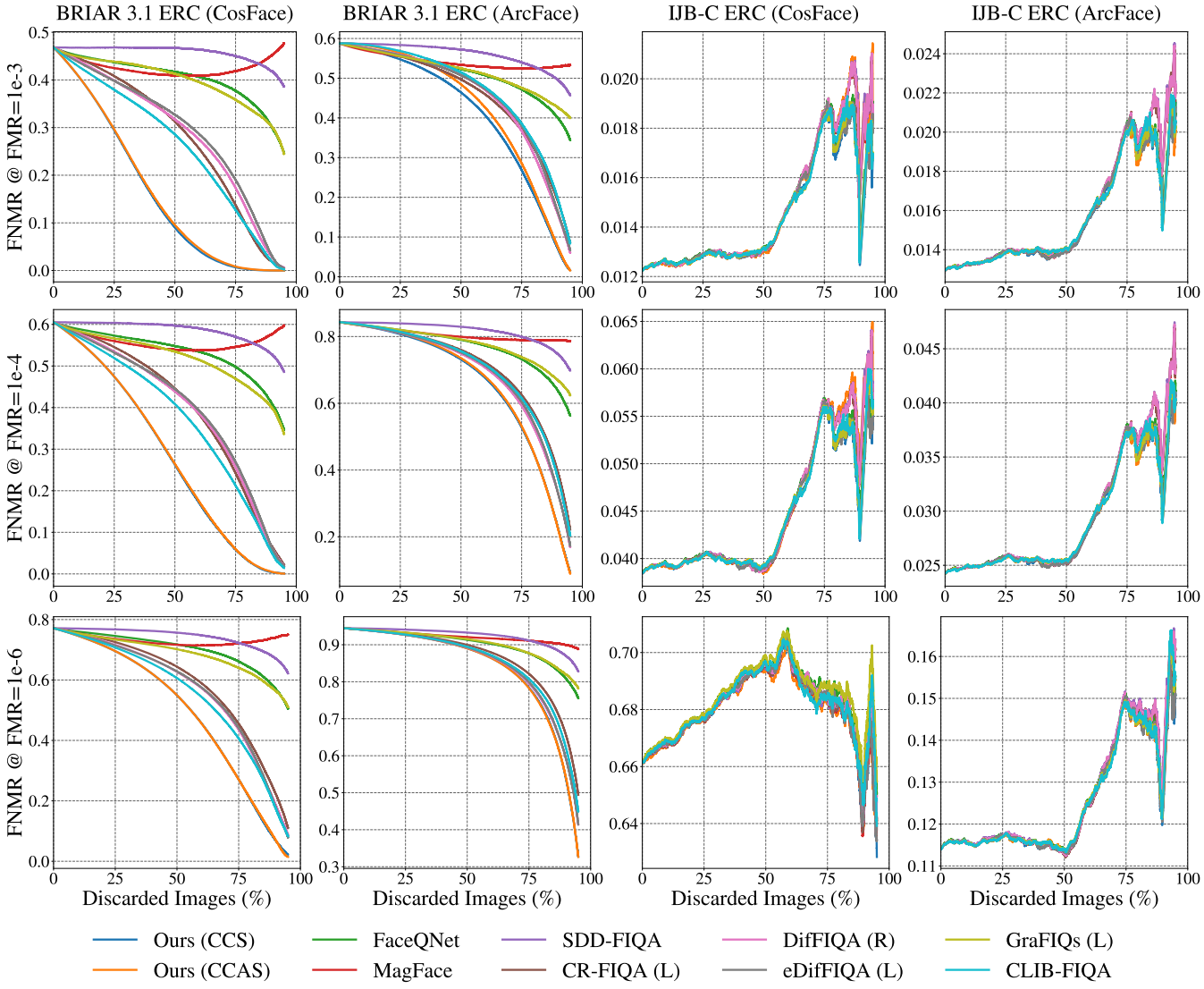

Fig. 6: **ERC analysis at multiple operating points.** Top: image-level ERC curves at a target FMR of 10^{-3}. Middle: curves at 10^{-4}. Bottom: curves at 10^{-6}. Curves closer to the bottom left indicate more effective trade-offs between discarding fraction and FNMR. Corresponding AUCs and Spearman correlations are reported in Table III. For clarity, only the strongest variant of each method is shown.