TRACE: TASK-DEFINED REPRESENTATION ALIGNMENT AND CALIBRATED EXCHANGE FOR CONTROLLABLE FINGERPRINT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable fingerprint generation offers a promising way to expand limited biometric datasets and support presentation-attack research. Beyond visual realism, its practical utility depends on modifying a specified factor while preserving identity-related ridge patterns, non-target attributes, and fine-grained image details. This is challenging because fingerprint content, presentation material, and acquisition batch are often correlated and expressed through shared, spatially distributed visual patterns. We propose **TRACE** (**T**ask-Defined **R**epresentation **A**lignment and **C**alibrated **E**xchange), a two-stage framework that separates factor-structured representation learning from generation-oriented refinement. In the first stage, Factor-Structured Information Extraction combines factor-wise supervised contrastive learning with image reconstruction to learn semantically aligned latent groups that retain the information needed for faithful generation. In the second stage, Structured Latent Distillation freezes the learned encoder and calibrates these groups through residual refinement and factor-specific spatial adaptation, enabling selective factor exchange while limiting unintended changes to non-target attributes. Experiments on a real-world fingerprint dataset and Shapes3D show that TRACE achieves state-of-the-art factor organization and controllable generation while maintaining competitive reconstruction quality. These results establish TRACE as an effective solution for task-defined controllable generation, combining accurate and interpretable factor control with faithful image synthesis.

1 INTRODUCTION

In recent years, deep generative models have made remarkable progress in image synthesis, particularly in terms of visual quality, realism, and diversity. However, for data generation intended to support downstream tasks, visual fidelity alone does not guarantee practical utility. This issue is especially important in fields such as medical imaging and biometrics, where data collection is costly and privacy-sensitive, often resulting in limited datasets with insufficient diversity. In these settings, image generation is valuable for data augmentation and dataset expansion, but the generated samples must accurately reflect specified attributes while preserving task-relevant structures (Hu et al., 2023; Zhang et al., 2023; Konz et al., 2024; Zhang et al., 2024). Fingerprint images exemplify this challenge, as information related to identity, liveness, and presentation conditions is encoded in local ridge patterns and their imaging characteristics. Therefore, task-oriented fingerprint generation should go beyond producing images that merely resemble fingerprints. It should enable explicit control over selected task-defined factors while keeping other critical attributes and fine-grained structures stable (Shoshan et al., 2024; Grosz & Jain, 2024).

Unlike natural images with relatively distinct semantic regions, fingerprints consist of fine-grained, continuous, and globally connected ridge structures (Shoshan et al., 2024; Grosz et al., 2020; Chugh & Jain, 2020). Their appearance is jointly shaped by fingerprint content, sensor characteristics, acquisition conditions, presentation material, collection batch, and geometric variation. These factors do not correspond to separate spatial regions; instead, they are expressed through shared and spatially distributed evidence, such as ridge clarity, ridge–valley contrast, local deformation, reflection, texture, and acquisition artifacts. As illustrated in Figure 1, material- and batch-related variations may be subtle in the image space despite being distinguishable in the feature space. They

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.

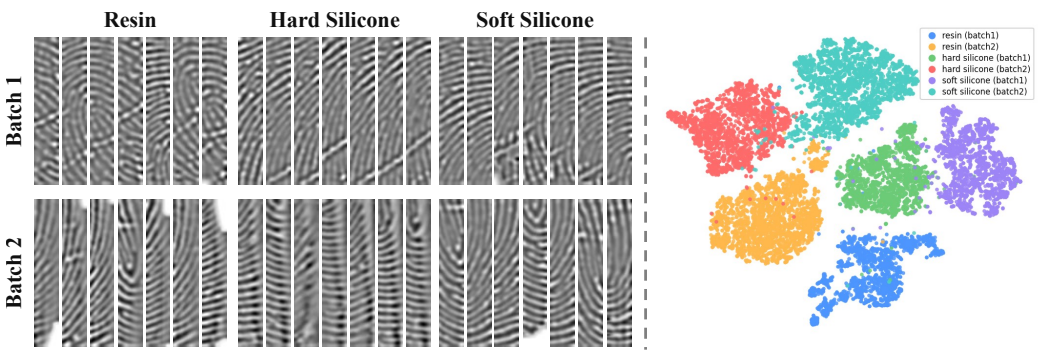


Figure 1: Comparison of visual appearance and feature space. Left: Representative fingerprints with different batches and materials. Right: t-SNE visualization of samples in the fingerprint dataset.

can also be statistically correlated, for example, a presentation material may occur only within a particular acquisition batch, encouraging models to rely on spurious shortcuts (Locatello et al., 2019; Liu et al., 2023). As a result, visually realistic generation does not necessarily imply correct factor transfer, preservation of non-target attributes, or generalization to rare and unseen factor combinations. Addressing these issues requires representations that support selective factor preservation, exchange, and recombination (Locatello et al., 2019; Liu et al., 2023; Grosz & Jain, 2024).

Existing disentangled generation methods use different generative backbones and forms of supervision to separate sources of image variation. VAE-based methods encourage latent factorization through distribution regularization (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018), while GAN- and diffusion-based methods support semantic manipulation and high-quality synthesis (Chen et al., 2016; Lee et al., 2020; Gabbay & Hoshen, 2021; Yang et al., 2023; 2024). However, their underlying representations do not necessarily provide the selective control required for fingerprint generation. First, conditional models may satisfy a target label while unintentionally altering ridge structures or non-target attributes. Second, unsupervised disentanglement offers no guarantee of alignment with the semantics of interest (Locatello et al., 2019), whereas weakly and fully supervised methods reduce this ambiguity through grouped observations, factor labels, causal structures, or domain knowledge (Bouchacourt et al., 2018; Locatello et al., 2020; Yap & Ng, 2024; Shen et al., 2022; Liu et al., 2023). Nevertheless, label predictability alone does not make a latent group selectively controllable: target information may remain distributed across other groups, the designated group may retain non-target information, and exchanging it may induce unintended changes. Moreover, enforcing strict statistical independence is also ill-suited to naturally correlated factors that share fine-grained visual evidence and may compromise reconstruction fidelity (Träuble et al., 2021). Therefore, a fingerprint generation framework that simultaneously achieves task-semantic alignment, detail preservation, and factor-selective generative control remains insufficiently explored.

Motivated by these challenges, we formulate this problem as *task-defined factorized fingerprint generation*. Given fingerprint images annotated with task-defined factors, the goal is to organize their latent representations into factor-aligned groups while preserving sufficient information for faithful reconstruction. Exchanging one group should transfer its designated factor from a reference image while retaining the source image's non-target attributes and instance-specific details. Rather than pursuing unconditional independence among correlated factors, we focus on selective and verifiable control in the generated images.

To this end, we propose TRACE, a two-stage framework that separates factor-structured representation learning from generation-oriented calibration. Stage I combines factor-wise supervised contrastive learning with image reconstruction to obtain factor-aligned latent groups that remain collectively sufficient for reconstruction. Because semantic alignment alone does not ensure selective exchange, Stage II freezes the learned encoder and applies structured latent distillation to calibrate each group for generation through residual refinement and factor-specific spatial adaptation. This separation preserves the factor structure established in Stage I while improving factor-selective control and supporting flexible factor exchange and recombination. Our contributions are threefold:

- We formulate task-defined factorized fingerprint generation as the joint problem of factor alignment, fine-grained detail preservation, and selective control under potentially correlated factors.
- We propose TRACE, a two-stage framework that decouples factor-structured representation learning from generation-oriented calibration, enabling selective factor exchange and recombination while preserving non-target content.
- We establish a calibrated evaluation protocol for factor exchange and conduct comprehensive experiments on a real-world fingerprint dataset and Shapes3D. Across both domains, TRACE achieves state-of-the-art factor organization and selective control while maintaining competitive reconstruction quality.

2 Related Work

Factor-Aligned Representation Learning. Disentangled representation learning organizes sources of variation into distinct latent components. Classical VAE-based methods, including β-VAE, FactorVAE, and β-TCVAE, promote factorization through KL regularization, total-correlation penalties, or latent-dependency decomposition (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018). However, statistical factorization may trade off reconstruction fidelity and offers no guarantee of alignment with factors of interest; without additional assumptions, unsupervised disentanglement is generally unidentifiable and remains sensitive to correlations in the data (Locatello et al., 2019; Träuble et al., 2021). Supervised and weakly supervised approaches reduce this ambiguity using grouped observations, factor annotations, or causal structure. ML-VAE and related methods exploit shared factors across observations (Bouchacourt et al., 2018; Locatello et al., 2020); Group-Wise Contrastive Bottleneck and XFactors align feature groups or predefined subspaces with semantic factors (Yap & Ng, 2024; Myara et al., 2026); and DEAR and C-Disentanglement explicitly model dependent or confounded factors (Shen et al., 2022; Liu et al., 2023). While these methods improve semantic organization, factor predictability alone does not ensure that the resulting representation retains sufficient information for faithful reconstruction.

Factorized Representations for Controllable Generation. Controllable generation seeks to translate latent semantics into selective and predictable image changes. InfoGAN learns interpretable controls through mutual-information maximization, while ID-GAN and OverLORD separate representation learning from synthesis to balance factorization and generation quality (Chen et al., 2016; Lee et al., 2020; Gabbay & Hoshen, 2021). Diffusion-based methods such as DisDiff and EncDiff introduce factor-conditioned denoising or concept-specific representations (Yang et al., 2023; 2024). In fingerprint generation, FPGAN-Control and GenPrint preserve identity under predefined acquisition or appearance conditions (Shoshan et al., 2024; Grosz & Jain, 2024). Nevertheless, a semantically aligned representation is not necessarily exchangeable: target information may remain distributed across groups, and swapping one group may unintentionally alter non-target attributes. This gap is particularly consequential for correlated factors that share fine-grained visual evidence.

3 Method

3.1 Task Formulation and Overview

Let $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^N$ denote the training set, where $x_i \in \mathbb{R}^{C\times H\times W}$ is an image and $\mathbf{y}_i = (\mathbf{y}_i^1, \ldots, \mathbf{y}_i^K)$ contains its labels for K user-defined factors. For factor k with n_k classes, $\mathbf{y}_i^k \in \{0,1\}^{n_k}$ is a one-hot label satisfying $\mathbf{1}^\top \mathbf{y}_i^k = 1$. Given a source image x_a, a reference image x_b, and a target factor k, our goal is to generate $\hat{x}_{a\leftarrow b}^{(k)}$ by transferring the k-th factor from x_b while preserving the non-target factors and instance-specific details of x_a.

As illustrated in Figure 2, TRACE follows a two-stage extract-and-reorganize framework. Stage I learns a latent representation structured by the user-defined factors. Stage II reorganizes the extracted information into more factor-specific groups for selective control. During training, the refined representation is decoded for reconstruction. At inference, factor k is transferred by replacing the corresponding Stage-I posterior-mean block of the source with that of the reference; the resulting mixed representation is then processed by all residual refinement modules and the Stage-II decoder.

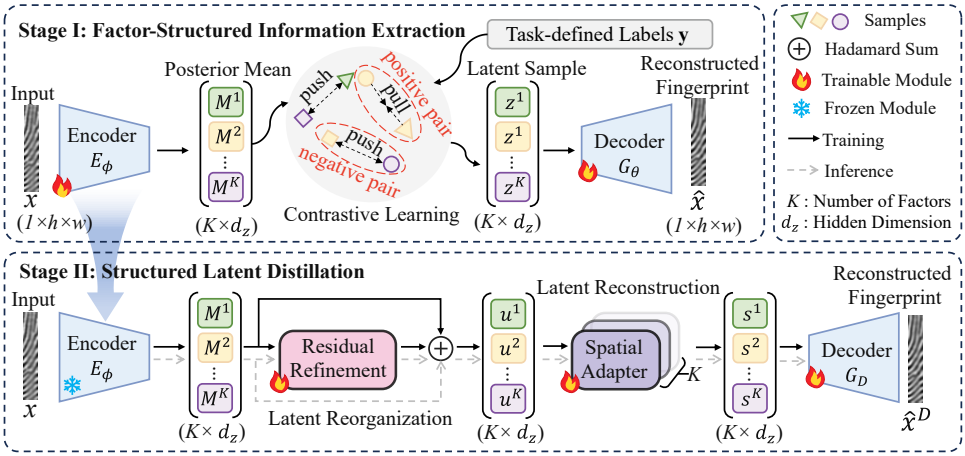


Figure 2: Overview of TRACE.

3.2 Factor-Structured Information Extraction

Factor-Structured Information Extraction (FSIE) learns a generatively sufficient representation organized around task-defined factors. Since image-level labels specify the factors of interest but not how their visual evidence should be separated, FSIE combines factor-wise semantic supervision with image reconstruction without enforcing strict inter-factor independence.

Factor-Structured Representation. We instantiate FSIE as a variational autoencoder (VAE) rather than a deterministic feature extractor. Given an image x_i, the encoder E_ϕ parameterizes

$$(\mu_i, \log \sigma_i^2) = E_\phi(x_i), \qquad q_\phi(z \mid x_i) = \mathcal{N}\big(\mu_i, \mathrm{diag}(\sigma_i^2)\big)\,.$$

At the random-variable level, let $M := \mu_\phi(X) = [M^1; \ldots; M^K]$ denote the posterior-mean representation, with $\mu_i = \mu_\phi(x_i)$ as its realization for sample x_i. We partition μ_i into K factor-specific blocks:

$$\mu_i = [\mu_i^1; \ldots; \mu_i^K], \qquad \mu_i^k \in \mathbb{R}^{d_k}, \tag{1}$$

where μ_i^k is assigned to factor label $\mathbf{y}_i^k$. This assignment specifies the intended semantics of each block without requiring it to exclusively encode that factor.

We establish these block–factor associations by applying supervised contrastive learning to each block separately. For factor k, the positive set of sample i is

$$\mathcal{P}_k(i) = \{p \neq i \mid \mathbf{y}_p^k = \mathbf{y}_i^k\}\,. \tag{2}$$

The corresponding factor-wise contrastive loss is

$$\mathcal{L}_{\mathrm{con}}^k = -\sum_i \frac{1}{|\mathcal{P}_k(i)|} \sum_{p\in\mathcal{P}_k(i)} \log \frac{\exp\big(\mathrm{sim}(h_i^k, h_p^k)/\tau\big)}{\sum_{a\neq i} \exp\big(\mathrm{sim}(h_i^k, h_a^k)/\tau\big)}, \tag{3}$$

where $h_i^k = \mathrm{norm}(\mu_i^k)$, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, and τ is the temperature. This objective organizes each mean block according to its designated factor.

Reconstruction instead uses a stochastic sample $z_i \sim q_\phi(z \mid x_i)$, obtained as

$$z_i = \mu_i + \sigma_i \odot \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, I).$$

The reconstruction objective is

$$\widehat{x}_i = G_\theta(z_i), \qquad \mathcal{L}_{\mathrm{rec}} = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \widehat{x}_i), \tag{4}$$

where $d(\cdot,\cdot)$ is the reconstruction distance. Thus, contrastive learning operates on individual mean blocks μ_i^k, whereas reconstruction uses the complete sampled latent z_i, allowing it to jointly retain factor-related and instance-specific visual information.

Structured Reconstructive Bottleneck. We term the combination of block-wise semantic supervision and full-latent reconstruction a *structured reconstructive bottleneck*. The classical Information

Bottleneck (IB) seeks a compact representation that preserves target-relevant information while suppressing unnecessary input information Tishby et al. (2000):

$$\max_{p(z|x)} I(Z;Y) - \beta I(X;Z). \tag{5}$$

In contrast, our bottleneck is architectural and representational: the decoder accesses the input only through a fixed-dimensional, block-partitioned latent whose mean blocks are individually anchored to designated factors. It does not rely on strong posterior compression or a prescribed information-capacity constraint. Strong compression may be undesirable here, as coarse or correlated labels may omit instance-specific information needed for faithful reconstruction. Rather than strongly minimizing $I(X;Z)$, FSIE therefore semantically anchors the mean blocks while maintaining the reconstructive sufficiency of the complete sampled latent.

Let $M = [M^1; \dots; M^K]$ denote the complete posterior mean, whose blocks are used for factor-wise contrastive learning, and let $Z \sim q_\phi(z \mid X)$ denote the complete stochastic latent used for reconstruction. The intended information-theoretic behavior is

$$\max_{\phi,\theta} \sum_{k=1}^{K} \alpha_k I(M^k; Y^k) - \lambda H_{q_\phi}(X \mid Z). \tag{6}$$

The first term is promoted by factor-wise contrastive learning, while the second is reduced through reconstruction. Specifically, the reconstruction negative log-likelihood satisfies

$$\begin{aligned} \mathcal{L}_{\text{rec}}(\phi,\theta) &= \mathbb{E}_{q_\phi(x,z)}\left[-\log p_\theta(x \mid z)\right] \\ &= H_{q_\phi}(X \mid Z) + \mathbb{E}_{q_\phi(z)}\left[D_{\text{KL}}\left(q_\phi(x \mid z) \,\|\, p_\theta(x \mid z)\right)\right] \\ &\geq H_{q_\phi}(X \mid Z), \end{aligned} \tag{7}$$

where $q_\phi(x,z) = p_{\text{data}}(x) q_\phi(z \mid x)$. The mean absolute error used in our implementation is, up to a positive scale and additive constant, the negative log-likelihood under a fixed-scale, factorized Laplace observation model. Thus, minimizing the reconstruction loss reduces a variational upper bound on the uncertainty of X given the complete latent, encouraging reconstructive sufficiency regardless of whether the posterior variance becomes small.

3.3 Structured Latent Distillation

FSIE produces factor-anchored but non-exclusive latent blocks: information about Y^k may be distributed across multiple blocks, and M^k may retain information about non-target factors. Structured Latent Distillation (SLD) improves controllability through *residual refinement modules* and *spatial adapters*. The former reorganize the FSIE blocks according to a *preserve–suppress–anchor* principle: each refined block preserves its designated factor, suppresses non-target factors, and remains anchored to its Stage-I counterpart. The latter map the refined vectors to spatial features for Stage-II decoding.

(1) Residual Refinement Module. Let $M = \mu_\phi(X) = [M^1; \dots; M^K]$ be the posterior mean produced by the frozen FSIE encoder. For each factor k, a residual module Δ_k, conditioned on the complete posterior mean, predicts a block-sized correction:

$$U^k = M^k + \Delta_k(M), \qquad u_i^k = \mu_i^k + \Delta_k(\mu_i), \tag{8}$$

where U^k has the same dimensionality as M^k, and u_i^k is its realization for sample i. Conditioning on the complete representation allows information about Y^k distributed across other blocks to be transferred into U^k.

(i) Factor Preservation. To preserve information about the designated factor, we introduce an own-factor classifier C_k for each refined block:

$$\mathcal{L}_{\text{pre}} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_i \left[\ell_{\text{CE}}\left(C_k(u_i^k), y_i^k\right)\right]. \tag{9}$$

Minimizing this loss reduces the uncertainty of Y^k given U^k, thereby promoting $I(U^k; Y^k)$.

(ii) Conditional Suppression. Factor preservation alone does not prevent U^k from encoding non-target factors. For each source block k, we therefore introduce a conditional adversary Q_k, with an output head $Q_{k\to\ell}$ for each $\ell \neq k$. These heads share a feature-extraction trunk and predict the

non-target factors from U^k, conditioned on the ground-truth own-factor label Y^k. The conditional suppression loss is

$$\mathcal{L}_{\text{sup}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{\substack{\ell=1\\ \ell\neq k}}^{K}\mathbb{E}_i\left[\ell_{\text{CE}}\left(Q_{k\to\ell}(u_i^k, y_i^k), y_i^\ell\right)\right]. \tag{10}$$

Conditioning on Y^k avoids penalizing dependencies inherent in the factor labels. The adversaries minimize this loss, while the refinement modules maximize it, suppressing the additional predictability of Y^ℓ from U^k beyond Y^k. With sufficiently expressive and well-optimized adversaries, this provides a surrogate for reducing $\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} I(U^k; Y^\ell \mid Y^k)$.

(iii) Latent Anchoring. Without an additional constraint, the residual transformations could replace the Stage-I blocks with unrestricted embeddings. We therefore penalize deviations from the frozen posterior-mean blocks:

$$\mathcal{L}_{\text{anc}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{d_k}\mathbb{E}_i\left[\left\|u_i^k - \mu_i^k\right\|_1\right], \tag{11}$$

where d_k normalizes for block dimensionality. This loss preserves the Stage-I factor allocation while permitting cross-block corrections required by preservation and suppression.

The three objectives above operate on the mean-refined blocks u_i^k. Reconstruction instead uses a reparameterized sample from the frozen FSIE posterior: $z_i = \mu_i + \sigma_i \odot \epsilon_i, \epsilon_i \sim \mathcal{N}(0, I)$, where both μ_i and σ_i are produced by the frozen Stage-I encoder. After partitioning $z_i = [z_i^1; \ldots; z_i^K]$, we apply the same residual refinement modules to the sampled latent:

$$\widetilde{u}_i^k = z_i^k + \Delta_k(z_i). \tag{12}$$

Thus, factor preservation, conditional suppression, and latent anchoring operate on the mean-refined blocks u_i^k, whereas reconstruction operates on the sampled-refined blocks $\widetilde{u}_i^k$.

(2) Spatial Adaptation. To retain the information required for image generation, each sampled-refined block is mapped to a spatial feature tensor by a factor-specific adapter A_k:

$$s_i^k = \text{reshape}\left(A_k(\widetilde{u}_i^k)\right). \tag{13}$$

The resulting tensors are concatenated as $s_i = [s_i^1; \ldots; s_i^K]$ and decoded by G_D:

$$\widehat{x}_i^D = G_D(s_i), \qquad \mathcal{L}_{\text{rec}}^D = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \widehat{x}_i^D). \tag{14}$$

This reconstruction objective ensures that reorganizing information across latent blocks does not compromise the generative sufficiency of the complete representation.

3.4 Two-Stage Optimization

The model is optimized sequentially because SLD requires a stable, factor-anchored representation as its refinement target. In Stage I, the FSIE encoder and decoder are trained using

$$\mathcal{L}_{\text{FSIE}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \beta_{\text{KL}}\mathcal{L}_{\text{KL}} + \frac{\lambda_{\text{NCE}}}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{con}}^k, \tag{15}$$

where λ_{rec}, β_{KL}, and λ_{NCE} control reconstruction, posterior regularization, and factor-wise semantic anchoring, respectively. In Stage II, the encoder E_ϕ and its posterior-parameter head are frozen so that SLD reorganizes, rather than alters, the Stage-I representation. Let Ω denote the parameters of the refinement modules, preservation classifiers, spatial adapters, and Stage-II decoder, and let Ψ denote those of the conditional adversaries. The adversaries are optimized by

$$\min_{\Psi}\ \mathcal{L}_{\text{sup}}, \tag{16}$$

while the remaining Stage-II components are optimized by

$$\min_{\Omega}\ \mathcal{L}_{\text{rec}}^D + \lambda_{\text{pre}}\mathcal{L}_{\text{pre}} - \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}}. \tag{17}$$

The opposing objectives form an adversarial game: the conditional adversaries predict non-target factors, while the refinement modules remove their additional predictability. This optimization can be implemented with alternating updates or a gradient-reversal layer.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use two datasets in this study. Our internal real-world fingerprint dataset contains 21,240 normalized grayscale images of size $1 \times 160 \times 26$, annotated with $K = 3$ task-defined factors: fingerprint content (78 classes), presentation material (5 classes), and acquisition batch (2 classes). We additionally use Shapes3D (Burgess & Kim, 2018) to evaluate generalization beyond the fingerprint domain. Further dataset details are included in Appendix A.1.

Implementation Details. FSIE and SLD adopt standard encoder–decoder architectures, detailed in Appendix A.2. We set $d_z = 224$, allocating 128, 64, and 32 dimensions to the content, material, and batch blocks for fingerprints, respectively, and to the content, color, and orientation blocks for Shapes3D. For FSIE, we set $\tau = 0.5$, $\beta_{\mathrm{KL}} = 10^{-6}$, and $\lambda_{\mathrm{NCE}} = 1$; for SLD, we use $\lambda_{\mathrm{pre}} = \lambda_{\mathrm{sup}} = 1$ and $\lambda_{\mathrm{anc}} = 0.025$. Both stages are trained for 80 epochs with a batch size of 256 using Adam at a learning rate of 3×10^{-4} on a single NVIDIA 4090 GPU.

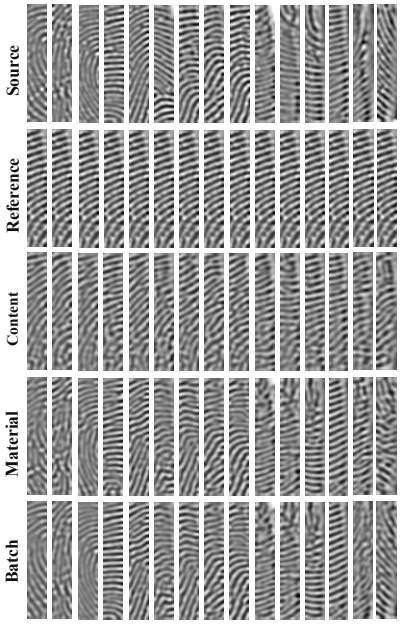


Figure 3: Factor-swapping on fingerprint dataset.

Evaluation Metrics. We evaluate reconstruction using MSE, LPIPS (Zhang et al., 2018), and SSIM (Wang et al., 2004), and latent organization using DCI disentanglement and completeness (Eastwood & Williams, 2018). As latent predictability does not guarantee factor-selective swaps, we propose the image-level *Calibrated Factor Swap Score* (CFSS) to evaluate swap effectiveness and specificity, particularly for complex fingerprints.

Let $p_t(x)$ be the output of a frozen factor-t classifier, $d_t(x, y) = \|p_t(x) - p_t(y)\|_1$, and $\hat{x}^{(k)}_{a \leftarrow b}$ the result of replacing source a's block k with reference b's. We use $R_{k,k}$ for target movement toward the reference and $R_{t,k}$ for calibrated non-target departure from the source, and define CFSS below; more details are included in Appendix A.3.

$$E_k = R_{k,k}, \quad P_k = 1 - \sqrt{\frac{\sum_{t \neq k} R_{t,k}^2}{K-1}}, \quad \mathrm{CFSS} = \left(\prod_{k=1}^{K} E_k P_k^{1/2} \right)^{1/K}. \tag{18}$$

4.2 MAIN RESULTS

Quantitative analysis. As shown in Table 1, TRACE achieves the strongest disentanglement and factor control on both datasets (all compared methods are trained with seed 0). On fingerprints, TRACE achieves 0.905 DCI-D, 0.252 DCI-C, and 0.3615 CFSS, outperforming the strongest baselines by 0.192 in DCI-D and 0.1166 in CFSS while also obtaining the highest DCI-C. Most baselines obtain substantially lower CFSS, highlighting the difficulty of selectively manipulating correlated factors that share fine-grained visual evidence. On Shapes3D, TRACE achieves 0.940 DCI-D and 0.981 CFSS, exceeding the next-best results by 0.087 and 0.196. While some baselines yield lower reconstruction errors, their weaker CFSS indicates that accurate reconstruction alone does not ensure selective control. Overall, TRACE provides the best balance between factor organization, controllability, and reconstruction quality. More qualitative results are included in the Appendix A.4.

Table 1: Comparison of factor organization, controllability, and reconstruction quality.

Model	Fingerprint					Shapes3D				
	$D\uparrow$	$C\uparrow$	CFSS↑	MSE↓	SSIM↑	$D\uparrow$	$C\uparrow$	CFSS↑	MSE↓	SSIM↑
Ada-ML-VAE (Locatello et al., 2020)	0.020	0.016	0.1057	0.0168	0.337	0.353	0.233	0.183	0.00286	0.842
HFS(Roth et al., 2023)	0.015	0.013	0.0900	0.0180	0.268	0.179	0.108	0.107	0.00421	0.785
Latent Quant (Hsu et al., 2023)	0.001	0.001	0.0407	0.0233	0.151	0.469	0.254	0.438	0.00055	0.952
Causal Disentangle (Liu et al., 2023)	0.399	0.122	0.0755	0.0215	0.158	0.138	0.047	0.278	0.00872	0.807
Flow Factorized (Song et al., 2023)	0.191	0.112	0.1441	0.0154	0.370	0.142	0.059	0.106	0.00118	0.981
DisDiff (Yang et al., 2023)	0.056	0.067	0.1006	0.0227	0.154	0.002	0.002	0.014	0.02555	0.619
InfoDiffusion (Wang et al., 2023)	0.089	0.048	0.0880	0.0201	0.167	0.066	0.035	≈ 0	**0.00001**	**0.999**
Ada-GVAE (Michlo et al., 2023)	0.048	0.043	0.1110	0.0167	0.334	0.353	0.226	0.156	0.00306	0.833
EncDiff (Yang et al., 2024)	0.036	0.042	0.0914	0.0228	0.153	0.537	**0.277**	0.352	0.00102	0.904
SoftTPR (Sun et al., 2024)	0.034	0.014	0.0947	0.0232	0.230	0.619	0.151	0.103	0.02503	0.710
FPGAN(Shoshan et al., 2024)	0.017	0.005	0.0673	0.0313	0.101	0.179	0.075	0.150	0.00349	0.833
FDAE (Wu & Zheng, 2024)	0.007	0.002	0.0673	0.0215	0.163	0.853	0.236	≈ 0	**0.00001**	**0.999**
DyGA (Jun et al., 2025)	0.114	0.222	0.0446	**0.0031**	**0.913**	0.387	0.240	0.785	0.00143	0.887
XFactors (Myara et al., 2026)	0.713	0.243	0.2449	0.0043	0.827	0.387	0.153	0.444	0.00029	0.971
TRACE (Ours)	**0.905**	**0.252**	**0.3615**	0.0132	0.597	**0.940**	0.232	**0.981**	0.00016	0.994

Factor-swapping Visualization. We visualize factor control by replacing the k-th block of an encoded source with its counterpart from a reference. As shown in Figure 3, the learned blocks have clear task-specific roles: content transfers global ridge structure, material alters local ridge texture, and batch controls overall appearance and fine-detail statistics, while non-target attributes are preserved. On Shapes3D, TRACE similarly modifies only the attribute associated with the swapped block (Figure 4). These results visually confirm the disentanglement and controllability of the learned factors.

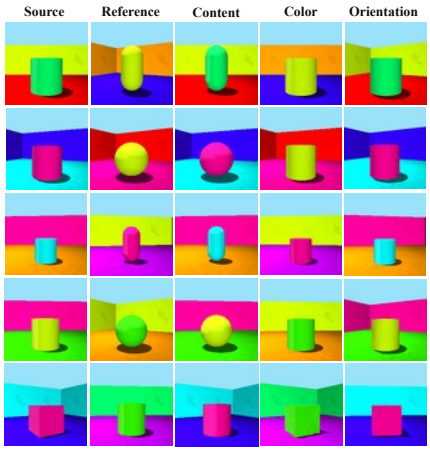


Figure 4: Factor-swapping on Shapes3D dataset.

Task-Defined Factor Interpretation. Figure 5 isolates each block by replacing the others with dataset-wide means. The reconstructions show that the content preserves the ridge structure, the material retains its local ridge appearance, and the batch captures broader texture statistics. These factor-specific but spatially overlapping effects demonstrate semantic and functional disentanglement rather than simple spatial partitioning.

4.3 Ablation Study

To assess the necessity and effectiveness of TRACE's key components and their contributions to controllable generation and reconstruction, we conduct loss and architecture ablations on the fingerprint dataset across 6 random seeds (0∼5). For loss functions, we progressively remove the SLD objectives and separately ablate the two FSIE objectives. For architecture, we compare the full model against FSIE-only training and joint optimization. Results are reported in Table 2, with additional ablations in Appendix A.5.

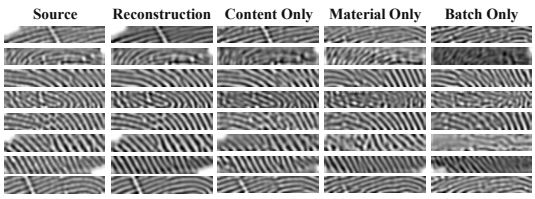


Figure 5: Factor-block isolation on fingerprints.

(1) Loss Functions. The loss ablations reveal a trade-off between factor controllability and reconstruction fidelity. In SLD, removing $\mathcal{L}_{\text{anc}}$ reduces CFSS from 0.3580 to 0.3254, decreases SSIM by 0.093, and increases LPIPS by 0.031. This suggests that latent anchoring limits deviation from the Stage-I representation, preserving both controllability and generative information. Further removing $\mathcal{L}_{\text{sup}}$ decreases CFSS by 0.0477 but improves reconstruction, confirming that conditional suppression reduces non-target information leakage at some cost to pixel-level fidelity. Removing all residual-refinement losses further lowers CFSS to 0.2838, with only a slight reconstruction improvement, demonstrating their joint role in reorganizing Stage-I representations for selective control.

Table 2: Ablation results for TRACE on the fingerprint dataset.

	Ablation Settings	CFSS ↑	MSE ↓	SSIM ↑	LPIPS ↓	D ↑
SLD	w/o $\mathcal{L}_{\text{anc}}$	0.3254 ± 0.0144	0.0156 ± 0.0006	0.514 ± 0.019	0.104 ± 0.008	0.925 ± 0.018
	w/o $\mathcal{L}_{\text{anc}}, \mathcal{L}_{\text{sup}}$	0.3103 ± 0.0087	0.0105 ± 0.0001	0.653 ± 0.001	0.063 ± 0.001	0.925 ± 0.018
	w/o $\mathcal{L}_{\text{anc}}, \mathcal{L}_{\text{sup}}, \mathcal{L}_{\text{pre}}$	0.2838 ± 0.0085	0.0101 ± 0.0001	0.665 ± 0.002	0.061 ± 0.001	0.925 ± 0.018
FSIE	w/o FSIE InfoNCE	0.0957 ± 0.0031	$\mathbf{0.0016 \pm 0.0001}$	$\mathbf{0.914 \pm 0.004}$	$\mathbf{0.016 \pm 0.001}$	0.020 ± 0.002
	w/o FSIE reconstruction	0.0490 ± 0.0164	0.0481 ± 0.0099	0.035 ± 0.006	0.753 ± 0.037	$\mathbf{0.964 \pm 0.000}$
Network	FSIE only (w/o SLD)	0.3378 ± 0.0141	0.0086 ± 0.0008	0.8288 ± 0.0037	0.068 ± 0.009	0.925 ± 0.018
	Joint optimization (unfrozen)	0.1540 ± 0.0089	0.0100 ± 0.0003	0.645 ± 0.008	0.070 ± 0.003	0.904 ± 0.013
	TRACE (full model)	$\mathbf{0.3580 \pm 0.0071}$	0.0127 ± 0.0003	0.607 ± 0.005	0.073 ± 0.001	0.925 ± 0.018

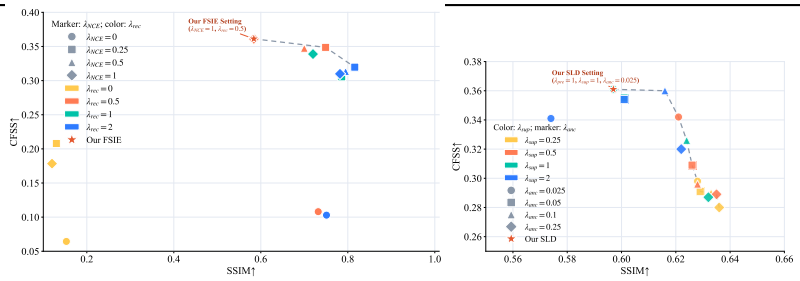


Figure 6: Sensitivity analysis of hyperparameters on FSIE (left) and SLD (right).

In FSIE, removing InfoNCE reduces CFSS by 0.2623 and D from 0.925 to 0.020, despite substantially improving reconstruction. Thus, reconstruction alone cannot align latent blocks with semantic factors. Conversely, removing the reconstruction objective severely degrades both reconstruction and CFSS despite yielding the highest D, indicating that contrastive supervision produces label-discriminative representations but fails to retain sufficient instance-specific information for controllable generation. Together, these results validate FSIE's structured reconstructive bottleneck: contrastive learning establishes factor–block associations, while reconstruction preserves the visual information needed for generation.

(2) Architecture. Table 2 validates the two-stage design. FSIE alone provides meaningful factor control, while SLD further improves selective transfer by reorganizing the factor-anchored representation. In contrast, joint optimization substantially degrades both controllability and disentanglement, indicating that freezing FSIE is important for preserving its semantic structure during Stage-II refinement.

4.4 Sensitivity Analysis

We conduct grid-based sensitivity analyses of the FSIE weights $(\lambda_{\text{NCE}}, \lambda_{\text{rec}})$ and SLD weights $(\lambda_{\text{pre}}, \lambda_{\text{sup}}, \lambda_{\text{anc}})$ under a fixed protocol. Figure 6 shows the results in the SSIM–CFSS plane, together with the empirical Pareto frontier. For FSIE, contrastive supervision primarily improves controllability, whereas reconstruction weighting favors fidelity; removing contrastive supervision degrades both. For SLD, preservation and suppression generally improve controllability at a fidelity cost, while anchoring supports reconstruction. Additional settings are reported in Appendix A.6. As controllability is our primary objective, we select the configuration with the highest CFSS, which provides the best overall balance across the evaluated metrics.

5 Conclusion

In this work, we presented TRACE, a two-stage framework for task-defined representation learning and controllable generation from coarsely annotated data. FSIE learns factor-anchored latent blocks while preserving the information required for faithful reconstruction, whereas SLD reorganizes these blocks through factor preservation, conditional suppression, and latent anchoring to enable selective control. Quantitative results and factor-swapping visualizations demonstrate that TRACE achieves strong factor organization and controllable generation while retaining competitive reconstruction quality. Ablation studies and sensitivity analyses further validate the proposed objectives, the two-stage optimization strategy, and the overall architecture. Overall, our findings highlight the importance of treating semantic representation learning and generation-oriented control as related but distinct objectives, providing a strong foundation for controllable generation under coarse and correlated factor annotations.

REFERENCES

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Tarang Chugh and Anil K Jain. Fingerprint spoof detector generalization. *IEEE Transactions on Information Forensics and Security*, 16:42–55, 2020.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

Aviv Gabbay and Yedid Hoshen. Scaling-up disentanglement for image translation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6763–6772. IEEE, 2021.

Steven A Grosz and Anil K Jain. Universal fingerprint generation: Controllable diffusion model with multimodal conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2): 1028–1041, 2024.

Steven A Grosz, Tarang Chugh, and Anil K Jain. Fingerprint presentation attack detection: A sensor and material agnostic approach. In *2020 IEEE international joint conference on biometrics (IJCB)*, pp. 1–10. IEEE, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36:45463–45488, 2023.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360. IEEE, 2023.

Youngjun Jun, Jiwoo Park, Kyobin Choo, Tae Eun Choi, and Seong Jae Hwang. Disentangling disentangled representations: Towards improved latent units via diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3559–3569. IEEE, 2025.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 88–98. Springer, 2024.

Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *European conference on computer vision*, pp. 157–174. Springer, 2020.

Xiaoyu Liu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. C-disentanglement: discovering causally-independent generative factors under an inductive bias of confounder. *Advances in Neural Information Processing Systems*, 36:39566–39581, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pp. 6348–6359. PMLR, 2020.

Nathan Michlo, Richard Klein, and Steven James. Overlooked implications of the reconstruction loss for vae disentanglement. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 4073–4081. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/453. URL https://doi.org/10.24963/ijcai.2023/453. Main Track.

Alexandre Myara, Nicolas Bourriez, Thomas Boyer, Thomas Lemercier, Ihab Bendidi, and Auguste Genovesio. Xfactors: Disentangled information bottleneck via contrastive supervision. *arXiv preprint arXiv:2601.21688*, 2026.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. *International conference on learning representations*, 2023.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23 (241):1–55, 2022.

Alon Shoshan, Nadav Bhonker, Emanuel Ben Baruch, Ori Nizan, Igor Kviatkovsky, Joshua Engelsma, Manoj Aggarwal, and Gérard Medioni. Fpgan-control: A controllable fingerprint generator for training with synthetic data. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6067–6076, 2024.

Yue Song, Andy Keller, Nicu Sebe, and Max Welling. Flow factorized representation learning. *Advances in Neural Information Processing Systems*, 36:49761–49782, 2023.

Bethia Sun, Maurice Pagnucco, and Yang Song. Fully distributed, flexible compositional visual representations via soft tensor products. *Advances in Neural Information Processing Systems*, 37: 139470–139560, 2024.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *International conference on machine learning*, pp. 10401–10412. PMLR, 2021.

Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *International conference on machine learning*, pp. 36336–36354. PMLR, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ancong Wu and Wei-Shi Zheng. Factorized diffusion autoencoder for unsupervised disentangled representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5930–5939, 2024.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2023.

Tao Yang, Cuiling Lan, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. *Advances in Neural Information Processing Systems*, 37: 82465–82492, 2024.

Boon Peng Yap and Beng Koon Ng. Group-wise contrastive bottleneck for weakly-supervised visual representation learning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2235–2244. IEEE, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Wei Zhang, Tik Ho Hui, Pui Ying Tse, Fraser Hill, Condon Lau, and Xinyue Li. High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 178–188. Springer, 2024.

AI USE STATEMENT

In this work, we used generative AI tools to polish the draft and check for formatting and typographical errors. We have not used generative AI tools to assist with translation, assist in the writing of proofs generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools to source and search for information and to identify relevant literature. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The internal real-world fingerprint dataset used in this study was collected and processed in compliance with applicable institutional requirements and relevant regulations. All data used in this study were obtained under appropriate authorization and approval for research purposes. To protect individual privacy, the dataset was carefully processed prior to use, and direct personal identifiers and unrelated metadata were removed or de-linked from the fingerprint samples prior to use. The dataset is used solely for scientific research, and access is restricted to authorized personnel under controlled conditions. We are committed to responsible data governance and we follow controlled data-governance procedures to minimize privacy risks and protect individual rights.

A APPENDIX

In the appendix, we first provide further details on the datasets, network architecture, and CFSS metric introduced in the main text. We then present additional quantitative and qualitative comparisons, ablation studies, and hyperparameter sensitivity analyses for TRACE. Finally, we discuss the limitations of this work.

A.1 DATASET DETAILS

In this section, we include detailed descriptions of the two datasets used in this study.

Fingerprint Dataset. We use a private fingerprint dataset containing 21,240 grayscale side-mounted fingerprint images at a resolution of 160×26. During training, we transform the image to a resolution of 160×32 to ensure stable down-sampling and up-sampling in the Encoder and Decoder. Each image

is annotated with three coarse factors: finger identity, spoof material, and acquisition batch, which we refer to as *content*, *material*, and *batch*, respectively. The content label identifies the physical finger from which an image is acquired and therefore primarily represents identity-dependent ridge structure. The dataset contains 78 content classes, 5 material classes including resin, gelatin, soft silicone, hard silicone, and wood glue, and 2 acquisition batches. We generate a single fixed data split using split seed 42 and store the resulting sample indices for reuse across all experiments. The split contains 17,040 training images, 2,100 validation images, and 2,100 test images. All compared methods, ablation variants, hyperparameter settings, and model training seeds use the same split.

Shapes3D Dataset. We use the public Shapes3D dataset, which contains 480,000 RGB images at a resolution of 64×64 with six generative factors: floor hue, wall hue, object hue, scale, shape, and orientation. For TRACE and all supervised baselines, the six original Shapes3D factors are grouped into three task-defined factors: content = shape × scale (4 × 8 = 32 classes), color = floor hue × wall hue × object hue (10 × 10 × 10 = 1000 classes), and orientation (15 classes). The same grouping is used for training, DCI evaluation, and factor swapping. To ensure reproducibility across heterogeneous baseline implementations, we construct a fixed subset of 100,000 source samples by sampling without replacement with seed 42 and sorting the selected source indices. We split this subset with split seed 2025 into 64,000 training, 16,000 validation, and 20,000 test images. All static baselines in the fair comparison are trained on the training split and evaluated on the same held-out test split. Reconstruction metrics are computed at 64×64 RGB resolution using MSE, LPIPS-Alex, and SSIM. DCI and controlled-swap evaluation use the fixed 100,000-sample latent population and the grouped content, color, and orientation labels defined in our evaluation protocol. The subset is used to remove hidden split and loader differences, prevent train/test overlap, control the evaluation population, and make training multiple heterogeneous baselines computationally feasible.

Table A1: FSIE encoder and posterior configuration.

Stage	Operation	Output size
Input	Grayscale fingerprint image	$1 \times 160 \times 32$
Encoder-1.1	5×5 Conv $1 \to 16$, BN, GELU	$16 \times 160 \times 32$
Encoder-1.2	$3\times$ ResBlock, $16 \to 16$, stride 1	$16 \times 160 \times 32$
Encoder-1.3	Downsampling ResBlock, $16 \to 32$, stride 2	$32 \times 80 \times 16$
Encoder-1.4	$2\times$ ResBlock, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Encoder-2.1	5×5 Conv $32 \to 32$, BN, GELU	$32 \times 80 \times 16$
Encoder-2.2	$3\times$ ResBlock, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Encoder-2.3	Downsampling ResBlock, $32 \to 64$, stride 2	$64 \times 40 \times 8$
Encoder-2.4	$2\times$ ResBlock, $64 \to 64$, stride 1	$64 \times 40 \times 8$
Encoder total	Two-stage residual encoder	$64 \times 40 \times 8$
Posterior	Flatten and Linear $20480 \to 448$, followed by an equal split into $\boldsymbol{\mu}_t$ and $\log \boldsymbol{\sigma}_t^2$	2×224
Factor split	Split each posterior mean vector into content, material, and batch groups	$128 + 64 + 32$

Table A2: Configuration of the custom `ResBlockTranspose` block used in the FSIE decoder. C_1 and C_2 denote the input and output channels, respectively.

Stage	Operation	Output size
Input	Input feature map	$C_1 \times H \times W$
Main branch	5×5 ConvTranspose $C_1 \to C_1$, stride s, BN, GELU; 5×5 ConvTranspose $C_1 \to C_2$, stride 1, BN, GELU	$C_2 \times sH \times sW$
Shortcut	Identity for same-resolution blocks; 1×1 ConvTranspose $C_1 \to C_2$, stride 2, BN for upsampling blocks	$C_2 \times sH \times sW$
Residual output	Shortcut + activated main branch	$C_2 \times sH \times sW$
Same-resolution ResT	$C_1 \to C_1$, stride 1	$C_1 \times H \times W$
Upsampling ResT	$C_1 \to C_2$, stride 2	$C_2 \times 2H \times 2W$

Table A3: FSIE decoder configuration. ResT denotes the custom `ResBlockTranspose` block defined in Table A2.

Stage	Operation	Output size
Latent	latent z_i	224
Decoder-FC	Linear $224 \to 20480$, reshape	$64 \times 40 \times 8$
Decoder-1.1	$3\times$ ResT, $64 \to 64$, stride 1	$64 \times 40 \times 8$
Decoder-1.2	Upsampling ResT, $64 \to 32$, stride 2	$32 \times 80 \times 16$
Decoder-1.3	$2\times$ ResT, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Decoder-2.1	$3\times$ ResT, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Decoder-2.2	Upsampling ResT, $32 \to 16$, stride 2	$16 \times 160 \times 32$
Decoder-2.3	$2\times$ ResT, $16 \to 16$, stride 1	$16 \times 160 \times 32$
Output	5×5 Conv $16 \to 1$, sigmoid	$1 \times 160 \times 32$
Decoder total	Fully connected projection, two-stage residual decoder	$1 \times 160 \times 32$

A.2 Detailed Network Architecture

This section lists the layer-wise configuration of the FSIE and SLD modules in Tables A1∼A6.

Table A4: SLD residual refinement and factor-supervision configuration. The same refinement modules are applied to the posterior mean for the factor-level objectives and to a posterior sample for reconstruction. CondHead denotes the conditional adversary that receives the refined block and a 32-dimensional embedding of its own-factor label.

Stage	Operation	Output size
FSIE statistics	Frozen posterior mean and log-variance	$224 + 224$
Sampling	$z = \mu + \exp(\frac{1}{2}\log\sigma^2) \odot \epsilon$	224
Refinement-c	LN, Linear $224 \to 128$, GELU, Linear $128 \to 128$, residual add	128
Refinement-m	LN, Linear $224 \to 128$, GELU, Linear $128 \to 64$, residual add	64
Refinement-b	LN, Linear $224 \to 128$, GELU, Linear $128 \to 32$, residual add	32
Own-factor heads	Three LN–Linear–GELU–Linear classifiers	$78/5/2$
Conditional heads	Three CondHeads for predicting the non-target factors	$7/80/83$
Refinement total	Three residual refiners and six factor-level heads	$128/64/32$

Table A5: SLD spatial-adapter configuration. Each adapter maps one refined vector-valued factor block to a factor-specific spatial feature tensor.

Stage	Operation	Output size
Adapter-c	LN, Linear $128 \to 20480$, GELU, Linear $20480 \to 10240$, reshape	$32 \times 40 \times 8$
Adapter-m	LN, Linear $64 \to 12800$, GELU, Linear $12800 \to 6400$, reshape	$20 \times 40 \times 8$
Adapter-b	LN, Linear $32 \to 7680$, GELU, Linear $7680 \to 3840$, reshape	$12 \times 40 \times 8$
Concatenation	Channel-wise concatenation of the three spatial tensors	$64 \times 40 \times 8$
Adapter total	Three factor-specific nonlinear spatial projections	$64 \times 40 \times 8$

Table A6: SLD decoder configuration. ResT denotes the custom `ResBlockTranspose` block defined in Table A2.

Stage	Operation	Output size
Bottleneck	concatenated latent factor s_i	$64 \times 40 \times 8$
Decoder-1.1	$3\times$ ResT, $64 \to 64$, stride 1	$64 \times 40 \times 8$
Decoder-1.2	Upsampling ResT, $64 \to 32$, stride 2	$32 \times 80 \times 16$
Decoder-1.3	$2\times$ ResT, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Decoder-2.1	$3\times$ ResT, $32 \to 32$, stride 1	$32 \times 80 \times 16$
Decoder-2.2	Upsampling ResT, $32 \to 16$, stride 2	$16 \times 160 \times 32$
Decoder-2.3	$2\times$ ResT, $16 \to 16$, stride 1	$16 \times 160 \times 32$
Output	5×5 Conv $16 \to 1$, sigmoid	$1 \times 160 \times 32$
Decoder total	Two-stage residual upsampling decoder without latent FC	$1 \times 160 \times 32$

A.3 CFSS FORMULATION AND CLASSIFIER SETTING

Let $k \in \{\text{content}, \text{material}, \text{batch}\}$ denote the target factor to be transferred. Given a source image a and a reference image b, we denote by $\hat{x}^{(k)}_{a \leftarrow b}$ the generated image obtained by replacing the k-th factor representation of a with the corresponding representation extracted from b, while retaining the non-target factor representations of a.

Table A7: Factor-classifier configuration. The content, material, and batch classifiers use the same lightweight CNN architecture but are trained independently with factor-specific output layers.

Stage	Operation	Output size
Input	Grayscale fingerprint image	$1 \times 160 \times 32$
Conv-1	3×3 Conv, $1 \to 32$, BatchNorm, ReLU	$32 \times 160 \times 32$
Conv-2	3×3 Conv, $32 \to 32$, BatchNorm, ReLU, 2×2 MaxPool	$32 \times 80 \times 16$
Conv-3	3×3 Conv, $32 \to 64$, BatchNorm, ReLU	$64 \times 80 \times 16$
Conv-4	3×3 Conv, $64 \to 64$, BatchNorm, ReLU, 2×2 MaxPool	$64 \times 40 \times 8$
Conv-5	3×3 Conv, $64 \to 128$, BatchNorm, ReLU	$128 \times 40 \times 8$
Global pooling	Adaptive average pooling to 1×1, flatten	128
Classification head	Linear $128 \to 128$, ReLU, Dropout ($p = 0.3$)	128
Output	Linear $128 \to$ `num_classes`	`num_classes`

For each annotated factor k, we independently train a factor classifier f_k on the training set. The three classifiers follow the architecture in Table A7 and differ only in their output dimensionality: 78 classes for content, 5 for material, and 2 for batch. Their parameters are learned independently and frozen during CFSS evaluation. The validation set is used only to estimate the calibration statistics described below.

Let $h_k(x)$ denote the softmax output of the frozen classifier f_k. The classifier-induced distance between two images is defined as

$$d_k(x_i, x_j) = d(h_k(x_i), h_k(x_j)) \,,$$

where the same representation and distance function d are used for both validation-set calibration and test-set CFSS evaluation. For each factor t, let n_t and s_t denote the average validation-set distances between image pairs with the same and different t-labels, respectively.

The factor-response matrix is defined as

$$R_{k,k} = \mathbb{E}_{(a,b)} \left[\text{clip}_{[0,1]} \left(\frac{d_k(a,b) - d_k\left(\hat{x}^{(k)}_{a \leftarrow b}, b\right)}{d_k(a,b) + \epsilon} \right) \right], \tag{A1}$$

$$R_{t,k} = \mathbb{E}_{(a,b)} \left[\text{clip}_{[0,1]} \left(\frac{d_t\left(a, \hat{x}^{(k)}_{a \leftarrow b}\right) - n_t}{s_t - n_t + \epsilon} \right) \right], \qquad t \neq k, \tag{A2}$$

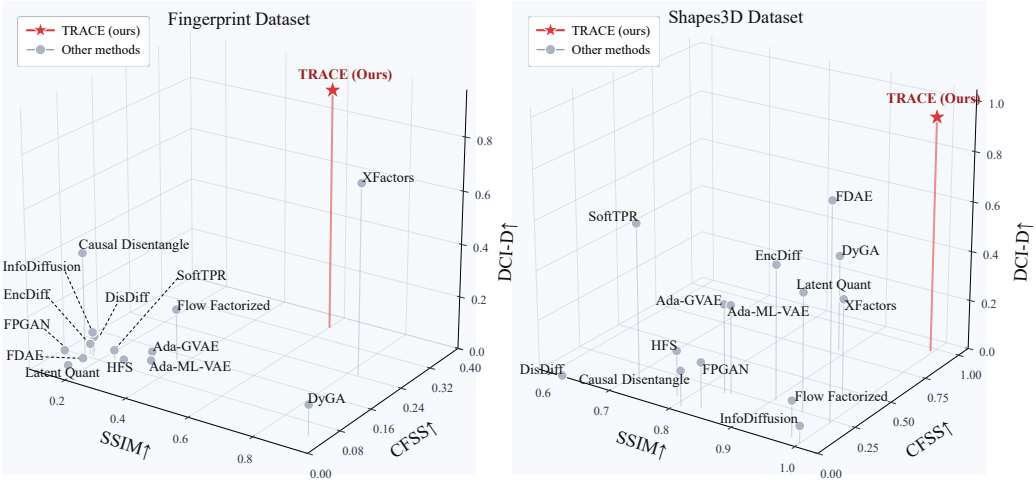


Figure A1: Quantitative comparisons among different methods.

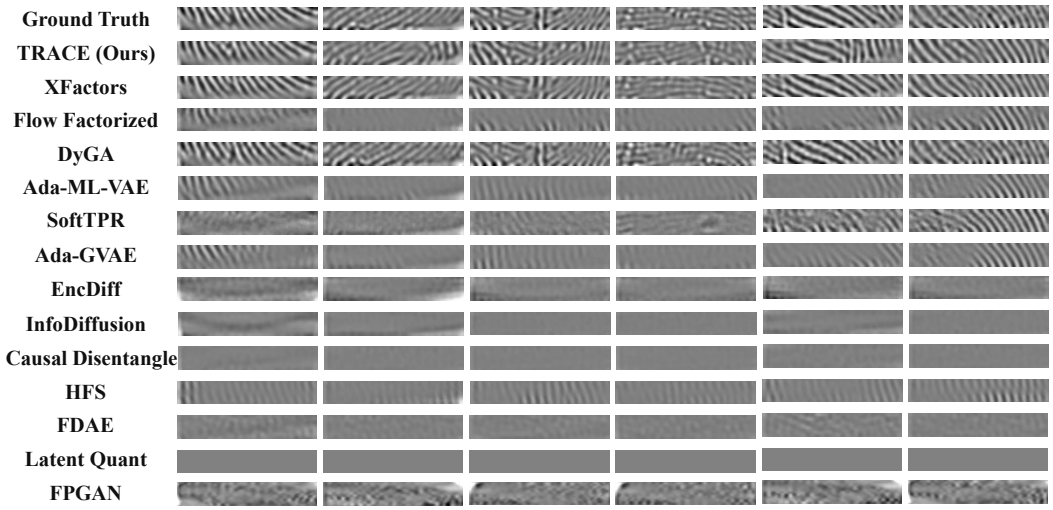


Figure A2: Qualitative visual comparisons on the Fingerprint dataset.

where $\epsilon > 0$ ensures numerical stability.

The diagonal entry $R_{k,k}$ measures the normalized reduction in distance to the reference image for the transferred factor. A value near zero indicates little or no movement toward the reference, whereas a value near one indicates that the generated image approaches the reference-level representation of factor k. The off-diagonal entry $R_{t,k}$ measures unintended changes in a non-target factor t. It is calibrated such that a value near zero corresponds to the typical distance between validation images sharing the same t-label, while a value near one corresponds to the typical distance between images with different t-labels.

For all CFSS evaluations, we use identical swap pairs, frozen factor classifiers, and calibration statistics estimated exclusively from the validation set.

A.4 Additional Quantitative and Qualitative Comparisons

In this section, we provide quantitative (Figure A1) and qualitative comparisons (Figures A2 and A3) with competing methods on the Fingerprint and Shapes3D datasets, further demonstrating the overall effectiveness of our proposed TRACE framework.

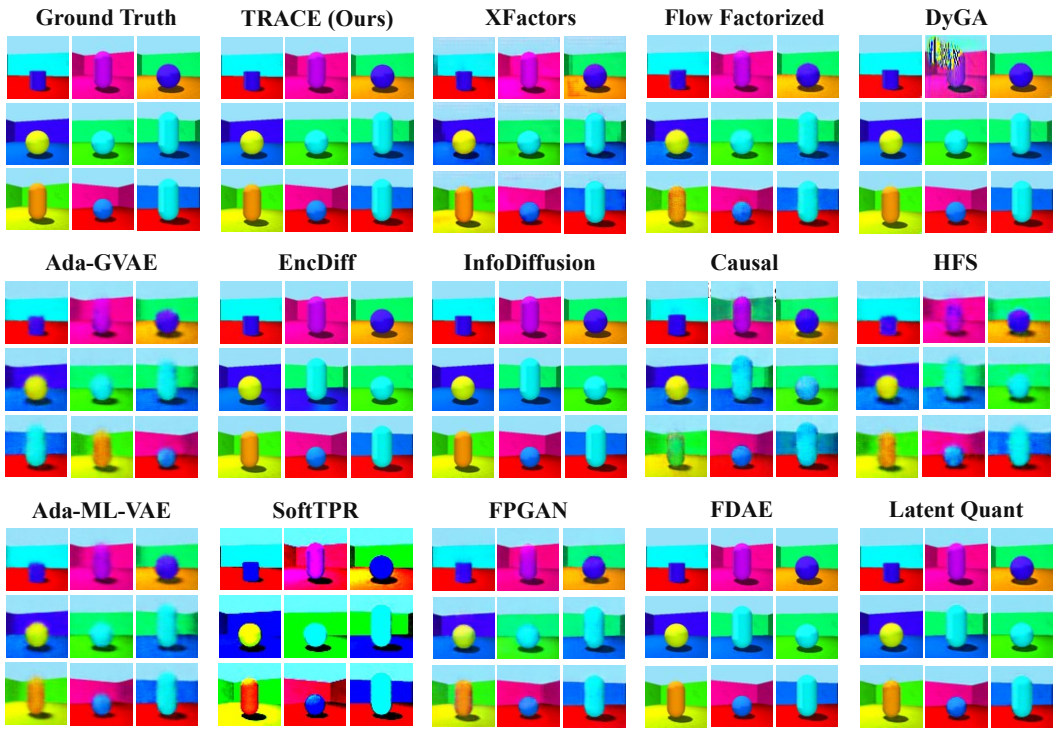


Figure A3: Qualitative visual comparisons on Shapes3D dataset.

Table A8: Additional ablation results for TRACE on the fingerprint dataset.

	Ablation Settings	CFSS ↑	MSE ↓	SSIM ↑	LPIPS ↓	D ↑
SLD	w/o $\mathcal{L}_{\text{sup}}$	0.2299 ± 0.0077	0.0108 ± 0.0001	0.644 ± 0.002	0.0653 ± 0.0010	0.925 ± 0.018
	w/o $\mathcal{L}_{\text{pre}}$	0.3493 ± 0.0073	0.0123 ± 0.0002	0.613 ± 0.005	0.0734 ± 0.0020	0.925 ± 0.018
	w/o $\mathcal{L}_{\text{anc}}, \mathcal{L}_{\text{pre}}$	0.1586 ± 0.0619	0.0205 ± 0.0030	0.315 ± 0.071	0.2330 ± 0.0629	0.925 ± 0.018
	w/o $\mathcal{L}_{\text{sup}}, \mathcal{L}_{\text{pre}}$	0.2533 ± 0.0098	0.0107 ± 0.0001	0.650 ± 0.003	0.0635 ± 0.0010	0.925 ± 0.018
Training	FSIE z_i for InfoNCE	0.3226 ± 0.0193	0.0084 ± 0.0002	0.701 ± 0.003	0.0649 ± 0.0010	0.935 ± 0.004
	SLD μ for reconstruction	0.3571 ± 0.0041	0.0127 ± 0.0002	0.606 ± 0.007	0.0731 ± 0.0022	0.925 ± 0.018
	No reconstruction gradient to residual refinement	0.3562 ± 0.0081	0.0127 ± 0.0002	0.609 ± 0.003	0.0730 ± 0.0008	0.925 ± 0.018
	TRACE (full model)	$\mathbf{0.3580 \pm 0.0071}$	0.0127 ± 0.0003	0.607 ± 0.005	0.0730 ± 0.0010	0.925 ± 0.018

A.5 Additional Ablations

Table A8 presents additional ablations of TRACE, covering different combinations of SLD losses and alternative training strategies. The loss ablations evaluate the contribution of each SLD objective to controllable generation, while the training-strategy ablations validate our design choices for reconstruction and residual refinement. Specifically, "FSIE z_i for InfoNCE" replaces μ_i with z_i in the InfoNCE objective, and "SLD μ_i for reconstruction" uses μ_i instead of z_i for final reconstruction. "No reconstruction gradient to residual refinement" blocks gradients from the reconstruction loss to the residual refinement module. The results show that the complete TRACE configuration achieves the best controllable-generation performance while maintaining competitive generation quality.

A.6 Additional Sensitivity Analysis

We conduct additional sensitivity analyses over multiple SLD parameter settings, with the corresponding Pareto plots shown in Figure A4. Across all settings, reconstruction fidelity and controllability exhibit a consistent trade-off, in line with the results in the main text. Our final configuration achieves the highest CFSS and offers the best overall balance between disentanglement and generation quality.

A.7 Discussion

TRACE learns selective factor exchange from image-level factor annotations without requiring paired supervision for the desired swapped outputs. This setting reflects our fingerprint data, where exact swap targets are unavailable; such supervision could nevertheless be incorporated when valid

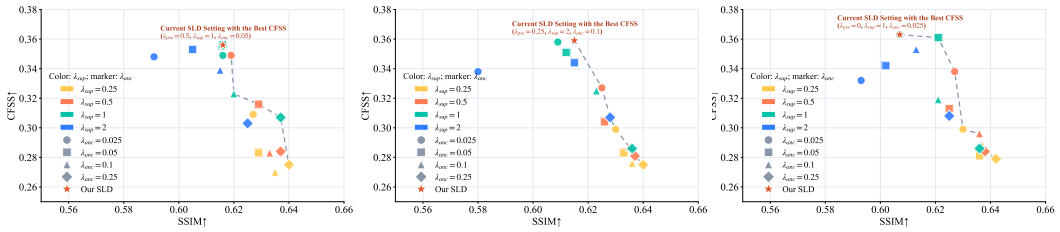


Figure A4: Additional sensitivity analysis of hyperparameters on SLD.

paired targets can be constructed. We use a VAE-based backbone to isolate the effects of factor organization and generation-oriented refinement. The proposed factor interface can also be integrated with higher-capacity generative backbones without changing the two-stage formulation.