FAMILIAR: CONJURING THE MANY FACES OF A VIRTUAL IDENTITY VIA FAMILIARIZED OCCASION PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Face synthesis conjures portraits of people who have never existed. Applications ask more of a virtual person than one portrait: a repertoire of appearances in which identity stays consistent. To create such repertoires, most existing generators treat each identity as a fixed point and add variation by a prescribed rule: perturbing its embedding, borrowing reference styles, or sampling predefined facial parameters. The form of an individual's variation is thus fixed by the method, and data at most fit its parameters. Cognitive research points to why this matters: familiarity with a face is built by learning how that face varies across encounters, and different faces vary along different dimensions. Following this insight, we introduce FAMILIAR, which learns person-specific variation instead of prescribing it. FAMILIAR represents a virtual identity as a *distribution over photographic occasions*. It first learns the occasion space through identity-conditioned image synthesis, without an attribute inventory, and then learns how that space is occupied for each individual: a conditional normalizing flow shared across identities predicts a full occasion distribution from a recognition embedding alone. A newly sampled identity therefore arrives with its own repertoire, without reference photographs. Occasion statistics differ more between people than within them, and with the generator fixed, another person's occasions lower recognition accuracy despite wider pose spread: useful variation is not interchangeable across identities. FAMILIAR therefore provides state-of-the-art synthetic data for face recognition and, by fixing a source occasion, de-identifies faces while preserving identity-irrelevant facial attributes.

1 INTRODUCTION

Contemporary face synthesis can generate photorealistic portraits of non-existent individuals (Kim et al., 2023; Wu et al., 2025a). Such faces matter where real ones are hard to use: web-scale face datasets were collected without consent, and their identity labels are noisy (Appendix B.9). Yet one portrait does not make an identity. Synthetic face-recognition datasets (Qiu et al., 2021; Bae et al., 2023) must show each virtual person many times, varying expression, pose, and illumination while identity stays fixed, since recognizers learn from this intra-class variation. Constructing them hinges on two decisions: enforcing inter-identity distinction (Otroshi Shahreza & Marcel, 2025; Kim et al., 2025) and structuring the intra-identity distribution (Sun et al., 2024). This study addresses the latter.

Prevailing generators structure this distribution in the same way. Each identity is modeled as a fixed point, typically a recognition embedding, with variation introduced via predefined rules: reference styles in DCFace (Kim et al., 2023), controlled embedding perturbations in Vec2Face and IDperturb (Wu et al., 2025a; Boutros et al., 2026), and staged conditioning in UIFace (Lin et al., 2025). MorphFace incorporates subject-awareness (Mi et al., 2025) by injecting 3D morphable attributes from class-specific Gaussian statistics, with facial means anchored to reference images. Even so, the space of variation remains constrained by the underlying 3D parametrization, and the functional dependence on identity remains fixed by design. Across all such methods, the structural form of identity variation is prescribed a priori, while empirical data merely fit its parameters.

Cognitive research suggests that this choice matters. In photo-sorting experiments, viewers unfamiliar with a person divide that person's photographs among several apparent identities, whereas

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

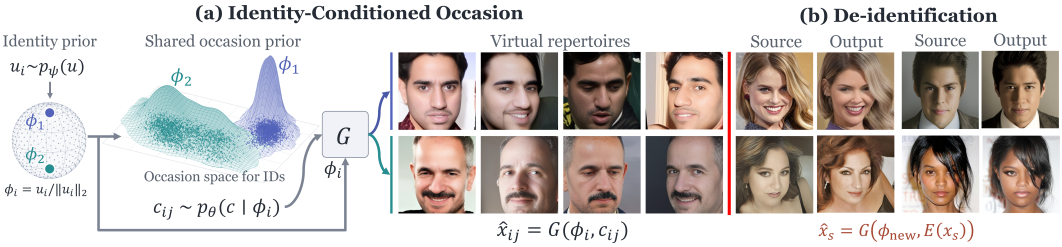

Figure 1: Virtual repertoires and De-ID using FAMILIAR. (a) An identity from the identity prior p_ψ conditions the occasion prior and generator. Each row holds one virtual identity. Projected densities illustrate conditional variation. (b) Each source/output pair keeps the occasion code $E(x_s)$ and replaces the identity.

familiar viewers group them correctly (Jenkins et al., 2011). Seeing more varied photographs of a newly learned face also improves recognition of photographs not yet seen (Ritchie & Burton, 2017). Accounts of face familiarity attribute this robustness to knowledge of how an individual characteristically varies (Burton et al., 2011; Kramer et al., 2018), and analyses of many photographs per person show that the dimensions describing one face's variability need not describe another's (Burton et al., 2016). These findings suggest a different approach for synthesis. A virtual identity can be described by the distribution of *occasions* in which it appears, instead of a point with prescribed variation, and because its form differs across people, it should be *learned* rather than *specified* (Figure 1(a)).

To realize this, we introduce **FAMILIAR**, a face synthesis framework. It models a photograph through three factors: *who the person is, the occasion of the photograph, and the rendering of that occasion into pixels*. Identity is given by a frozen recognition embedding ϕ (Deng et al., 2019), and rendering by latent rectified flow (Liu et al., 2023). The intermediate factor, the distribution of occasions given an identity, has so far been prescribed by design in most existing generators. FAMILIAR learns it as an identity-conditioned occasion prior, which raises two problems.

The first is what counts as an occasion. We let the generative model decide. We use an occasion encoder to compress each photograph into a code c, which the generator must use, alongside the identity input, to predict the photograph's latent. Keeping this code compact and noised during training pushes c to carry the image-specific appearance that identity leaves unexplained, such as co-occurring expression and pose; this divides roles without making c independent of ϕ. The second is how occasions depend on the person. With the encoder and generator frozen, a conditional normalizing flow with affine coupling (Dinh et al., 2017) learns $p_\theta(c \mid \phi)$ across all training identities. One set of parameters maps any recognition embedding to a full distribution over occasion codes, with its own location, spread, and dependencies, and an exact model likelihood. A virtual identity sampled from an identity prior thus comes with *its own repertoire*, without predefined style. Freezing the generator also makes person-specific variation testable: the occasion prior can be made identity-blind, conditioned on other photographs, or reassigned to another identity (Section 4.3).

These interventions yield our central finding: *useful variation is not interchangeable across identities*, as it belongs mainly to the individual rather than a single image. Sampling each identity's own occasions yields state-of-the-art synthetic face-recognition data. Conversely, holding the occasion constant while changing identity achieves de-identification (De-ID) with preserved attributes (Figure 1(b)). Recognition needs identity to persist as occasions change, and De-ID needs occasions to persist as identity changes; FAMILIAR serves both with one fixed model. Our contributions are:

- **Variation as a learned, person-specific distribution.** Following face-familiarity accounts, we represent a virtual identity as a distribution over photographic occasions, not a point with prescribed variation: synthesis defines the occasion space, and one shared conditional flow predicts each identity's occasion distribution from its embedding alone.
- **Whose variation matters, not only how much.** With the generator fixed, giving each identity another person's occasions lowers recognition accuracy from 94.29% to 92.86% despite wider pose spread. Priors conditioned on other photographs of the same person still fit held-out occasions far better than another person's (105.3 versus 119.8 nats): most of the learned dependence belongs to the person rather than the photograph.
- **Two complementary uses of FAMILIAR.** Sampling occasions for new identities gives state-of-the-art synthetic face recognition; fixing a source occasion and replacing identity gives de-identification with little identity leakage and near-original attribute utility.

2 Related Work

Familiarity, variation, and identity in face synthesis. Cognitive accounts separate invariant aspects of a face, used for recognition, from changeable aspects such as expression and gaze (Haxby et al., 2000). Familiarity builds across encounters: a person-specific average stabilizes with exposure (Burton et al., 2005; 2011) and is updated as new photographs arrive (Koca & Oriet, 2023), while variability across photographs is itself informative about identity (Kramer et al., 2018) and can follow person-specific directions (Burton et al., 2016). FAMILIAR borrows this two-part structure without modeling it: the identity embedding stands in for the stable description, and the conditional occasion prior for the characteristic variability. For identity, DiscoFaceGAN controls facial factors through 3D priors (Deng et al., 2020b), and Arc2Face conditions diffusion on recognition embeddings (Paraperas Papantoniou et al., 2024), as FAMILIAR does. HyperFace packs virtual identities on the embedding hypersphere (Otroshi Shahreza & Marcel, 2025), and VIGFace learns virtual identity prototypes (Kim et al., 2025). These methods decide which identities exist; FAMILIAR uses a simple identity prior for this step and concentrates on the images each identity receives.

Recent methods study where variation comes from. Some move the identity condition itself: Vec2Face and IDperturb perturb the recognition embedding (Wu et al., 2025a; Boutros et al., 2026), Vec2Face+ adds attribute and pose control while stressing identity consistency (Wu et al., 2025b), CemiFace targets chosen similarities to the identity center (Sun et al., 2024), and VariFace conditions on a divergence score that trades identity preservation against diversity (Yeung et al., 2025). Others add a second input: DCFace pairs an identity image with a real style reference (Kim et al., 2023), ID^3 conditions on age and pose (Xu et al., 2024), and UIFace switches from an empty to an identity context during sampling (Lin et al., 2025). MorphFace samples 3DMM attributes from class-specific distributions so that subjects vary differently (Mi et al., 2025). SteerFace discourages the generator from relying on appearance cues during training (Mi et al., 2026). Each of these methods fixes by design how variation depends on identity. FAMILIAR instead learns it: one conditional density shared across identities predicts each person's distribution over a learned occasion space.

Face de-identification. De-ID replaces identifying information while keeping useful attributes, and de-identified faces are usually evaluated along privacy, utility, and quality (Wei et al., 2026). CIAGAN combines landmark-guided synthesis with identity control (Maximov et al., 2020), Disguise uses mixture-of-experts supervision to preserve utility (Cai et al., 2024), and FLUID optimizes latent displacements in a pretrained diffusion model (Park et al., 2026). These methods are built for De-ID; FAMILIAR reuses its repertoire model, fixing the source occasion code and replacing only identity. This reverse intervention tests the factorization from the other side: a code encoding the source identity would leave matching high, and one missing image-specific appearance would lose age and expression. Among methods that keep source matching low, FAMILIAR retains far more age and expression utility than CIAGAN (Table 4); resampling the occasion degrades both (Table 10).

3 Method

FAMILIAR builds on a finding from face-familiarity research: how a person's photographs vary is itself specific to that person (Burton et al., 2011; 2016). We therefore learn this variation per identity instead of prescribing it, and model the photographs of an identity as arising from three sources: the identity ϕ, an occasion c that captures the image-specific appearance of one photograph, and rendering noise ξ. For an identity embedding ϕ, the model's images follow

$$p_{\theta,\gamma}(x \mid \phi) = \int p_\gamma(x \mid \phi, c)\, p_\theta(c \mid \phi)\, \mathrm{d}c, \tag{1}$$

where $p_\gamma(x \mid \phi, c)$ is the distribution of the generated image $G(\phi, c; \xi)$ over $\xi \sim \mathcal{N}(0, I)$, and $p_\theta(c \mid \phi)$ is the *identity-conditioned occasion prior*. At generation time, ϕ is a normalized draw from a learned identity prior p_ψ. Existing generators fix the second factor by design, as a neighborhood of ϕ (Wu et al., 2025a; Boutros et al., 2026), real style references (Kim et al., 2023), or class-specific Gaussians over 3DMM parameters (Mi et al., 2025). FAMILIAR learns both the occasion space (Section 3.1) and each identity's distribution over it (Section 3.2), so the form of variation, not only its scale, can differ between people. The occasion encoder and generator are trained jointly and frozen before the priors are fitted (Figure 2); with the generator fixed, the occasion prior can be replaced by identity-blind or simpler priors, or by those of other identities.

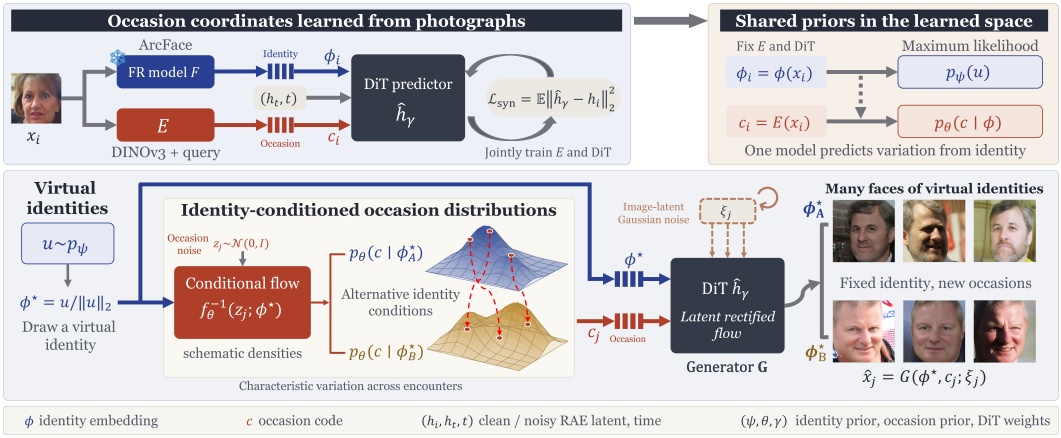

Figure 2: FAMILIAR learns occasions from photographs and predicts their distribution from identity. Top left: frozen ArcFace supplies identity; occasion encoder E and the DiT of G are jointly trained to predict clean latents. Top right: with both fixed, maximum likelihood fits an identity prior and one occasion prior shared across identities. Bottom: a normalized identity draw conditions the inverse occasion flow and the generator. Independent draws of z_j and ξ_j yield many faces of that identity with diverse attributes.

3.1 Learning an Occasion Space through Synthesis

A frozen ArcFace encoder F (Deng et al., 2019) supplies identity as the 512-dimensional embedding $\phi(x) = F(x)/\|F(x)\|_2$. It is a continuous condition rather than a person index, and different photographs of one person need not share an embedding. The occasion encoder E, with parameters η, pools fine-tuned DINOv3 features (Siméoni et al., 2025) with a single learned attention query (Vaswani et al., 2017) and projects the result to a 256-dimensional code $c = E(x)$. No attribute labels define c; its meaning comes from the role it plays in synthesis.

The generator operates in the latent space of a representation autoencoder (RAE) (Zheng et al., 2026), with a frozen encoder A and a decoder D adapted to face images (Appendix B.1). A diffusion transformer (Peebles & Xie, 2023) with parameters γ predicts the clean latent $h = A(x)$ from the rectified-flow state $h_t = (1-t)\xi + th$, where $\xi \sim \mathcal{N}(0, I)$ and time runs from noise at $t = 0$ to data at $t = 1$ (Liu et al., 2023). Identity and occasion enter through separate groups of condition tokens,

$$\mathcal{T}_a = U_a + \Pi_a(\bar{a}), \qquad a \in \{\phi, c\}, \tag{2}$$

where U_a is a learned token array and Π_a is a linear layer whose output is reshaped to that array. During training, $\bar{\phi} = \phi$ and $\bar{c} = \tilde{c} = E(x) + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and each condition is independently replaced by its own learned null vector, written $\perp$, with probability 0.1. The encoder and transformer are trained jointly with an unweighted clean-latent objective without auxiliary alignment,

$$\mathcal{L}_{\text{syn}}(\eta, \gamma) = \mathbb{E}_{x,t,\xi,\epsilon,\bar{\phi},\bar{c}} \left\| \widehat{h}_\gamma(h_t, t; \bar{\phi}, \bar{c}) - h \right\|_2^2. \tag{3}$$

This objective defines what an occasion is. Identity already reaches the transformer through its own tokens, so gradients through $\tilde{c}$ train the encoder to keep whatever else is needed to predict the image, such as co-occurring expression, pose, and illumination. The compact code and additive noise regularize this channel, and the noise ($\sigma = 0.1$) exposes the generator to codes near, but not at, encoder outputs, as sampled codes will be. The result is a division of roles, not statistical independence: the prior in Section 3.2 models the dependence of c on ϕ that remains. Density fitting and source encoding use the deterministic code $E(x)$. At inference, the clean prediction defines a velocity for $t < 1$ (Appendix A.1), and guidance nests the conditions from none, to identity, to both (Brooks et al., 2023), so w_ϕ weights the identity term and w_c the occasion term given identity:

$$\begin{aligned} v_\gamma(h_t, t; \bar{\phi}, \bar{c}) &= \frac{\widehat{h}_\gamma(h_t, t; \bar{\phi}, \bar{c}) - h_t}{1 - t}, \\ v^{\mathbf{w}}_\gamma(h_t, t; \phi, c) &= v_{\perp\perp} + w_\phi \left(v_{\phi\perp} - v_{\perp\perp}\right) + w_c \left(v_{\phi c} - v_{\phi\perp}\right), \end{aligned} \tag{4}$$

where $v_{ab} = v_\gamma(h_t, t; a, b)$ with $a \in \{\phi, \perp\}$, $b \in \{c, \perp\}$, and $\mathbf{w} = (w_\phi, w_c)$. Integrating $v^{\mathbf{w}}_\gamma$ from $h_0 = \xi$ and decoding with D defines the generator $G_{\mathbf{w}}(\phi, c; \xi)$. Unit weights recover the jointly conditioned velocity $v_{\phi c}$; repertoires use $\mathbf{w} = (1, 1)$, and G in Eq. (1) denotes $G_{(1,1)}$.

3.2 AN IDENTITY-CONDITIONED OCCASION PRIOR

The encoder fixes what an occasion can express, but a new virtual identity has no photographs from which its variation could be estimated. A separate density per person is therefore unavailable, and a single identity-blind prior would discard what distinguishes one person's variation from another's. We instead predict the occasion distribution from the identity embedding with one set of parameters shared across all individuals. With E and the generator frozen, we collect pairs $(\phi_i, c_i) = (\phi(x_i), E(x_i))$ for the N training photographs and fit θ by maximum likelihood:

$$\mathcal{L}_{\text{prior}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(c_i \mid \phi_i),$$
$$\log p_\theta(c \mid \phi) = \log \mathcal{N}\big(f_\theta(c;\phi); 0, I\big) + \log\Big|\det \frac{\partial f_\theta(c;\phi)}{\partial c}\Big|. \tag{5}$$

Because θ is shared, photographs of real individuals inform the occasion distribution assigned to any embedding, including embeddings that no photograph has produced. We parameterize $f_\theta(\cdot;\phi)$ as a conditional normalizing flow (Dinh et al., 2017): after a fixed standardization, affine coupling layers transform the code, and each layer receives a learned feature of ϕ (Appendix A.2). Identity can therefore shape not only where a person's occasions lie but also their spread and the dependencies among coordinates, mirroring the finding that different faces vary along different dimensions (Burton et al., 2016). Conditional Gaussian and Gaussian-mixture priors restrict this dependence to Gaussian or finite-mixture forms, and Section 4.3 compares them under the same frozen generator.

The flow gives exact log-likelihoods of held-out codes under the learned model, which makes priors comparable within a fixed representation; gaps between conditional and unconditional priors are not exact estimates of mutual information. It also samples exactly: $c = f_\theta^{-1}(z;\phi)$ with $z \sim \mathcal{N}(0, I)$ draws from $p_\theta(c \mid \phi)$, whereas scaling z by a temperature T changes the distribution (Section 3.3).

Each training pair comes from a single photograph, and recognition embeddings retain appearance cues (Mi et al., 2026), so part of the fitted dependence may reflect the photograph rather than the person. To test this, we refit the prior with ϕ computed from another photograph of the same person, or from the normalized mean embedding of five other photographs (Section 4.3); generation uses the per-image prior. MorphFace also assigns each subject its own distribution, but over 3DMM parameters (Mi et al., 2025). FAMILIAR instead learns, through synthesis, the coordinates in which the distribution is defined. One shared model then predicts a non-Gaussian joint density in those coordinates from a recognition embedding, without a reference photograph.

3.3 GENERATING VIRTUAL REPERTOIRES

In FAMILIAR, virtual identities come from an unconditional normalizing flow $p_\psi(u)$ fitted by maximum likelihood to normalized training embeddings. We use it only as a sampler: each draw is projected to the unit sphere, and the ambient likelihood is not a density on the sphere. Each accepted embedding ϕ^* defines one synthetic class, whose images are

$$\widehat{x}_j = G\big(\phi^*, f_\theta^{-1}(z_j;\phi^*);\xi_j\big), \qquad \phi^* = \frac{u}{\|u\|_2},\ u \sim p_\psi,\ z_j, \xi_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{6}$$

with no reference photograph at generation time. Released datasets additionally follow a *release policy*: identity proposals too similar to a gallery of real embeddings (cosine ≥ 0.3) are rejected, 20% of the rendered candidates are drawn at $T = 1.4$, and the 50 of 65 candidates closest to ϕ^* in the conditioning embedding space are retained (Appendix A.3). The policy changes the released distribution but not the fitted prior, so the controlled comparisons of priors use $T = 1$ without ranking. Gallery screening is an empirical similarity check, not a privacy guarantee.

3.4 DE-IDENTIFICATION WITH A FIXED SOURCE OCCASION

De-identification applies the reverse intervention. Instead of sampling occasions for a fixed identity, we fix the occasion of a source photograph x_s and replace its identity (Figure 3):

$$\widehat{x}_s = G_{\mathbf{w}}\big(\phi_{\text{new}}, c_s; \xi\big), \qquad c_s = E(x_s),\ \phi_{\text{new}} = \frac{u}{\|u\|_2},\ u \sim p_\psi,\ \xi \sim \mathcal{N}(0, I). \tag{7}$$

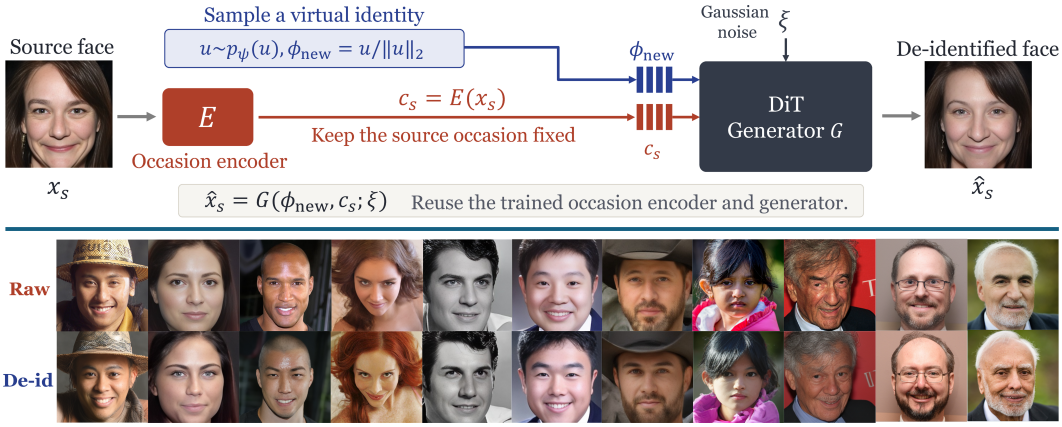

Figure 3: Using FAMILIAR for face De-ID with preserved facial attributes. Examples use reference-derived identities and candidate selection (Appendix C); Section 4.4 evaluates one sampled identity per source.

The encoder and generator are reused without retraining. The source code c_s is computed deterministically; resampling it from $p_\theta(c \mid \phi_{\text{new}})$ would replace the source occasion as well. The guidance weights $\mathbf{w}$, which can strengthen the replacement identity, are selected on validation data (Appendix B.5), and a reference-derived embedding can take the place of ϕ_{new}, as in the displayed examples (Appendix C). Because c is not independent of ϕ, c_s can carry residual source-identity cues as well as expression, pose, and other appearance. Section 4.4 therefore measures source matching and attribute utility together, under three recognizers and three attribute estimators.

4 Experiments

4.1 Implementation Details

Models. Identity embeddings come from a frozen ArcFace (Deng et al., 2019). The occasion encoder is a pretrained DINOv3-L, fine-tuned jointly with the generator. We implement FAMILIAR with the RAE (Zheng et al., 2026) generative framework. Images are generated with 25 Euler steps and unit guidance, except for $\mathbf{w} = (1.5, 1)$ in de-identification. Appendix A gives the remaining architecture and the sampling procedure, Appendix B.1 the WebFace-21M training schedule, optimization, and release policy, and Appendix B.5 the de-identification protocol.

Data and recognition protocol. We follow the evaluation protocols of Wu et al. (2025a) and Kim et al. (2023). FAMILIAR is trained on WebFace-21M, the restored WebFace42M subset (Zhu et al., 2021) used to train Arc2Face (Paraperas Papantoniou et al., 2024) (Appendix B.1). We use FAMILIAR to build a synthetic dataset containing 10,000 identities with 50 images each, and a 1.2M-image dataset with 24,000 identities for the second common scale, both generated with the release policy of Section 3.3. On each dataset, we train an IR-50 from scratch with ArcFace loss under the conventional settings (Mi et al., 2025; Lin et al., 2025). We report mean verification accuracy on LFW (Huang et al., 2007), CFP-FP (Sengupta et al., 2016), CPLFW (Zheng & Deng, 2018), AgeDB-30 (Moschoglou et al., 2017), and CALFW (Zheng et al., 2017), together with IJB-C (Maze et al., 2018) template verification. A recognizer trained on the CASIA-WebFace training split (Yi et al., 2014) serves as a reference. Our FR results are means over 5 independent runs.

Separable conditions. Figure 4(a, b) measures how well the separate token groups (Eq. 2) divide the roles of ϕ and c. Joint and occasion-only renders start equal at random initialization and separate by 0.57, the latter staying near 0.18: occasion-only renders reconstruct the person far less often than joint renders (15.19% versus 96.31%; Table 10), as de-identification requires.

4.2 Face Recognition with Synthetic Repertoires

Results. In Table 1, where baseline accuracies are those reported by the original papers (Appendix B.1), FAMILIAR reaches 94.79% average accuracy (SD 0.06), 0.74 points above VariFace and 1.07 above SteerFace. Without its release policy ($T = 1$, no ranking; Table 5), it still reaches 94.29%, above every baseline generator, whereas an identity-blind prior with the same generator

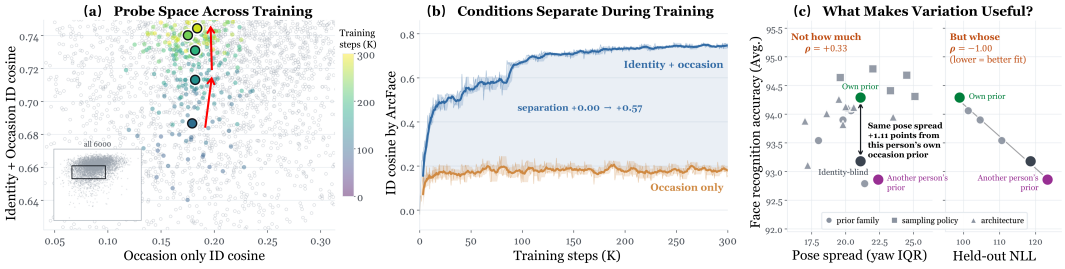

Figure 4: Separable conditions and useful variation. (a, b) ArcFace cosine between a render and its source identity ϕ. (a) With ϕ dropped (x) or kept (y); rings: 6,000 photographs at the final checkpoint; dots: probe events, colored by training step; large markers: five temporal-block means. (b) Both start equal and separate by 0.57; occasion-only renders stay near 0.18. (c) Accuracy of 21 configurations (Tables 2, 5, 11) against pose spread (left; Spearman $\rho = +0.33$; $+0.27$ at a fixed generator) and, for the six priors with comparable NLL, against held-out NLL (right; $\rho = -1.00$). Appendix C gives details and further analysis.

Table 1: Face recognition at 0.5M and 1.2M images. *: 1.2M columns at 1.0M. IJB-C: 0.5M datasets.

	0.5M images (10K × 50)						1.2M images						IJB-C TAR	
Method	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.	10^{-5}	10^{-4}
DCFace	98.55	85.33	82.62	89.70	91.60	89.56	98.58	88.61	85.07	90.97	92.82	91.21	—	—
Arc2Face	98.81	91.87	85.16	90.18	92.63	91.73	98.92	94.58	86.45	92.45	93.33	93.14	—	—
CemiFace	99.03	91.06	87.65	91.33	92.42	92.30	99.22	92.84	88.86	92.13	93.03	93.22	—	—
MorphFace	99.25	94.11	88.73	91.80	92.73	93.32	99.35	94.77	90.07	93.27	93.40	94.17	—	—
UIFace*	99.27	94.29	89.58	90.95	92.25	93.27	99.22	95.03	90.42	92.45	93.18	94.06	—	—
Vec2Face+*	99.35	93.56	88.03	94.33	94.17	93.89	99.40	94.23	89.10	95.58	94.67	94.60	—	85.75
IDperturb*	99.40	93.61	88.37	93.20	93.50	93.62	**99.48**	95.01	90.01	94.03	93.85	94.48	82.28	89.49
VariFace	99.38	94.60	88.82	94.38	93.05	94.05	99.45	95.61	90.63	94.75	94.13	94.92	—	—
SteerFace	99.40	94.67	88.85	92.73	92.95	93.72	99.35	95.27	90.66	93.12	93.43	94.37	—	—
FAMILIAR	**99.42**	**95.41**	**89.87**	**94.99**	**94.27**	**94.79**	99.42	**96.33**	**91.35**	**95.60**	**94.85**	**95.51**	**87.84**	**92.93**
Real data	99.53	95.97	90.00	94.64	93.43	94.71	—	—	—	—	—	—	92.15	95.76

reaches only 93.18%, below six of them (Table 2). At 1.2M images, FAMILIAR reaches 95.51%, 0.59 points above VariFace and 1.14 above SteerFace. At 0.5M, the gains come from the pose- and age-variant benchmarks (CFP-FP, CPLFW, AgeDB-30, and CALFW) rather than the nearly saturated LFW. The average also slightly exceeds that of the CASIA-WebFace real-data reference (94.71%), owing to AgeDB-30 and CALFW, whereas real data remain ahead on IJB-C. Among the generators with IJB-C results, FAMILIAR leads at both FARs, reaching 92.93% TAR at FAR 10^{-4}. Most baseline generators are trained on CASIA-WebFace; Arc2Face, trained on the same WebFace-21M subset as FAMILIAR, trails it by 3.06 points at 0.5M and 2.37 at 1.2M. Appendices B.7–B.9 study scaling to 6M images, demographic and occlusion robustness, and label fidelity.

4.3 Evaluating the Identity-Conditioned Occasion Prior

Controlling the generator. To isolate the occasion prior, we fix all other components, render exactly 50 images per identity, and vary only the distribution of occasion codes: an empirical bank of training codes, an unconditional flow, conditional diagonal and rank-32-plus-diagonal Gaussians, an eight-component conditional Gaussian mixture (Bishop, 1994), and the conditional flow of Section 3.2. All are fitted to the same training codes for the same number of passes, with baseline capacity selected by validation NLL; the conditional Gaussians take the class-specific form that MorphFace assumes for 3DMM attributes (Mi et al., 2025), but in the learned occasion space. A wrong-identity control draws each identity's occasions from another identity's prior, $c \sim p_\theta(c \mid \phi_{\pi(i)})$ for a derangement π, while the generator still receives ϕ_i. Varying the prior changes the factor that matters: with identity and generator fixed, the occasion code, not rendering noise, carries most within-identity variation (72.7% versus 11.5% of the pose-response energy; Appendix B.3).

Whose variation, not how much. The wrong-identity control is the most direct test of the premise. With the same flow, generator, and identities, it lowers accuracy from 94.29% to 92.86%, below even the unconditional flow (93.18%), although its yaw spread is wider (22.4 versus 21.1; Table 2). Its renders are also less identity-consistent (0.563 versus 0.616). *Useful within-identity variation is not interchangeable across identities: it helps recognition when it belongs to the identity being rendered, not merely when there is more of it.* Figure 4(c) extends this to 21 configurations: yaw spread orders accuracy only weakly (Spearman $\rho = +0.33$), whereas held-out NLL orders it exactly for

Table 2: Fixed-generator occasion priors.

Occasion source	NLL ↓	ID	Yaw	FR ↑	IJB ↑
Empirical bank	—	0.573	21.4	92.79±0.06	89.24
Uncond. flow	118.6±0.8	0.588	21.1	93.18±0.06	89.83
Cond. diag. Gaussian	110.6±0.6	0.612	18.0	93.54±0.07	90.31
Cond. rank-32 Gaussian	104.6±0.4	0.605	19.8	93.90±0.12	90.94
Cond. 8-comp. mixture	101.3±0.4	0.606	20.4	94.06±0.06	91.44
Cond. flow	**98.9±0.3**	**0.616**	21.1	**94.29±0.07**	**92.07**
Wrong-ID flow	123.3±0.3	0.563	22.4	92.86±0.07	89.17

ID: within-class cosine under AdaFace-IR101. Yaw: within-class IQR (°); FR: five-benchmark mean (%); IJB: TAR at FAR 10^{-4} (%); ±: SD.

Table 3: Cross-photograph conditioning.

ϕ from	NLL ↓	SWD ↓	Cov. ↓
Same photo	98.8±0.6	0.239±0.002	0.334±0.005
Other photo	103.4±0.4	0.260±0.011	0.345±0.008
Mean of 5	105.3±0.2	0.269±0.008	0.358±0.005
Other person	119.8±0.3	0.372±0.010	0.428±0.013

Each prior is retrained for its conditioning construction and evaluated on 200 held-out identities with 20 query photographs each. NLL: nats per code; SWD and Cov.: sliced Wasserstein and covariance distances between sampled and query codes (Appendix B.2). ±: SD (standard deviation) across replicates.

Table 4: Identity suppression and attribute preservation under FDeID-Toolbox.

		AgeDB TAR ↓			Attribute utility		
Method	Type	ArcFace	CosFace	AdaFace	Age ↓	Expr. ↑	NME ↓
Original	—	95.82	95.62	93.57	10.18	70.54	0.3601
Blur (σ=20)	Naive	48.63	57.65	30.81	14.18	24.56	0.3620
Pixelation (8×8)	Naive	4.11	2.54	28.38	15.26	14.79	0.3615
k-Same-Sel. (Gross et al., 2006)	k-Same	0.33	0.43	0.40	17.81	15.55	0.3973
k-Same-Fur. (Meng & Sun, 2014)	k-Same	0.20	0.20	0.17	18.38	15.35	0.3974
TIP-IM (Yang et al., 2021)	Adv.	61.58	70.56	75.90	10.95	68.90	0.3630
CIAGAN (Maximov et al., 2020)	Gen.	2.27	2.10	2.03	18.47	28.99	0.4242
AMT-GAN (Hu et al., 2022)	Gen.	60.74	54.17	46.80	16.95	66.28	0.3595
G2Face (Yang et al., 2024)	Gen.	62.98	55.57	52.77	11.82	63.27	0.3629
WeakenDiff (Salar et al., 2025)	Gen.	54.66	39.28	38.23	11.46	65.36	0.3603
FAMILIAR	Gen.	5.62	5.79	5.47	**10.84**	**69.21**	**0.3553**

TAR (%) at FAR 10^{-3}; selected settings (all in Table 9); bold: best utility among methods shown. Age: MAE (years) on AgeDB; Expr.: POSTER accuracy (%) on AffectNet; NME: HRNet landmark error on CelebA-HQ; Adv./Gen.: adversarial/generative methods.

the six priors with comparable likelihoods ($\rho = -1.00$, $p = 0.003$, also for IJB-C). The conditional flow lowers NLL from the unconditional 118.6 to 98.9 nats per code; a conditional diagonal Gaussian recovers only part of this gain, and correlated coordinates, mixtures, and the flow add more, consistent with *identity shaping how occasions co-vary, not only where they lie*.

Cross-photograph conditioning. To test whether the prior reads the photograph rather than the person (Section 3.2), we condition it on other photographs of the same person, retraining it for each construction (Table 3). Another photograph gives 103.4 nats and the mean of five others 105.3, against 98.8 for the same photograph and 119.8 for another person. The photograph contributes, but 14.5 of the 21.0 nats that separate the two survive without it. Sliced Wasserstein and covariance distances order the four conditions the same way. The dependence is also person-specific in shape: occasion-code covariances differ more between people than within them (split-half distances 0.495 versus 0.428), whereas isotropic and shared-covariance controls with matched sample sizes show no such gap (Table 6). Predicted occasions also help recognize unseen real people: from one enrollment photograph, they raise rank-1 accuracy from 72.66% to 77.21%, above image augmentation (74.59%) and both unconditional (74.91%) and wrong-identity (73.78%) occasions (Appendix B.4).

4.4 De-identification and Attribute Preservation

Protocol. We evaluate de-identification with FDeID-Toolbox (Wei et al., 2026). Privacy is measured on AgeDB (Moschoglou et al., 2017) as the TAR at FAR 10^{-3} under ArcFace-R100, CosFace-R50, and AdaFace-R50 (Deng et al., 2019; Wang et al., 2018; Kim et al., 2022). The toolbox's verification protocol compares de-identified faces with original faces; the *Original* row applies the same protocol to unmodified images. Utility is age MAE on AgeDB, POSTER (Zheng et al., 2023) expression accuracy on AffectNet (Mollahosseini et al., 2019), and HRNet (Wang et al., 2021) landmark NME on CelebA-HQ (Karras et al., 2018). Baselines use the toolbox's settings (Table 9); FAMILIAR's guidance weights are selected on validation data for at most 6% matching.

Two failure modes. Table 4 separates methods that remove identity from those that preserve attributes. Naive and k-Same methods lose the attributes: k-Same-Furthest lowers matching to 0.17–0.20%, but expression accuracy falls to 15.35% and age error rises from 10.18 to 18.38 years. Gaussian blur ($\sigma = 20$) still leaves 30.81–57.65% of sources matched at 24.56% expression accuracy. The adversarial method and most generative methods show the opposite failure. TIP-IM, AMT-GAN, G2Face, and WeakenDiff keep expression accuracy between 63.27% and 68.90%, yet 38.23–75.90% of sources remain matched. CIAGAN is the only generative baseline with low matching (2.03–2.27%), and its age error (18.47 years) is as high as that of k-Same.

Privacy with utility. FAMILIAR keeps matching at 5.47–5.79% under all three recognizers while retaining 69.21% expression accuracy and an age error of 10.84 years, 1.33 points and 0.66 years from the unmodified images. Among the other methods below 6% under every recognizer (Table 9), the highest expression accuracy is 28.99% (CIAGAN) and the lowest age error 15.91 years (k-Same, $k = 5$). FAMILIAR's utility instead exceeds that of the generative methods that leave 38–63% of sources matched: its age error is 0.62 years below WeakenDiff's and its expression accuracy 2.93 points above AMT-GAN's, while both methods match sources at about 7–11 times FAMILIAR's rates. Its landmark NME (0.3553) is also the lowest in the benchmark, although NME separates most methods only slightly. At validated guidance weights (Appendix B.5), FAMILIAR reaches an operating point that no other evaluated setting attains: matching below 6% under every recognizer, with attribute utility at the level of generative methods that leave far more sources matched. Matching single source images with fixed recognizers is a limited adversary, not a privacy guarantee.

Two interventions, one factorization. Recognition and de-identification are complementary tests of the factorization in Eq. (1). A repertoire fixes ϕ and draws $c \sim p_\theta(c \mid \phi)$, so identity must follow ϕ as the occasion varies; de-identification fixes $c_s = E(x_s)$ and replaces ϕ, so the occasion must follow c_s as identity changes. Both require G to take identity from ϕ and image-specific appearance from c. Any remaining dependence of c on ϕ is a residual source-identity cue for de-identification but, for synthesis, it is the person-specific structure that $p_\theta(c \mid \phi)$ is designed to capture (Section 4.3). Internal controls in Table 10 test the de-identification side one input at a time on RAF-DB (Li et al., 2017) and AgeDB. Resampling the occasion lowers expression accuracy from 82.36% to 55.78% and raises age error from 5.94 to 10.19 years at similar source matching.

4.5 Ablation Studies

Separating the prior from policy. Table 5 evaluates the temperature mixing with candidate ranking. The two effects are additive: ranking adds 0.35–0.38 points with or without higher-temperature proposals, and drawing 20% of candidates at $T = 1.4$ adds 0.12–0.15 points with or without ranking. They act on opposite sides of the identity–variation trade-off: ranking raises ID-agreement but narrows yaw spread, whereas higher-temperature proposals widen it. The combined policy has both a wider yaw spread and higher ID-agreement than unranked $T = 1$ sampling. Ranking also rejects higher-temperature candidates more often, lowering their retained share. About 59% of them survive at $T = 1.4$, against 81% of $T = 1$ candidates, and only 42% survive at $T = 1.8$, where accuracy falls back to that of unranked $T = 1$ sampling (94.31% versus 94.29%). Occasions drawn farther from the prior's high-density region thus *break identity more often*, and the policy works by *exploring slightly beyond the prior and filtering this drift*.

Table 5: Temperature and ranking. ID cos* uses AdaFace-IR101, not the conditioning ArcFace.

Policy	FR Avg. ↑	ID cos*	Yaw IQR	% at $T>1$
T=1; no rank	94.29	0.616	21.1	—
T=1; rank	94.64	0.648	19.6	—
20% T=1.4; no rank	94.41	0.601	23.3	20.0
20% T=1.4; rank	94.79	0.637	22.0	15.3
40% T=1.4; rank	94.68	0.617	24.5	32.3
20% T=1.8; rank	94.31	0.604	25.1	10.8

Occasion representation. Freezing the occasion backbone has the largest effect (94.29% to 93.11%), suggesting that synthesis, not the pretrained features alone, defines the occasion space. Query pooling outperforms mean pooling, and the conditional-prior gain persists under both (1.11 and 0.86 points). Additive noise is best at $\sigma = 0.1$: removing it raises identity agreement but narrows yaw spread; $\sigma = 0.2$ does the reverse. Appendix B.6 gives full results (Tables 11–13).

5 Conclusion

This paper presents FAMILIAR, which treats a virtual identity not as a point with variation added by a fixed rule, but as a learned distribution over the occasions in which that person is photographed. Synthesis defines the occasion space, and one conditional flow shared across identities predicts how each identity occupies it. With the generator fixed, the central finding is that useful variation is not interchangeable across identities. Another person's occasions lower recognition accuracy despite wider pose spread, and occasion priors conditioned on other photographs of the same person fit far better than another person's. Across six priors, the prior that better fits held-out occasions always trains the better recognizer. FAMILIAR gives state-of-the-art synthetic data for face recognition when each identity receives its own occasions. It also de-identifies a photograph by keeping its occasion and giving it a new identity, with little identity leakage and near-original attribute utility.

AI USE STATEMENT

We used generative AI tools to assist with polishing the manuscript text, formatting tables, and finding typos. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

ETHICS STATEMENT

This work trains and evaluates on face images of real people. FAMILIAR is trained on WebFace-21M, a subset of WebFace42M (Zhu et al., 2021); the real-data reference uses CASIA-WebFace (Yi et al., 2014); and evaluation uses public face benchmarks (Section 4). These collections were gathered from the web without individual consent, and we use them only for research under their terms of use. Synthetic identities are screened against a gallery of real-face embeddings (Section 3.3); this is an empirical similarity check against a finite gallery under one recognizer, not a privacy guarantee, and it cannot exclude resemblance to every real person. Our de-identification results measure how often a single de-identified image is matched to its source by fixed recognizers. They do not establish protection against multi-image galleries or retrained recognizers, and because the occasion code is not independent of identity, outputs can retain residual source cues (Table 10). Identity replacement could also be misused to alter photographs of real people. RFW (Wang et al., 2019) categories do not establish biological distinctions, attribute estimators may carry demographic biases, and further limitations are listed in Appendix B.10.

REPRODUCIBILITY STATEMENT

We will release the code, datasets, and model weights. Section 3 and Appendix A specify the model, its conditioning, the flow densities, and the sampling and selection procedure. Appendix B.1 gives the data partitions and seeds, the training schedule and optimization of every component, the release policy, and the downstream recognition protocol. Appendices B.2–B.9 describe the protocol behind each further analysis, including the prior baselines, the de-identification benchmark and its operating-point selection, scaling, robustness, and label fidelity; Appendix B.10 reports generation cost.

REFERENCES

Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. DigiFace-1M: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3526–3535, 2023.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

Fadi Boutros, Eduarda Caldeira, Tahar Chettaoui, and Naser Damer. IDperturb: Enhancing variation in synthetic face generation via angular perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40119–40129, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

A. Mike Burton, Rob Jenkins, Peter J. B. Hancock, and David White. Robust representations for face recognition: The power of averages. *Cognitive Psychology*, 51(3):256–284, 2005.

A. Mike Burton, Rob Jenkins, and Stefan R. Schweinberger. Mental representations of familiar faces. *British Journal of Psychology*, 102(4):943–958, 2011.

A. Mike Burton, Robin S. S. Kramer, Kay L. Ritchie, and Rob Jenkins. Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1):202–223, 2016.

Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M. Salman Asif, and Ziyan Wu. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 918–926, 2024.

Ka-Ho Chow, Sihao Hu, Tiansheng Huang, and Ling Liu. Personalized privacy protection mask against unauthorized facial recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15140 of *Lecture Notes in Computer Science*, pp. 434–450. Springer, 2024.

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

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212, 2020a.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5154–5163, 2020b.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4312–4321, 2019.

Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2021.

Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies: 5th International Workshop, PET 2005, Revised Selected Papers*, volume 3856 of *Lecture Notes in Computer Science*, pp. 227–242. Springer, 2006.

James V. Haxby, Elizabeth A. Hoffman, and M. Ida Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233, 2000.

Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15014–15023, 2022.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Rob Jenkins, David White, Xandra Van Montfort, and A. Mike Burton. Variability in photos of the same face. *Cognition*, 121(3):313–323, 2011.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. DCFace: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12715–12725, 2023.

Minsoo Kim, Min-Cheol Sagong, Gi Pyo Nam, Junghyun Cho, and Ig-Jae Kim. VIGFace: Virtual identity generation for privacy-free face recognition dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10043–10053, 2025.

Yaren Koca and Chris Oriet. From pictures to the people in them: Averaging within-person variability leads to face familiarization. *Psychological Science*, 34(2):252–264, 2023.

Robin S. S. Kramer, Andrew W. Young, and A. Mike Burton. Understanding face familiarity. *Cognition*, 172:46–58, 2018.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2852–2861, 2017.

Xiao Lin, Yuge Huang, Jianqing Xu, Yuxi Mi, Shuigeng Zhou, and Shouhong Ding. UIFace: Unleashing inherent model capabilities to enhance intra-class diversity in synthetic face recognition. In *International Conference on Learning Representations (ICLR)*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5447–5456, 2020.

Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA Janus Benchmark - C: Face dataset and protocol. In *International Conference on Biometrics (ICB)*, pp. 158–165. IEEE, 2018.

Lily Meng and Zongji Sun. Face de-identification with perfect privacy protection. In *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1234–1239. IEEE, 2014.

Yuxi Mi, Zhizhou Zhong, Yuge Huang, Qiuyang Yuan, Xuan Zhao, Jianqing Xu, Shouhong Ding, Shaoming Wang, Rizen Guo, and Shuigeng Zhou. Data synthesis with diverse styles for face recognition via 3DMM-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21203–21214, 2025.

Yuxi Mi, Qiuyang Yuan, Jianqing Xu, Yichun Zhou, Xuan Zhao, Jun Wang, Rizen Guo, and Shuigeng Zhou. SteerFace: Debiasing synthetic face generation via adaptive residue perturbation. *arXiv preprint arXiv:2605.30894*, 2026.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 51–59, 2017.

Elaine M. Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

Hatef Otroshi Shahreza and Sébastien Marcel. HyperFace: Generating synthetic face recognition datasets by exploring face embedding hypersphere. In *International Conference on Learning Representations (ICLR)*, 2025.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2Face: A foundation model for ID-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Jinhyeong Park, Muhammad Shaheryar, Seangmin Lee, Jong Taek Lee, and Soon Ki Jung. FLUID: Training-free face de-identification via latent identity substitution. *arXiv preprint arXiv:2511.17005*, 2026. Version 2.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. SynFace: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10880–10890, 2021.

Kay L. Ritchie and A. Mike Burton. Learning faces from variability. *Quarterly Journal of Experimental Psychology*, 70(5):897–905, 2017.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2–4): 144–157, 2018.

Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. Enhancing facial privacy protection via weakening diffusion purification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8235–8244, 2025.

Marko Radisa Savic and Guoying Zhao. De-identification of facial videos while preserving remote physiological utility. In *British Machine Vision Conference (BMVC)*, 2023.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9. IEEE, 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. arXiv:2508.10104.

Zhonglin Sun, Siyang Song, Ioannis Patras, and Georgios Tzimiropoulos. CemiFace: Center-based semi-hard synthetic face generation for face recognition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35612–35638, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial Faces in the Wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 692–702, 2019.

Hui Wei, Hao Yu, and Guoying Zhao. FDeID-Toolbox: Face de-identification toolbox. *arXiv preprint arXiv:2603.13121*, 2026.

Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin W. Bowyer. Vec2Face: Scaling face dataset generation with loosely constrained vectors. In *International Conference on Learning Representations (ICLR)*, 2025a.

Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin W. Bowyer. Vec2Face+ for face dataset generation. *arXiv preprint arXiv:2507.17192*, 2025b.

Jianqing Xu, Shen Li, Jiaying Wu, Miao Xiong, Ailin Deng, Jiazhen Ji, Yuge Huang, Guodong Mu, Wenjie Feng, Shouhong Ding, and Bryan Hooi. ID^3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77777–77798, 2024.

Haoxin Yang, Xuemiao Xu, Cheng Xu, Huaidong Zhang, Jing Qin, Yi Wang, Pheng-Ann Heng, and Shengfeng He. G^2Face: High-fidelity reversible face anonymization via generative and geometric priors. *IEEE Transactions on Information Forensics and Security*, 19:8773–8785, 2024.

Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3897–3907, 2021.

Michael Yeung, Toya Teramoto, Songtao Wu, Tatsuo Fujiwara, Kenji Suzuki, and Tamaki Kojima. VariFace: Fair and diverse synthetic dataset generation for face recognition. *arXiv preprint arXiv:2412.06235*, 2025. Version 2, revised April 17, 2025.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-Makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1252–1258, 2021.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

Ce Zheng, Matias Mendieta, and Chen Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3146–3155, 2023.

Tianyue Zheng and Weihong Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, 2018.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

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

A METHOD DETAILS

Representation and conditioning. The occasion encoder uses DINOv3 ViT-L/16 with 448×448 inputs and eight attention heads for query pooling, producing a 256-dimensional code. The RAE encoder uses the same backbone family, but the two outputs serve different roles: the occasion code provides a global condition, while the RAE latent preserves the spatial structure used for image synthesis.

RAE. The encoder A is a frozen DINOv3 ViT-L/16 applied to 256×256 images, giving a 16×16 grid of 1024-dimensional patch tokens. The latent averages the patch tokens of seven blocks, each passed through the final layer normalization with its affine parameters removed, and adds the spatial mean of the last of these blocks; each channel is then standardized with fixed statistics, $(z-\mu)/\sqrt{\sigma^2+10^{-5}}$. The decoder D is a ViT in the ViT-XL configuration. A linear layer lifts each latent token from 1024 to 1152 channels, 28 transformer blocks with 16 heads and feed-forward width 4096 process the 256 tokens with frozen positional embeddings, and a final linear layer predicts the $16 \times 16 \times 3$ pixels of each patch, which are reassembled into the image. The decoder has 415.65M parameters.

Generator. The generator follows the DiT^{DH} design of RAE (Zheng et al., 2026): a transformer encoder followed by a shallow, wide head. Since the latent grid is already 16×16, both use patch size 1, giving 256 signal tokens. The encoder appends 16 condition tokens to them, four for time, four for identity, and eight for occasion, and processes the 272 tokens of width 768 with 12 self-attention blocks (12 heads, RMSNorm, SwiGLU feed-forward width 2048). In the encoder, identity and occasion act only through self-attention, without cross-attention or adaptive normalization. Two-dimensional rotary position embeddings are applied to the signal tokens, while the condition tokens receive zero rotation and carry no grid position. Each condition group follows Eq. (2): U_a is initialized with standard deviation $768^{-1/2}$ and Π_a with zeros, so training starts from an unconditional model, and condition dropout passes the learned null vector through the same projection. Time is encoded by Gaussian Fourier features with 256 fixed random frequencies, mapped by an MLP to width 768 and added to four learned time tokens. The head re-embeds the noisy latent h_t with its own patch embedding and applies two blocks of width 1024 (16 heads, SwiGLU width 2730), whose adaptive layer normalization is modulated token by token by the encoder's signal tokens after a SiLU and a linear map from 768 to 1024 channels; a final adaptively normalized linear layer outputs $\widehat{h}_\gamma$. The head thus receives ϕ and c only through the encoder output. The generator has 133.55M parameters, 85.05M of them in the encoder blocks and 37.78M in the head blocks.

Components and training stages. Figures 2 and 3 draw the generator as a single block. It has three parts: the frozen RAE encoder A, used only in training to map photographs to latents $h = A(x)$; the diffusion transformer, which predicts $\widehat{h}_\gamma$ from h_t, t, and the identity and occasion tokens; and the RAE decoder D, which maps the integrated latent to pixels. Training has three stages. The decoder is first adapted to face images with A frozen. The occasion encoder and the transformer are then trained jointly with Eq. (3), while A, D, and the ArcFace encoder stay fixed. Finally, with all of these frozen, p_ψ and p_θ are fitted by maximum likelihood to the embeddings and deterministic codes of training photographs. Generation integrates Eq. (4) from $h_0 = \xi$ with 25 Euler steps and decodes the result with D; de-identification replaces the sampled code by $E(x_s)$ and uses $\mathbf{w} = (1.5, 1)$.

Identity information in the occasion code. Condition dropout gives the occasion code an incentive to carry identity: in the 9% of training steps where ϕ is dropped and $\widetilde{c}$ is kept, the transformer must predict the whole latent from the code. Two measurements show how much identity the code carries in practice. Renders from the code alone stay near an ArcFace cosine of 0.18 to the source identity, 0.57 below joint renders (Figure 4(b)), and they match their source in 15.19% of cases, against 96.31% for joint renders (Table 10). Dropout is needed for the nested guidance of Eq. (4), and removing it changes accuracy only from 94.29% to 94.13% (Table 11). The additive noise also remains effective: raising σ from 0 through 0.05 and 0.1 to 0.2 lowers identity agreement from 0.635 to 0.590 and widens yaw spread from 17.0 to 23.5, both monotonically.

A.1 Clean-Latent Prediction and Rectified-Flow Velocity

We use $h_t = (1-t)\xi + th$, with noise at $t = 0$ and data at $t = 1$. The pathwise target velocity is $h - \xi$. The clean prediction in Eq. (4) satisfies

$$\begin{aligned} v_\gamma(h_t, t; \phi, c) - (h - \xi) &= \frac{\widehat{h}_\gamma - h_t - (1-t)(h-\xi)}{1-t} \\ &= \frac{\widehat{h}_\gamma - h}{1-t}. \end{aligned} \tag{8}$$

Hence,

$$\|\widehat{h}_\gamma - h\|_2^2 = (1-t)^2 \|v_\gamma - (h-\xi)\|_2^2. \tag{9}$$

The clean-latent prediction objective is unweighted, while the equivalent velocity loss is weighted by $(1-t)^2$. The velocity expression applies for $t < 1$, with the data endpoint treated as the terminal limit. Reversing the time convention changes the signs and endpoint factors but leaves the interpolation unchanged.

A.2 Conditional Couplings and Exact Model Density

Let $d_c = 256$ denote the occasion-code dimension and $y^{(0)} = (c - \mu) \oslash \tau$, where μ and τ are fixed coordinate-wise training statistics and $\oslash$ denotes elementwise division. Standard deviations are clamped below by 10^{-6}. In coupling layer ℓ, a binary mask m_ℓ selects the coordinates left unchanged. A multilayer perceptron takes these coordinates and the identity feature $r_\theta(\phi)$ to predict a shift b_ℓ and bounded log-scale s_ℓ:

$$\begin{aligned} (b_\ell, \widetilde{s}_\ell) &= \mathrm{MLP}_\ell\Big(m_\ell \odot y^{(\ell-1)}, r_\theta(\phi)\Big), & s_\ell &= 2\tanh(\widetilde{s}_\ell), \\ y^{(\ell)} &= m_\ell \odot y^{(\ell-1)} + (1 - m_\ell) \odot \Big(y^{(\ell-1)} \odot e^{s_\ell} + b_\ell\Big). \end{aligned} \tag{10}$$

Masks alternate between odd and even coordinates. Each layer is inverted by subtracting the shift and dividing by the scale on the transformed coordinates, since the conditioning coordinates remain unchanged. With $z = y^{(L)}$, the complete log-density is

$$\log p_\theta(c \mid \phi) = -\frac{\|z\|_2^2}{2} - \frac{d_c}{2}\log(2\pi) + \sum_{\ell=1}^{L}\sum_{k=1}^{d_c}(1 - m_{\ell,k})s_{\ell,k} - \sum_{k=1}^{d_c}\log\tau_k. \tag{11}$$

The final term accounts for standardization and expresses the density in the original occasion coordinates. We use $L = 12$ coupling layers, hidden width 512, and an identity-conditioning network of width 512. Each coupling conditioner has two SiLU hidden layers and a zero-initialized output layer. The unconditional identity prior uses the same construction without identity conditioning. Normalizing its samples induces a distribution on the unit sphere; the ambient flow likelihood does not give the density of this normalized distribution.

A.3 Sampling and Candidate Selection

Let $\mathcal{R}$ denote the gallery of normalized real-face embeddings. We accept a proposed virtual embedding ϕ^* when

$$\max_{r \in \mathcal{R}} \langle \phi^*, r \rangle < \tau_{\mathrm{id}}. \tag{12}$$

This check is specific to the gallery, embedding model, and threshold τ_{id}.

For each accepted identity, we generate $M \geq m$ candidates and retain m images. The occasion policy assigns $\mathrm{round}(\alpha M)$ candidates to a temperature $T > 1$ and the remainder to temperature 1. A temperature-T occasion is sampled as

$$c = f_\theta^{-1}(Tz; \phi^*), \qquad z \sim \mathcal{N}(0, I_{d_c}). \tag{13}$$

At $T = 1$, this samples from $p_\theta(c \mid \phi^*)$. Other temperatures change the Gaussian base and the induced occasion distribution, without selecting an exact density tail or guaranteeing a particular attribute. Each candidate receives the score $s_j = \langle \phi(\widehat{x}_j), \phi^* \rangle$, and we retain the m highest-scoring images. Ranking favors agreement with the assigned identity but imposes no minimum similarity. Temperature allocation and ranking both alter the final training distribution, so the retained higher-temperature fraction need not equal α.

B Extended Experimental Protocols and Results

B.1 Data Boundaries, Training Budgets, and Model Selection

Training data. FAMILIAR is trained on WebFace-21M, the restored subset of WebFace42M (Zhu et al., 2021) that was used to train Arc2Face (Paraperas Papantoniou et al., 2024); our copy contains 20.7M images of 1.10M identities. Validation identities for model selection and held-out NLL, and the analysis identities used for prior evaluation and few-shot enrollment (Appendices B.2 and B.4) are held out from WebFace-21M training. The CASIA-WebFace split starts from our copy of the dataset (Yi et al., 2014), which contains 490,592 images and 10,572 identities. Identities are sorted by canonical label and shuffled with seed 1234, then divided into 500 validation, 500 held-out, and 9,572 training identities. Every photograph inherits its identity's partition. The inventory count describes the starting pool, not the training subset. The real-data reference is trained on the training identities. Total real-data exposure also includes the pretrained components used by each method.

Pretraining exposure. FAMILIAR uses public DINOv3 initialization and a frozen WebFace600K ArcFace-R50 identity encoder. Other methods retain their prescribed pretrained image and identity components. The primary comparison matches synthetic dataset size (except the 1.0M entries marked * in Table 1), but not the real data used to train each generator, downstream recognition training, pretraining, or total compute. Whether benchmark identities appear in WebFace-21M or in external pretraining data is not established. "Unseen identity" therefore refers to identities held out from FAMILIAR's training data.

Published results. Baseline verification accuracies in Table 1 are the values reported in the original papers for 10K identities with 50 images each. In the 1.2M column, DCFace, Arc2Face, and CemiFace use 20K identities with 50 images plus 40K with 5, VariFace 60K identities with 20 images, and the other methods 24K with 50; UIFace, Vec2Face+, and IDperturb report no 1.2M result and are shown at 1.0M (20K identities with 50 images). Their generators are trained on CASIA-WebFace, except Vec2Face+ (WebFace4M) and Arc2Face (the WebFace-21M subset also used by FAMILIAR), and their recognizers follow each paper's backbone, loss, and schedule, so these differences combine generator, training data, and recognition protocol. IJB-C values for Vec2Face+ and IDperturb in Table 1 are likewise those reported for their 0.5M-image datasets ($10\text{K} \times 50$, the same scale as ours), each under its own recognizer and IJB-C protocol; Vec2Face+ evaluates still images only and reports TAR only at FAR 10^{-4}. VariFace, SteerFace, and MorphFace release neither data nor code, so they are compared only through their reported values. VariFace also reports 95.67% at 6M images with 60K identities and 100 images each (Yeung et al., 2025). The ranking in Section 4.2 concerns the listed methods at these dataset sizes.

Complete FAMILIAR schedule. The schedule adapts the RAE decoder on WebFace-21M for 80 epochs on eight B300 GPUs with effective batch size 256, pixel MSE, LPIPS weight 1, and latent-noise standard deviation 0.2. An adversarial loss with weight 0.2 is activated after three epochs. The decoder learning rate decays from 2×10^{-5} to 2×10^{-6}, while the encoder remains frozen. Joint training then updates the occasion encoder and generator on WebFace-21M for 300,000 steps with global batch size 192 on four H200 GPUs. It uses AdamW with $\beta = (0.9, 0.95)$ and zero weight decay, bfloat16 mixed precision, and gradient-norm clipping at 1.0; the learning rate warms up to 2×10^{-4} over 3,000 updates and then follows a cosine decay to 2×10^{-5}. The occasion branch starts from DINOv3, the generator starts from random initialization, and the decoder remains fixed. Training uses the alignment-free objective in Eq. (3); alignment is added only in its designated ablation. An EMA with decay 0.9998 is maintained for diagnostics, while final non-EMA weights are used consistently for code extraction and generation.

Both flows are fitted to the codes and embeddings of WebFace-21M training images and optimized with AdamW ($\beta = (0.9, 0.99)$) for 200 epochs with batch size 4,096, 500 warm-up updates, zero weight decay, and cosine learning-rate decay from 10^{-3} to 10^{-5}. Normalization statistics use training codes only. All reported NLLs include the Jacobian of this fixed normalization.

Numerical time convention. Our implementation uses noise fraction $s = 1 - t$. Given $a \sim \mathcal{N}(0, 1)$ and $r = \mathrm{sigmoid}(a)$, training samples $s = 8r/(1 + 7r)$ and sets $t = 1 - s$. During generation, the denominator $1 - t$ is clamped below by 0.05 when converting clean predictions to velocities. Euler integration uses a uniform grid without the training-time shift. This clamp approximates the unclipped field derived in Appendix A.1.

Table 6: Finite-sample covariance controls.

Source	Within person	Between people	Effective rank
Occasion codes	$0.428_{\pm 0.005}$	$0.495_{\pm 0.002}$	$7.9_{\pm 0.1}$
Isotropic	$0.444_{\pm 0.005}$	$0.445_{\pm 0.004}$	$14.6_{\pm 0.4}$
Shared cov.	$0.385_{\pm 0.002}$	$0.384_{\pm 0.003}$	$8.1_{\pm 0.2}$

200 identities, 20 observations, fixed 32-D projection. Split-half covariance distances; effective rank ≤ 19. Shared cov.: shared-covariance Gaussian. $\pm$: SD across replicates.

Downstream training. Primary recognizers use IR-50 with ArcFace scale 64 and margin 0.5, batch size 512, SGD momentum 0.9, and weight decay 5×10^{-4}. The learning rate starts at 0.1 and decreases tenfold at epochs 20, 30, and 36 of the 40-epoch schedule. Images are landmark-aligned to 112×112 and converted to RGB in $[-1, 1]$. Augmentation includes horizontal flipping with probability 0.5 and random resized cropping with probability 0.2, using area 0.9–1.0 and aspect ratio 0.95–1.05. Synthetic-data training contains no real images. Sampling is uniform over images, with equal class sizes in each synthetic dataset. The final epoch is evaluated without benchmark-based early stopping.

LFW, CPLFW, AgeDB-30, and CALFW use their official 6,000 pairs; CFP-FP uses its official 7,000 pairs. Thresholds follow the prescribed folds. For IJB-C, normalized features from each image and its horizontal flip are averaged. Frame features are then averaged within each medium, and media features are averaged and normalized to form a template. TAR is computed from the full official template pair list at the specified FARs. Test scores do not select hyperparameters. The conditioning network is not used to evaluate downstream recognizers. Each of the five replicates uses a new dataset-generation seed and downstream training seed with a fixed generator; recognition $\pm$ values are SDs across these replicates and exclude generator-training variability.

Release policy. Normalized identity proposals are accepted when their maximum cosine similarity to a fixed 200,000-image training gallery is below 0.3. No pairwise uniqueness constraint is imposed on accepted virtual identities. For each identity, 65 candidates are rendered: 13 at $T = 1.4$ and 52 at $T = 1$. The 50 candidates with the highest conditioning-ArcFace similarity are retained, without a minimum-similarity requirement. Prior ablations reuse the gallery and identity proposals but generate exactly 50 images per identity without ranking. Competing methods retain their own generation policies, so equal retained dataset sizes need not imply equal generation costs. Our released datasets contain 200K, 0.5M, 1.2M, and 10M images and are generated under this policy.

B.2 Prior Families and Cross-Photograph Conditioning

Baseline definitions. The empirical baseline samples training codes with replacement, independently of the generated identity. Conditional Gaussian baselines use an MLP to predict the mean and positive diagonal covariance from the identity embedding; the rank-32 variant adds $UU^\top$ to that covariance. The mixture model predicts eight weights, means, and diagonal covariances, with density evaluated by log-sum-exp. Gaussian scale floors are selected from $10^{-3}, 10^{-2}, 10^{-1}$ in standardized coordinates. Validation NLL selects hidden width from 256, 512, and 1,024, depth from two and four, and learning rate from 3×10^{-4} and 10^{-3}. The flow uses the architecture in Appendix A.2. All models receive the same number of passes over the training cache, although their parameter counts may differ. The discrete empirical bank has no continuous density for query-code NLL evaluation, so that entry is omitted.

Support–query evaluation. The evaluation selects 200 analysis identities with at least 25 usable images, based only on image counts. Five randomly selected support images provide a normalized mean identity embedding, while 20 disjoint query images provide occasion observations. Conditional models are retrained using the corresponding constructions on training identities, avoiding a change of conditioning only at test time. In the same-person other-photo variant, the conditioning photograph always differs from the target photograph. Per-image NLL instead uses each query's own embedding. Samples from the per-image variant form an equal mixture over the 20 query-conditioned priors and therefore use information unavailable to the support-mean model. Because these NLLs are computed on the query sets of the analysis identities, the per-image value in Table 3 need not equal the held-out NLL in Table 2. Generation uses the per-image prior because its condi-

Table 7: Occasion and rendering-noise responses.

Response	Occasion	Noise	Interaction
Occasion re-encoding	84.68 ± 0.19	7.16 ± 0.11	8.17 ± 0.17
Pose probe	72.66 ± 0.27	11.48 ± 0.33	15.86 ± 0.26
Expression probe	67.63 ± 0.25	14.36 ± 0.19	18.01 ± 0.29
Independent FR	48.86 ± 0.24	28.51 ± 0.41	22.63 ± 0.36

$128 \times 6 \times 6$ identity–occasion–noise design. Energy shares (%) sum to 100 before rounding; $\pm$: SD across five replicate tensors. Re-encoding uses FAMILIAR; other probes are independent.

tioning matches both the generator, which is trained on $\phi(x)$ and $E(x)$ of the same photograph, and the identity prior, which is fitted to per-image embeddings.

For each identity, 20 occasions are drawn in each of 50 sampling repetitions to match the query-set size. Sliced Wasserstein distance (SWD) averages one-dimensional Wasserstein-1 distances over 128 fixed random directions in training-standardized coordinates. Covariance distance is the Frobenius distance between trace-normalized covariance matrices after projection onto 32 training-set principal components. Distances are averaged across sampling repetitions and then equally across identities; NLL also weights identities equally. Permuted support conditions use a derangement within each analysis partition.

The ordering in Table 3 illustrates contributions from both person-associated and photograph-specific information. Finite sample size limits the interpretation of low rank: with 20 query images, sample covariance has rank at most 19, including after projection to 32 dimensions. Table 6 therefore includes isotropic and shared-covariance Gaussian controls with matched sample sizes. The shared covariance is estimated from training-set within-person residuals with diagonal shrinkage coefficient 0.1. This control can produce low effective rank without a within–between covariance gap. Distances compare disjoint ten-image halves; between-person comparisons use a derangement with the same half-sample sizes. Effective rank is $\exp[-\sum_k q_k \log q_k]$, where q_k are eigenvalues normalized to sum to one and $0 \log 0 = 0$. These controls separate finite-sample effects and shared anisotropy from person-associated covariance structure. The NLL ordering of the Gaussian families in Table 2 reflects both correlation among coordinates and its dependence on identity; Table 6 isolates the latter, since the shared-covariance control has correlated coordinates but no between-person gap (0.385 versus 0.384), whereas occasion codes show one (0.428 versus 0.495).

Independent generation diagnostics. Identity agreement is defined as mean pairwise cosine similarity within each generated class under frozen AdaFace-IR101, averaged equally across identities. It is distinct from similarity to the ArcFace conditioning vector, which lies in a different embedding space. Pose variation is the within-class interquartile range of signed yaw from a frozen pose estimator, averaged across classes. A wider yaw range alone establishes neither realism nor useful variation; downstream recognition evaluates its training value.

Reading the wrong-identity control. Pairs (ϕ_i, c) with $c \sim p_\theta(c \mid \phi_{\pi(i)})$ lie outside the joint distribution on which the generator was trained, and their lower identity agreement (0.563 versus 0.616; Table 2) shows that another person's occasions partly conflict with the rendered identity. This conflict is part of what non-interchangeability means. Tables 3 and 6 avoid this conflict: they compare densities and covariances of codes without generating any image. The unconditional flow, which is equally blind to ϕ_i, matches the conditional flow's yaw spread (21.1) yet trains a recognizer 1.11 points worse. In enrollment (Table 8), wrong-identity renders still improve on real supports alone (73.78% versus 72.66%), so they are usable images, but they stay below conditional occasions (77.21%).

B.3 Occasion Variation and Rendering Noise

Initial latent noise ξ already introduces variation into generation. To separate these sources of variation, we cross six independently sampled occasion codes with six initial latent noises for each of 128 virtual identities, producing 4,608 images per replicate. Identity, generator weights, guidance, and the 25-step solver remain fixed. For each probe, a two-factor decomposition separates centered squared response into occasion, noise, and interaction terms, with normalized shares averaged across identities.

Table 8: Recognition from limited real supports.

Enrollment policy	1 real	2 real	4 real	8 real
Real supports only	72.66 ± 1.08	77.82 ± 1.19	84.30 ± 0.87	90.63 ± 0.39
Image augmentation	74.59 ± 0.90	79.61 ± 1.03	85.41 ± 0.78	91.24 ± 0.48
Unconditional occasions	74.91 ± 0.95	79.92 ± 0.91	85.56 ± 0.83	91.38 ± 0.46
Wrong-identity occasions	73.78 ± 0.96	78.76 ± 1.17	84.85 ± 0.83	90.92 ± 0.48
Conditional occasions	77.21 ± 0.90	82.05 ± 0.68	87.14 ± 0.66	92.17 ± 0.61

200 identities; 2,000 queries; 16 added supports per augmentation policy. Rank-1 accuracy (%); $\pm$: SD across five replicate episodes. Independent frozen query encoder; no recognizer fine-tuning.

Different sources of change. In Table 7, occasion accounts for 72.7% of the pose-response energy, with a similar ordering for expression. Independent probes are used for both responses. Occasion re-encoding shows a larger share, but reuses FAMILIAR's encoder and therefore provides an internal consistency check rather than an independent measurement. The recognition response also varies with occasion. This indicates imperfect invariance of the generated-image embedding and does not, by itself, establish identity information in the occasion code.

B.4 Few-Shot Enrollment with Predicted Occasions

To connect repertoire modeling with recognition from limited observations, the enrollment task augments a small real support set for each unseen identity. It selects 200 eligible identities from the held-out analysis partition, with ten query photographs and $s \in \{1, 2, 4, 8\}$ disjoint support photographs per person. The normalized mean support embedding conditions both the generator and a support-conditioned occasion prior. Query photographs are excluded from generation and prior fitting. For each support size s, the support-conditioned prior is retrained on training identities with s-image support means, so that enrollment uses the conditioning seen in training (Appendix B.2).

Each augmentation policy adds 16 enrollment images per identity. A frozen, independent AdaFace encoder maps real and generated supports into an enrollment space. Their mean embeddings receive weights 1 and 0.25, respectively, before final normalization. The synthetic weight is selected on separate validation identities and shared by image augmentation and all occasion policies. The recognition network is not fine-tuned on the 200 evaluation identities.

A larger repertoire helps most when observations are scarce. With one real support, conditional occasions increase rank-1 accuracy by 4.55 percentage points over real supports alone (Table 8). The difference decreases as more real photographs become available. Unconditional and wrong-identity occasions give smaller gains while retaining the same rendered identity condition. This enrollment task evaluates a recognition use of predicted variation rather than a behavioral model of human familiarity.

B.5 De-identification Protocols and Interface Controls

Toolbox benchmark. Tables 4 and 9 follow FDeID-Toolbox (Wei et al., 2026). Faces are detected and aligned with RetinaFace (Deng et al., 2020a) using five landmarks. Privacy is evaluated on AgeDB with three frozen recognizers, ArcFace-R100, CosFace-R50, and AdaFace-R50, and reported as TAR at FAR 10^{-3} under the toolbox's verification protocol, which compares de-identified faces with original faces; the Original row applies the same protocol to unmodified images. Utility is age MAE on AgeDB, seven-class POSTER expression accuracy on AffectNet, and HRNet landmark NME on CelebA-HQ at 256×256. Adversarial baselines use an ℓ_∞ budget $\epsilon = 8/255$, step size $2/255$, and 20 iterations with ArcFace as the surrogate. All baseline values are the toolbox's re-implementations at its reported settings; each adversarial and generative baseline is evaluated at a single setting. The privacy recognizers differ from FAMILIAR's conditioning encoder, an ArcFace-R50 trained on WebFace600K, and no utility network is used for FAMILIAR generation, ranking, or test-time optimization. Replacement identities are drawn from p_ψ independently of source appearance. Matching against the source image evaluates a limited adversary and does not establish protection against multi-image galleries or retrained recognizers.

Operating point. Generation settings are selected on a validation split disjoint from all test images, maximizing expression accuracy subject to source matching of at most 6% under all three recognizers. For FAMILIAR, the search uses $w_\phi \in \{0.75, 1, 1.25, 1.5, 2\}$ and $w_c \in \{0.6, 0.8, 1\}$. The

Table 9: Complete FDeID-Toolbox benchmark.

Method	Type	AgeDB TAR ↓			Attribute utility		
		ArcFace	CosFace	AdaFace	Age ↓	Expr. ↑	NME ↓
Original	—	95.82	95.62	93.57	10.18	70.54	0.3601
Blur (σ=10)	Naive	86.57	84.03	83.03	12.00	54.09	0.3609
Blur (σ=20)	Naive	48.63	57.65	30.81	14.18	24.56	0.3620
Pixelation (8×8)	Naive	4.11	2.54	28.38	15.26	14.79	0.3615
Pixelation (16×16)	Naive	0.00	0.00	0.00	18.93	15.55	0.3772
Black mask	Naive	0.30	0.43	0.63	18.31	14.63	0.8733
k-Same (k=5) (Newton et al., 2005)	k-Same	0.93	0.87	0.50	15.91	18.01	0.3751
k-Same (k=10) (Newton et al., 2005)	k-Same	0.73	1.13	1.33	16.00	20.27	0.3802
k-Same-Sel. (k=5) (Gross et al., 2006)	k-Same	0.70	0.70	0.57	17.64	15.87	0.3923
k-Same-Sel. (k=10) (Gross et al., 2006)	k-Same	0.33	0.43	0.40	17.81	15.55	0.3973
k-Same-Fur. (k=10) (Meng & Sun, 2014)	k-Same	0.20	0.20	0.17	18.38	15.35	0.3974
MI-FGSM (Dong et al., 2018)	Adv.	82.52	85.42	82.12	10.29	69.26	0.3601
PGD (Madry et al., 2018)	Adv.	88.76	88.39	85.79	10.04	69.91	0.3599
TI-DIM (Dong et al., 2019)	Adv.	67.99	69.33	77.24	10.91	68.70	0.3628
TIP-IM (Yang et al., 2021)	Adv.	61.58	70.56	75.90	10.95	68.90	0.3630
Chameleon (Chow et al., 2024)	Adv.	89.89	88.73	89.26	9.95	70.30	0.3599
Adv-Makeup (Yin et al., 2021)	Gen.	95.23	94.76	93.53	10.28	70.48	0.3613
CIAGAN (Maximov et al., 2020)	Gen.	2.27	2.10	2.03	18.47	28.99	0.4242
AMT-GAN (Hu et al., 2022)	Gen.	60.74	54.17	46.80	16.95	66.28	0.3595
DeID-rPPG (Savic & Zhao, 2023)	Gen.	95.10	95.00	93.66	10.78	69.38	0.3627
G2Face (Yang et al., 2024)	Gen.	62.98	55.57	52.77	11.82	63.27	0.3629
WeakenDiff (Salar et al., 2025)	Gen.	54.66	39.28	38.23	11.46	65.36	0.3603
FAMILIAR	Gen.	5.62	5.79	5.47	10.84	69.21	0.3553

Metrics and types as in Table 4. Adversarial methods: $\epsilon = 8/255$. Baseline values from Wei et al. (2026), Tables 2 and 3.

Table 10: Identity and occasion controls.

Identity, occasion	Match ↓	Age ↓	Expr. ↑	Pose ↓
Same, source	$96.31_{\pm 0.08}$	$5.20_{\pm 0.08}$	$85.69_{\pm 0.26}$	$2.27_{\pm 0.03}$
New, source	$4.42_{\pm 0.04}$	$5.94_{\pm 0.04}$	$82.36_{\pm 0.13}$	$3.61_{\pm 0.08}$
New, resampled	$4.08_{\pm 0.08}$	$10.19_{\pm 0.11}$	$55.78_{\pm 0.08}$	$13.86_{\pm 0.21}$
Null, source	$15.19_{\pm 0.13}$	$5.68_{\pm 0.02}$	$83.11_{\pm 0.22}$	$3.42_{\pm 0.08}$

Internal protocol (Appendix B.5.1), not comparable to Table 4. Match: ArcFace-IR50 source matching on RAF-DB (%); age: MAE (years) on AgeDB; expr.: RAF-DB accuracy (%); pose: change (°). "New, source" is the FAMILIAR configuration of Table 4. ±: SD across replicates.

selected setting, $w_\phi = 1.5$ and $w_c = 1.0$, overrides the unit-guidance synthesis default and is used for every de-identification result. No evaluated setting has a maximum matching rate between FAMILIAR's 5.79% and the 28.38% of 8×8 pixelation (Table 9), so any limit in this range selects the same set of low-matching methods.

Complete benchmark. Table 9 lists every toolbox setting. Methods that leave matching near the unmodified level, such as PGD, Chameleon, Adv-Makeup, and DeID-rPPG, also keep utility near that level, whereas every setting below 6% matching under all three recognizers other than FAMILIAR has expression accuracy of at most 28.99% and age error of at least 15.91 years.

B.5.1 Source-Occasion and Identity Controls

Table 10 varies one input at a time under an internal protocol with its own estimators, so its values are not comparable to Table 4. Expression uses the 3,068-image RAF-DB basic-expression test split and a frozen seven-class POSTER classifier trained on its training partition. Age uses 5,000 AgeDB images whose identities are disjoint from a 1,000-image validation subset, with a frozen ResNet-50 scalar regressor trained on IMDB-WIKI (Rothe et al., 2018). The ArcFace-IR50 source-match threshold is calibrated once at FAR 10^{-3} on 100,000 real impostor pairs and then held fixed. Pose change is the mean absolute difference in estimated yaw, pitch, and roll, rather than error against ground-truth 3D pose. Each source produces one image; invalid outputs count as incorrect expression predictions and privacy failures and receive penalties of 100 years for age error and 90 degrees for pose change.

Table 11: Representation sensitivity.

Variant	FR Avg.	ID cosine	Yaw IQR
Mean pooling	$93.82_{\pm 0.08}$	0.602	19.8
Frozen backbone	$93.11_{\pm 0.09}$	0.606	17.2
No additive noise	$93.88_{\pm 0.09}$	0.635	17.0
Noise 0.05	$94.26_{\pm 0.04}$	0.630	19.5
Noise 0.10	$94.29_{\pm 0.07}$	0.616	21.1
Noise 0.20	$93.95_{\pm 0.10}$	0.590	23.5
Code dim. 128	$94.01_{\pm 0.05}$	0.611	18.6
Code dim. 512	$94.11_{\pm 0.13}$	0.632	20.2
Alignment loss	$94.12_{\pm 0.10}$	0.626	20.6
No cond. dropout	$94.13_{\pm 0.08}$	0.622	20.0

$T = 1$, no ranking. Default: query pooling, trainable backbone, $d_c = 256$, noise 0.10, no alignment. One factor changes per row; ID/yaw follow Table 2. $\pm$: SD across replicate seeds.

Table 12: Pooling and prior conditioning.

Pooling	Uncond. prior	Cond. prior
Mean	$92.96_{\pm 0.17}$	$93.82_{\pm 0.08}$
Learned query	$93.18_{\pm 0.06}$	$94.29_{\pm 0.07}$

FR accuracy (%); $T = 1$, no ranking. Each representation has separately fitted priors.

Table 13: Identity proposal distributions.

Identity source	Accept. (%)	Nearest cosine	FR Avg.
Isotropic	$99.68_{\pm 0.08}$	0.168	$93.81_{\pm 0.10}$
PCA Gaussian	$96.22_{\pm 0.13}$	0.277	$94.64_{\pm 0.09}$
Identity flow	$95.43_{\pm 0.12}$	0.289	$94.79_{\pm 0.06}$

Same occasion flow, gallery and release policy; 10K accepted identities. Isotropic: normalized Gaussian. Nearest cosine is measured among accepted identities in conditioning space.

Source-identity reconstruction retains both source conditions and resamples only rendering noise, so it retains the source identity. Identity replacement with the source occasion is the FAMILIAR configuration of Table 4. The resampled-occasion control replaces $E(x_s)$ with a draw from the new identity's prior. The null-identity control retains $E(x_s)$ but drops the identity condition, which is a different operation from sampling a virtual identity. These comparisons examine whether preserving source appearance retains useful attributes alongside identifying information. The resampled-occasion control shares no input with its source, so its 4.08% matching, which counts invalid outputs as privacy failures, is the floor of this protocol: keeping the source occasion raises matching only to 4.42%, whereas dropping the identity condition raises it to 15.19%. Age depends on both inputs, because identity and occasion embeddings can each encode age-related information.

B.6 Additional Ablations

Both pooling variants in Table 12 are trained end-to-end with otherwise identical architectures. Conditional and unconditional flows are then fitted separately in each code space with the same optimization budget. The factorial compares recognition accuracy, since raw NLLs across different learned representations are not directly comparable. Table 11 gives the complete single-factor sensitivity study. Each variant retrains the generator and refits the priors (code dimension also resizes the conditioning projections), so none is a fixed-generator intervention.

The alignment variant adds an MSE loss with coefficient 0.5 between the output of the sixth encoder block, linearly projected from 768 to 1024 channels, and the fixed clean RAE tokens. The protocol selects the main-model setting on validation data; the test table reports sensitivity without using test results for tuning.

The PCA Gaussian and ambient identity flow in Table 13 use the same training embeddings, and all proposals are normalized before screening. Normalized isotropic Gaussian samples are uniform on the unit sphere. Since the accepted identity sets differ across rows, this comparison evaluates identity generation rather than the occasion prior. The small difference between PCA and flow does not explain the larger conditional-occasion gain. Low similarity among proposed embeddings alone also does not establish that the generator can realize those identities.

B.7 Dataset Scale and Training Compute

Table 14 holds the FAMILIAR generator fixed and increases the number of virtual identities, retaining 50 images per identity. With 40 recognition epochs, larger datasets also entail more image exposures and updates. Table 15 controls this difference by fixing the update budget to the 0.5M-image, 40-epoch run. It uses uniformly shuffled samples and a learning-rate schedule indexed by the fraction of completed updates. Larger datasets consequently receive fewer effective passes.

Table 14: Scaling with a fixed epoch budget.

Images (identities × 50)	VariFace Yeung et al. (2025)	SteerFace Mi et al. (2026)	FAMILIAR	FAMILIAR IJB-C
0.125M (2,500 × 50)	—	—	92.84±0.10	90.02
0.25M (5,000 × 50)	—	—	93.85±0.08	91.52
0.5M (10,000 × 50)	94.05	93.72	94.79±0.06	92.93
1.2M (24,000 × 50)	—	94.37	95.51±0.09	94.18
3M (60,000 × 50)	95.48	—	96.08±0.05	94.85
6M (120,000 × 50)	—	—	96.24±0.03	95.12

40 recognition epochs for FAMILIAR; baseline entries are values reported in the original papers for the same number of identities and images per identity, under each paper's recognition protocol (—: not reported). 50 images per identity. Average verification accuracy (%); FAMILIAR IJB-C: TAR at FAR 10^{-4} (%). Larger datasets require more updates.

Table 15: Fixed recognition compute.

Synthetic images	FR Avg.
0.5M	$94.79_{\pm 0.06}$
1.2M	$95.50_{\pm 0.08}$
3M	$95.61_{\pm 0.07}$
6M	$95.57_{\pm 0.05}$

Update budget equals 0.5M images × 40 epochs. Generator, augmentation and optimizer fixed.

Table 16: Fixed image budget.

Identities	Images per identity	FR Avg.
5,000	100	$93.65_{\pm 0.05}$
10,000	50	$94.29_{\pm 0.07}$
20,000	25	$94.09_{\pm 0.06}$
25,000	20	$93.93_{\pm 0.08}$

0.5M images; matched recognition updates; $T = 1$, no ranking.

Table 17: Occlusion and demographic robustness.

	RFW verification						ROF rank-1	
Method	African	Asian	Caucasian	Indian	Avg.	Gap ↓	Masked	Sunglasses
VariFace (Yeung et al., 2025)	88.95	87.33	92.95	89.88	89.78	5.62	—	—
FAMILIAR	90.03	89.23	94.44	90.92	**91.15**	**5.21**	60.17	73.62

Accuracy (%). Gap: largest minus smallest RFW group accuracy (pp). ROF: masked and sunglasses probes, rank-1 identification (%), reported separately from verification. VariFace values are from its paper, which does not report ROF; FAMILIAR entries are our evaluations with our recognition protocol (Appendix B.1).

Table 18: Transfer across recognition models.

Dataset	IR-50 / ArcFace Avg.	IR-50 / AdaFace Avg.	IR-100 / ArcFace Avg.	IR-100 IJB-C
FAMILIAR	$94.79_{\pm 0.06}$	$94.91_{\pm 0.12}$	$95.13_{\pm 0.07}$	$93.75_{\pm 0.20}$

0.5M images; accuracy (%). IR-50/ArcFace is the default recognizer of Table 1. Recognition losses: AdaFace (Kim et al., 2022) and ArcFace (Deng et al., 2019). IJB-C: TAR at FAR 10^{-4}.

Table 16 reallocates a fixed image budget between identity coverage and observations per identity, using $T = 1$ without ranking. Its 10K-by-50 entry is shared with Table 2. It differs from the release-policy 0.5M entry in Table 1 (94.29% versus 94.79%).

B.8 Robustness and Recognition Transfer

On RFW (Wang et al., 2019), FAMILIAR has a higher average accuracy and a smaller between-group gap than VariFace (Table 17). On ROF (Erakın et al., 2021), it reaches 60.17% rank-1 identification for masked probes and 73.62% for sunglasses; VariFace does not report ROF.

For Table 17, the RFW average is the unweighted mean of the four groups, and ROF evaluates masked and sunglasses probes under its own gallery protocol. Both rows use 0.5M images (10K × 50).

For Table 18, AdaFace (Kim et al., 2022) uses $m = 0.4$, $h = 0.333$, $s = 64$, and $t_\alpha = 0.01$, with the same IR-50 optimizer and schedule. The IR-100 run retains ArcFace and the 40-epoch schedule.

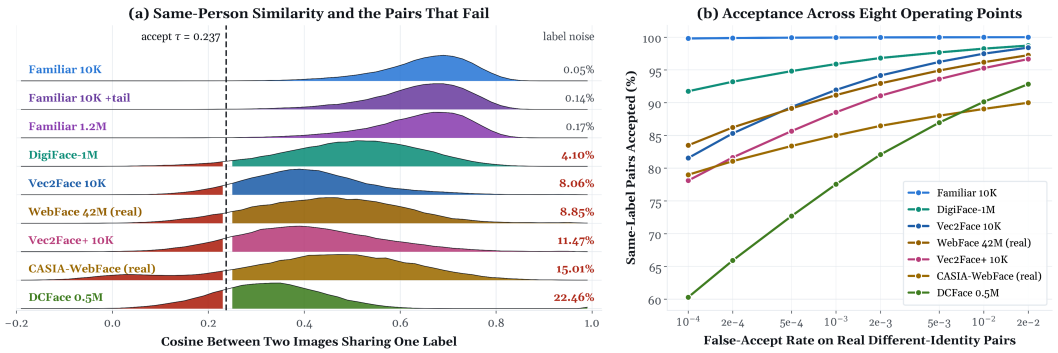

Figure 5: Label fidelity under a fixed real-data threshold. (a) Same-label cosine distributions for 1,200 identities per dataset under a separate audit recognizer. The dashed line is the threshold $\tau = 0.237$, calibrated at FAR 10^{-3} on different-identity pairs of real WebFace-21M images; red mass marks rejected same-label pairs, and the right column gives $1 - \mathrm{TAR}$. "WebFace 42M (real)" denotes this 20.7M-image WebFace subset. (b) Acceptance across eight operating points: same-label TAR at FARs from 10^{-4} to 2×10^{-2}, each calibrated in the same way, for seven of the nine datasets. In (b), the upper brown curve is WebFace.

B.9 Label Fidelity of Synthetic and Real Datasets

Setup. Figure 5 audits each dataset against its own identity labels under one fixed protocol. From every dataset we draw 1,200 identities with at least 32 images and sample exactly 32 images from each (seed 20260923), so that no dataset is favored by having more images per identity. All embeddings come from a frozen `buffalo_sc/w600k_mbf` recognizer, a MobileFaceNet that differs from the `w600k_r50` encoder on which FAMILIAR conditions. For each dataset we form all within-identity pairs over 12 of the 32 images per identity (79,200 pairs) and 200,000 random pairs with same-label draws discarded (about 199,800 cross-identity pairs), and histogram their cosine similarities into 60 bins over $[-0.2, 1.0]$.

Threshold and metric. The accept threshold is calibrated once, on different-identity pairs drawn in the same way from the real WebFace-21M corpus (20.7M images of 1.10M identities), and then applied unchanged to every dataset, so an operating point has the same meaning for all rows and no dataset is scored with a threshold fitted to itself. This gives $\tau = 0.237$ at FAR 10^{-3}, 0.294 at 10^{-4}, and 0.176 at 10^{-2}, with linear interpolation inside the histogram bin that straddles the target rate. At 10^{-4} only about 20 impostor pairs exceed τ, so that threshold is less precise than the others. For each dataset, TAR is the fraction of same-label pairs whose similarity clears τ, with the straddling bin split proportionally. We call $1 - \mathrm{TAR}$ label noise in an operational sense: the share of same-label pairs that the audit recognizer does not accept as the same person.

Results. At FAR 10^{-3}, the FAMILIAR datasets reject at most 0.17% of same-label pairs: 0.05% for 10K identities at $T = 1$, 0.14% for the 10K dataset of Table 1 ("+tail", with higher-temperature proposals), and 0.17% at 1.2M images. Prior synthetic datasets reject between 4.10% (DigiFace-1M) and 22.46% (DCFace), and the two real corpora reject 8.85% (WebFace-21M) and 15.01% (CASIA-WebFace). The distributions also differ in location and shape. The mean same-label cosine is 0.64–0.65 for FAMILIAR, 0.43–0.45 for the real corpora, and 0.33–0.50 for the prior synthetic datasets. These means use the audit recognizer on sampled pairs, so they are not on the same scale as the ID columns of Tables 2 and 5, which average all within-class pairs under AdaFace-IR101. CASIA-WebFace shows a secondary mode near zero similarity, which is consistent with distinct people filed under one label rather than degraded photographs of the same person.

Operating points. Panel (b) repeats the measurement at eight FARs from 10^{-4} to 2×10^{-2}. Among the seven datasets plotted, FAMILIAR (10K, $T = 1$) ranks first at every operating point, with TAR of 99.80% even at FAR 10^{-4}, and DigiFace-1M ranks second throughout. The order among the remaining datasets depends on the threshold: DCFace has the lowest TAR at 10^{-4} (60.29%) but exceeds CASIA-WebFace at 10^{-2} (90.13% versus 89.04%).

Scope. No recognizer is trained in this analysis, so it does not depend on the downstream recipes, which differ across the compared papers in backbone, loss, and schedule (Appendix B.1). The measurement concerns agreement with the assigned label, not intra-class diversity: a dataset can reach a high TAR by varying less, and FAMILIAR's same-label similarities are higher than those of

Table 19: Generation cost and recognition accuracy.

Euler steps	Seconds/ candidate	GPU-hours/ 1M retained	FR Avg.
10	0.0191	6.90	94.15 ± 0.05
15	0.0247	8.92	94.52 ± 0.07
25	0.0361	13.04	94.79 ± 0.06
40	0.0533	19.25	94.90 ± 0.08
60	0.0765	27.62	94.88 ± 0.07

GPU-hours include 65/50 candidate overhead; training excluded.

real photographs. The rejected fraction also combines mislabeled images with genuine same-person pairs that the recognizer fails to match. The audit recognizer and the conditioning encoder differ in architecture but are both trained on WebFace600K, and the release policy ranks candidates with the conditioning encoder (Appendix A.3), so the agreement is not fully independent of the generation pipeline. Downstream recognition (Section 4.2) remains the test of training value.

B.10 Efficiency and Limitations

Table 19 uses an on-device runtime model for a nominal single H200 with bfloat16 precision. Candidate cost includes prior sampling, latent integration, decoding, scoring, and amortized output writing, with fixed overhead. Training is excluded. Accounting for 65 candidates per 50 retained images gives $10^6(65/50)t_{\text{candidate}}/3600$ GPU-hours per million retained images. At 25 steps, prior sampling accounts for 0.4% of candidate latency, so the identity-conditioned prior adds almost no generation cost over an identity-blind one. Training costs are separately modeled as 1,120 H200 GPU-hours for joint training and 12 for both priors.

Further limitations. Because the occasion code is not independent of identity, de-identified images can retain residual source cues. Finite-gallery screening cannot exclude similarity to every real person. Higher-temperature samples need not depict rare attributes or remain realistic. Differences in real training data and pretraining exposure across generators remain, and attribute estimators may carry demographic biases. Source-image matching also represents a weaker adversary than multi-image recognition. Because the generator and the occasion prior accept any normalized embedding, identity samplers built for inter-class separation, such as the hypersphere packing of HyperFace (Otroshi Shahreza & Marcel, 2025), could replace p_ψ in the same embedding space without retraining either model.

C Figure Provenance

Geometry and selected examples. The sphere in Figure 1 depicts identity normalization schematically. The surfaces in Figures 1 and 2 illustrate two identity conditions of one shared prior. Neither visualization establishes high-dimensional separation or an intrinsic manifold. Throughout the text, G includes RAE decoding; in Figures 1(b) and 3 it stands for $G_{\mathbf{w}}$ of Eq. (7).

Photographic panels contain selected examples. The curated virtual-identity archive groups images by identical conditioning embeddings. The identity-replacement archive (Figures 1(b) and 3) uses reference-derived embeddings and candidate selection; the diagrams also illustrate the supported identity-prior option. Replacement identities need not coincide with the displayed virtual repertoires. The one-output-per-source benchmark in Appendix B.5 is a separate evaluation.

Training diagnostic. Panels (a) and (b) of Figure 4 re-encode the generator's renders with ArcFace and measure their cosine to the identity embedding ϕ of the source photograph, for held-out photographs with occasion code $c = E(x)$. Joint generation renders $G(\phi, c)$; occasion-only generation replaces ϕ with its null vector. Panel (a) shows 6,000 photographs rendered at the final checkpoint (rings; the inset shows all of them and marks the plotted window) and the 265 later probe events that fall inside this window (dots), each averaging a fixed batch of eight photographs, so the dots vary less than the single-photograph rings; large markers are the means of five equal temporal blocks of these events, and arrows mark the direction of training. Panel (b) plots all 394 logged probe events over the 300K training steps; lines are nine-event trailing means and bands the same window's 10th

and 90th percentiles, an empirical range rather than a confidence interval. Both branches start at cosine -0.019 at random initialization, and their gap reaches 0.567, while occasion-only renders settle near 0.18 early in training. A guidance sweep at the final checkpoint, not plotted, renders 96 identities with 16 images each; consistency is the mean cosine to ϕ and diversity one minus the mean pairwise identity cosine within an identity. Sweeping $w_\phi \in [0, 2]$ moves consistency by 0.587, while for $w_\phi \geq 1$ sweeping $w_c \in [0.5, 1.5]$ moves it by at most 0.004. Perturbing ϕ with isotropic noise of scale β, the mechanism behind embedding-perturbation generators rather than any released implementation, gives up 0.547 consistency for 0.254 diversity at its first step from the same starting point. The sweep thus characterizes the conditioning interface of one checkpoint on one identity set and measures consistency with the recognizer that defines ϕ; the recognition results of Section 4.2 use independent recognizers.

Variation and accuracy. Panel (c) pools every configuration for which the paper reports both a within-class yaw IQR and a five-benchmark accuracy: the prior families of Table 2 (circles), the sampling policies of Table 5 (squares), and the representation variants of Table 11 (triangles, "architecture" in the legend), counting once the default configuration that all three tables share. The resulting 21 points span yaw IQR from 17.0 to 25.1 and accuracy from 92.79% to 94.79%. Correlations are Spearman coefficients with average ranks for ties. Yaw IQR and accuracy give $\rho = +0.33$ over all 21 points and $+0.27$ over the twelve that share one generator (Tables 2 and 5); the representation variants retrain the encoder and generator, so only the twelve-point value holds the generator fixed. The conditional and unconditional flows share a yaw IQR of 21.1 under the same generator, identities, and image count, and differ by 1.11 points; the widest spread, 25.1 at $T = 1.8$, trains a recognizer 0.48 points below the best configuration. The right panel is restricted to the six priors of Table 2 with a held-out NLL: the empirical bank has no continuous density, sampling policies leave the fitted prior unchanged, and NLLs in different code spaces are not comparable (Appendix B.6). Across these six, held-out NLL is perfectly anti-ordered with both accuracy and IJB-C TAR ($\rho = -1.00$; exact two-sided permutation $p = 0.003$). Held-out NLL is a property of the fitted prior and is computed without rendering any image. Within-class identity agreement tracks accuracy more closely than yaw IQR ($\rho = +0.78$ over the same twelve configurations), so the weak correlation is specific to the amount of variation.