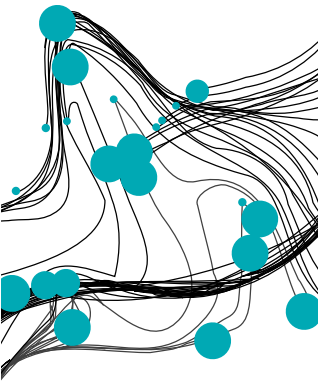
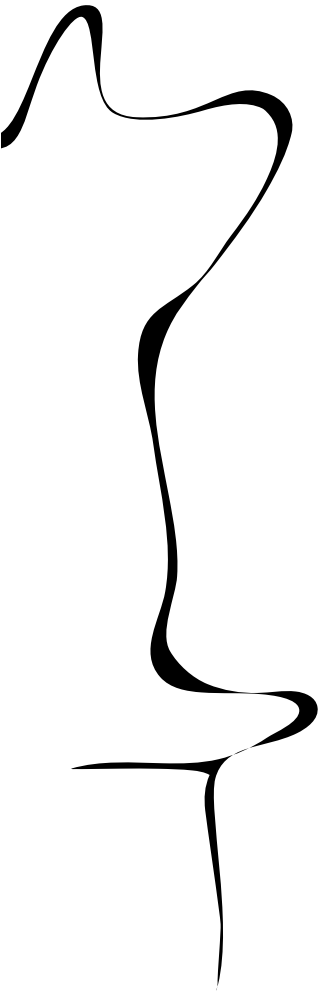


.13351

DMB

DATA MANAGEMENT AND BIOMETRICS



EVALUATING THE EFFICACY OF A NEW SYNTHETIC AI-GENERATED DATASET FOR TRAINING FACE RECOGNITION MODELS

Adham Elhabashy

BACHELOR'S ASSIGNMENT

Committee:
dr.ir. L.J. Spreeuwers
R. H. O. Ismayilov

July, 2025

Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Evaluating the Efficacy of a New Synthetic AI-Generated Dataset for Training Face Recognition Models

ADHAM ELHABASHY, University of Twente, The Netherlands

The increasing demand for large-scale facial datasets to train deep learning-based face recognition (FR) systems has raised critical concerns regarding privacy, consent, and data collection ethics. Synthetic face datasets, particularly those generated by diffusion models have proven to be a promising solution by offering diversity, scalability, and identity control without relying on real individuals. This study investigates the effectiveness of FLUXSynID—a high-resolution, diffusion-based document-style synthetic dataset for training face recognition models. We conducted a comprehensive evaluation under three experimental scenarios: full-data training, sequential learning in data-scarce settings, and hybrid training with mixed real and synthetic data. The results show that models trained on synthetic faces can match or exceed the performance of models trained on real data, particularly in high-security verification tasks and expressive test conditions. Moreover, strategically combining synthetic and real data improves generalization and bridges performance gaps caused by data imbalance. These findings highlight the viability of synthetic data as both a privacy-preserving alternative and a valuable complement to real-world datasets for modern face recognition.

1 INTRODUCTION

Over the past decade, face recognition (FR) systems have achieved remarkable advancements driven by deep learning and the availability of large-scale real-world facial datasets, such as VGGFace2[1] and MS-Celeb-1M [3]. These models have attained impressive performance in both identification and verification tasks, particularly when trained on large, diverse datasets of real human faces [7, 15]. However, collecting and using real facial data at scale raises significant ethical, legal, and privacy concerns, and it is also very difficult to gather comprehensive real datasets. Consequently, there is a growing interest in using synthetic face datasets generated by generative models as a scalable, privacy-friendly alternative [8]. Synthetic data offer the ability to produce large quantities of training images without violating the privacy of real people, as well as an easy alternative to gather diverse experiment-specific data.

Early efforts in synthetic face generation were dominated by Generative Adversarial Networks (GANs), particularly StyleGAN, which demonstrated the capability to produce high-quality, photo-realistic face images [5]. Recently, diffusion models have made a significant advancement in generative AI, enabling the creation of highly diverse and realistic facial images [2, 4]. The advancement of diffusion models presents an opportunity to re-evaluate the current state-of-the-art training of FR models on synthetic data, potentially overcoming the limitations observed in current GAN-based datasets. In this research, we utilize *FLUXSynID* [24], a new synthetic dataset generated using the FLUX.1 diffusion model[10]. FLUXSynID offers a variety of a high-resolution synthetic identities and is originally developed to support research in Morph Attack Detection (MAD)

TScIT 43, July 4, 2025, Enschede, The Netherlands

2 PROBLEM STATEMENT

Despite significant advancements in face recognition driven by large-scale real-world datasets, ethical and privacy constraints greatly limit the collection and use of such data. Synthetic datasets offer a promising alternative for training face recognition models, as they provide a scalable and privacy-preserving substitute for real faces. However, prior work has shown a performance gap, when training FR models exclusively on GAN-generated images. In this study, we attempt to evaluate this gap systematically by comparing FR models trained on the FLUXSynID diffusion-based synthetic dataset against those trained on equally sized subsets of real data. By evaluating both scenarios of full datasets with maximum identities available and limited-data scenarios, we aim to evaluate how synthetic and real data influence recognition performance and draw actionable conclusions for privacy-preserving, data-efficient FR model development.

3 RELATED WORK

3.1 Synthetic vs. Real Data in Face Recognition

Current efforts to use fully synthetic datasets for face recognition have revealed a notable performance gap compared to real datasets. Qiu et al. [32] reported that models trained on GAN-generated faces (SynFace) achieved around 92% accuracy on LFW, whereas models trained on real datasets gave 99% [7]. Similar scenarios were observed with other GAN-based datasets, such as DigiFace-1M [9] and SFace [33], which is mainly due to limited intra-class variation and subtle generative artifacts.

Recent advances in generative modeling narrowed this gap. StyleGAN2 [25] enables high-fidelity identity-preserving face image synthesis, and diffusion-based models have since improved realism and diversity further [2, 4]. Notably, DCFace [26] introduced dual conditioning to control identity and facial attributes, while IDiff-Face [11] achieved up to 98.0% verification accuracy on LFW. Xu et al. [36] extended this with ID^3, enhancing intra-class diversity and identity consistency.

Furthermore, SynthDistill [34] uses a real-pretrained teacher to train models on purely synthetic faces, reaching 99.5% LFW accuracy. Such methods demonstrate that modern synthetic datasets, coupled with effective training strategies, can rival real-data performance. However, this requires altering the standard face recognition model architecture to work with synthetic data.

3.2 Hybrid Training: Mixing Real and Synthetic Data

Recent literature shows that hybrid training combining synthetic and real data is a practical strategy to achieve high performance. Qiu et al. [32] found that adding a small real subset to synthetic data

greatly improved performance. Conversely, Bae et al. [9] showed that supplementing real datasets with synthetic subsets enabled performance that is comparable to a full real dataset.

This approach gained traction in benchmarks (e.g., FRCSyn 2024 (CVPR Workshops) [12]), where top-performing models used hybrid training with diffusion-generated datasets [11, 26]. Key factors influencing success include the synthetic-to-real ratio, integration strategy, and domain alignment.

Strategies such as progressive mixing, loss weighting, consistency regularization, and domain adversarial training have been proposed to mitigate the domain shift. SynthDistill [34] is a notable example that uses a real-data teacher model to guide hybrid learning.

4 RESEARCH QUESTIONS

To evaluate the efficacy of using the FLUXSynID datasets to train face recognition models. The research is guided by the following three key research questions:

- **RQ1** Under identical training conditions, how do face recognition models trained on the synthetic (FLUXSynID) dataset compare to those trained on the real-world (VGGFace2) dataset in terms of generalization performance and robustness to expression and pose variations on unseen test sets?
- **RQ2** In low-data regimes with a limited number of identities, how does the shared similarity between a synthetic dataset (FLUXSynID) and a real-world dataset (DemorphDB-FRGC) influence on face recognition model performance under fair comparison conditions?
- **RQ3** How does combining varying ratios of real and synthetic data (Hybrid Training) impact the face recognition model performance, generalization, and ability to address characteristic variations like expression, pose, and lighting across diverse testing environments?

5 DATASETS

5.1 VGGFace2 (Real Data Subset)

VGGFace2 is a large-scale face recognition dataset originally containing 3.31 million images of 9,131 identities, with high diversity in pose, age, illumination, and background [1]. For the experiments, a balanced subset of 9,130 identities from VGGFace2 was extracted,selecting three images per identity (27,393 images in total) to ensure fairness when comparing to synthetic data. The selection prioritized frontal or near-frontal poses and well-aligned faces (see Section 6.1 for preprocessing details) to get as similar images as possible to the FLUXSynID dataset. This subset of VGGFace2 serves as the "real" data in experiments.

5.2 FLUXSynID (Synthetic Data Subset)

FLUXSynID contains 14,889 unique synthetic identities, each composed of a document-style image and three live-capture variants generated through different post-processing pipelines (LivePortrait [18], PuLID [19], and Arc2Face [29]). These identities are created using the FLUX.1 diffusion model, guided by prompts enabling high fidelity and control over identity attributes such as gender, age, and region of origin.

Compared to traditional GAN-based datasets such as StyleGAN2 or even other traditional diffusion-based datasets such as ONOT, FLUXSynID offers more realistic and varied identity representations. As shown in Figure 3, a t-SNE [6] visualization of facial features reveals that FLUXSynID identities are spread across a broader area and they overlap heavily with real datasets (FRLL [13] and CFD [28]). In contrast, GAN-based identities tend to form denser clusters with less distribution. This broader distribution highlights the diversity and realism of FLUXSynID, making it a strong candidate for training face recognition models.

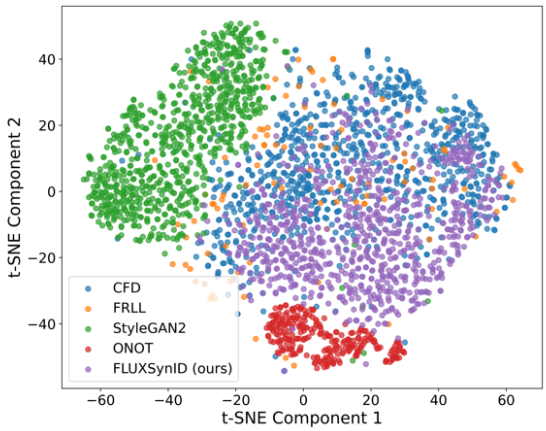


Fig. 1. t-SNE visualization showing the distribution of FLUXSynID identities, figure copied from [24]

Fig. 2. Sample of the FLUXSynID, cropped from the figure in [24]

To match the VGGFace2 subset, a subset of 9131 identities from the FLUXSynID dataset is extracted. However, to maintain diversity and reduce redundancy, we used one of the provided filtered identity lists, which contains identities retained after removing overly similar faces based on ArcFace similarity [15, 24] scoring under a

certain threshold (**0.4969**). This allowed us to select more varied and discriminative identities for training within the restricted number of identities we have to match the VGGFace2 with.

From each selected synthetic identity, we chose the three live capture images (excluding the document-style photo) to mirror the structure of the VGGFace2 subset, where three images per identity were also selected. This selection was done explicitly to ensure consistency in dataset size and per-identity image count, allowing for a fair one-to-one comparison in the experiments between real and synthetic datasets. The chosen synthetic images reflect natural variations in facial expression, pose, and lighting similar to what is observed in VGGFace2 while maintaining identity consistency. All images were originally 1024 × 1024 pixels and were resized to 112 × 112 during preprocessing (Section 6.1) to match the input requirements of the face recognition model.

5.3 DemorphDB-FRGC (FRGC Images Filtered for Document-Like Quality)

DemorphDB aggregates five facial-image collections (FRGC[38], EURECOM-IST [39], Utrecht ECVP [40], CFD [28], and FRLL [13]) and applies strict filters to retain only "document-like" images with frontal poses, removing images with closed eyes and extreme expressions.

For our experiments, we primarily used identities from the FRGC subset of DemorphDB. Following the same three-crop preprocessing as in Section 6.1, we selected all identities with at least three available images. This resulted in up to 489 genuine (unmorphed) identities, mostly from FRGC, with a small number from EURECOM-IST. These non-FRGC samples met the same quality criteria and were visually similar in terms of pose and lighting conditions.

5.4 Unseen Test Sets

To evaluate the generalization ability of the face recognition models trained on synthetic vs real data, we selected two face datasets(CFD and FRLL) as unseen test sets. These datasets were not used during training and serve to assess how well the learned facial embeddings perform on previously unseen identities.

While CFD and FRLL are not standard benchmarks for state-of-the-art face recognition (such as LFW (Labelled Faces In Wild) [23], they are highly compatible with the characteristics of our training data in terms of resolution, pose, and controlled capture conditions. This makes them suitable for isolating and comparing the effects of training on real versus synthetic data.

Furthermore, given that the primary goal of this work is not to achieve the highest possible recognition accuracy, but rather to evaluate data-driven model behavior, these test sets serve as meaningful benchmarks. They allow us to focus on comparability and performance consistency across training conditions rather than absolute accuracy on challenging public state-of-the-art scores.

- **CFD (Chicago Face Database):** A high-resolution dataset of 597 individuals photographed in a studio with consistent lighting, frontal poses, and different expressions. Images are demographically varied across ethnicity and gender. CFD provides a controlled yet demographically rich testing environment. Since facial accessories and background clutter are minimal, it evaluates how well a model performs in recognizing subtle facial distinctions. Unfortunately due to the privacy restrictions, images from the dataset cannot be showed, but it is important to highlight that CFD has various expressions for some identities not just neutral and smiling (e.g. happy, angry, excited).To further increase the dataset's demographic diversity, two diversity extensions were incorporated: the CFD-MR extension, which adds 88 multiracial individuals to enhance Middle-Eastern and North African representation, as well as the CFD-India extension, which adds 142 Indian individuals to augment South Asian coverage. These additional sets increase the total number of unique identities to 827.

  Note: In DemorphDB's CFD subset each identity has only a single image, so these were filtered out and not included in our DemorphDB-FRGC training set. Consequently, we use the full CFD collection separately as a test set.

- **FRLL (Face Research Lab London Set):** Similar to CFD, a studio based dataset of 102 adult subjects, each photographed under standardized conditions with neutral and smiling expressions. The high-resolution images are well aligned, with clean backgrounds and frontal poses. Although limited in size, FRLL's document-style format make it highly relevant for assessing performance in biometric verification scenarios such as identity verification in airports.

Fig. 3. Sample of the FRLL [13] dataset

The following table summarizes key visual attributes across the training and test datasets:

Attribute	VGGFace2	CFD	FRLL	FLUXSynID	DemorphDB-FRGC
Quality	Medium	High	High	High	High
Lighting	Varied	Controlled	Controlled	Varied	Controlled
Expression	Varied	Mostly Neutral	Neutral/Smiling	Varied	Neutral
Background	Varied	Controlled	Controlled	Varied	Controlled
Pose	Mostly Frontal	Frontal	Frontal	Frontal	Frontal

Table 1. Characteristics of training and test datasets.

6 METHODOLOGY

6.1 Preprocessing and Augmentations

To make sure all input images both real and synthetic are handled in a consistent and fair way, the same preprocessing steps is applied

across all datasets. It starts by detecting faces using the Multi-Task Cascaded Convolutional Networks (MTCNN) detector [37]. This method locates the face in the image and then crops a tight region around the face to focus on the relevant interest section. Later all images are resized to 112×112 pixels, which is a standard input size for many popular face recognition models [15].

To ensure fairness, improve generalization and increase data variation due to limited data available, identical augmentation pipelines were applied during training to both real and synthetic training images. The augmentations introduced variation while preserving facial identities:

- **Horizontal Flip:** Simulates mirrored poses.
- **Color Jitter:** Imitates changes in lighting and camera settings.
- **Affine Transforms:** Adds slight rotation, scaling, and translation.
- **Random Erasing:** Mimics occlusions like glasses or accessories.

All augmentations were applied during training using torchvision [30]. Parameters were carefully chosen to avoid excessive distortion as the data is already limited, and a fixed random seed is used to ensure consistent augmentation across runs.

6.2 Model Architecture and Loss Function

Due to the limited data available for the experiments and to ensure fairness and comparability, the experiments were conducted on a fine-tuned ResNet-50 [21] backbone pre-trained on ImageNet [14]. This choice was made instead of fine-tuning a pre-trained face model to not introduce any bias towards real data.

Backbone Architecture: We adopt a deep convolutional network based on ResNet-50 [21] as our face embedding model. ResNet-50 is a 50-layer residual network known for its strong performance on ImageNet [14]. We initialize the network with weights pre-trained on the ImageNet 1,000-class dataset, which provides a robust baseline for feature extraction. The original 1000-class classification layer is removed and replaced with a fully connected layer of size 512. Then ℓ_2 normalization is applied to the embedding. The 512-D embedding was chosen to align with the state-of-the-art face recognition frameworks (for example ArcFace also use embeddings 512-D). The ℓ_2 normalization ensures that similarity can be measured by simple dot product or cosine similarity, and it stabilizes training when using a distance-based loss. In a nutshell, the backbone network processes 112×112 RGB images and produces compact 512-D normalized embeddings.

Loss Function: Different margin-based loss functions such as ArcFace [15], CosFace [35], and SphereFace [27], are widely used in face recognition as well as metric learning losses such as triplet loss [7] and contrastive loss [20]. Margin-based softmax losses are widely used as the state-of-the-art for face recognition and serve as the standard benchmark for evaluating synthetic FR models. However, they require a large number of images per identity to learn stable class-specific representations and prevent overfitting [15, 35].

Given our dataset contains only three images per identity, such losses tend to be less stable and prone to overfitting. In contrast, triplet loss with batch-hard mining [22] does not rely on maintaining a separate classifier weight for each identity and is better suited for this scenario.

Triplet loss operates on sets of 2 images: an *anchor* face, a *positive* face of the same identity (as the anchor), and a *negative* face from a different identity. The loss works by encouraging the model to make the embedding of the anchor closer to that of a positive image (same person) than to a negative image (different person) by at least a margin α. Following FaceNet [7], we set this margin to $\alpha = 0.2$ in our experiments. In other words, for each triplet, the network tries to ensure:

$$||f(x_{\text{anchor}}) - f(x_{\text{positive}})||_2^2 + 0.2 < ||f(x_{\text{anchor}}) - f(x_{\text{negative}})||_2^2,$$

where $f(x)$ represents the embedding of image x. By applying this rule, the model learns to "pull" embeddings of the same person closer together, while simultaneously "pushing" embeddings of different people farther apart.

The batch-hard triplet mining method is used [22]; for each training batch it picks the hardest (most challenging) positives and negatives for every anchor. This way, the model focuses on the most challenging cases, which speeds up learning and makes the training more effective especially since we have only three images per identity.

6.3 Training Protocols

Experiment 1: Full Supervised Training (Flux vs. VGG)
The first experiment evaluates the efficacy of fully synthetic data versus real data for training a face recognition model from scratch (finetuning the backbone). We train two separate models under identical settings. One on the real VGGFace2 subset and one on the synthetic FLUXSynID subset, and then compare their performance on the unseen test sets. Fairness and consistency are maintained: both models use the same ResNet-50 architecture, initialization, and hyperparameters, and see the exact same number of images (9,130 identities ×3 images/identity). The only difference is the nature of the data (real vs AI-generated). This one-to-one comparison allows us to isolate the impact of synthetic training data on model effectiveness.

For each dataset (VGGFace2-real and Flux-synthetic), we train the model for the same number of epochs, ensuring both models have equivalent training iterations. We also maintain the same batch size and triplet mining strategy. This means if a batch contains for example 30 identities with 3 images each, that structure is the same whether those images are real or synthetic.

We monitored the training loss and fixed the learning rate of $1e{-}4$ for both runs. To ensure scientific reproducibility, each training run was executed with a fixed random seed for weight initialization and data shuffling. After training, models are and evaluated on CFD and FRLL. This experiment addresses the core question: can a model trained purely on synthetic faces achieve performance compared to one trained on real faces? By keeping everything else consistent, the aim is to provide a fair and a comparable result.

Experiment 2: Sequential Learning on FLUX-Style Real Data (Incremental IDs: 50 → 489)

Experiment 2 is designed to evaluate the competence of the synthetic training specifically within a domain that is really similar in the characteristics of the FLUXSynID dataset, but is constrained in the number of available real identities.Accordingly, we use DemorphDB-FRGC for comparison. The primary objective is to examine whether models maintain a performance difference as we incrementally expose them to increasing amounts of data. To ensure that any observed improvements follow a clear and explainable trend. Four pairs of comparisons are tested: **50 identities, 100 identities, 250 identities, and the full set of 489 identities.**

At each stage, models are trained independently using either FLUXSynID synthetic or DemorphDB-FRGC real identities. To ensure comparability and stability given the small data regime, we adopt a smaller learning rate of $3e{-}5$ and train with the default explained triplet loss using a ResNet-50 backbone.

To accommodate the limited number of classes per stage, we implement a custom batch sampler that draws a fixed number of identities per batch. Each stage is then trained for a fixed number of 15 epochs, providing the same total number of training steps across all data samples while validation performance is monitored to ensure this training protocol does not lead to overfitting.

The key goal of this experiment is to observe the trend of performance across the increase of identities available and determine whether synthetic pretraining retains an advantage or if real data eventually overtakes it.

Experiment 3: Hybrid Training with Mixed Datasets

This experiment investigates whether combining real and synthetic face data in different proportions improves face recognition performance by leveraging the strengths of both sources. Unlike pretraining and fine-tuning setups, the training dataset here is constructed by mixing FLUXSynID (synthetic) and VGGFace2 (real) at varying ratios from the start.

The rationale is that real images offer natural texture, noise, and expression diversity, while synthetic images contribute well-balanced and demographically controlled samples. By training on combined datasets, the model is exposed to a wider distribution of facial variations in terms of diversity such as demographics as well as visual characteristics (e.g lighting and poses), potentially improving generalization.

To ensure comparability with the models trained in Experiment 1, the training protocol is kept identical: the same ResNet-50 backbone, initialization, batch size, triplet mining strategy, epochs and learning rate. Each hybrid model sees the same total number of images and follows the same data augmentation and evaluation procedure.

The following mixing ratios of synthetic to real data were tested:

- 10% synthetic + 90% real
- 25% synthetic + 75% real
- 50% synthetic + 50% real
- 75% synthetic + 25% real
- 90% synthetic + 10% real

Each ratio was trained using the same architecture and augmentation pipeline. This design directly addresses RQ3 by examining how different proportions affect performance using a fixed learning rate of $1e{-}4$ and whether mixed datasets help capture a broader range of facial characteristics.

7 EVALUATION PROTOCOL

Model performance is evaluated using standard biometric verification metrics: **Equal Error Rate (EER)** and **True Match Rate (TMR)** at specified False Match Rate (FMR) thresholds (1% and 0.1%). These metrics are widely used in academic benchmarks, such as the NIST's Face Recognition Vendor Test (FRVT) [17], and are considered robust indicators of a system's practical verification capabilities [31].

To define these core metrics, consider the following:

- **False Match Rate (FMR)** is the proportion of impostor pairs incorrectly matched:

$$FMR = \frac{FP}{FP + TN} \tag{1}$$

- **False Non-Match Rate (FNMR)** is the proportion of genuine pairs incorrectly false-matched:

$$FNMR = \frac{FN}{TP + FN} \tag{2}$$

  Where TP, FN, FP, and TN represent True Positives, False Negatives, False Positives, and True Negatives.

The **Equal Error Rate (EER)** is the specific rate at which $FMR = FNMR$. This point on the ROC curve highlights the optimal balance between false acceptances and false rejections, providing a single overall measure of system performance. A lower EER indicates a higher likelihood of better performance.

True Match Rate (TMR) indicates the proportion of genuine pairs correctly matched:

$$TMR = \frac{TP}{TP + FN} \tag{3}$$

The **TMR at a specified FMR** (e.g., TMR@FMR = 1%) is simply the percentage of genuine matches you obtain when the decision threshold is set so that at most α% of impostors are accepted. It is especially useful in applications that require very low false-acceptance rates, such as airport security checks.

To compute these metrics, all identity pairs in the evaluation set are compared using cosine similarity between the embedding vectors. Pairs are labeled as either *genuine* (same identity) or *impostor* (different identity), and a Receiver Operating Characteristic (ROC) curve is generated. From this curve, both the EER and TMR@FMR values are derived.

For completeness, overall **accuracy** at the EER threshold is also reported. While accuracy offers an intuitive understanding of general performance, it is less informative in this case, as a model may achieve high accuracy by favoring the majority class [31]. Therefore, accuracy is presented as a supplementary metric rather than a primary evaluation criterion.

8 RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the three core experiments conducted to evaluate the efficacy of the synthetic FLUXSynID dataset in comparison to real data. Each experiment addresses one of the key research questions (RQ1–RQ3) related to

full-data training, low-data regimes sequential training, and hybrid dataset mixing. For each experiment, you can find the relevant plot summary in Appendix C

8.1 RQ1: Full Training Results (FLUXSynID vs. VGGFace2)

Table 2. Performance of models trained exclusively on VGGFace2 vs. FLUXSynID across FRLL and CFD.

Trained On	Test Dataset	EER	Accuracy	TMR@FMR=1e-2	TMR@FMR=1e-3
VGGFace2	FRLL	**0.0048**	**0.9952**	**0.9902**	0.7941
FLUXSynID	FRLL	0.0070	0.9930	**0.9902**	**0.9412**
VGGFace2	CFD	0.0535	0.9465	0.7082	0.3760
FLUXSynID	CFD	**0.0405**	**0.9595**	**0.8607**	**0.6048**

Table 2 shows the results from experiment 1 regarding RQ1 by comparing models trained exclusively on FLUXSynID or VGGFace2. Results on FRLL demonstrate that both datasets support high accuracy and low error rates with VGGFace2 achieving lower EER. However, FLUXSynID achieves notably higher at FMR=1e-3, indicating improved precision in higher security thresholds.

On the more expressive CFD dataset, FLUXSynID outperforms VGGFace2 across all metrics. This suggests that the inherent expression diversity within FLUXSynID provides better generalization to expressive, real-world faces. The alignment in pose and image quality between FLUXSynID and the test datasets further supports its strong performance.

These findings reveal that synthetic data, when generated with modern diffusion models, can compete with real-world "in-the-wild" datasets in full-data training scenarios. However, this does not show much difference as it is on a higher level and there are several factors related to datasets difference that may cause this. Therefore to narrow the scope the following experiment should provide a more solid conclusion regarding performance difference.

8.2 RQ2: Sequential Learning (DemorphDB-FRGC vs. FLUXSynID)

Table 3. TMR@FMR=1e-2 performance of sequential learning across DemorphDB-FRGC and FLUXSynID at varying identity counts.

Metric	ID Count	Test Dataset	DemorphDB-FRGC	FLUXSynID
TMR@FMR=1e-2	50	FRLL	0.5980	**0.6275**
	100	FRLL	0.6275	**0.6667**
	250	FRLL	**0.7549**	0.7255
	489	FRLL	**0.8333**	0.8137
	50	CFD	0.1963	**0.4244**
	100	CFD	0.2182	**0.4841**
	250	CFD	0.4058	**0.5391**
	489	CFD	0.4668	**0.5517**

This experiment investigates the performance of synthetic data on a smaller scope (RQ2), comparing FLUXSynID with the smaller, high-quality DemorphDB-FRGC dataset. Both datasets are comparable in terms of diversity and visual characteristics. Identity counts of 50, 100, 250, and 489 were evaluated.

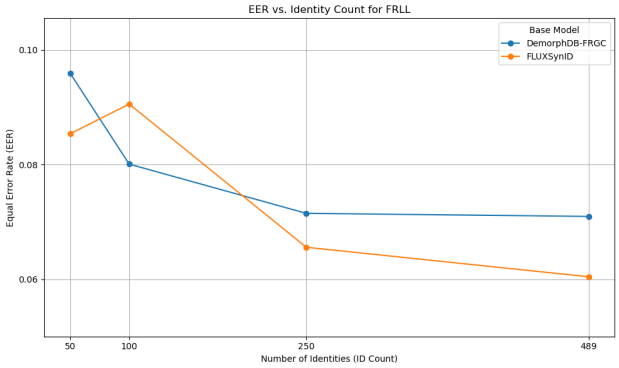


Fig. 4. FRLL EER Performance Plot for Experiment 2

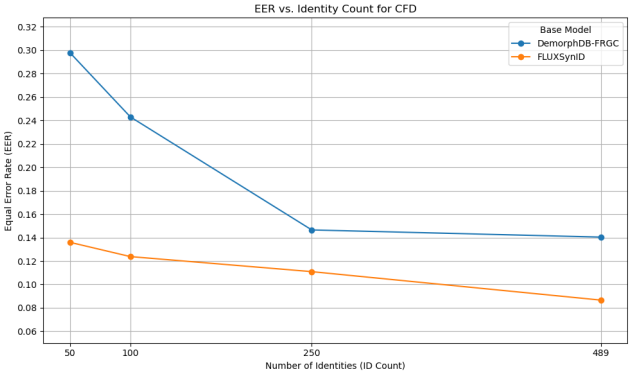


Fig. 5. CFD EER Performance Plot for Experiment 2

Across the FRLL dataset, both datasets show improved performance with increasing identity count, though FLUXSynID exhibits slightly lower EERs at 250 and 489 identities. On CFD, the synthetic model outperforms DemorphDB-FRGC at every stage.

In terms of TMR@FMR = 1e-2 as shown in Table 3, when evaluated on the FRLL dataset the performance is close between both datasets, however DemorphDB-FRGC slightly outperforms performs better. On CFD models trained on FLUXSynID consistently outperform DemorphDB-FRGC.

Plots 4 & 5 show the trend in the EER as identities increase. In CFD the results are consistent showing how the models trained on FLUXSynID perform better overall. On the other hand, the noisy plot for FRLL correlates with the inconsistent results of the TMR@FMR, which can be due to the overall poor performance of the models combined with the limited evaluation identities in the FRLL datasets.

Overall. the results are explainable due to the attribute similarity of the datasets. DemorphDB lacks the variation of expressions present in CFD, while FLUXSynID maintains such variation, the opposite is true for FRLL.

8.3 RQ3: Hybrid Mixing Ratios (Synthetic + Real Data)

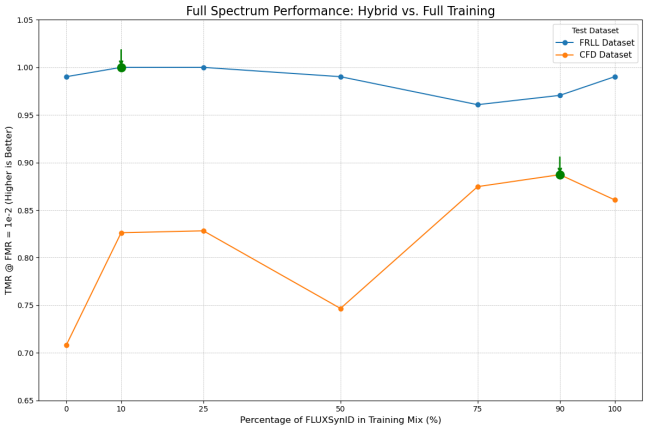


Fig. 6. Optimal Combination Ratio Plot for Experiment 3

To address RQ3, we evaluate hybrid training strategies using varied mixing ratios of FLUXSynID and VGGFace2. Figure 6 presents the full overview of the performance of the different ratios for the two evaluation datasets, highlighting the optimal performing ratio with the green dot. It also combines the baseline full models trained exclusively on VGGFace2 (100% real) and FLUXSynID (100% synthetic) from Experiment 1 to show a complete comparison. Additionally, the detailed performance metrics for all mixing ratios and evaluation thresholds are summarized in the Appendix (see Table A).

On the FRLL dataset, the best overall performance is observed at 25% synthetic data. This model achieves the lowest EER (0.0032), the highest accuracy (0.9969), and a TMR@FMR=1e-3 of 0.9510. These results suggest that the embedding of synthetic images provides valuable addition to the diversity of the dataset such as expression and lighting variations without introducing a disruptive shift in the model domain.

On the other hand, results on the CFD dataset are less noisy and steadily improve as the synthetic embedding ratio increases, from 75% up to a full model the results are very close with minor differences, these results also support the fact that the hybrid datasets reduce that gap difference between datasets with mismatches in feature similarity.

These findings validate hybrid training as a robust strategy. When carefully balanced, synthetic data not only complements real data by filling diversity gap but also helps improve the performance beyond what either training set achieves alone. Moreover, incorporation of synthetic samples can match or even exceed the effectiveness of purely real training, particularly when the real dataset lacks variability.

9 LIMITATIONS AND FUTURE WORK

9.1 Limitations

- **Limited Data Availability:** Due to the unique style of the FLUXSynID dataset, it is challenging to get similar datasets with close shot style and number of identities. Accordingly none of the benchmark datasets (e.g., LFW, IJB) could be used, reducing the dependability of the results.
- **Controlled Domain Bias:** All datasets (FRLL, CFD, FRGC) are captured under uniform studio conditions, which gives some sort of bias towards either real or synthetic datasets based on the common features.
- **Architectural Dependence:** We evaluated only a ResNet-50 backbone with triplet loss, which is considered a poor architecture for FR models.
- **Modality Constraints:** FLUXSynID lacks multi-view content, limiting insights into the models robustness under pose and motion variations.

9.2 Future Work

Building on these findings, several directions require further exploration. First, extending FLUXSynID with more data to match global state-of-the-art FR training datasets will definitely be beneficial to further experiment the capabilities of utilizing the FLUXSynID dataset on a bigger scale and gain insight on whether the conclusion of this paper can be generalized or not. Second, evaluating synthetic training across a wider variety of architectures (e.g., attention-based backbones, lightweight mobile networks) could reveal model-specific benefits or limitations. Third, integrating domain-adaptation techniques such as adversarial alignment or style transfer could help close the gap between synthetic and real data gap when training on highly diverse datasets. Finally, investigating multi-view synthetic data generation would also help in generalization and open the door to utilize the dataset for more complex face recognition scenarios. Collectively, these recommendations will help solidify the role of diffusion-based synthetic dataset "FLUXSynID" as a foundation for next-generation, privacy-preserving biometric systems.

10 CONCLUSION

In this work, we conducted a systematic evaluation of the efficacy of the FLUXSynID diffusion-based synthetic dataset for training face recognition models under three complementary scenarios directly addressing our research questions. In Experiment 1 (Full Datasets Training), which addressed RQ1, we found that a model trained purely on FLUXSynID can match, and at high-security thresholds even exceed the performance of a model trained on the real VGGFace2 dataset, particularly on the expression-rich CFD evaluation dataset. Experiment 2 (Sequential Learning) addressed RQ2 and demonstrated that synthetic training maintains an advantage over small real datasets (DemorphDB) across increasing identity counts, especially under challenging expressive conditions. Finally, in Experiment 3 (Hybrid Training), addressing RQ3, we found that mixing a ratio of 25% FLUXSynID with 75% real data yields the lowest EER and highest TMR@FMR on FRLL, while higher synthetic ratios benefit performance on CFD.

To conclude, these results confirm that the new diffusion-based synthetic data (FLUXSynID) can serve not only as a privacy-friendly stand-in for real faces but also as a valuable addition that bridges diversity gaps when limited real data is available and enhances generalization. However, given the controlled conditions and limited

dataset scope, further work is needed to validate these findings on in-the-wild benchmarks and with more advanced face recognition models.

REFERENCES

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, 2018.

[2] P. Dhariwal and P.; Nichol, A. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8780–8794.

[3] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *European Conf. Computer Vision*, 2016.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[8] F. Boutros, V. Struc, J. Fierrez, and N. Damer, "Synthetic data for face recognition: Current state and future prospects," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 26–39, 2023.

[9] G. Bae, M. de La Gorce, T. Baltrusaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "DigiFace-1M: 1 Million Digital Face Images for Face Recognition," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, 2023.

[10] Black Forest Labs, "Announcing FLUX.1," August 2024. [Online]. Available: https://bfl.ai/announcements/flux-1-kontext

[11] F. Boutros, J. Grebe, A. Kuijper, and N. Damer, "IDiff-Face: Synthetic-based Face Recognition through Fizzy Identity-Conditioned Diffusion Models," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19593–19604.

[12] I. DeAndres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fierrez, J. Ortega-Garcia, Z. Zhong, Y. Huang, Y. Mi, S. Ding, and S. Zhou, "Second Edition FRCSyn Challenge at CVPR 2024: Face Recognition Challenge in the Era of Synthetic Data," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[13] L. DeBruine and B. Jones, "Face Research Lab London Set," *figshare*, 2017. DOI: 10.6084/m9.figshare.5047666.v3.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[16] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "ONOT: a High-Quality ICAO-compliant Synthetic Mugshot Dataset," in *The 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–6.

[17] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT), Part 2: Identification," *NIST Interagency/Internal Report (NISTIR) 8238*, National Institute of Standards and Technology, 2018.

[18] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control," *arXiv preprint arXiv:2407.03168*, 2024.

[19] Z. Guo, Y. Wu, Z. Chen, L. Chen, P. Zhang, and Q. He, "PuLID: Pure and Lightning ID Customization via Contrastive Alignment," *arXiv preprint arXiv:2404.16022*, 2024. (Accepted to NeurIPS 2024)

[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[23] 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," University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[24] R. Ismayilov, L. Spreeuwers, and D. Sero, "FLUXSynID: A Framework for Identity-Controlled Synthetic Face Generation with Document and Live Images," *arXiv preprint arXiv:2505.07530*, 2025.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[26] M. Kim, F. Liu, A. K. Jain, and X. Liu, "DCFace: Synthetic Face Generation with Dual Condition Diffusion Model," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 23256–23267.

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2017, pp. 212–220.

[28] D. S. Ma, J. Correll, and B. K. Wittenbrink, "The Chicago Face Database: A free stimulus set of faces and norming data," *Behavior Research Methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[29] F. Paraperas Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2Face: A Foundation Model for ID-Consistent Human Faces," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. (*arXiv preprint arXiv:2403.11641*, 2024)

[30] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8026–8037.

[31] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction to evaluating biometric systems," *Computer*, vol. 33, no. 2, pp. 56–63, 2011.

[32] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "SynFace: Face Recognition with Synthetic Data," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 10880–10890.

[33] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data," *arXiv preprint arXiv:2206.10520*, 2022.

[34] H. O. Shahreza, A. George, and S. Marcel, "SynthDistill: Face Recognition with Knowledge Distillation from Synthetic Data," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2023.

[35] Y. Wang, J. Deng, X. Li, and S. Zafeiriou, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[36] J. Xu, S. Li, J. Wu, M. Xiong, A. Deng, J. Ji, Y. Huang, G. Mu, W. Feng, S. Ding, and B. Hooi, "ID^3: Identity-Preserving-yet-Diversified Diffusion Models for Synthetic Face Recognition," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[38] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The Face Recognition Grand Challenge (FRGC) Dataset," National Institute of Standards and Technology, Tech. Rep. FRGC2005, 2005.

[39] A. Hadid and I. Essa, "EURECOM-IST Face Database," EURECOM, Tech. Rep. EURECOM-IST-DB, 2011.

[40] J. Koenderink and A. van Doorn, "The Utrecht ECVP Face Appearance Database," ECVP Workshop, 2010.

A EXTENDED RESULTS

Table 4. Performance of hybrid and full models trained with varying percentages of FLUX on FRLL and CFD datasets.

Training Mix (% FLUX)	Test Dataset	EER	Accuracy	TMR@FMR=1e-2	TMR@FMR=1e-3
0% (VGGFace2 only)	FRLL	0.0048	0.9952	0.9902	0.7941
10%	FRLL	0.0041	0.9959	1.0000	0.9314
25%	FRLL	**0.0032**	**0.9969**	**1.0000**	**0.9510**
50%	FRLL	0.0085	0.9915	0.9902	0.7549
75%	FRLL	0.0186	0.9814	0.9608	0.8431
90%	FRLL	0.0269	0.9731	0.9706	0.7843
100% (FLUX only)	FRLL	0.0070	0.9930	0.9902	0.9412
0% (VGGFace2 only)	CFD	0.0535	0.9465	0.7082	0.3760
10%	CFD	0.0479	0.9521	0.8263	0.5869
25%	CFD	0.0451	0.9549	0.8282	0.5577
50%	CFD	0.0411	0.9589	0.7467	0.3369
75%	CFD	**0.0350**	**0.9650**	0.8747	0.5544
90%	CFD	0.0364	0.9636	**0.8873**	0.6015
100% (FLUX only)	CFD	0.0405	0.9595	0.8607	**0.6048**

B AI ASSISTANCE DISCLOSURE

Assistance from AI tools such as ChatGPT and Paperpal were used in writing this paper. These tools were used to improve the flow and enhance the expression of ideas for clarity. All content was manually reviewed before usage.

C ADDITIONAL FIGURES

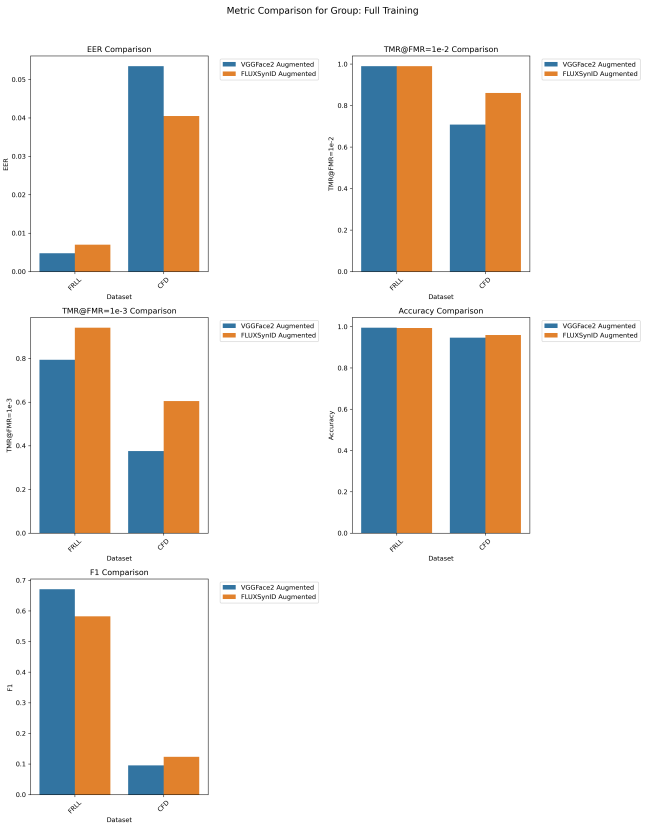


Fig. 7. Performance metrics for the full models.

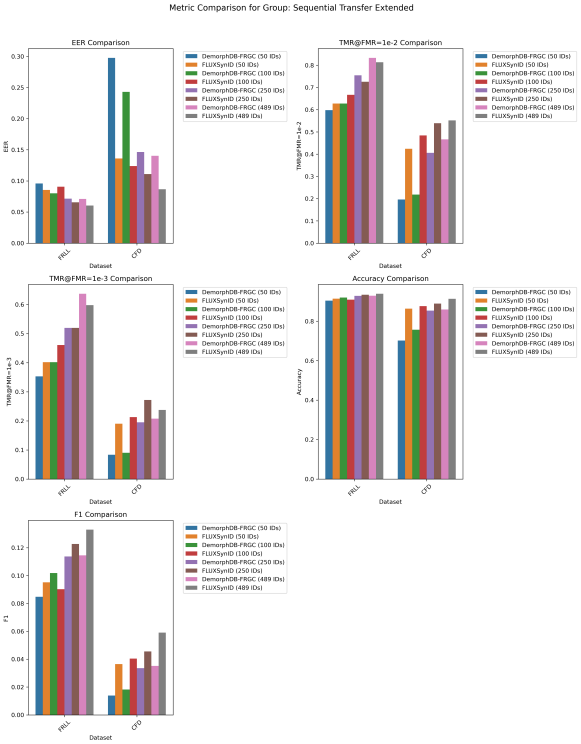


Fig. 8. Metrics for sequential training.

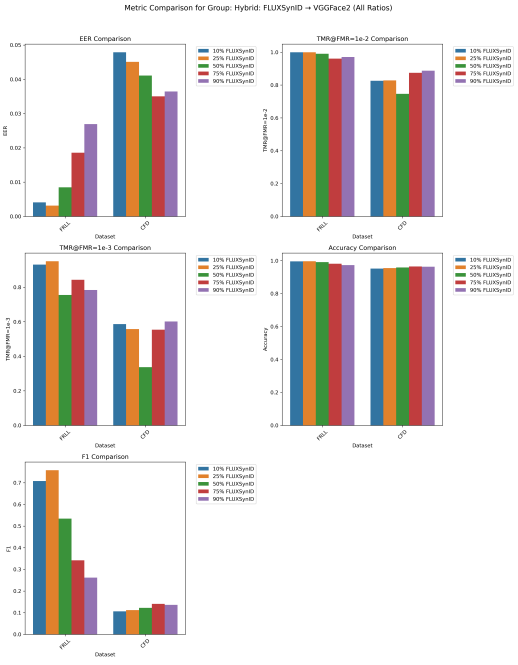


Fig. 9. Hybrid training performance for different FLUX ratios.