NECESSARY AND SUFFICIENT CONDITIONS FOR AUTOENCODER-BASED STYLE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

While style transfer and controllable generation have achieved strong empirical performance across audio, vision, and multimodal learning, most existing methods remain largely heuristic and offer limited theoretical guarantees. This paper studies autoencoder-based attribute transformation from a theoretical perspective and characterizes when exact conversion is achievable. We introduce a formal framework that isolates a task-relevant factor from a controllable attribute and ask when a learned encoder–decoder pair can implement the desired transformation while preserving all non-target information. Our main contributions are theoretical: under minimal and interpretable assumptions on the data-generating process, we establish both sufficient and necessary conditions for exact attribute conversion. We further move beyond idealized settings by deriving explicit error bounds that quantify how imperfect reconstruction and residual dependence in the learned representation translate into transfer degradation. Overall, our analysis shifts the focus from latent recovery to synthesis-oriented guarantees, providing a principled foundation for understanding when reliable task-driven style transfer is possible.

1 INTRODUCTION

Understanding and controlling structured variability in high-dimensional observations is a central problem in machine learning. Across domains such as speech, vision, text, and multimodal learning, observed data arise from the interaction of multiple latent factors, including semantic content, style or identity, and other controllable attributes. The ability to selectively manipulate a designated latent factor while preserving all remaining factors underlies a wide range of controllable generation systems, including voice conversion and emotion transfer in speech (Qi et al., 2025; Kim et al., 2025), controlled image synthesis and style manipulation in vision (Wang et al., 2023; Jin, 2025), and task-oriented text generation and writing style transfer (John et al., 2019; Liang et al., 2024; Ashok & Poczos, 2024).

These applications motivate the study of latent representations that are stable, controllable, and recoverable from data, particularly in regimes where direct supervision of the underlying factors is unavailable. A common empirical approach relies on autoencoder-based architectures, defined as pairs of functions, an encoder and a decoder, trained jointly to represent observations in a latent space from which the input can be reconstructed or transformed. While such models demonstrate strong empirical performance, e.g., (Wang et al., 2024), their success alone does not explain when or why the learned representations correspond to the underlying generative factors. Consequently, the structural conditions under which an encoder-decoder pair can be guaranteed to recover and manipulate task-relevant latent variables remain insufficiently understood.

From a theoretical perspective, this problem is closely related to identifiability in latent variable models. In particular, Independent Component Analysis (ICA) studies conditions under which latent sources can be recovered from observed data (Comon, 1994). Classical ICA theory focuses on full source recovery, establishing identifiability of all latent components up to admissible transformations. More recent work has extended these guarantees to certain classes of nonlinear ICA under assumptions on the data-generating process and the availability of auxiliary variables (Hyvarinen et al., 2019; Khemakhem et al., 2020). Our work is aligned with this line of research but departs from it in scope and objective. Rather than requiring recovery of all latent components, we consider a setting in which only a single task-relevant latent factor must be recovered, while the remaining factors are treated as

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.

observed or conditioned upon. Within this framework, we study when a learned encoder–decoder pair can implement an exact *attribute conversion map*, preserving the relevant latent factor while modifying the conditioning variable.

Recent work by Shaham et al. (2022) provides theoretical guarantees for latent component *analysis* in related settings, establishing sufficient conditions for recovering latent variables from observed data. We extend this line of work by shifting the focus from analysis to *synthesis*. Rather than asking when latent components can be identified, we ask when they can be preserved up to an invertible transformation and reliably *used* to generate target-conditioned samples. This synthesis-oriented perspective enables a principled theoretical characterization of controllable attribute conversion. We introduce sufficient and necessary conditions under which correct style conversion is guaranteed as a synthesis task: the model must preserve the recovered latent factor up to an invertible transformation and combine it with the manipulated observed attribute to generate the desired target output.

2 NOTATIONS

Let $\mathbf{X}$ be a random variable with support $\mathrm{supp}(\mathbf{X}) = \mathbb{X}$, and let $\mathbf{x} \in \mathbb{X}$ denote a realization of $\mathbf{X}$. $\mathcal{P}_{\mathbf{X}}$ is a distribution of the random variable $\mathbf{X}$, $p_{\mathbf{X}}$ is the density function of $\mathcal{P}_{\mathbf{X}}$. If $\mathbf{V} := \pi(\mathbf{U})$ and π is measurable, then $\pi_{\#}\mathcal{P}_{\mathbf{U}}$ is the distribution of $\mathbf{V}$. $\mathrm{Im}(f)$ denotes the image of map f. The operator $\otimes$ denotes the product of probability measures. In particular, if $\mathbf{U}$ and $\mathbf{V}$ are independent random variables with distributions $\mathcal{P}_{\mathbf{U}}$ and $\mathcal{P}_{\mathbf{V}}$, then their joint distribution is given by the product measure $\mathcal{P}_{\mathbf{U}} \otimes \mathcal{P}_{\mathbf{V}}$, whose density (when it exists) is $p_{\mathbf{U}} \otimes p_{\mathbf{V}}$. The operator $\times$ denotes Cartesian product of maps, e.g., if $f : \mathbb{X} \to \mathbb{U}$, and $g : \mathbb{Y} \to \mathbb{V}$, then $f \times g : \mathbb{X} \times \mathbb{Y} \to \mathbb{U} \times \mathbb{V}$.

3 PROBLEM FORMULATION

Let $\mathbf{S} \in \mathcal{S}$ be a discrete latent random variable representing task-relevant information that should be preserved under conversion. Let $\mathbf{C} \in \mathcal{C}$ be a discrete observed random variable representing controllable style attributes, such as speaker identity in speech, writing style in text, or pose angle in head image generation. We denote by $\mathbb{S} = \mathrm{supp}(\mathbf{S}) \subseteq \mathcal{S}$ and $\mathbb{C} = \mathrm{supp}(\mathbf{C}) \subseteq \mathcal{C}$ the supports of $\mathbf{S}$ and $\mathbf{C}$, respectively.

Let $\mathcal{P}$ be a probability distribution over $\mathbb{S} \times \mathbb{C}$ having density p. We assume that $\mathbf{S}$ and $\mathbf{C}$ are independent, i.e., p is an outer product of the marginal densities $p = p_{\mathbf{S}} \otimes p_{\mathbf{C}}$. This independence reflects the standard assumption that task-relevant content $\mathbf{S}$ is statistically decoupled from controllable style variables $\mathbf{C}$; all necessity results below are conditioned on this assumption.

Let $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$ be a random variable that is generated by an unknown deterministic function f. Our task is *attribute conversion*: given an observation $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$ and a target attribute $\mathbf{c}'$, synthesize a sample $\mathbf{x}' = f(\mathbf{s}, \mathbf{c}')$ that preserves $\mathbf{s}$ while replacing $\mathbf{c}$ by $\mathbf{c}'$.

4 LEARNING FRAMEWORK

We formalize attribute conversion as the problem of learning a structured factorization of the unknown generative process f through an autoencoder architecture. Since observations are generated as $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$, successful conversion requires isolating the task-relevant latent variable $\mathbf{S}$ from the observed sample $\mathbf{X}$ while allowing the controllable attribute $\mathbf{C}$ to be modified. Formally, we model this process using a conditional autoencoder, which we now define.

Let $\mathcal{A}$ be a hypothesis class of neural network–based conditional autoencoders $g : \mathbb{X} \times \mathbb{C} \to \mathbb{X}$ and representable as

$$g(\mathbf{x}, \mathbf{c}) = d(e(\mathbf{x}), \mathbf{c}),$$

where the target condition $\mathbf{c}$ is provided to the network at the layer that defines the start of the decoder $d : \hat{\mathbb{S}} \times \mathbb{C} \to \mathbb{X}$ map. We assume that the hypothesis class $\mathcal{A}$ is sufficiently expressive to represent nonlinear measurable functions on the support of variable $\mathbf{X}$ (Hornik, 1993). Moreover, we require that models in $\mathcal{A}$ satisfy two main properties that are essential for reliable attribute conversion. In Section 5, we show that these two properties constitute a minimal characterization of exact conversion:

they are sufficient, and they are necessary either up to an equivalent model, or for the given model under a minimal information bottleneck principle.

Property 1 (Perfect reconstruction) *For a given trained model that is represented by the map $g(\mathbf{x})$, it holds that:* $g(\mathbf{x}, \mathbf{c}) = \mathbf{x}$ *for all* $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$ *such that* $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$.

Perfect reconstruction ensures that the decoder can faithfully reproduce any observed sample when conditioned on its true attribute. Exact conversion additionally requires that the latent representation supplied to the decoder be invariant to the original attribute. Any residual dependence would prevent correct counterfactual synthesis under a different conditioning value.

Property 2 (Perfect independence) *For the learned autoencoder g, there exists a layer defining an encoder map $e : \mathbb{X} \to \hat{\mathbb{S}}$, such that the induced latent random variable $\hat{\mathbf{S}} := e(\mathbf{X})$ is statistically independent of the observed attribute $\mathbf{C}$, that is, $\hat{\mathbf{S}} \perp\!\!\!\perp \mathbf{C}$.*

Given realizations of the observed variables $\mathbf{X}$ and $\mathbf{C}$, such that for each pair of realizations $\{(\mathbf{x}, \mathbf{c})\}$ there exists unobserved $\mathbf{s} \in \mathbf{S}$ such that $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$, our goal is to learn $g \in \mathcal{A}$ such that the resulting conversion map

$$g(\mathbf{x}, \mathbf{c}') = d(e(\mathbf{x}), \mathbf{c}') \tag{1}$$

implements the desired conversion: the output of $g(\mathbf{x}, \mathbf{c}')$ should match the sample $f(\mathbf{s}, \mathbf{c}')$. Our following theoretical analysis focuses on the next *conversion correctness* property.

Definition 1 (Conversion correctness) *We say g achieves* **exact conversion** *if for all* $\mathbf{s} \in \mathrm{supp}(\mathbf{S})$ *and* $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$*:* $g(f(\mathbf{s}, \mathbf{c}), \mathbf{c}') = f(\mathbf{s}, \mathbf{c}')$.

In the following section, we present five main results that characterize exact attribute conversion in autoencoder-based models: **(1)** Exact conversion is shown to require neither full disentanglement nor exact recovery of the latent variable, but only preservation of the task-relevant component up to an invertible, attribute-invariant transformation. **(2)** Any encoder–decoder pair that achieves exact conversion is characterized by two fundamental structural properties: (i) perfect reconstruction of the observed input under its original conditioning variable $\mathbf{C}$, and (ii) existence of independence between the latent variable learned by the model and the controllable attribute $\mathbf{C}$. **(3)** Exact conversion implies the existence of an internal representation whose distribution is independent of the attribute $\mathbf{C}$. **(4)** Moreover, under a minimal information bottleneck principle, both perfect reconstruction and independence from $\mathbf{C}$ are necessary properties of any model that achieves exact conversion, even in the absence of an explicitly constructed attribute-invariant encoder. **(5)** Relaxing the idealized assumptions yields explicit error bounds that quantify how reconstruction error and residual dependence translate into degradation of synthesis accuracy.

5 Theoretical Analysis of Conversion Task

5.1 Data Generation Assumptions

We state minimal regularity conditions on the assumed model under which the conversion is guaranteed.

Assumption 1 (Generative model) *The map $f : \mathbb{S} \times \mathbb{C} \to \mathbb{X}$ is **injective**.*

Similar injectivity or invertibility assumptions on the generative map are standard in the ICA literature, where they are required to ensure identifiability of the latent variables, particularly in nonlinear settings (Hyvarinen et al., 2019; Khemakhem et al., 2020). Our second assumption related to the nature of the latent variable $\mathbf{S}$.

Assumption 2 (Distributional asymmetry of S) *The only measure-preserving automorphism π : supp$(\mathbf{S})$ $\to$ supp$(\mathbf{S})$ with pushforward distribution $\pi_{\#}\mathcal{P}_{\mathbf{S}} = \mathcal{P}_{\mathbf{S}}$ is the **identity**.*

Assumption 2 plays a role analogous to the variability assumption in nonlinear ICA (Hyvarinen et al., 2019), which enforces that auxiliary variables induce a sufficiently strong and diverse effect to break

latent symmetries. While Hyvarinen et al. (2019) impose variability through changes in conditional distributions, our assumption achieves a similar symmetry-breaking effect directly at the level of the latent prior, by ruling out nontrivial measure-preserving automorphisms of $\mathbf{S}$.

Example 1 (Discrete latent variable with asymmetric law) *Let* $\mathbf{S}$ *be a discrete random variable with probability mass function* $p_{\mathbf{S}}$*. Assume that*

$$p_{\mathbf{S}}(\mathbf{s}_1) \neq p_{\mathbf{S}}(\mathbf{s}_2) \qquad \text{for all distinct } \mathbf{s}_1, \mathbf{s}_2 \in \operatorname{supp}(\mathbf{S}).$$

Then the only bijection $\pi : \operatorname{supp}(\mathbf{S}) \to \operatorname{supp}(\mathbf{S})$ *satisfying* $\pi_{\#}\mathcal{P}_{\mathbf{S}} = \mathcal{P}_{\mathbf{S}}$ *is the identity map. Consequently,* $\mathbf{S}$ *satisfies Assumption 2. In Appendix E, we present empirical observations consistent with this example.*

5.2 Sufficiency

At a high level, attribute conversion asks whether a learned encoder–decoder pair can implement the *counterfactual map* $(\mathbf{x}, \mathbf{c}') \mapsto f(\mathbf{s}, \mathbf{c}')$, given only an observed sample $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$ and a target attribute $\mathbf{c}'$. For this to be possible, the model must implicitly retain enough information about the underlying task-relevant factor $\mathbf{s}$ while discarding any dependence on the original attribute $\mathbf{c}$.

Our first key observation is that exact conversion does *not* require recovering all latent variables or fully disentangling the generative process. Instead, it suffices that the encoder preserves the task-relevant component $\mathbf{s}$ up to an invertible transformation.

We first show that Properties 1 and 2 are sufficient for correct attribute conversion. Importantly, Property 2 only constrains how the learned representation relates to the attribute variable $\mathbf{C}$; it does not require recovering ground-truth content variable $\mathbf{S}$. In Section 6 we derive an explicit error bound that quantifies how conversion degrades when reconstruction is imperfect and the representation is only approximately independent of the attribute. The proof for the next Lemma is provided in Appendix A.

Lemma 1 (Existence of a $\mathbf{C}$-invariant latent reparameterization) *Assume that Assumptions 1 and 2 hold for the generative model* $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$*, and that the model* g *satisfies Property 1 and Property 2. For each* $\mathbf{c} \in \operatorname{supp}(\mathbf{C})$*, define the map*

$$T_{\mathbf{c}} := e \circ f(\cdot, \mathbf{c}) : \operatorname{supp}(\mathbf{S}) \to \operatorname{supp}(\hat{\mathbf{S}}), \qquad \hat{\mathbf{S}} = e(\mathbf{X}).$$

Then there exists a **single bijection** $T : \operatorname{supp}(\mathbf{S}) \to \operatorname{supp}(\hat{\mathbf{S}})$ *such that*

$$T_{\mathbf{c}}(\mathbf{s}) = T(\mathbf{s}) \quad \text{for all } \mathbf{s} \in \operatorname{supp}(\mathbf{S}) \text{ and all } \mathbf{c} \in \operatorname{supp}(\mathbf{C}).$$

That is, although the learned representation $e(\mathbf{x})$ need not equal $\mathbf{s}$ itself, it must be a one-to-one reparameterization of $\mathbf{s}$ that is invariant to $\mathbf{c}$. This reflects a synthesis-oriented notion of identifiability: we only require that the information needed to generate $f(\mathbf{s}, \mathbf{c}')$ is retained.

Assuming perfect reconstruction, the encoder–decoder pair behaves like a conditional inverse of the generative process: for a fixed attribute $\mathbf{c}$, the encoder must map distinct values of $\mathbf{s}$ to distinct latent representations, otherwise reconstruction would be impossible. This implies that, for each $\mathbf{c}$, the composed map $e \circ f(\cdot, \mathbf{c})$ is injective. If, in addition, the learned latent representation is statistically independent of $\mathbf{c}$, then the distribution of $e(f(\mathbf{s}, \mathbf{c}))$ must be the same for all $\mathbf{c}$. Together, these two properties force all such maps to coincide (up to measure-preserving symmetries of $\mathbf{s}$), yielding a single latent reparameterization $T(\mathbf{s})$ that is invariant to $\mathbf{c}$.

Once such a $\mathbf{c}$-invariant bijection exists, conversion becomes straightforward: the decoder can be interpreted as first inverting this reparameterization to recover $\mathbf{s}$, and then reapplying the generative mechanism with the desired target attribute $\mathbf{c}'$. This establishes sufficiency: perfect reconstruction and independence guarantee exact conversion.

Proposition 1 (Sufficiency) *Assume that Assumptions 1, 2 hold and the model* g *satisfies the Properties 1,2. Then correct conversion (Definition 1) holds.*

The proof is provided in the Appendix B. Proposition 1 guarantees that preserving the task-relevant latent variable $\mathbf{S}$ up to an invertible transformation T, while replacing the values of the observed attribute $\mathbf{C}$, yields a valid attribute conversion of $\mathbf{X}$ to the target style.

5.3 NECESSITY

In the opposite direction, we study which structural properties are *implied* by exact conversion. We first show that any model that achieves exact conversion necessarily induces an internal representation that is invariant, in distribution, to the attribute $\mathbf{C}$. This representation can be characterized abstractly through the *orbit* of the conversion operator and exists independently of how the model is parameterized. We then show that, under a minimal information bottleneck principle, this invariance must be realized by the encoder $e(\mathbf{X})$, implying that both reconstruction and independence from $\mathbf{C}$ are necessary properties of the model achieving exact conversion.

5.3.1 NECESSITY UNDER EQUIVALENT AUTOENCODER

If exact conversion holds for all $(\mathbf{s}, \mathbf{c}, \mathbf{c}')$, then reconstruction must hold as a special case (by setting $\mathbf{c}' = \mathbf{c}$). The following proposition shows that if a model g achieves exact conversion, then it is possible to construct an equivalent model g' that admits an autoencoder-like factorization. Importantly, this factorization does not correspond to a classical encoder–decoder architecture; rather, it is an abstract representation induced by the conversion capability of g.

Proposition 2 (Necessity under Equivalent Autoencoder) *Assume that the conversion is correct, i.e.*

$$g(f(\mathbf{s}, \mathbf{c}), \mathbf{c}') = f(\mathbf{s}, \mathbf{c}') \quad \forall \mathbf{s} \in \mathrm{supp}(\mathbf{S}),\ \mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C}). \tag{2}$$

Then g admits an equivalent realization $g'(x, c) = d'(e'(x), c)$ that satisfies Properties 1, 2.

Proof 1 ***(1) Reconstruction.*** *Setting $\mathbf{c}' = \mathbf{c}$ in equation 2 yields $g(f(\mathbf{s}, \mathbf{c}), \mathbf{c}) = f(\mathbf{s}, \mathbf{c})$, which implies that the realization $g'(x, c) = d'(e'(x), c)$ defined below satisfies perfect reconstruction on $\mathrm{Im}(f)$.*

(2) Independence. *For each $\mathbf{x} \in \mathbb{X}$, define its* orbit *under g as the function*

$$\phi(\mathbf{x}) : \mathbb{C} \to \mathbb{X}, \qquad \phi(\mathbf{x})(\mathbf{c}') := g(\mathbf{x}, \mathbf{c}').$$

If $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$, then by equation 2, $\phi(f(\mathbf{s}, \mathbf{c}))(\mathbf{c}') = f(\mathbf{s}, \mathbf{c}'), \forall \mathbf{c}' \in \mathbb{C}$. Hence $\phi(f(\mathbf{s}, \mathbf{c}))$ depends only on $\mathbf{s}$. We may therefore define a map $T : \mathbb{S} \to \mathrm{Im}(\phi)$ by $T(\mathbf{s}) := \phi(f(\mathbf{s}, \mathbf{c}))$, which is independent of the choice of $\mathbf{c}$. Let $\hat{\mathbf{S}} := T(\mathbf{S})$. Since $\mathbf{S} \perp\!\!\!\perp \mathbf{C}$ and $\hat{\mathbf{S}}$ is a deterministic function of $\mathbf{S}$, the joint distribution of $(\hat{\mathbf{S}}, \mathbf{C})$ satisfies $\mathcal{P}_{\hat{\mathbf{S}},\mathbf{C}} = (T \times \mathrm{id})_{\#}(\mathcal{P}_{\mathbf{S}} \otimes \mathcal{P}_{\mathbf{C}}) = \mathcal{P}_{\hat{\mathbf{S}}} \otimes \mathcal{P}_{\mathbf{C}}$, and thus $\hat{\mathbf{S}} \perp\!\!\!\perp \mathbf{C}$. Finally, define

$$e'(\mathbf{x}) := \phi(\mathbf{x}), d'(e'(\mathbf{x}), \mathbf{c}') := e'(\mathbf{x})(\mathbf{c}').$$

By construction,

$$d'(e'(\mathbf{x}), \mathbf{c}') = g(\mathbf{x}, \mathbf{c}'),$$

so $g'(x, c) = d'(e'(x), c)$ is equivalent to g and satisfies both Property 1 and Property 2.

To ensure that the independence property is realized directly by the representation computed within g, specifically, at the internal layers that define the encoder component of g, we require an additional *minimality* assumption. Under this assumption, any functionally equivalent realization that achieves exact conversion must coincide with g itself.

5.3.2 NECESSITY UNDER INFORMATION BOTTLENECK

In Proposition 2, necessity was established for a derived model g'. We now show that, when standard *minimality* principles are imposed on the autoencoder, independence necessarily emerges for the original autoencoder g as well.

In the proof of Proposition 2, we avoid imposing additional structural assumptions on the encoder or decoder. However, in practice, autoencoder-based models are often implicitly constrained by a minimality requirement on the latent representation $\hat{\mathbf{S}}$, motivated by the Information Bottleneck (IB) principle introduced by Tishby et al. (2000). Within the IB framework, a representation is considered optimal if it preserves exactly the information required for a prescribed task, while discarding all other sources of variability. In our setting, the relevant task is exact attribute conversion of $\mathbf{X}$ given the side information $\mathbf{C}$. Imposing this minimality constraint alongside perfect reconstruction yields an

alternative necessity argument, whereby any valid solution is forced to encode only the information required for the reconstruction task and discard all irrelevant variability. For simplicity of discussion and to use entropy terminology, we assume that $\mathbf{S}, \mathbf{C}, \hat{\mathbf{S}}$ are discrete random variables[1].

Assumption 3 (Minimality of $\hat{\mathbf{S}}$) *The representation learned by the encoder $\hat{\mathbf{S}} = e(\mathbf{X})$ attains the minimal entropy, i.e.*

$$e \in \arg\min_{\tilde{e}} H(\tilde{e}(\mathbf{X})) \quad \textit{subject to} \quad \exists\, \tilde{d} \textit{ s.t. } \tilde{d}(\tilde{e}(\mathbf{x}), \mathbf{c}) = \mathbf{x} \ \ \forall\, \mathbf{x}, \mathbf{c}.$$

Since the encoder is deterministic and $\hat{\mathbf{S}}$ is discrete, minimizing $H(\hat{\mathbf{S}})$ is equivalent to minimizing $I(\hat{\mathbf{S}}; \mathbf{X})$. The minimization is taken over all deterministic encoders $\tilde{e}$ for which there exists a (possibly encoder-dependent) decoder $\tilde{d}$ achieving exact reconstruction.

Proposition 3 (Necessity under Bottleneck Information) *Assume that the model g satisfies the Assumptions 1, 3 and conversion is correct, i.e.*

$$g(f(\mathbf{s}, \mathbf{c}), \mathbf{c}') = f(\mathbf{s}, \mathbf{c}') \quad \forall \mathbf{s} \in \mathrm{supp}(\mathbf{S}), \textit{ and any } \mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C}), \tag{3}$$

then Properties 1 and 2 must hold.

In contrast to Proposition 2, which establishes necessity only up to an equivalent realization, Proposition 3 shows that under a minimal information bottleneck principle, reconstruction and independence are necessary properties of the given model g itself. The detailed proof is provided in Appendix. C

5.4 Characterization of Exact Attribute Conversion

We summarize the preceding propositions and intermediate results in the following theorem, which provides a characterization of exact attribute conversion.

Theorem 1 (Characterization of exact attribute conversion) *Assume that the generative model $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$ satisfies Assumptions 1 and 2. Then the following statements hold:*

1. *If g satisfies Properties 1 and 2, then g achieves exact conversion (Definition 1).*
2. *If g achieves exact conversion, then g satisfies Property 1, and there exists g' that is equivalent to g on $\mathrm{supp}(\mathbf{X}) \times \mathrm{supp}(\mathbf{C})$ and satisfies both Properties 1 and 2.*
3. *If g achieves exact conversion under Assumption 3, then g satisfies Properties 1 and 2.*

6 Conversion Error Bound

In practice, exact reconstruction and perfect independence in the learned representation are rarely achievable due to finite data, non-optimal model configuration, and optimization error. Consequently, attribute conversion is typically only approximate, even when the ideal conditions are satisfied in expectation under the true data distribution. In this section, we analyze how deviations from the sufficient conditions for exact conversion translate into conversion error. We derive explicit bounds on the expected conversion error that quantify the impact of reconstruction error and residual dependence in the latent space. These bounds characterize the stability of the conversion map and clarify how close one must be to the ideal conditions in order to achieve reliable conversion.

Our analysis proceeds in three steps. Motivated by prior work on Lipschitz regularity and stability of neural networks (Virmaux & Scaman, 2018; Gouk et al., 2021), we first assume that the decoder d is bounded, i.e. is L-Lipschitz with respect to its latent input. This regularity enables us to show that deviations from reconstruction accuracy and latent independence lead to controlled, bounded conversion error. We then assume uniform bounds on the reconstruction error and on the degree of dependence between the learned representation and the conditioning variable. Finally, we combine these assumptions to obtain an explicit bound on the conversion error, revealing how reconstruction accuracy and residual dependence jointly determine the quality of style conversion.

[1]This does not limit our analysis, which applies also in the continuous case, replacing entropy with the limiting density of discrete points.

Assumption 4 (L-Lipschitz decoder) *The decoder d is L-Lipschitz in its latent input. That is, there exists a constant $L \geq 0$ such for $\mathbf{x}_1 = d(\hat{\mathbf{s}}_1, \mathbf{c})$, $\mathbf{x}_2 = d(\hat{\mathbf{s}}_2, \mathbf{c})$:*

$$||\mathbf{x}_1 - \mathbf{x}_2||_2 \leq L||\hat{\mathbf{s}}_1 - \hat{\mathbf{s}}_2||_2 \qquad \forall \hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2 \in \hat{\mathbb{S}},\ \forall \mathbf{c} \in \mathbb{C}.$$

Assumption 4 imposes an L-Lipschitz condition on the decoder with respect to its latent input, which enforces stability: small perturbations in the latent space cannot produce arbitrarily large deviations in the reconstructed signal. This smoothness is crucial for generalization and for interpreting the latent space as a structured representation of content.

Assumption 5 (Uniform Reconstruction Error Bound)

$$\exists \epsilon \geq 0 : ||d(e(\mathbf{x}), \mathbf{c}) - \mathbf{x}||_2^2 \leq \epsilon, \forall \mathbf{x} \in \mathbb{X}, \textit{ s.t. } \mathbf{x} = f(\mathbf{s}, \mathbf{c}).$$

Assumption 5 establishes a uniform bound on reconstruction error, ensuring that the autoencoder approximates the data-generating process with controlled fidelity across all samples. This guarantees that the representations are informative enough to recover the inputs up to a small error ϵ.

The following assumption captures a mild relaxation of the Information Bottleneck minimality discussed in Section 5.3.2, allowing for small deviations from exact $\mathbf{C}$-invariance in the learned latent representation. It formalizes the notion of $\mathbf{C}$-invariance in the encoder: for a fixed task-relevant factor $\mathbf{s}$, changing the controllable attribute from $\mathbf{c}$ to $\mathbf{c}'$ produces latent representations that remain within distance ϵ'.

Assumption 6 ($\mathbf{C}$-invariance Uniform Bound) *For any $\mathbf{x}, \mathbf{x}'$ such that $\mathbf{x} = f(\mathbf{s}, \mathbf{c}), \mathbf{x}' = f(\mathbf{s}, \mathbf{c}')$:*

$$\exists \epsilon' \geq 0 : ||e(\mathbf{x}) - e(\mathbf{x}')||_2^2 \leq \epsilon'.$$

Together, these assumptions ensure that the representation is approximately invariant to $\mathbf{C}$ while retaining sufficient information for reconstruction, and that the decoder is stable in the sense that small perturbations in the latent space lead to controlled changes in the output.

During conversion, assuming we have a parallel validation dataset, for given two conditions $\mathbf{c}, \mathbf{c}' \in \mathbb{C}$, we aim to convert the sample $\mathbf{x}_0 \in \mathbb{X}$, by applying target condition $\mathbf{c}'$, to the target sample $\mathbf{x}' = d(e(\mathbf{x}_0), \mathbf{c}') \in \mathbb{X}$. We denote the converted sample by $\hat{\mathbf{x}}' = d(e(\mathbf{x}_0), \mathbf{c}')$. In this setup, we derive the error bound for the converted sample.

Lemma 2 (Conversion Error Bound) *Let $\epsilon_{conv} = ||\mathbf{x}' - \hat{\mathbf{x}}'||_2^2$ be a conversion error. Then*

$$\epsilon_{conv} \leq 2(L^2\epsilon' + \epsilon).$$

The detailed technical proof is provided in Appendix D. Note that the conversion error is reduced by training a model with a smoother decoder (a lower Lipschitz constant) and pushing the reconstruction and independence errors toward zero.

7 Remark on Empirical Objective

To obtain g, defined in Section 4, we consider the learning objective

$$\min_{(e,d)\in\mathcal{A}} \mathcal{L}_{\text{rec}}(e, d) + \lambda\, \mathcal{L}_{\text{dep}}(e), \qquad \lambda > 0, \tag{4}$$

where $\mathcal{L}_{\text{rec}}$ measures reconstruction accuracy and $\mathcal{L}_{\text{dep}}$ measures statistical dependence between the learned representation and the conditioning variable. Concretely, we define $\mathcal{L}_{\text{rec}}(e, d) := \mathbb{E}_{(\mathbf{X},\mathbf{C})}\Big[\mathcal{R}\big(\mathbf{X},\ d(e(\mathbf{X}), \mathbf{C})\big)\Big]$, for a discrepancy metric $\mathcal{R}$, and $\mathcal{L}_{\text{dep}}(e) := \mathbb{E}_{\mathbf{X},\mathbf{C}}\big[\mathcal{I}\big(\mathbf{C},\ e(\mathbf{X})\big)\big]$, for a dependence metric $\mathcal{I}$ satisfying $\mathcal{I}(\mathbf{U}, \mathbf{V}) = 0$ iff $\mathbf{U} \perp\!\!\!\perp \mathbf{V}$. The reconstruction term encourages $d(e(\mathbf{X}), \mathbf{C}) \approx \mathbf{X}$, while the dependence term encourages $e(\mathbf{X})$ to be invariant to $\mathbf{C}$. In practice, $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{dep}}$ are replaced by their empirical counterparts computed on N samples, leading to an empirical risk minimization problem of the same form as equation 4. Statistical conditions under which the empirical objective approximates the population objective are discussed in Section 6.

Several prior works propose different instantiations of the dependence penalty $\mathcal{L}_{\text{dep}}$ to enforce independence, typically alongside a standard reconstruction loss $\mathcal{R}$ (e.g., an ℓ_2 or ℓ_1 discrepancy in the data space). We briefly review representative examples and relate them to our setting.

Hilbert-Schmidt Independence Criterion Ma et al. (2020) proposed to use an empirical estimate of the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) objective, which measures the statistical dependence between two random variables using kernel methods. Hilbert-Schmidt norm of the cross-covariance operator between the distributions in the Reproducing Kernel Hilbert Space (RKHS) defined by

$$(N-1)^{-2}\text{tr}(\mathbf{K}_{\hat{\mathbf{S}}}\mathbf{H}\mathbf{K}_{\mathbf{C}}\mathbf{H}), \tag{5}$$

where $\mathbf{K}_{\hat{\mathbf{S}}}, \mathbf{K}_{\mathbf{C}} \in \mathbb{R}^{N\times N}$ are kernel matrices computed over the set of variables $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N\}$ using a positive-definite kernel function (e.g., Gaussian/RBF), $\mathbf{H} \in \mathbb{R}^{N\times N}$ is the centering matrix defined by $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ to ensure that the kernels are computed on zero-mean features and N is the number of samples. Intuitively, HSIC measures the covariance between features in two RKHSs induced by the kernels on $\hat{\mathbf{S}}$ and $\mathbf{C}$. If the variables are statistically independent, the cross-covariance operator vanishes, and thus the objective in equation 5 approaches zero.

Contrastive Log-ratio Upper Bound Cheng et al. (2020) proposed a contrastive log-ratio upper bound (vCLUB) of mutual information for a variational autoencoder, which is trained with the next objective function:

$$\min_{\theta}\left[\mathbb{E}_{\hat{\mathbf{S}},\mathbf{C}}[\log q_\theta(\mathbf{C}|\hat{\mathbf{S}})] - \mathbb{E}_{\hat{\mathbf{S}}}\mathbb{E}_{\mathbf{C}}[\log q_\theta(\mathbf{C}|\hat{\mathbf{S}})]\right],$$

where $q_\theta(\mathbf{C}|\hat{\mathbf{S}})$ is a variational distribution parametrized by θ that approximates $\mathcal{P}(\mathbf{C}|\hat{\mathbf{S}})$. The intuition behind the vCLUB objective is to estimate the mutual information between two variables by contrasting how well a variational model can predict the true paired samples versus mismatched (independent) samples. The first term encourages a high likelihood of true pairs $(\mathbf{S}, \mathbf{C})$. In contrast, the second term penalizes a high likelihood of randomly paired $\mathbf{S}$ and $\mathbf{C}$, sampled independently from their marginals. The difference between these two expectations provides an upper bound on the mutual information, which can be minimized to encourage independence between the variables.

Adversarial Independence Another approach was proposed by Shaham et al. (2022) where the model is trained in an adversarial way, and a discriminator $\delta(e(\mathbf{x}))$ is trained to predict a condition $\mathbf{c}$ from the latent $\hat{\mathbf{s}} = e(\mathbf{x})$ by maximizing the objective $-\mathcal{I}(\delta(\hat{\mathbf{s}}), \mathbf{c})$, e.g., minimizing cross entropy loss term $\sum_{i=1}^{K} c_i \log(\delta(\hat{\mathbf{s}_\mathbf{i}}))$ between condition $\mathbf{c}$ and discriminator prediction $\delta(\hat{\mathbf{s}})$. Following the logic of the Hirschfeld–Gebelein–Rényi (HGR) theorem, the autoencoder is trained to confuse the discriminator by minimizing $-\mathcal{I}(\delta(e(\mathbf{x})), \mathbf{c})$. Assuming a problem with a single conditional source $\mathbf{C}$, the training objective becomes:

$$\min_{e,d}\max_{\delta}[\mathcal{R}(\mathbf{x}, d(\hat{\mathbf{s}}, \mathbf{c})) - \lambda\mathcal{I}(\delta(\hat{\mathbf{s}}, \mathbf{c})].$$

8 Related Work

Our theoretical framework provides a principled lens for controllable generation, identifying when encoder–decoder models can provably implement attribute conversion and how errors propagate when the ideal assumptions are only approximately met. Unlike much of the prior theory that targets latent recovery or identifiability for its own sake, our guarantees are directly aligned with the synthesis objective: producing the correct transformed output while preserving task-relevant information. This perspective yields crisp necessary-and-sufficient conditions, clarifies the fundamental limits of conversion, and offers interpretable guidance on what training objectives and architectural choices must enforce. In the related work below, we position these guarantees relative to four main domains: theoretical foundations of nonlinear ICA and identifiability, deterministic (non-probabilistic) autoencoders, VAE-based identifiable disentanglement, and practical style-conversion methodologies in speech and vision.

Theoretical Foundations of Nonlinear ICA Nonlinear ICA historically faced an "impossibility" phenomenon: without additional structure, the latent sources are unidentifiable, since infinitely many nonlinear transforms can yield independent components that are arbitrary mixtures of the true

sources (Hyvärinen & Pajunen, 1999). Recent theory restores identifiability by incorporating auxiliary structure, e.g., conditioning on time indices or class labels (Hyvarinen et al., 2019; Khemakhem et al., 2020), exploiting non-stationarity or temporal dependencies (Hälvä & Hyvarinen, 2020), or restricting the mixing map via geometric constraints (Hälvä et al., 2021; Horan et al., 2021). Our work is inspired by these principles but shifts the emphasis from *source recovery* to *interventional correctness*: we characterize when an encoder–decoder model provably implements the conversion map $f(\mathbf{s}, \mathbf{c}) \mapsto f(\mathbf{s}, \mathbf{c}')$, and how conversion degrades when the ideal conditions are only approximately met.

Non-probabilistic Autoencoders Deterministic autoencoders are widely used for high-fidelity synthesis (e.g., voice conversion, image-to-image translation), where VAE stochasticity can blur fine structure (Wang et al., 2024). Empirically, many methods encourage the latent code $e(\mathbf{X})$ to be independent of the attribute $\mathbf{C}$ via mutual-information penalties, adversarial objectives, or invariance criteria (John et al., 2019; Qian et al., 2019); existing theory in deterministic settings typically targets latent recovery (Shaham et al., 2022). In contrast, we provide task-aligned guarantees: we show that **Properties 1 and 2** are sufficient for correct conversion, without requiring recovery of a designated ground-truth content variable, and we derive explicit error bounds when these properties hold only approximately.

Variational Autoencoder (VAE) Based Works VAEs offer a natural route to identifiable disentanglement. In particular, iVAE connects VAEs to nonlinear ICA by conditioning the latent prior on an auxiliary variable and proves identifiability under suitable distributional assumptions (e.g., exponential-family priors) (Khemakhem et al., 2020). Heuristic variants such as β-VAE (Higgins et al., 2017) and β-TCVAE (Chen et al., 2018) encourage factorization through stronger regularization but generally lack rigorous identifiability guarantees. Our perspective is complementary: rather than focusing on latent identification, we study deterministic encoder–decoder conversion through the lens of correctness and robustness of the induced attribute conversion.

Style Conversion Methodologies Style conversion aims to implement the transformation $f(\mathbf{s}, \mathbf{c}) \to f(\mathbf{s}, \mathbf{c}')$. In the **vision** domain, several recent high-performing systems adopt an autoencoder factorization of content and style. For example, CSGO (Xing et al., 2025) employs separate content and style encoders trained jointly, while leveraging a pre-trained diffusion-based decoder for high-fidelity synthesis; related autoencoder-based designs have also been explored in more recent frameworks (Lei et al., 2025). In the **speech** domain, Zhang et al. (2025) propose a VQ-VAE–based conversion model, aligning naturally with the encoder–decoder paradigm through discrete content representations. Other works explicitly target content-style separation: Zhu et al. (2025) integrate an autoencoder objective with prompt-based style extraction while disentangling speaker timbre in the prompt encoder to reduce identity leakage, and Kim et al. (2025) combine an encoder–decoder conversion pipeline with codebook-based content representations designed to be insensitive to speaker style. In the **text** domain, Dai et al. (2019) introduce a Style Transformer trained with reconstruction and adversarial objectives that encourage separation between content and style. More recently, Kocak et al. (2025) apply an autoencoder on top of a pretrained language model to extract content representations and then debias harmful attributes (or exploit beneficial correlations) according to the downstream task.

9 Conclusion

This work develops a synthesis-oriented theory of attribute conversion for deterministic encoder–decoder models under the generative assumption $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$. We provide a sharp characterization of *when* exact conversion is achievable: perfect reconstruction and independence between the learned representation and the controllable attribute are sufficient for exact conversion, and are necessary up to an equivalent representation for arbitrary models, and for the given model itself under a minimal information bottleneck principle. These conditions formalize the functional roles of the encoder and decoder, ensuring that the task-relevant factor is preserved up to an invertible reparameterization and can be consistently recombined with a target attribute, and they clarify the fundamental limits of controllable generation beyond heuristic architectural choices. Moving beyond the idealized regime, we further derive explicit conversion error bounds that quantify how reconstruction imperfections and residual dependence translate into degradation of conversion quality. Together, our results connect common training objectives to provable task-level guarantees and offer principled guidance for the design of reliable attribute conversion systems.

AI USE STATEMENT

In this work, we used generative AI tools only to improve the writing style, clarity, and grammar of the manuscript. We did not use generative AI tools to formulate scientific claims, derive theoretical results, construct proofs, analyze experimental results, or draw conclusions. All scientific claims, mathematical derivations, proofs, experimental design, analyses, and conclusions are the authors' own work.

We have reviewed all AI-assisted edits and verified that they preserve the intended technical meaning and do not introduce unsupported statements or modifications to the scientific content. We take full responsibility for the final content of this work, including any text revised with the aid of generative AI.

REFERENCES

Dhananjay Ashok and Barnabas Poczos. Controllable text generation in the instruction-tuning era. *arXiv preprint arXiv:2405.01490*, 2024.

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.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuan-Jing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5997–6007, 2019.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

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.

Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34:5150–5161, 2021.

Kurt Hornik. Some new results on neural network approximation. *Neural networks*, 6(8):1069–1072, 1993.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. Pmlr, 2019.

Haoran Jin. Controllable generation with disentangled representative learning of multiple perspectives in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 388–395, October 2025.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 424–434, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Jaehun Kim, Ji-Hoon Kim, Yeunju Choi, Tan Dat Nguyen, Seongkyu Mun, and Joon Son Chung. Adaptvc: High quality voice conversion with adaptive learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Aysenur Kocak, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. Cure: Controlled unlearning for robust embeddings-mitigating conceptual shortcuts in pre-trained language models. In *EMNLP (Findings)*, pp. 14283–14297, 2025.

Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23443–23452. IEEE, 2025.

Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. Controlled text generation for large language model with dynamic attribute graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5797–5814, 2024.

Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5085–5092, 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Tianhua Qi, Shiyan Wang, Cheng Lu, Tengfei Song, Hao Yang, Zhanglin Wu, and Wenming Zheng. PromptEVC: Controllable Emotional Voice Conversion with Natural Language Prompts. In *Interspeech 2025*, pp. 4588–4592, 2025. doi: 10.21437/Interspeech.2025-948.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pp. 5210–5219. PMLR, 2019.

Uri Shaham, Jonathan Svirsky, Ori Katz, and Ronen Talmon. Discovery of single independent latent variable. *Advances in Neural Information Processing Systems*, 35:25251–25263, 2022.

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

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in neural information processing systems*, 31, 2018.

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17979–17989. IEEE, 2023.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696, 2024.

Peng Xing, Haofan Wang, Yanpeng Sun, wangqixun, Baixu, Hao Ai, Jen-Yuan Huang, and Zechao Li. CSGO: Content-style composition in text-to-image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=n4nmiIq3qj.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *International Conference on Learning Representations*, volume 2025, pp. 58436–58459, 2025.

Xinfa Zhu, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. Zsvc: Zero-shot style voice conversion with disentangled latent diffusion models and adversarial training. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

A Proof of Lemma 1

Proof 2 *We first show that T_c is injective. Fix $\mathbf{c}$. Suppose $T_c(\mathbf{s}_1) = T_c(\mathbf{s}_2)$. Using Property 1 and Assumption 1, $f(\mathbf{s}_1, \mathbf{c}) = d(T_c(\mathbf{s}_1), \mathbf{c}) = d(T_c(\mathbf{s}_2), \mathbf{c}) = f(\mathbf{s}_2, \mathbf{c}) \Rightarrow \mathbf{s}_1 = \mathbf{s}_2$, so T_c is injective. By Property 2, the distribution of $\hat{\mathbf{S}} = e(\mathbf{X})$ does not depend on $\mathbf{C}$, hence for any $\mathbf{C} = \mathbf{c}$ it yields:*

$$(T_c)_{\#}\mathcal{P}_{\mathbf{S}} = \mathcal{P}_{e(f(\mathbf{S},\mathbf{C}=\mathbf{c}))} = \mathcal{P}_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}} = \mathcal{P}_{\hat{\mathbf{S}}},$$

which is independent of $\mathbf{C}$. Hence all $(T_c)_{\#}\mathcal{P}_{\mathbf{S}}$ coincide. To show uniqueness of T, assume we are given a distinct pair of samples $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$. Since T_c and $T_{c'}$ are both injective maps from $\mathrm{supp}(\mathbf{S})$ into $\mathrm{supp}(\hat{\mathbf{S}})$, and they induce the same distribution *of $\hat{\mathbf{S}}$: $(T_c)_{\#}\mathcal{P}_{\mathbf{S}} = \mathcal{P}_{\hat{\mathbf{S}}} = (T_{c'})_{\#}\mathcal{P}_{\mathbf{S}}$, hence, the images of T_c and $T_{c'}$ coincide up to $\mathcal{P}_{\mathbf{S}}$-null sets, and there exists a bijection*

$$\pi_{c,c'} := T_{c'}^{-1} \circ T_c : \mathrm{supp}(\mathbf{S}) \to \mathrm{supp}(\mathbf{S})$$

that satisfies $(\pi_{c,c'})_{\#}\mathcal{P}_{\mathbf{S}} = \mathcal{P}_{\mathbf{S}}$, $T_c = T_{c'} \circ \pi_{c,c'}$. That is, each T_c differs from $T_{c'}$ only by a measure-preserving bijection $\pi_{c,c'}$ of the variable $\mathbf{S}$. By Assumption 2, every measure-preserving automorphism of $\mathcal{P}_{\mathbf{S}}$ is the identity. Thus $\pi_{c,c'}$ is identity for all $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(C)$, and $T_c = T_{c'}$ for all $\mathbf{c}, \mathbf{c}'$. Define $T := T_c$. Now we only need to show the ***surjectivity*** *of T onto $\mathrm{supp}(\hat{\mathbf{S}})$. Let $\hat{\mathbf{s}} \in \mathrm{supp}(\hat{\mathbf{S}})$ be arbitrary. By definition of $\hat{\mathbf{S}} = e(\mathbf{X})$, there exists $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$ and some $\mathbf{c}$ such that $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$ and $\hat{\mathbf{s}} = e(\mathbf{x})$. Hence, there exists $\mathbf{s} \in \mathrm{supp}(\mathbf{S})$ satisfying $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$. Applying e to both sides and using the definition of T gives*

$$\hat{\mathbf{s}} = e(\mathbf{x}) = e(f(\mathbf{s}, \mathbf{c})) = T(\mathbf{s}).$$

Therefore, every $\hat{\mathbf{s}} \in \mathrm{supp}(\hat{\mathbf{S}})$ can be written as $\hat{\mathbf{s}} = T(\mathbf{s})$ for some $\mathbf{s}$, implying $\mathrm{Im}(T) = \mathrm{supp}(\hat{\mathbf{S}})$. Since T is already known to be injective, it follows that $T : \mathrm{supp}(\mathbf{S}) \to \mathrm{supp}(\hat{S})$ is a ***bijection****.*

B Proof of Proposition 1

Proof 3 *Let $T : \mathbb{S} \to \hat{\mathbb{S}}$ be the bijection established in Lemma 1. Since T is a bijection between $\mathbb{S}$ and $\hat{\mathbb{S}}$, its inverse T^{-1} exists. By Property 1, for every $\mathbf{s}$ and $\mathbf{c}$ we have $f(\mathbf{s}, \mathbf{c}) = d(e(f(\mathbf{s}, \mathbf{c})), \mathbf{c})$. Substituting $e(f(\mathbf{s}, \mathbf{c})) = T(\mathbf{s})$ (Lemma 1) yields $f(\mathbf{s}, \mathbf{c}) = d(T(\mathbf{s}), \mathbf{c})$. Because T is bijective, for any $\hat{\mathbf{s}} = T(\mathbf{s})$ we can rewrite this equality as $d(\hat{\mathbf{s}}, \mathbf{c}) = f(T^{-1}(\hat{\mathbf{s}}), \mathbf{c})$, which holds for all $\hat{\mathbf{s}}$ and $\mathbf{c}$. Finally, for any $(\mathbf{s}, \mathbf{c}, \mathbf{c}')$:*

$$d(e(f(\mathbf{s}, \mathbf{c})), \mathbf{c}') = d(T(\mathbf{s}), \mathbf{c}') = f(T^{-1}(T(\mathbf{s})), \mathbf{c}') = f(\mathbf{s}, \mathbf{c}'),$$

proving the conversion property by Definition 1.

C Proof of Proposition 3

Proof 4 ***(1)*** *Property 1 follows from correct conversion, as in the proof of Proposition 2 (by setting $\mathbf{c}' = \mathbf{c}$ in equation 3).* ***(2)*** *Since f is injective (Assumption 1), there exists a left-inverse $h : \mathbb{X} \to \mathbb{S} \times \mathbb{C}$ defined on $\mathrm{Im}(f)$ such that $h(f(\mathbf{s}, \mathbf{c})) = (\mathbf{s}, \mathbf{c})$ for all $(\mathbf{s}, \mathbf{c}) \in \mathbb{S} \times \mathbb{C}$. Under the generative model $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$, we have $\mathbf{X} \in \mathrm{Im}(f)$, hence $(\mathbf{S}, \mathbf{C}) = h(\mathbf{X})$. Using Property 1, i.e. $\mathbf{X} = d(\hat{\mathbf{S}}, \mathbf{C})$, we obtain $(\mathbf{S}, \mathbf{C}) = h(\mathbf{X}) = h\big(d(\hat{\mathbf{S}}, \mathbf{C})\big)$. Taking the first coordinate yields that there exists a deterministic map η such that*

$$\mathbf{S} = \eta(\hat{\mathbf{S}}, \mathbf{C}) \tag{6}$$

and therefore $H(\mathbf{S} \mid \hat{\mathbf{S}}, \mathbf{C}) = 0$. By combining the definition of mutual information and equation 6, $I(\mathbf{S}; \hat{\mathbf{S}} \mid \mathbf{C}) = H(\mathbf{S} \mid \mathbf{C}) - H(\mathbf{S} \mid \hat{\mathbf{S}}, \mathbf{C}) = H(\mathbf{S} \mid \mathbf{C})$. Since $\mathbf{S} \perp\!\!\!\perp \mathbf{C}$, $H(\mathbf{S} \mid \mathbf{C}) = H(\mathbf{S})$, hence $I(\mathbf{S}; \hat{\mathbf{S}} \mid \mathbf{C}) = H(\mathbf{S})$. On the other hand, for discrete random variables, $I(\mathbf{S}; \hat{\mathbf{S}} \mid \mathbf{C}) \leq H(\hat{\mathbf{S}} \mid \mathbf{C}) \leq H(\hat{\mathbf{S}})$, which implies $H(\hat{\mathbf{S}}) \geq H(\mathbf{S})$. Consider the candidate representation $\tilde{\mathbf{S}} := \mathbf{S}$ and define $\tilde{d}(\tilde{\mathbf{s}}, \mathbf{c}) := f(\tilde{\mathbf{s}}, \mathbf{c})$. Then $\tilde{d}(\tilde{\mathbf{S}}, \mathbf{C}) = f(\mathbf{S}, \mathbf{C}) = \mathbf{X}$. Thus $\tilde{\mathbf{S}}$ is feasible for Assumption 3 and satisfies $H(\tilde{\mathbf{S}}) = H(\mathbf{S})$. By minimality (Assumption 3), $H(\hat{\mathbf{S}}) \leq H(\tilde{\mathbf{S}}) = H(\mathbf{S})$. Together with $H(\hat{\mathbf{S}}) \geq H(\mathbf{S})$ we conclude that $H(\hat{\mathbf{S}}) = H(\mathbf{S})$. Finally, since

$$H(\mathbf{S}) = I(\mathbf{S}; \hat{\mathbf{S}} \mid \mathbf{C}) \leq H(\hat{\mathbf{S}} \mid \mathbf{C}) \leq H(\hat{\mathbf{S}}) = H(\mathbf{S}),$$

all inequalities are tight. In particular, $H(\hat{\mathbf{S}} \mid \mathbf{C}) = H(\hat{\mathbf{S}})$, which for discrete random variables is equivalent to $\hat{\mathbf{S}} \perp\!\!\!\perp \mathbf{C}$.

D PROOF OF LEMMA 2

Proof 5 *Let $\mathbf{x}_0 \in \mathbb{X}$ be a source sample with attribute $\mathbf{c}$, and let $\mathbf{x}' \in \mathbb{X}$ denote the corresponding target sample with the same underlying $\mathbf{s}$ and attribute $\mathbf{c}'$ (i.e., $\mathbf{x}_0 = f(\mathbf{s}, \mathbf{c})$ and $\mathbf{x}' = f(\mathbf{s}, \mathbf{c}')$). Define the model's converted output*

$$\hat{\mathbf{x}}' \;:=\; d(e(\mathbf{x}_0), \mathbf{c}').$$

Then the conversion error is

$$\epsilon_{conv} \;=\; \|\mathbf{x}' - \hat{\mathbf{x}}'\|_2^2 \;=\; \|\mathbf{x}' - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2. \tag{7}$$

By Assumption 5 applied to the target sample $\mathbf{x}'$ (which lies in $\mathbb{X}$), we have

$$\|d(e(\mathbf{x}'), \mathbf{c}') - \mathbf{x}'\|_2^2 \le \epsilon.$$

Define the reconstruction residual

$$\mathbf{r} \;:=\; \mathbf{x}' - d(e(\mathbf{x}'), \mathbf{c}'), \qquad \textit{so that} \qquad \|\mathbf{r}\|_2^2 \le \epsilon \;\textit{ and }\; \mathbf{x}' = d(e(\mathbf{x}'), \mathbf{c}') + \mathbf{r}.$$

Substituting into equation 7 yields

$$\epsilon_{conv} = \|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}') + \mathbf{r}\|_2^2.$$

Using $\|a+b\|_2^2 \le 2(\|a\|_2^2 + \|b\|_2^2)$, we obtain

$$\epsilon_{conv} \le 2\Big(\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2 + \|\mathbf{r}\|_2^2\Big) \le 2\Big(\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2 + \epsilon\Big).$$

By Assumption 4, the decoder $d(\cdot, \mathbf{c}')$ is L-Lipschitz in its latent argument, hence

$$\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2 \le L\|e(\mathbf{x}') - e(\mathbf{x}_0)\|_2,$$

and therefore

$$\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2 \le L^2\|e(\mathbf{x}') - e(\mathbf{x}_0)\|_2^2.$$

Substituting this bound gives

$$\epsilon_{conv} \le 2\Big(L^2\|e(\mathbf{x}') - e(\mathbf{x}_0)\|_2^2 + \epsilon\Big).$$

Finally, by Assumption 6 (the $\mathbf{c}$-invariance uniform bound),

$$\|e(\mathbf{x}') - e(\mathbf{x}_0)\|_2^2 \le \epsilon',$$

and we conclude that

$$\epsilon_{conv} \;\le\; 2\big(L^2\epsilon' + \epsilon\big).$$

E EXAMPLE OF $\mathbf{S}$ WITH ASYMMETRIC LAW

To provide empirical evidence that Assumption 2 holds in real-world data, we analyze the variable $\tilde{\mathbf{S}}$, that approximates $\mathbf{S}$, constructed from the K-means labels of samples $\{\mathbf{x}_i\}_{i=1}^N$. The samples are obtained from the self-supervised pre-trained speech model WavLM (Chen et al., 2022) as continuous embeddings of the input waveform samples (audios from the LibriSpeech-clean-100 dataset (Panayotov et al., 2015)). We set $K = 100$ and estimate the prior probability of each category by measuring the empirical frequency of the corresponding label in the training dataset,

$$p_k \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\tilde{s}_i = k\}, \qquad k \in \{1, \dots, K\}.$$

From the resulting probability vector $\mathbf{p} = [p_1, p_2, \dots, p_K]$, we compute the pairwise square-root l_1 distance matrix $\mathbf{D} \;=\; \big|\,\mathbf{p}\mathbf{1}^\top - \mathbf{1}\mathbf{p}^\top\,\big|^{\frac{1}{2}}$, where $\mathbf{1} \in \mathbb{R}^K$ denotes the all-ones column vector.

Figure 1 visualizes the matrix $\mathbf{D}$. All off-diagonal entries are non-zero and clearly separated from the diagonal, indicating that the empirical prior probabilities of the speech unit categories are distinct. Since the unobserved content variable $\mathbf{S}$ is approximated by the proxy $\tilde{\mathbf{S}}$, which captures phonetic structure, this analysis provides empirical evidence that Assumption 2 is approximately satisfied across real-world speech datasets.

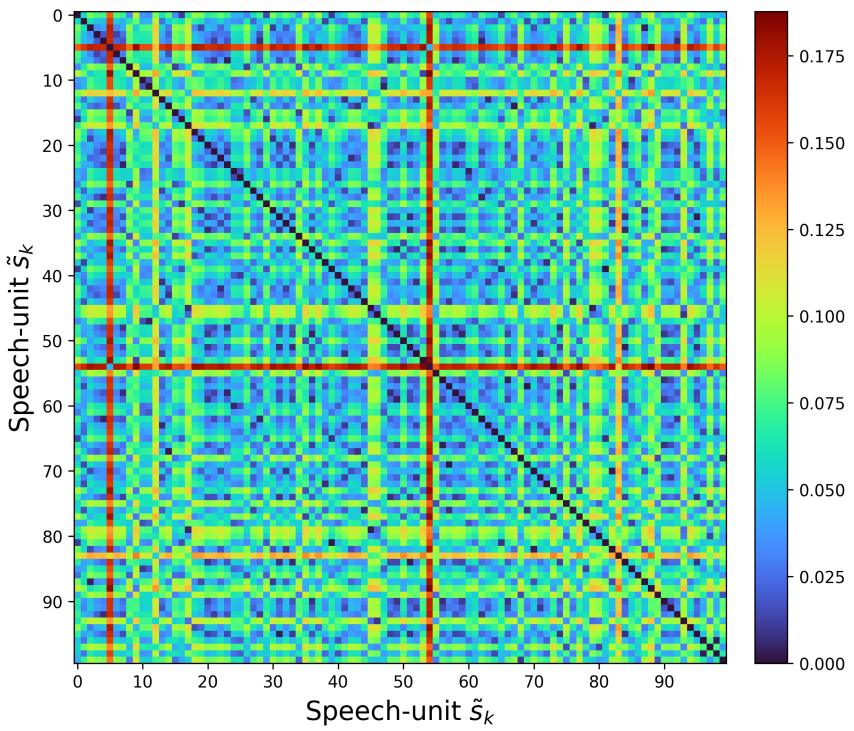


Figure 1: Empirical analysis supporting Assumption 2. We compute the pairwise distance matrix $\mathbf{D}$ over the empirical prior probabilities p_k of the speech unit categories $\tilde{s}_k$, for $k \in \{1, \ldots, K\}$. Since $\tilde{\mathbf{S}}$ serves as a proxy for the latent content variable $\mathbf{S}$ and $\mathbf{D}_{ii} = 0$ for all i while $\mathbf{D}_{ij} \neq 0$ for all $i \neq j$, the results indicate non-uniform category priors in real-world speech data.

F Limitations

As with most theoretical analyses, our results rely on idealized assumptions such as perfect optimization and convergence, as well as structural conditions on the data-generating process. While these assumptions may not hold exactly in practice, our framework explicitly accounts for such deviations by quantifying how reconstruction errors and residual dependence affect the quality of attribute conversion. Nevertheless, extending the analysis to more realistic settings, including stochastic generation and imperfect independence, remains an important direction for future work.

G Extension to Stochastic Generation

We now consider a stochastic generative process represented through an exogenous random variable $\mathbf{Z}$:

$$\mathbf{X} = f(\mathbf{S}, \mathbf{C}, \mathbf{Z}), \qquad \mathbf{Z} \perp\!\!\!\perp (\mathbf{S}, \mathbf{C}),$$

where $\mathbf{Z}$ has distribution $\mathcal{P}_{\mathbf{Z}}$ and finite second moment. For fixed $\mathbf{s} \in \mathrm{supp}(\mathbf{S})$ and $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$, define

$$\mathbf{x}_0(\mathbf{z}) := f(\mathbf{s}, \mathbf{c}, \mathbf{z}), \qquad \mathbf{x}'(\mathbf{z}) := f(\mathbf{s}, \mathbf{c}', \mathbf{z}),$$

and

$$\hat{\mathbf{x}}'(\mathbf{z}) := d(e(\mathbf{x}_0(\mathbf{z})), \mathbf{c}').$$

The use of the same $\mathbf{z}$ in $\mathbf{x}_0(\mathbf{z})$ and $\mathbf{x}'(\mathbf{z})$ defines a coupling between the source and target distributions. This coupling is used only for the theoretical analysis and does not require paired stochastic realizations in the observed data.

Assumption 7 (Expected stochastic reconstruction error) *There exists* $\epsilon \geq 0$ *such that*

$$\mathbb{E}_{\mathbf{Z}}\left[\|d(e(\mathbf{x}'(\mathbf{Z})), \mathbf{c}') - \mathbf{x}'(\mathbf{Z})\|_2^2\right] \leq \epsilon$$

for every $\mathbf{s} \in \mathrm{supp}(\mathbf{S})$ *and* $\mathbf{c}' \in \mathrm{supp}(\mathbf{C})$.

Assumption 8 (Expected stochastic C-invariance) *There exists* $\epsilon' \geq 0$ *such that*

$$\mathbb{E}_{\mathbf{Z}}\left[\|e(\mathbf{x}_0(\mathbf{Z})) - e(\mathbf{x}'(\mathbf{Z}))\|_2^2\right] \leq \epsilon'$$

for every $\mathbf{s} \in \mathrm{supp}(\mathbf{S})$ *and* $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$.

Lemma 3 (Stochastic Conversion Error Bound) *Suppose that Assumption 4 and Assumptions 7 and 8 hold. Then*

$$\mathbb{E}_{\mathbf{Z}}\left[\|\mathbf{x}'(\mathbf{Z}) - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] \leq 2\left(L^2\epsilon' + \epsilon\right).$$

Moreover, let

$$\mathcal{P}^{\star}_{\mathbf{s},\mathbf{c}'} := \mathbf{x}'_{\#}\mathcal{P}_{\mathbf{Z}}$$

denote the true target distribution and let

$$\widehat{\mathcal{P}}_{\mathbf{s},\mathbf{c}\to\mathbf{c}'} := \hat{\mathbf{x}}'_{\#}\mathcal{P}_{\mathbf{Z}}$$

denote the converted distribution. Then

$$W_2^2\left(\widehat{\mathcal{P}}_{\mathbf{s},\mathbf{c}\to\mathbf{c}'}, \mathcal{P}^{\star}_{\mathbf{s},\mathbf{c}'}\right) \leq 2\left(L^2\epsilon' + \epsilon\right),$$

where W_2 *denotes the* 2*-Wasserstein distance.*

Proof 6 *For a realization* $\mathbf{z}$*, the reconstruction is defined by*

$$\tilde{\mathbf{x}}'(\mathbf{z}) := d(e(\mathbf{x}'(\mathbf{z})), \mathbf{c}').$$

Then

$$\begin{aligned}
&\|\mathbf{x}'(\mathbf{z}) - \hat{\mathbf{x}}'(\mathbf{z})\|_2^2 \\
&\quad = \|\mathbf{x}'(\mathbf{z}) - \tilde{\mathbf{x}}'(\mathbf{z}) + \tilde{\mathbf{x}}'(\mathbf{z}) - \hat{\mathbf{x}}'(\mathbf{z})\|_2^2 \\
&\quad \leq 2\|\mathbf{x}'(\mathbf{z}) - \tilde{\mathbf{x}}'(\mathbf{z})\|_2^2 + 2\|\tilde{\mathbf{x}}'(\mathbf{z}) - \hat{\mathbf{x}}'(\mathbf{z})\|_2^2.
\end{aligned}$$

By Assumption 4,

$$\|\tilde{\mathbf{x}}'(\mathbf{z}) - \hat{\mathbf{x}}'(\mathbf{z})\|_2^2 \leq L^2\|e(\mathbf{x}'(\mathbf{z})) - e(\mathbf{x}_0(\mathbf{z}))\|_2^2.$$

Taking expectation over $\mathbf{Z}$ *and applying Assumptions 7 and 8 gives*

$$\mathbb{E}_{\mathbf{Z}}\left[\|\mathbf{x}'(\mathbf{Z}) - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] \leq 2\left(\epsilon + L^2\epsilon'\right).$$

The joint random pair

$$(\hat{\mathbf{x}}'(\mathbf{Z}), \mathbf{x}'(\mathbf{Z}))$$

has marginals $\widehat{\mathcal{P}}_{\mathbf{s},\mathbf{c}\to\mathbf{c}'}$ *and* $\mathcal{P}^{\star}_{\mathbf{s},\mathbf{c}'}$*. Since the squared* 2*-Wasserstein distance is the infimum of the expected squared distance over all couplings* $\Pi\left(\widehat{\mathcal{P}}_{\mathbf{s},\mathbf{c}\to\mathbf{c}'}, \mathcal{P}^{\star}_{\mathbf{s},\mathbf{c}'}\right)$*,*

$$W_2^2\left(\widehat{\mathcal{P}}_{\mathbf{s},\mathbf{c}\to\mathbf{c}'}, \mathcal{P}^{\star}_{\mathbf{s},\mathbf{c}'}\right) \leq \mathbb{E}_{\mathbf{Z}}\left[\|\mathbf{x}'(\mathbf{Z}) - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right],$$

which proves the result.

The preceding result evaluates conversion under a shared-noise coupling, where the source sample, the target sample, and the converted output are associated with the same realization of the exogenous variable $\mathbf{Z}$. In practice, however, a reference target may be generated using an independent realization $\mathbf{Z}'$. In this case, the converted output and the reference target share the same task-relevant factor $\mathbf{s}$ and target attribute $\mathbf{c}'$, but may differ because of intrinsic stochastic variability. Consequently, their expected squared distance reflects not only the model's conversion error, but also the discrepancy between two independent valid samples from the target distribution. The following Lemma makes this additional irreducible term explicit.

Lemma 4 (Error against an independently sampled target) *Let* $\mathbf{Z}$ *and* $\mathbf{Z}'$ *be independent random variables with distribution* $\mathcal{P}_{\mathbf{Z}}$*, and define*

$$\mathbf{x}'_{\text{ind}} := f(\mathbf{s}, \mathbf{c}', \mathbf{Z}').$$

Let

$$\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} := \mathbb{E}_{\mathbf{Z}}\left[f(\mathbf{s}, \mathbf{c}', \mathbf{Z})\right]$$

and define the intrinsic stochastic variance

$$\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}') := \mathbb{E}_{\mathbf{Z}}\left[\|f(\mathbf{s}, \mathbf{c}', \mathbf{Z}) - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'}\|_2^2\right].$$

Writing

$$M := L^2\epsilon' + \epsilon,$$

the expected squared error against an independently sampled target satisfies

$$\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] \le 2\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + 2M + 2\sqrt{2\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}')M}.$$

Proof 7 *Because* $\mathbf{Z}'$ *is independent of* $\mathbf{Z}$*,* $\mathbf{x}'_{\text{ind}}$ *is conditionally independent of* $\hat{\mathbf{x}}'(\mathbf{Z})$ *given* $(\mathbf{s}, \mathbf{c}')$*. Therefore,*

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] = \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} + \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] = \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'}\|_2^2\right] + \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] + \\
&\quad + 2\,\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\langle \mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'},\, \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\rangle\right].
\end{aligned}$$

Since

$$\mathbf{x}'_{\text{ind}} = f(\mathbf{s}, \mathbf{c}', \mathbf{Z}')$$

depends only on $\mathbf{Z}'$*, whereas* $\hat{\mathbf{x}}'(\mathbf{Z})$ *depends only on* $\mathbf{Z}$*, and* $\mathbf{Z}' \perp\!\!\!\perp \mathbf{Z}$*, the two terms inside the inner product are independent. Moreover,*

$$\mathbb{E}_{\mathbf{Z}'}\left[\mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'}\right] = \mathbf{0}.$$

Therefore,

$$\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\langle \mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'},\, \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\rangle\right] = \left\langle \mathbb{E}_{\mathbf{Z}'}\left[\mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'}\right],\, \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\right]\right\rangle = 0.$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] = \mathbb{E}_{\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'}\|_2^2\right] + \mathbb{E}_{\mathbf{Z}}\left[\|\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] = \\
&= \sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + \mathbb{E}_{\mathbf{Z}}\left[\|\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right].
\end{aligned}$$

By Minkowski's inequality,

$$\begin{aligned}
&\left(\mathbb{E}\left[\|\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right]\right)^{1/2} \le \left(\mathbb{E}\left[\|\boldsymbol{\mu}_{\mathbf{s},\mathbf{c}'} - \mathbf{x}'(\mathbf{Z})\|_2^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\mathbf{x}'(\mathbf{Z}) - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right]\right)^{1/2} \le \\
&\le \sigma_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + \sqrt{2M}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\|\mathbf{x}'_{\text{ind}} - \hat{\mathbf{x}}'(\mathbf{Z})\|_2^2\right] \le \sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + \left(\sigma_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + \sqrt{2M}\right)^2 = \\
&= 2\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}') + 2M + 2\sqrt{2\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}')M}.
\end{aligned}$$

The stochastic bound shows that evaluation against an independently sampled target contains an irreducible term, even with perfect conversion, the expected error equals

$$2\sigma^2_{\text{stoch}}(\mathbf{s}, \mathbf{c}').$$

Therefore, a large sample-wise MSE does not necessarily imply poor conversion.

Second, the controllable model-dependent terms remain the same as in the deterministic setting: $L^2\epsilon' + \epsilon$. Thus, practitioners should improve reconstruction accuracy, enforce stronger attribute invariance in the encoder, and control the decoder's Lipschitz constant.

Finally, when paired targets with matched stochastic factors are unavailable, evaluation should use multiple target samples or distribution-level criteria rather than relying only on one-to-one squared error.

H Numerical Evaluation of the Conversion Error Bound

Lemma 2 provides the worst-case guarantee

$$\epsilon_{\text{conv}} \leq 2\left(L^2\epsilon' + \epsilon\right),$$

where ϵ is the reconstruction error, ϵ' measures the $\mathbf{C}$-invariance error of the learned representation, and L is the Lipschitz constant of the decoder with respect to its latent input. In this section, we describe a controlled numerical experiment for evaluating the informativeness and tightness of this bound.

We consider a deterministic injective generative model

$$\mathbf{X} = f(\mathbf{S}, \mathbf{C}),$$

where $\mathbf{S} \in \mathbb{R}^2$ is the task-relevant latent variable and $\mathbf{C} \in \{1, \ldots, K\}$ is the controllable attribute. The synthetic generator is chosen such that paired observations

$$f(\mathbf{s}, \mathbf{c}) \qquad \text{and} \qquad f(\mathbf{s}, \mathbf{c}')$$

can be generated for the same realization $\mathbf{s}$ under any pair of conditions $\mathbf{c}, \mathbf{c}'$. This provides access to the ground-truth target sample required to evaluate conversion error directly.

We train a conditional autoencoder

$$g(\mathbf{x}, \mathbf{c}) = d(e(\mathbf{x}), \mathbf{c})$$

using the objective

$$\min_{e,d} \; \mathcal{L}_{\text{rec}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_L\mathcal{L}_{\text{spec}},$$

where

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{S},\mathbf{C}}\left[\|d(e(f(\mathbf{S}, \mathbf{C})), \mathbf{C}) - f(\mathbf{S}, \mathbf{C})\|_2^2\right]$$

is the reconstruction loss, and

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_{\mathbf{S},\mathbf{C},\mathbf{C}'}\left[\|e(f(\mathbf{S}, \mathbf{C})) - e(f(\mathbf{S}, \mathbf{C}'))\|_2^2\right]$$

encourages invariance of the learned representation to the controllable attribute. The term $\mathcal{L}_{\text{spec}}$ penalizes the spectral norms of the decoder weight matrices, and λ_L controls the strength of decoder regularization.

On a held-out test set, we compute empirical analogues of the uniform quantities appearing in Lemma 2. The empirical reconstruction error is

$$\widehat{\epsilon} := \max_{\mathbf{s},\mathbf{c}} \|d(e(f(\mathbf{s}, \mathbf{c})), \mathbf{c}) - f(\mathbf{s}, \mathbf{c})\|_2^2 .$$

The empirical $\mathbf{C}$-invariance error is

$$\widehat{\epsilon}' := \max_{\mathbf{s},\mathbf{c},\mathbf{c}'} \|e(f(\mathbf{s}, \mathbf{c})) - e(f(\mathbf{s}, \mathbf{c}'))\|_2^2 .$$

The empirical conversion error is

$$\widehat{\epsilon}_{\text{conv}} := \max_{\mathbf{s},\mathbf{c},\mathbf{c}'} \|d(e(f(\mathbf{s},\mathbf{c})),\mathbf{c}') - f(\mathbf{s},\mathbf{c}')\|_2^2 .$$

Suppose that the decoder is an m-layer feedforward network with weight matrices $W_1,\dots,W_m$ and 1-Lipschitz activation functions, such as ReLU. For a fixed condition $\mathbf{c}$, a global upper bound on the Lipschitz constant of the decoder with respect to its latent input is

$$\widehat{L}_{\text{arch}} := \prod_{\ell=1}^{m} \|W_\ell\|_2,$$

where $\|W_\ell\|_2$ denotes the spectral norm of W_ℓ. Indeed, for any $\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2 \in \hat{\mathbb{S}}$,

$$\|d(\hat{\mathbf{s}}_1,\mathbf{c}) - d(\hat{\mathbf{s}}_2,\mathbf{c})\|_2 \le \left(\prod_{\ell=1}^{m} \|W_\ell\|_2\right) \|\hat{\mathbf{s}}_1 - \hat{\mathbf{s}}_2\|_2 .$$

The corresponding architecture-level empirical bound is

$$\widehat{B}_{\text{arch}} := 2\left(\widehat{L}_{\text{arch}}^2 \widehat{\epsilon}' + \widehat{\epsilon}\right).$$

Although $\widehat{L}_{\text{arch}}$ is a valid global upper bound for the decoder architecture, products of layerwise spectral norms can be highly conservative. In particular, this quantity may substantially exceed the decoder sensitivity observed on latent representations produced by the encoder.

To quantify decoder sensitivity near the learned data manifold, we also compute:

$$\widehat{L}_{\text{local}} := \max_{\substack{i\neq j \\ \mathbf{c}\in\mathbb{C}}} \frac{\|d(\hat{\mathbf{s}}_i,\mathbf{c}) - d(\hat{\mathbf{s}}_j,\mathbf{c})\|_2}{\|\hat{\mathbf{s}}_i - \hat{\mathbf{s}}_j\|_2}, \qquad \hat{\mathbf{s}}_i = e(\mathbf{x}_i),$$

where the maximum is approximated over randomly sampled pairs of test representations.

The associated local estimate of the conversion bound is

$$\widehat{B}_{\text{local}} := 2\left(\widehat{L}_{\text{local}}^2 \widehat{\epsilon}' + \widehat{\epsilon}\right).$$

Unlike $\widehat{L}_{\text{arch}}$, the quantity $\widehat{L}_{\text{local}}$ is not a global Lipschitz upper bound. It measures only the largest decoder variation observed among the sampled latent pairs. Consequently,

$$\widehat{\epsilon}_{\text{conv}} \le \widehat{B}_{\text{local}}$$

is not theoretically guaranteed to hold. Nevertheless, $\widehat{L}_{\text{local}}$ provides a useful diagnostic of the effective decoder smoothness on the learned latent manifold.

To assess the numerical slack of the theoretical bound, we report the ratios

$$R_{\text{arch}} := \frac{\widehat{\epsilon}_{\text{conv}}}{\widehat{B}_{\text{arch}}}, \quad R_{\text{local}} := \frac{\widehat{\epsilon}_{\text{conv}}}{\widehat{B}_{\text{local}}}.$$

A value of R_{arch} close to one indicates that the architecture-level bound is relatively tight on the evaluated test set, whereas a small value indicates that the bound is conservative. The local ratio R_{local} measures the agreement between the observed conversion error and the estimate obtained using data-local decoder sensitivity. Because $\widehat{L}_{\text{local}}$ is not a global upper bound, R_{local} may exceed one.

We evaluate these quantities across different values of the spectral regularization coefficient λ_L. This sweep reveals the trade-off between decoder smoothness and reconstruction accuracy. Increasing λ_L is expected to reduce $\widehat{L}_{\text{arch}}$ and potentially $\widehat{L}_{\text{local}}$, but overly strong regularization may increase $\widehat{\epsilon}$. Hence, spectral regularization improves the bound only when the reduction in $\widehat{L}^2\widehat{\epsilon}'$ is not offset by an increase in reconstruction error.

Table 1: Empirical evaluation of the conversion-error bound under different spectral regularization strengths. $\widehat{\epsilon}$ is the maximum reconstruction error, $\widehat{\epsilon}'$ is the maximum latent invariance error, $\widehat{\epsilon}_{\mathrm{conv}}$ is the maximum conversion error, $\widehat{L}_{\mathrm{arch}}$ is the architecture-level Lipschitz upper bound, and $\widehat{L}_{\mathrm{local}}$ is an empirical local Lipschitz estimate. The corresponding bounds are $\widehat{B}_{\mathrm{arch}} = 2(\widehat{L}^2_{\mathrm{arch}}\widehat{\epsilon}' + \widehat{\epsilon})$ and $\widehat{B}_{\mathrm{local}} = 2(\widehat{L}^2_{\mathrm{local}}\widehat{\epsilon}' + \widehat{\epsilon})$. The tightness ratios are defined as $\widehat{\epsilon}_{\mathrm{conv}}/\widehat{B}$.

λ_L	$\widehat{\epsilon}$	$\widehat{\epsilon}'$	$\widehat{\epsilon}_{\mathrm{conv}}$	$\widehat{L}_{\mathrm{arch}}$	$\widehat{L}_{\mathrm{local}}$	$\widehat{B}_{\mathrm{arch}}$	$\widehat{B}_{\mathrm{local}}$	R_{arch}	R_{local}
0	2.6581×10^{-3}	1.3268×10^{-4}	2.7130×10^{-3}	8.1907	1.9276	2.3119×10^{-2}	6.3022×10^{-3}	0.1174	0.4305
10^{-5}	3.4025×10^{-3}	3.0815×10^{-4}	3.4530×10^{-3}	5.3304	1.7693	2.4316×10^{-2}	8.7341×10^{-3}	0.1420	0.3953
10^{-4}	7.0523×10^{-3}	1.1795×10^{-4}	6.9050×10^{-3}	3.6326	1.6331	1.7218×10^{-2}	1.4734×10^{-2}	0.4010	0.4687
10^{-3}	1.8723×10^{-2}	2.4019×10^{-4}	1.8723×10^{-2}	2.5628	1.1585	4.0601×10^{-2}	3.8090×10^{-2}	0.4611	0.4915
10^{-2}	1.2174×10^{-1}	6.8784×10^{-4}	1.2206×10^{-1}	2.1554	0.9093	2.4988×10^{-1}	2.4462×10^{-1}	0.4885	0.4990

This experiment is not intended to establish that Lemma 2 is uniformly tight. Rather, it evaluates whether the bound is numerically informative in a controlled setting and separates two potential sources of looseness. The first is the use of the inequality

$$\|\mathbf{a}+\mathbf{b}\|_2^2 \le 2\left(\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2\right)$$

in the proof of Lemma 2. The second, and potentially more significant, source is the use of a global decoder Lipschitz constant.

If

$$\widehat{L}_{\mathrm{arch}} \gg \widehat{L}_{\mathrm{local}},$$

then a large value of $\widehat{B}_{\mathrm{arch}}$ may primarily reflect the conservativeness of the architecture-level Lipschitz estimate rather than large decoder sensitivity on the data manifold. Conversely, if both Lipschitz quantities are large, the result indicates that small deviations from latent $\mathbf{C}$-invariance can be substantially amplified by the decoder.

The experiment therefore tests the qualitative implication of Lemma 2: reliable conversion requires jointly controlling reconstruction error, latent attribute invariance, and decoder smoothness.

We conduct the experiment on a controlled synthetic dataset for which the ground-truth latent variable and all counterfactual target samples are known. Specifically, we sample

$$\mathbf{S} \sim \mathrm{Unif}([-1,1]^2)$$

and

$$\mathbf{C} \sim \mathrm{Unif}\{1,\dots,K\}, \qquad K = 3.$$

Each observation is generated by a deterministic injective map

$$\mathbf{X} = f(\mathbf{S},\mathbf{C}) \in \mathbb{R}^6.$$

The generator preserves the two coordinates of $\mathbf{S}$ explicitly and augments them with condition-dependent rotations, translations, and nonlinear features. Consequently, f is injective with respect to $(\mathbf{S},\mathbf{C})$. The use of synthetic data allows us to generate, for every latent realization $\mathbf{s}$, the complete set of paired observations

$$\{f(\mathbf{s},\mathbf{c}) : \mathbf{c} \in \{1,\dots,K\}\}.$$

Thus, given a source observation $f(\mathbf{s},\mathbf{c})$ and a target condition $\mathbf{c}'$, the exact target sample $f(\mathbf{s},\mathbf{c}')$ is available. This makes it possible to measure reconstruction, latent invariance, and conversion errors without requiring an estimated or independently sampled reference target.

Table 1 shows that the architecture-level bound is non-vacuous for every spectral regularization strength considered:

$$\widehat{\epsilon}_{\mathrm{conv}} < \widehat{B}_{\mathrm{arch}} = 2\left(\widehat{L}^2_{\mathrm{arch}}\widehat{\epsilon}' + \widehat{\epsilon}\right).$$

For the unregularized model, the observed conversion error is

$$\widehat{\epsilon}_{\mathrm{conv}} = 2.7130 \times 10^{-3},$$

while the architecture-level bound is

$$\widehat{B}_{\mathrm{arch}} = 2.3119 \times 10^{-2},$$

giving a tightness ratio of

$$R_{\mathrm{arch}} = 0.1174.$$

Moderate spectral regularization substantially reduces the architecture-level Lipschitz estimate and improves the tightness of the bound. In particular, at $\lambda_L = 10^{-4}$, the architecture-level bound reaches its smallest value, $\widehat{B}_{\mathrm{arch}} = 1.7218 \times 10^{-2}$, with $R_{\mathrm{arch}} = 0.4010$. However, stronger regularization introduces a clear trade-off: although $\widehat{L}_{\mathrm{arch}}$ decreases from 8.1907 to 2.1554, the reconstruction error increases from 2.6581×10^{-3} to 1.2174×10^{-1}. Consequently, for $\lambda_L = 10^{-2}$, the reconstruction term dominates the bound and both the bound and the observed conversion error become substantially larger.

Table 1 also shows that the empirical local Lipschitz estimate is consistently much smaller than the architecture-level upper bound. For example, in the unregularized case,

$$\widehat{L}_{\mathrm{arch}} = 8.1907, \qquad \widehat{L}_{\mathrm{local}} = 1.9276.$$

This explains much of the conservatism of the bound. The local diagnostic gives

$$\widehat{B}_{\mathrm{local}} = 6.3022 \times 10^{-3},$$

which is substantially closer to the observed conversion error. Since $\widehat{L}_{\mathrm{local}}$ is estimated only over sampled latent pairs, however, $\widehat{B}_{\mathrm{local}}$ should be interpreted as an empirical diagnostic rather than an upper bound. Overall, the experiment confirms that the theoretical bound can be informative in a trained model, while illustrating that its tightness depends on balancing decoder smoothness against reconstruction accuracy.

I Reproducibility Statement

Our work is primarily theoretical. In addition, we perform experimental validations of the proposed theory using the setups and datasets described in the experimental section. The code used for these experimental validations, including the relevant training and evaluation procedures, will be released upon publication to facilitate reproduction of the reported results.