# FROM MARGINAL MATCHING TO SAMPLER COMPATIBILITY: AUDITING INDEPENDENT LATENT RECOMBINATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Split-latent generators reconstruct an observation from content and style codes encoded from that same observation, yet generate by sampling the two codes independently. We formalize this train–generation shift as *sampler compatibility* between the encoded conditional joint and the declared sampling law. An exact expected-conditional KL identity splits incompatibility into four clauses—style-prior mismatch, condition information in style, content-prior mismatch, and within-condition code dependence—so aggregate prior matching certifies only the first. We turn the identity into an audit that separates latent compatibility, decoded transport, and generation competence. An exact-marginal counterexample and audits of image, cross-family, and known-factor models show that aggregate matching coexists with leakage that the decoder actually uses, and a power analysis shows that equality tests flag leakage long before it becomes recoverable. The audit then guides repair. Refitting the sampler to the encoded class-conditional codes raises generation accuracy from 33.9% to 96.4% on MNIST, from 46.5% to 93.0% on Fashion-MNIST, and from 26.6% to 45.8% on CIFAR-10, with recall close to the declared sampler's. A shrunk class-style shortcut instead buys accuracy by collapsing diversity. A capacity-matched ablation shows that repairing the first three clauses suffices for these models: within-condition dependence is detected but decoder-irrelevant, so content and style can stay independent within a class. The tested encoder-side contract training lowers the latent proxies it optimizes, yet worsens decoded transport on MNIST and Fashion-MNIST and helps only modestly on CIFAR-10.

## 1 INTRODUCTION

Many split-latent generators are trained and sampled in different ways. During reconstruction, content and style codes come from the same observation. During generation, the two codes are drawn or recombined independently. If the decoder has learned dependencies that occur only in paired codes, generation queries it under a latent distribution it did not observe during reconstruction.

Let $Q_y = q(z_c, z_s \mid y)$ denote the encoded pair conditional on the requested condition $y$, and let $P_y = p(z_c \mid y)p(z_s)$ denote the independently sampled decoder-input law. We call $Q_y = P_y$ *sampler compatibility*. Throughout, "content" means the code intended to carry the requested condition and "style" means the designated residual code; these are operational roles, not universal perceptual definitions. Our tests assume a categorical condition with repeated samples per class; continuous or instance-level conditions are outside the empirical scope. The contract applies only when a model uses independent recombination. A generator that models and samples $Q_y$ directly makes no such distributional substitution.

Aggregate prior matching checks only one projection of this joint claim. It may make $q(z_s)$ close to a Gaussian without ensuring that style is invariant to $y$, that encoded content follows its sampling prior, or that the two codes are independent within each condition. Good reconstruction and a small global prior discrepancy can therefore coexist with $Q_y \neq P_y$.

A failure of sampler compatibility is not, by itself, a quality verdict. A decoder can be insensitive to the mismatched directions, while matched latent laws do not guarantee realistic or task-correct out-



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.

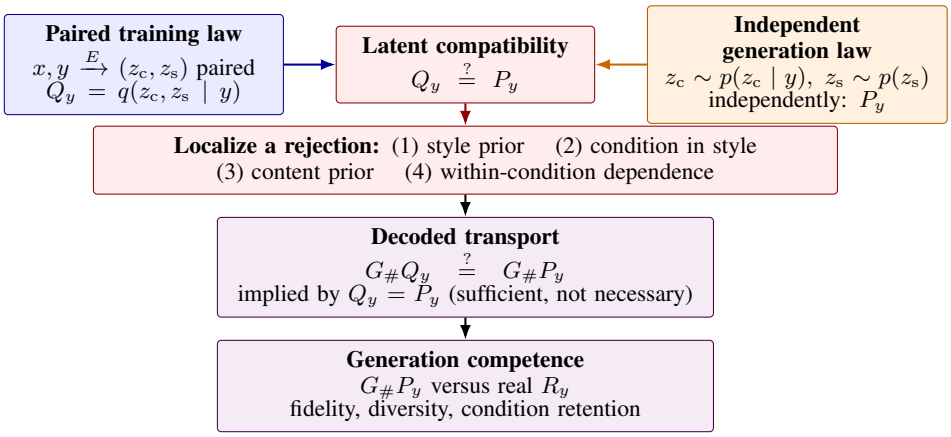


Figure 1: The audit starts from the model's actual training and generation laws. Latent equality implies decoded equality for a fixed decoder; neither implies agreement with real data, so transport and competence require separate evidence.

puts. We therefore distinguish three questions (Figure 1): latent *compatibility* ($Q_y = P_y$), *transport compatibility* ($G_\# Q_y = G_\# P_y$), and *generation competence* relative to the real conditional data law and intended task.

Our central contribution is a sampler-specific reference joint and the audit it induces. The decomposition follows from standard identities; what is new is using the declared generation law as the reference joint, which casts familiar marginal diagnostics as necessary but individually insufficient conditions, adds content-prior matching and within-condition independence, and separates latent incompatibility from its decoded consequences. We ask whether aggregate matching can conceal incompatibility, whether a latent mismatch affects decoding, and which components must be addressed to improve generation. Our findings are:

- **F1. Aggregate matching constrains one clause.** In a controlled construction, exact marginal matching coexists with over 90% label-recovery accuracy, and the sampler-compatibility null is rejected for every audited model (Sections 5.1–5.2).
- **F2. Statistical rejection is not practical severity.** Equality tests can reach 80% power while a held-out probe remains near chance, and the same diagnostics separate strong mismatch, broad prior mismatch, and near-null regimes (Sections 5.1, 5.3).
- **F3. Latent mismatch has decoded consequences.** Swap interventions make decoded identity follow the style donor, and known-factor audits locate misplaced factors (Sections 5.4–5.5).
- **F4. Three clauses suffice for sampler-side repair.** Independent class-conditional mixtures match a joint mixture on three datasets, so the detected within-condition dependence need not be modelled here (Section 6).
- **F5. The tested encoder-side objective is not a substitute.** It lowers the proxies it optimizes, but worsens decoded transport on MNIST and Fashion-MNIST and helps only modestly on CIFAR-10 (Section 6).

## 2 Related Work

**Aggregate matching and latent decomposition.** Aggregate-prior metrics examine one marginal; leakage probes examine whether a designated representation predicts a condition; factor scores describe selected cross-loadings. Each is useful, but none alone tests the conditional joint substituted at generation. This distinction connects several familiar objectives. VAEs (Kingma & Welling, 2013) penalize per-example KL, whereas WAEs (Tolstikhin et al., 2017) match aggregated posteriors with objectives such as MMD (Gretton et al., 2012). The averaged VAE KL separates information capacity from aggregate-prior mismatch (Hoffman & Johnson, 2016); $\beta$-VAE, $\beta$-TCVAE, FactorVAE,

and DIP-VAE further target capacity, unconditional total correlation, or aggregate moments (Higgins et al., 2017; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018). Our reference is instead the conditional product law induced by the declared sampler, which additionally requires a content prior and block independence given $y$. Aggregate matching is therefore not incorrect; it certifies one marginal of a larger joint claim. Aggregate-posterior "holes" in unconditional autoencoders motivate sampling from a density fitted to encoded codes after training (Rosca et al., 2018; Dai & Wipf, 2019; Ghosh et al., 2020); we test the conditional split-latent analogue of this sampler-side repair.

**Learning content–style invariance and preventing leakage.** Split-latent models use grouped observations, labels, paired supervision, or latent optimization to separate specified and residual variation (Mathieu et al., 2016; Bouchacourt et al., 2018; Jha et al., 2018; Ilse et al., 2020; Gabbay & Hoshen, 2020). Leakage filtering adds an auxiliary loss so the decoder cannot exploit style information left in the content code, or the reverse (Ridgeway & Mozer, 2018); fair representation, adversarial, Fader, and information-based objectives suppress designated nuisance signals (Louizos et al., 2015; Ganin & Lempitsky, 2015; Lample et al., 2017; Moyer et al., 2018). Recent diffusion and autoregressive systems explicitly suppress visual-prompt or style-code content leakage, including StyleKeeper and CSD-VAR (Jeong et al., 2025; Nguyen et al., 2025), as well as feature masking and variance-invariance approaches (Zhu et al., 2025; Wu et al., 2025). These methods target leakage or synthesis quality. Here condition information in style is one possible source of $Q_y \neq P_y$; it can vanish while content-prior mismatch or conditional code dependence remains. More broadly, unsupervised factor identification requires assumptions on the model or data (Locatello et al., 2019).

**Identifiability, editability, and known factors.** Recent theory treats content–style separation and editability as complementary and gives identifiability results for diffusion models under structural assumptions (Wang et al., 2025); known-factor datasets such as Shapes3D evaluate whether codes track generative variables (Higgins et al., 2017; Locatello et al., 2019; Google DeepMind, 2018). We prove neither identifiability nor editability: for a fixed encoder, decoder, and declared sampler we test whether the paired decoder-input law is the law used for recombination, and we use known factors to test a designated split in both directions.

**Recombination and its audit.** MUNIT and DRIT combine content with sampled style (Huang et al., 2018; Lee et al., 2018); recent methods use variance patterns or invertible merging to obtain content–style structure (Wu et al., 2025; Ma et al., 2025). These methods evaluate decoded quality and diversity under recombination. Recombining factors queries a decoder off its training support, where compositional generalization is empirically fragile (Schott et al., 2022) and provably needs support and architectural conditions (Wiedemer et al., 2023). Our complementary question is whether the encoded conditional joint equals the particular input law used by that recombination, which clauses differ if it does not, and whether the difference survives decoding.

## 3 Factorized Sampling and Sampler Compatibility

Let $q(y)$ denote the empirical distribution of an externally specified condition and let an encoder induce $q(z_c, z_s \mid y)$. We condition on the requested $y$ throughout; matching a separate sampling law $p(y)$ is outside this contract and would add $\mathrm{KL}(q(y)\|p(y))$ to a full-joint comparison. A factorized class-conditional sampler draws

$$z_c \sim p(z_c \mid y), \qquad z_s \sim p(z_s), \tag{1}$$

independently. In our case $p(z_s) = \mathcal{N}(0, I)$. Many aggregate objectives constrain only

$$\mathcal{L}_{\text{style}} = D(q(z_s), p(z_s)), \quad q(z_s) = \sum_y q(y) q(z_s \mid y), \tag{2}$$

For the remainder, write

$$Q_y := q(z_c, z_s \mid y), \qquad P_y := p(z_c \mid y)p(z_s), \qquad \boxed{\text{sampler compatibility: } Q_y = P_y \quad q(y)\text{-a.e.}} \tag{3}$$

Equation 1 covers generators that draw both codes from learned priors. Translation systems that deliberately retain content from an input instantiate the same analysis with that encoded-content law

in place of $p(z_c \mid y)$; term 3 is therefore sampler-specific rather than silently assumed. Our direct term-3 experiment concerns F-CS-WAE (Factorized Class-Structured Spherical Cauchy Wasserstein Auto-Encoder; Section 4), whose stated sampler does draw $p(z_c \mid y)$.

### 3.1 Compatibility, transport, and competence

The relevant object is not either marginal in isolation, but the discrepancy between $Q_y$ and $P_y$. For a possibly stochastic decoder Markov kernel $G(dx \mid z_c, z_s)$, *transport compatibility* is $G_\# Q_y = G_\# P_y$. Let $R_y = p_{\text{data}}(x \mid y)$ denote the real conditional data law. *Generation competence* asks whether $G_\# P_y$ agrees with $R_y$ under task-relevant fidelity, diversity, and condition-retention metrics. We test the latent equality first, then transport and competence separately.

**Theorem 1** (Factorized-sampling contract decomposition)**.** *Assume the displayed regular conditional distributions exist. Define the following extended non-negative quantity:*

$$M_{\text{fact}} := \mathbb{E}_{q(y)}[\mathrm{KL}(q(z_c, z_s \mid y) \,\|\, p(z_c \mid y)p(z_s))]\,. \tag{4}$$

*Then the following identity holds in* $[0, \infty]$*:*

$$\begin{aligned} M_{\text{fact}} = {} & \underbrace{\mathrm{KL}(q(z_s)\|p(z_s))}_{\text{(1) style-prior mismatch}} + \underbrace{I_q(z_s; y)}_{\text{(2) condition leakage}} \\ & + \underbrace{\mathbb{E}_{q(y)}\mathrm{KL}(q(z_c \mid y)\|p(z_c \mid y))}_{\text{(3) content-prior mismatch}} + \underbrace{I_q(z_c; z_s \mid y)}_{\text{(4) within-condition code dependence}}\,. \end{aligned} \tag{5}$$

*All four terms are non-negative. If* $M_{\text{fact}} < \infty$*, then all four terms are finite and Eq. 5 is an ordinary real-valued equality.*

The decomposition is an identity in KL, whereas our audit tests the zero-set of each term rather than estimating its numerical magnitude. The theorem allows discrete or continuous $y$; our class-stratified MMD and within-class HSIC implementation targets categorical conditions with repeated samples.

**Corollary 1** (Sampler compatibility implies transport compatibility)**.** $M_{\text{fact}} = 0$ *iff* $q(z_c, z_s \mid y) = p(z_c \mid y)p(z_s)$ *for* $q(y)$*-almost every* $y$*. Moreover, for any fixed decoder Markov kernel* $G$*,*

$$M_{\text{dec}} := \mathbb{E}_{q(y)}\mathrm{KL}(G_\# q(\cdot \mid y) \,\|\, G_\#[p(z_c \mid y)p(z_s)]) \le M_{\text{fact}}. \tag{6}$$

The implication is one-way. A decoder can discard mismatched latent directions, so $M_{\text{fact}} = 0$ is sufficient but not necessary for $M_{\text{dec}} = 0$. Moreover, neither latent nor decoder transport compatibility establishes agreement with real data. Competence metrics are therefore complementary to the latent decomposition and decoder interventions.

Exact equality is a falsifiable reference null, not by itself a practical severity threshold. For a pre-declared distribution-separating discrepancy $d$ on the relevant representation space, define

$$\Delta_d := \mathbb{E}_{q(y)} d(Q_y, P_y), \qquad (d, \epsilon)\text{-compatibility: } \Delta_d \le \epsilon. \tag{7}$$

The tolerance is necessarily metric-, scale-, and task-specific; we do not claim a universal $\epsilon$. Equality tests establish whether the zero anchor is tenable, while raw statistics, null calibration, relative effects, and decoder/task evidence assess practical magnitude.

Equation 2 constrains only term 1. The four terms are non-negative but not interchangeable, and Table 1 maps each to held-out evidence.

### 3.2 Why one clause cannot certify the contract

The following minimal counterexample isolates why exact satisfaction of term 1 leaves term 2—and therefore the joint contract—unconstrained.

**Proposition 1** (No clause-1 certificate for clause 2)**.** *Let* $z_s \sim \mathcal{N}(0, I)$ *and let* $\{A_k\}_{k=1}^K$ *be a measurable partition with* $P(z_s \in A_k) = 1/K$*. Define* $y = k$ *iff* $z_s \in A_k$*. Then* $q(z_s) = \mathcal{N}(0, I)$ *exactly for every marginal divergence* $D$*, while* $I(z_s; y) = H(y) = \log K$*.*

Table 1: The four latent requirements and their held-out evidence. These kernel statistics are calibrated proxies for distinct equalities, not additive estimates of the four KL terms.

| Term | Failure source | Required equality | Evidence used here |
|---|---|---|---|
| 1 | Style-prior mismatch | $q(z_{\mathrm{s}}) = p(z_{\mathrm{s}})$ | global MMD |
| 2 | Condition leakage | $q(z_{\mathrm{s}} \mid y) = q(z_{\mathrm{s}})$ | probe, HSIC, $\Delta_{\mathrm{inter}}$ |
| 3 | Content-prior mismatch | $q(z_{\mathrm{c}} \mid y) = p(z_{\mathrm{c}} \mid y)$ | conditional content MMD |
| 4 | Within-condition dependence | $q(z_{\mathrm{c}}, z_{\mathrm{s}} \mid y) = q(z_{\mathrm{c}} \mid y)q(z_{\mathrm{s}} \mid y)$ | within-class permutation HSIC |

The construction is intentionally extremal, but the issue is not restricted to discontinuous partitions. Let $z \sim \mathcal{N}(0, 1)$ and $P(y = 1 \mid z) = \sigma(az)$. Symmetry leaves the marginal of $z$ exactly Gaussian, while $q(z \mid y = 1) = 2\phi(z)\sigma(az)$ and $q(z \mid y = 0) = 2\phi(z)\sigma(-az)$ are smooth, full-support, and increasingly distinguishable as $a$ grows. Section 5.1 instantiates this construction only to establish clause-wise insufficiency; the trained-model experiments instead evaluate whether the complete audit distinguishes model regimes. Proofs are in Appendix A; all appendices are in the supplementary material.

# 4 Term-Aligned Audit and Case Study

## 4.1 Contract-first, term-aligned diagnostics

The audit follows diagnosis order: first test $Q_y = P_y$, and only after rejection localize the source. Independently sampled encoded and prior banks test the complete replacement without attribution:

$$T_{\mathrm{joint}} = \sum_{k \in \mathcal{K}} \omega_k \mathrm{MMD}^2(q(z_{\mathrm{c}}, z_{\mathrm{s}} \mid k), p(z_{\mathrm{c}} \mid k)p(z_{\mathrm{s}}))\,, \tag{8}$$

using a product content–style kernel and within-class pooled-label permutations. It is applied to frozen checkpoints and test populations, without retraining or model selection. If the null is rejected, the tests in Table 1 localize the difference; with a discrete condition, per-class HSIC with Holm correction over the classwise family tests conditional independence directly; kernel tests for general conditions are due to Zhang et al. (2011). These statistics test the corresponding equality rather than estimating the KL or mutual-information term. We also report classwise style-to-prior MMD, a composite diagnostic that can reflect terms 1 and 2 together. The fixed audit uses 2,048 test examples and a stratified 60/20/20 probe split with standardization fit on probe-training data only; kernels, bandwidths, permutation counts, and RNG separation are in Appendix C.

## 4.2 Controlled stress-test instrument

We use F-CS-WAE as a fully instrumented stress-test model, not as a proposed state-of-the-art generator. Its declared sampler exposes all four contract clauses; cross-family and known-factor experiments test whether the diagnostics distinguish behavior beyond this instrument. Its modified ResNet-18 encoder produces hyperspherical content $z_{\mathrm{c}} \in \mathbb{S}^{63}$ and Gaussian style $z_{\mathrm{s}} \in \mathbb{R}^{128}$ for a residual decoder. Class-specific Spherical Cauchy priors shape content; global MMD matches sampled style to $\mathcal{N}(0, I)$. Reconstruction, supervised content-prior matching, aggregate content matching, and a content classifier complete the phased 300-epoch objective. Appendix B specifies the sampler, architecture, coefficients, and schedule.

The historical per-class loss $K^{-1}\sum_k \mathrm{MMD}^2(q(z_{\mathrm{s}} \mid k), \mathcal{N}(0, I))$, with weight $\delta$ (Appendix D), targets conditional style matching but leaves term 4 outside the intervention. Its observed instability motivates the encoder-side repair compared with sampler-side repair in Section 6.

# 5 Diagnosing Incompatibility

**Reporting rules.** The image audits cover MNIST, Fashion-MNIST, CIFAR-10, paired $-30°/+30°$ Rotated-MNIST, and Shapes3D. All headline sampling-contract diagnostics use stochastic posterior samples and held-out splits.

Unbiased $\mathrm{MMD}^2$ estimates may be negative under the null and are never clipped. Non-rejection is interpreted at the available test power. HSIC uses plus-one permutation calibration, with domain-level decisions rather than averaged $p$-values. Shapes3D applies Holm correction separately within the six content tests and six style tests. Final epochs and audit populations are fixed without selecting checkpoints on leakage metrics.

### 5.1 Can aggregate matching hide joint incompatibility?

The decomposition makes aggregate matching a clause rather than a certificate. We first isolate that logical point by instantiating the smooth construction from Proposition 1 in 128 dimensions:

$$z \sim \mathcal{N}(0, I), \qquad p(y = 1 \mid z) = \sigma(2\kappa z_1). \tag{9}$$

The latent is sampled before the label, so $q_\kappa(z) = \mathcal{N}(0, I)$ exactly for every $\kappa$: term 1 is zero by the data-generating process, not because a test fails to reject. We use 50 replications of 2,048 samples, 499 MMD and HSIC permutations, and the same stratified 60/20/20 linear-probe protocol as the checkpoint audit.

Probe accuracy rises from $49.58 \pm 1.86\%$ at $\kappa = 0$ to $92.30 \pm 1.51\%$ at $\kappa = 16$. The independence null is rejected in $100\%$ of replications for every $\kappa > 0$, against a $2\%$ rejection rate at $\kappa = 0$. Mean conditional MMD rises from $6.16 \times 10^{-6}$ to $4.78 \times 10^{-4}$. These results support the intended existential claim: exact marginal equality can coexist with more than $90\%$ recoverable label information. The complete acceptance audit and sample-size sensitivity are reported in Appendix F. This is an exact-marginal counterexample, not a trained-model result.

Global-MMD rejection stays at its nominal 5–6% for $n = 512$–$4096$ (Table S8). A post-hoc power analysis at small $\kappa$ (Table S9) shows the label test reaching 80% power at $\kappa = 0.15$ for $n = 2048$, where a held-out linear probe is only 51.2% accurate: rejection detects leakage far below practical recoverability, so we report effect sizes and decoded consequences next to every test.

### 5.2 Does the complete contract fail in trained models?

On six F-CS-WAE seed–domain audits, the direct joint test rejects the sampler-compatibility null in every aggregate test and in all 60 digit-level tests, and the content-prior test in 56 of 60 (statistics $0.0194 \pm 0.0024$ and $0.0081 \pm 0.0035$; all 12 aggregate rejections survive Bonferroni correction; Appendix I). The joint test is omnibus; the clause diagnostics show that content-prior mismatch contributes alongside condition leakage and within-condition dependence.

### 5.3 Does the audit distinguish model-family regimes?

We next ask whether diagnostics shared across model families report leakage indiscriminately or separate practically different regimes. We train native-family, common-backbone adaptations of F-CS-WAE, DRIT (Lee et al., 2018), and DIVA (Ilse et al., 2020) for 100 epochs on the same two-domain Rotated-MNIST population. The audited style variables are F-CS-WAE $z_s$, DRIT's domain-specific attribute code, and DIVA's residual $z_x$. DRIT domains are audited separately because their attribute spaces are domain-specific.

Table 2: Cross-model audit on a common Rotated-MNIST backbone, mean $\pm$ sample standard deviation over three seeds. Accuracies and retention are percentages; values are rounded to three significant digits (full precision in Appendix H). Prior retention: how often an independent classifier recovers the source digit after decoding its encoded content with a style drawn from the family's own prior. Raw MMD values describe the fixed audit geometry, not a model-quality ranking; an unbiased estimate may be negative under its null.

| Family | Global MMD | Style–prior MMD | Style probe | Content acc. | Recon. L1 | Prior retention |
|---|---|---|---|---|---|---|
| F-CS-WAE | $0.00333 \pm 0.00014$ | $0.0323 \pm 0.0012$ | $91.9 \pm 2.4$ | $99.4 \pm 0.1$ | $0.0151 \pm 0.0004$ | $12.4 \pm 1.2$ |
| DRIT | $0.2553 \pm 0.0085$ | $0.2724 \pm 0.0019$ | $34.4 \pm 12.6$ | $88.3 \pm 1.5$ | $0.0428 \pm 0.0021$ | $9.9 \pm 0.1$ |
| DIVA | $(-1.8 \pm 2.7) \times 10^{-5}$ | $(1.8 \pm 1.3) \times 10^{-4}$ | $10.7 \pm 1.0$ | $96.7 \pm 0.8$ | $0.0345 \pm 0.0001$ | $89.5 \pm 0.8$ |

The cross-family subset separates three qualitatively different outcomes. F-CS-WAE has a larger conditional than global discrepancy and high, seed-stable label recovery. DRIT's global discrepancy is already large, so its label-bearing attribute code is a broad prior-sampler mismatch rather than a matched-marginal paradox. DIVA has near-null global discrepancy, chance-level probes, and competent prior-style retention. The conditional-independence null is rejected in all six seed–domain tests for F-CS-WAE and DRIT ($p = 1/201 \approx 0.005$). Two of six raw DIVA tests fall below 0.05, but none survives Holm correction within the DIVA family and the probe pattern does not repeat. DIVA therefore provides a near-null regime at the available power.

All three pass the pre-specified content-representation gate, so a collapsed content code does not explain the contrast (Appendix H).

### 5.4 Does the mismatch survive decoding?

**Decoded semantics.** The data-processing inequality leaves open whether a latent mismatch affects outputs. We therefore pair the fixed MNIST, Fashion-MNIST, and CIFAR-10 audit with prior-sampling and donor-swap interventions. The mismatch is detectable on both datasets: conditional MMD is 0.016 on MNIST and 0.010 on Fashion-MNIST, and every seedwise test rejects the independence null ($p = 0.005$). It also survives decoding. Replacing global Gaussian style by a post-hoc class-diagonal style law (shrunk to $\tau = 0.25$, which reduces diversity) raises MNIST Gen-ACC from 10.47% to 100%, showing decoder sensitivity rather than better generation; deterministic mean swaps similarly make decoded identity follow the style donor in 99.28% of MNIST and 63.92% of CIFAR-10 cases (a per-class style penalty lowers this to 23.37% and 35.92%; Appendix E).

The variance head of F-CS-WAE often reaches its numerical floor, but a variance-floor intervention (full sweep on seed 0; selected levels confirmed over three seeds) does not fully explain the effect: increasing posterior noise improves independent generation while substantial MNIST label recovery remains. The complete intervention, CIFAR-10 contrast, legacy classwise-style result, qualitative grids, and per-level statistics are reported in Appendices G and D.

### 5.5 Does the audit recover known generative factors?

On an IID Shapes3D split, every model passes reconstruction/content gates, but F-CS-WAE style predicts shape at 99.0% and content predicts nuisance factors at 87.0% on average. The Holm-corrected pattern repeats over three seeds: for F-CS-WAE the independence null is rejected for all six factors in both views, whereas for a learned content–style VAE only the expected complementary factors are rejected. That model also preserves shape at 99.7% under prior decoding and 99.8% under swap decoding, versus 89.0% and 88.0% for F-CS-WAE. Appendix J reports the second near-null control, oracle scope, and per-factor results.

## 6 Repairing a Failed Contract

The decomposition names which clauses fail; it does not say which side of the contract to change. Either the encoder is trained until $Q_y$ matches the declared $P_y$, or the sampler is replaced by one that matches the encoded $Q_y$. Corollary 1 is agnostic to which side moves: any sampler $P'_y = Q_y$ makes $G_\# P'_y = G_\# Q_y$ exactly. Encoder-side compatibility is achievable by design (DIVA is near-null in Table 2); our question is narrower: for a trained model that fails the audit, which post-hoc repair works, and which clauses must it cover?

**Protocol.** We retrain 3 matched pairs per dataset on MNIST, Fashion-MNIST, and CIFAR-10: a schedule control and a factorized-contract training (FACT) model per seed, with identical 40-epoch arguments frozen before training and never re-tuned per dataset (Appendix M). The MNIST controls reproduce the archived checkpoints of Section 5.2. Keeping encoder and decoder fixed, we fit samplers on training-split posterior samples and evaluate on the test split. They differ in the clauses of Theorem 1 they repair and in capacity: S0 is the declared $p(z_c \mid y)\mathcal{N}(0, I)$; S1 is the class-diagonal style law shrunk to $\tau = 0.25$ (the diagnostic of Section 5.4); S2 draws style from a full-covariance Gaussian fit of $q(z_s \mid y)$ (terms 1–2); S5 fits one Gaussian to the joint $(z_c, z_s) \mid y$ (terms 1–4); S4 fits independent ten-component mixtures to $q(z_c \mid y)$ and $q(z_s \mid y)$ (terms 1–3); and S3 fits a ten-component mixture to the joint (terms 1–4), in the spirit of ex-post density

Table 3: Sampler-side repair on fixed encoders and decoders (controls; means over 3 seeds). Gen-ACC in %; FD in each dataset's evaluator features (↓). Terms: sampler-compatibility conditions modelled by each sampler (*shrunk scale); capacity: unimodal (1) or ten-component mixture (10). Decoded $Q_y$ is the paired-code reference, not an upper bound. Per-seed values, precision, and memorization ratios are in Tables S26, S29, and S27.

| | | | | MNIST | | | Fashion-MNIST | | | CIFAR-10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sampler | Terms | Capacity | Acc | FD | Recall | Acc | FD | Recall | Acc | FD | Recall |
| S0 | declared | – | – | 33.9 | 1811 | 0.76 | 46.5 | 564 | 0.66 | 26.6 | 18.5 | 0.87 |
| S1 | class diag. | 1–2* | 1 | 99.0 | 426 | 0.13 | 98.0 | 331 | 0.09 | 29.4 | 22.3 | 0.11 |
| S2 | cond. style | 1–2 | 1 | 83.4 | 401 | 0.87 | 87.8 | 155 | 0.67 | 20.7 | 23.1 | 0.37 |
| S5 | joint Gaussian | 1–4 | 1 | 83.8 | 389 | 0.87 | 88.0 | 156 | 0.66 | 20.6 | 23.1 | 0.43 |
| S4 | indep. mixtures | 1–3 | 10 | **96.6** | **57** | 0.90 | **93.3** | **68** | 0.63 | **44.0** | **13.0** | 0.71 |
| S3 | joint mixture | 1–4 | 10 | **96.4** | **61** | 0.89 | **93.0** | **71** | 0.64 | **45.8** | **12.4** | 0.78 |
| | decoded $Q_y$ | – | – | 99.4 | 5.5 | | 90.5 | 52.1 | | 63.5 | 6.4 | |

estimation (Ghosh et al., 2020; Dai & Wipf, 2019). S4 and S5 form a post-hoc, capacity-matched ablation. An independent real-image classifier gives Gen-ACC; Fréchet distance (FD) and $k$-NN precision/recall (Kynkäänniemi et al., 2019) are computed in its penultimate features against real test images of the requested class.

**Refitting the class-conditional marginals improves generation.** On all three datasets the declared sampler falls well short of the paired-code reference (Table 3). The joint mixture S3 improves Gen-ACC and FD over S0 while keeping recall at or above S1 in 3/3, 3/3, and 3/3 controls, the pre-registered criterion, closing 97%, 96%, and 50% of the FD gap to that reference. Its memorization ratios (1.00, 0.95, 0.95; one means as far from the training set as real test images) show no increase under this metric. Accuracy alone favours the shrunk shortcut S1 on MNIST and Fashion-MNIST, despite its recall collapse (0.13, 0.09, 0.11); diversity-aware metrics are necessary to judge a repair. Training length is not the cause: on the 4 canonical 300-epoch checkpoints finished before the cut-off, S0 gives 57.1±36.2% at recall 0.55 and S3 97.2±0.5% at recall 0.86, closing 95% of the gap (4/4; Appendix M).

**Three clauses suffice for the evaluated models.** The capacity-matched ablation separates clause coverage from sampler capacity. Independent mixtures for content and style (S4, terms 1–3) match the joint mixture (S3) on every dataset, whereas a unimodal joint fit (S5, terms 1–4) performs like the style-only S2, and on CIFAR-10 both unimodal fits fall below the declared sampler. Within-condition dependence (term 4) is detected in all 18 audited checkpoints ($p = 0.005$) yet leaves the decoder unaffected here: a latent failure without decoded consequence, exactly the gap between compatibility and transport that Corollary 1 allows. We do not conclude that term 4 is generally irrelevant, only that it need not be modelled for these checkpoints. The repair can therefore keep content and style independent within each class; what it must give up is class-independent style.

**The tested encoder-side objective.** Our FACT objective replaces classwise penalties by null-calibrated MMD and HSIC constraints on the diagnosed clauses, plus a label adversary (Appendix L). On MNIST and Fashion-MNIST it does what an encoder-side optimizer is asked to do: the joint statistic falls and the style probe drops to chance (Table 4). Yet there the decoded discrepancy between $G_{\#}Q_y$ and $G_{\#}P_y$ grows in every pair, declared-sampler accuracy falls, and the models become harder to repair. On CIFAR-10 the pattern reverses: FACT moves the joint statistic inconsistently but lowers the decoded discrepancy and raises declared-sampler accuracy by 4.9–6.7 points, still far below the sampler repair. Corollary 1 bounds decoded KL by latent KL, not by the finite-sample MMD/HSIC proxies that FACT optimizes, so neither sign is guaranteed. The archived pairs show the MNIST pattern with additional swap-attribution evidence (Appendix L).

Table 4: Matched control/FACT pairs (3 seeds per dataset). $\Delta T_J$: change of the joint statistic; $\Delta\text{FD}_{QP}$: change of the decoded discrepancy between $G_{\#}Q_y$ and $G_{\#}P_y$ (ranges over pairs; negative is better); Gen-ACC (%) of the declared (S0) and joint-mixture (S3) samplers, control → FACT.

| Dataset | $\Delta T_J$ (%) | $\Delta\text{FD}_{QP}$ (%) | Worse pairs | S0 Gen-ACC | S3 Gen-ACC |
|---|---|---|---|---|---|
| MNIST | $-38$ to $-27$ | $+7$ to $+20$ | 3/3 | 33.9 → 28.9 | 96.4 → 83.7 |
| Fashion-MNIST | $-54$ to $-12$ | $+29$ to $+53$ | 3/3 | 46.5 → 32.6 | 93.0 → 72.7 |
| CIFAR-10 | $-43$ to $+10$ | $-37$ to $-14$ | 0/3 | 26.6 → 32.6 | 45.8 → 50.0 |

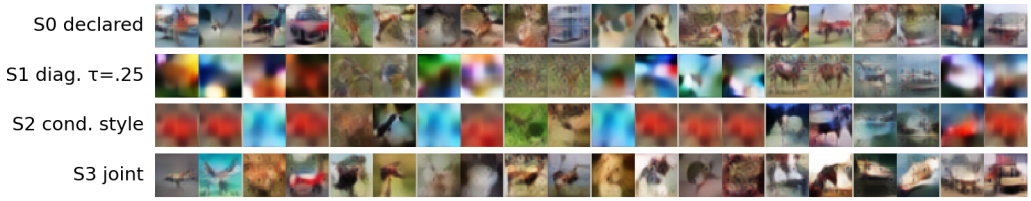


Figure 2: CIFAR-10 control (seed 0), requested classes airplane to truck with two unselected draws each; encoder and decoder are identical across rows. The shrunk shortcut S1 and the unimodal style fit S2 yield near-duplicate or off-class blobs; the joint mixture S3 yields varied, class-consistent samples (Table 3).

## 7 DISCUSSION

**A protocol for practitioners.** The findings suggest five steps for any split-latent generator that samples its codes independently. (1) Write down the declared sampler; it, not a generic prior, is the reference law. (2) Test $Q_y = P_y$ jointly. (3) If it fails, localize the failure with the four clauses. (4) Measure the decoded consequence and generation competence with diversity-aware metrics, and report effect sizes next to every rejection: rejection alone does not imply severity (F2). (5) Repair on the side that preserves the needed capability: refit the class-conditional marginals with enough capacity for class-conditional generation (F4); change the encoder or architecture for cross-class recombination, since contract penalties were unreliable (F5).

**What the sampler repair does and does not buy.** The repair keeps content and style independent within a class, but draws style given the label, so cross-class style transfer is not repaired; the audited models lacked that capability under the declared sampler as well. Deployment claims should predeclare $(d, \epsilon)$, report $\Delta_d$ with decoded effect sizes, and never certify a model with its own training target.

**Limitations.** Our evidence covers low-resolution datasets and split-latent autoencoding or translation architectures: Shapes3D is IID, DRIT and DIVA are common-backbone adaptations, SCFlow is a CLIP-space boundary case (Appendix K), and the CIFAR-10 paired-code reference itself reaches only 63.5% Gen-ACC. The repair uses finite mixtures, the encoder-side comparison covers one objective, and memorization is checked in pixel space only. The pre-registered population-level test of predictive validity did not finish before our cut-off (Appendix M). Released high-resolution models, continuous conditions, and flow or diffusion priors over $(z_c, z_s)$ remain open.

## 8 CONCLUSION

Independent latent recombination silently substitutes a factorized sampling law for the paired law a decoder was trained on. Treating that substitution as a contract makes its failure testable: aggregate prior matching constrains only one of four clauses, the remaining clauses are violated in the evaluated models, and some violations have measurable decoded consequences. The decomposition also guides repair: on three datasets, refitting the flagged class-conditional marginals improves generation without retraining, while detectable within-condition dependence needs no modelling here; the tested contract penalties do not reliably help. Separating compatibility, transport, and competence, and reporting decoded effect sizes with every rejection, should become standard practice.

## REPRODUCIBILITY STATEMENT

Section 4 specifies the latent views, preprocessing, tests, permutation calibration, and split discipline. Sections 5.1– 6 state setting-specific populations, seeds, analysis status, and fixed reporting rules. The appendices, provided as supplementary material, contain complete proofs, architectures, per-seed results, multiplicity rules, competence gates, and limitations for each external audit. The anonymous supplementary package ties unrounded headline rows, figures, split/protocol metadata, and critical source files to a content manifest; checkpoint hashes are retained without claiming that the large checkpoint files are packaged. The retrained population of Section 6 follows a protocol frozen before training, released with its deviation log, audit and analysis scripts, per-checkpoint result files, and a generator that writes every reported number of that section into the manuscript source. Appendix N collects the evidence roots and boundaries.

## ETHICS STATEMENT

This work uses standard public computer-vision benchmarks and existing model checkpoints, collects no new data, and involves no human subjects or sensitive personal information. Its purpose is to identify and measure a failure mode in factorized generative models. Generated examples are low-resolution benchmark images; nevertheless, the broader recommendation—auditing whether a decoder uses unintended latent information—may also help identify representation leakage in higher-stakes applications.

## AI USE STATEMENT

Generative AI tools assisted with searching and summarizing literature, and with writing and debugging code. They were not used to generate datasets or produce experimental measurements. The authors reviewed the cited primary literature, verified every mathematical claim and every reported value against its artifact, ran all experiments, and take responsibility for the final content, claims, code, and artifacts.

## 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, pp. 2095–2102, 2018. doi: 10.1609/aaai.v32i1.11867.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations*, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189, 2015.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020.

Google DeepMind. 3d shapes dataset. `https://github.com/google-deepmind/3d-shapes`, 2018. Accessed 2026-08-12.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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. *International Conference on Learning Representations*, 2017.

Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: Yet another way to carve up the variational evidence lower bound. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pp. 172–189, 2018.

Maximilian Ilse, Jakub M. Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pp. 322–348, 2020.

Jaeseok Jeong, Junho Kim, Gayoung Lee, Yunjey Choi, and Youngjung Uh. StyleKeeper: Prevent content leakage using negative visual query guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15760–15769, 2025.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision*, pp. 805–820, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, 2017.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, pp. 35–51, 2018.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Pingchuan Ma, Xiaopei Yang, Yusong Li, Ming Gui, Felix Krause, Johannes Schusterbauer, and Björn Ommer. SCFlow: Implicitly learning style and content disentanglement with flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14919–14929, 2025.

Michael F. Mathieu, Junbo Jake Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pp. 9102–9111, 2018.

Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. CSD-VAR: Content-style decomposition in visual autoregressive models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17013–17023, 2025. doi: 10.1109/ICCV51701.2025.01580.

Karl Ridgeway and Michael C. Mozer. Open-ended content-style recombination via leakage filtering. *arXiv preprint arXiv:1810.00110*, 2018.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Liming Wang, Muhammad Jehanzeb Mirza, Yishu Gong, Yuan Gong, Jiaqi Zhang, Brian H. Tracey, Katerina Placek, Marco Vilela, and James R. Glass. Can diffusion models disentangle? a theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0015.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, 2023.

Yuxuan Wu, Ziyu Wang, Bhiksha Raj, and Gus Xia. Unsupervised disentanglement of content and style via variance-invariance constraints. In *International Conference on Learning Representations*, 2025.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

Lin Zhu, Xinbing Wang, Chenghu Zhou, Qinying Gu, and Nanyang Ye. Less is more: Masking elements in image condition features avoids content leakages in style transfer diffusion models. In *International Conference on Learning Representations*, 2025.