

---

# Can Diffusion Models Disentangle? A Theoretical Perspective

---

Liming Wang<sup>1</sup>   Muhammad Jehanzeb Mirza<sup>1</sup>   Yishu Gong<sup>2</sup>   Yuan Gong<sup>1\*</sup>  
Jiaqi Zhang<sup>1</sup>   Brian H. Tracey<sup>2</sup>   Katerina Placek<sup>2</sup>   Marco Vilela<sup>2</sup>

James R. Glass<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology, <sup>2</sup>Takeda  
limingw@csail.mit.edu

## Abstract

This paper presents a novel theoretical framework for understanding how diffusion models can learn disentangled representations with commonly used weak supervision such as partial labels and multiple views. Within this framework, we establish identifiability conditions for diffusion models to disentangle latent variable models with *stochastic, non-invertible* mixing processes. We also prove *finite-sample global convergence* for diffusion models to disentangle independent subspace models. To validate our theory, we conduct extensive disentanglement experiments on subspace recovery in latent subspace Gaussian mixture models, image colorization, denoising, and voice conversion for speech classification. Our experiments show that training strategies inspired by our theory, such as style guidance regularization, consistently enhance disentanglement performance.

## 1 Introduction

Extracting hidden structure from raw sensory data is fundamental to progress in multimodal perception [1–5], scientific discovery [6–15], AI-assisted content creation [16–20], and many more. Autonomous vehicles must localize objects and auditory events while suppressing background noise to navigate safely in open-world conditions, and data-driven drug-discovery systems need to group and recombine functionally related chemical components to propose therapies for emerging diseases. Deep learning-based creative tools likewise hinge on isolating user-specified factors (e.g., speaker style or lighting) while leaving other aspects untouched.

Many latent factors are *disentangled* in the sense that they vary independently. For example, speech content persists regardless of the speaker, and object shape remains consistent under different lighting conditions. This intuition inspires breakthroughs in linear [21] and non-linear [22] independent component analysis (ICA), modern deep learning-based disentanglement [23, 24] and causal representation learning [25]. However, an impossibility result shows that fully *unsupervised* disentanglement is unattainable in general [26]. Recent work on disentanglement therefore relies on weak labels [27, 28] or multi-view supervision [29, 30].

Applications such as video editing and drug discovery often require both latent factor extraction and controlled synthesis of novel samples from the latent factors. Diffusion models (DMs) [31–33] based on learning score functions [34] of probability distributions, excel at generation and power state-of-the-art

---

\*Work done at MIT, now with xAI

editors and simulators [16, 17, 19]. However, standard DMs learn only the data *marginal*, encoding latent factors *implicitly*. To control data generation using the latent structure, *conditional* DM (CDM) inject side information into the score function [35–37] and achieve impressive empirical success in disentanglement tasks such as voice conversion and image editing [38, 39, 20, 35, 36]. Yet a principled understanding of *when* these models learn identifiably disentangled representations is still missing.

In our work, we broaden the disentanglement theory to diffusion models and provide the first learning-theoretic framework for DM-based disentanglement, which poses unique challenges. First, a sample from DM is generated by integrating a stochastic differential equation (SDE), so the generated sample becomes a *stochastic, non-invertible* mapping of the latent variables. The lack of an analytic inverse precludes the change-of-variables calculus and ICA-style arguments that underpin classical disentanglement theory [26, 40]. Further, the extra uncertainty injected by the stochastic drift can make it more difficult to leverage the commonly used weak supervision. We tackle these challenges by (i) proposing notions of approximate disentanglement applicable to stochastic, non-invertible settings, (ii) marrying information-regularised score matching with recent finite-sample analyses of score-based models [41]. The resulting framework lays the groundwork for the following contributions:

1. We show that, under mild Lipschitz assumptions, DMs can recover approximately disentangled representations of two latent factors (e.g., content and style), when given either partial supervision or multi-view inputs. For independent subspace models, we further prove global convergence in the finite-sample regime using gradient descent.
2. Building on the theory, we introduce a novel style-guided score matching loss that attains a global optimum for the independent-subspace case and improves disentanglement in practice.
3. Extensive experiments on Gaussian mixture models, image editing, and voice conversion for speech classification demonstrate that our theory-inspired training strategies consistently enhance disentanglement quality and downstream classification accuracy.

The rest of this paper is organized as follows: In Section 2, we provide the background on diffusion models. Section 3 formalizes the problem of content-style disentanglement, and later in Section 4 we present the main theoretical results and Section 5 details empirical evaluations on synthetic, image and speech data, demonstrating and supporting the theoretical findings. Finally, Section 7 concludes the paper.

## 2 Background: diffusion models

Diffusion models (DMs) [31–33] approximate the pdf  $p_X =: p_0$  of an r.v.  $X \sim p_0$  via a two-stage process: *noising* and *denoising*. In the noising stage, data is progressively corrupted using an SDE:

$$dX_t = \mu(X_t, t)dt + \xi(t)dB_t, \quad X_0 \sim p_0, \quad (1)$$

where  $B_t$  is a Brownian motion. We adopt the choices  $\mu(X_t, t) := -X_t$  and  $\xi(t) \equiv \sqrt{2}$ , leading to the Ornstein–Uhlenbeck (OU) process:

$$dX_t = -X_t dt + \sqrt{2}dB_t, \quad X_0 \sim p_0, \quad (2)$$

which converges to a standard Gaussian [42, 41]. Let  $p_t$  denote the marginal of  $X_t$  and  $p_{t|s}$  the conditional pdf of  $X_t$  given  $X_s$ . In the denoising stage, the goal is to recover  $X$  from noisy versions  $X_t, t \geq t_0$  by simulating the time-reversed process  $X_t^\leftarrow := X_{T-t}$ :

$$dX_t^\leftarrow = [X_t^\leftarrow + 2\nabla_x \log p_{T-t}(X_t^\leftarrow)]dt + \sqrt{2}dB_t^\leftarrow, \quad X_0^\leftarrow \sim p_T, \quad (3)$$

which converges back to  $p_0$  [43]. Since the score function  $\nabla_x \log p_t(X_t)$  is unknown, DM learns a *score estimator*  $s_\theta: \mathbb{R}^{d_X} \times [0, T] \mapsto \mathbb{R}^{d_X}$  by minimizing the *score matching* objective:

$$L(\theta) := \mathbb{E}_{t, p_t} \|s_\theta(X_t, t) - \nabla_x \log p_t(X_t)\|^2,$$

This objective is equivalent to a *conditional* score matching objective involving  $p_{t|0}$ :

$$L_c(\theta) := \mathbb{E}_{t, p_0 p_{t|0}} \|s_\theta(X_t, t) - \nabla_x \log p_{t|0}(X_t|X)\|^2 = \mathbb{E}_{t, p_0 p_{t|0}} \left\| s_\theta(X_t, t) + \frac{N_t}{\sigma(t)} \right\|^2, \quad (4)$$

where  $N_t$  is standard Gaussian and  $\sigma(t) := \sqrt{1 - \exp(-2t)}$ . During inference, new samples are generated by simulating an estimated SDE, with  $\hat{X}_0^\leftarrow \sim p_T$  and discretized time steps  $t \in [k\eta, (k+1)\eta]$ :

$$d\hat{X}_t^\leftarrow = [\hat{X}_t^\leftarrow + 2s_\theta(\hat{X}_{k\eta}^\leftarrow, T - k\eta)]dt + \sqrt{2}dB_t^\leftarrow. \quad (5)$$

### 3 Content-style disentanglement

The task of content-style disentanglement can be formalized through a *latent variable model* (LVM). For clarity, we refer to all random entities – scalars, vectors, or matrices – as random variables (r.v.'s), and focus on the *continuous* case. For any r.v.  $X$ , let  $p_X$  denote its probability density function (pdf). We assume two latent factors: the *content*  $Z \sim p_Z$  taking values in  $\mathbb{R}^{d_Z}$  and the *style*  $G \sim p_G$  taking values in  $\mathbb{R}^{d_G}$ , which jointly generate an observable *sample*  $X \in \mathbb{R}^{d_X}$  through the noisy, non-invertible and nonlinear transformation with an invertible *mixing function*  $f: \mathbb{R}^{d_Z} \times \mathbb{R}^{d_G} \mapsto \mathbb{R}^{d_X}$ :

$$X = \sqrt{1 - \delta^2} f(Z, G) + \delta N, \quad (6)$$

where  $\delta$  is the noise level and  $N$  is an independent, standard Gaussian noise. We assume that  $Z$  and  $G$  are statistically independent, as is common in disentanglement [24, 44, 26] and ICA literature [21, 40]. Although exact independence between  $Z$  and  $G$  is assumed for clarity, our framework naturally extends to settings where independence holds only approximately, as will be discussed subsequently.

This partition of latent factors arises in several settings. In *controllable generation*,  $Z$  encodes persistent attributes (e.g., object identity) while  $G$  governs editable factors (e.g., pose or lighting). In *self-supervised learning* (SSL),  $Z$  captures invariant content across views or modalities, and  $G$  captures modality- or augmentation-specific variations. In the low-noise regime  $\delta \rightarrow 0$ , our goal is to recover the latent variables  $Z$  and  $G$  from observations of  $X$ . A common notion of recovery is *block identifiability* [45–48, 29, 30], which ensures that subgroups of scalar latent variables can be recovered up to an invertible transformation. However, exact block identifiability is not achievable in the presence of noise  $\delta > 0$  due to the non-invertibility of the mixing process, motivating a need for *approximate disentanglement*. To this end, we propose two complementary criteria: (1) approximate information-theoretic disentanglement and (2) editability.

**$(\epsilon, \nu)$ -disentanglement.** To quantify how well the learned content and style representations are separated and informative, we define an information-theoretic notion of approximate disentanglement.

**Definition 3.1** ( $(\epsilon, \nu)$ -disentanglement). Let  $(\hat{Z}, \hat{G})$  be content and style encodings inferred from an observed sample  $X$ . They are  $(\epsilon, \nu)$ -disentangled if, for some  $\epsilon, \nu \geq 0$ , (i)  $I(\hat{Z}; \hat{G}) \leq \epsilon$ ; (ii)  $I(\hat{Z}, \hat{G}; X) \geq I(Z, G; X) - \nu$ , where  $I(A; B)$  denotes the mutual information (MI) between r.v.'s  $A$  and  $B$ .

These conditions ensure that the learned latent factors are (i) nearly independent (as the true  $Z$  and  $G$  are), and (ii) retain most of the information about the observed data  $X$ . The definition remains meaningful even as  $\delta \rightarrow 0$  and  $I(Z, G; X) \rightarrow \infty$ , since it is based on a bounded *difference* in mutual information.

**$\epsilon$ -editability.** In many applications, it is desirable to modify style while preserving content. This motivates the following notion of editability based on conditional sample generation.

**Definition 3.2.** ( $\epsilon$ -editability) Let  $(\hat{Z}, \hat{G})$  be encodings inferred from  $X$ , and let  $\hat{G}' \sim p_{\hat{G}}$  be an independent style encoding.  $(\hat{Z}, \hat{G})$  are  $\epsilon$ -editable if there exists a generative model  $q(\cdot | \hat{Z}, \hat{G}')$  such that the generated sample  $\hat{X} \sim q(\cdot | \hat{Z}, \hat{G}')$  satisfies  $\mathbb{E}_{Z \sim p_Z} d_{\text{TV}}(p_{\hat{X}|Z}, p_{X|Z}) \leq \epsilon$ , where  $\hat{X} := \sqrt{1 - \epsilon^2} X + \epsilon N$  is a smoothed version of  $X$ , and  $d_{\text{TV}}$  is the total variation distance.

This definition captures the ability to recombine content and style encodings to generate new samples that are consistent with the original content. For example, in a facial image editing task, the encoding  $\hat{Z}$  may represent identity while  $\hat{G}$  captures facial expression. By swapping  $\hat{G}$  with a new expression encoding  $\hat{G}'$ , we can generate a new image that preserves identity but alters the expression. Deterministic decoders are allowed as a special case with  $q(\cdot | \hat{Z}, \hat{G}) = \delta_{\hat{f}(\hat{Z}, \hat{G})}$  for some function  $\hat{f}$ .

The notions of  $(\epsilon, \nu)$ -disentanglement and  $\epsilon'$ -editability are complimentary but not equivalent. In particular, disentanglement in the mutual information sense does not guarantee editability. The following example illustrates this by constructing encodings that are perfectly disentangled (i.e.,  $(0, 0)$ -disentangled) yet fail the editability criterion due to a hidden ambiguity introduced during recombination. The proof is provided in Appendix A.

**Example 3.1.** Suppose content  $Z \sim \mathcal{N}(0,1)$  and  $G \sim \mathcal{N}(0,1)$  are independent standard Gaussian r.v.'s and consider the noiseless setting with  $\delta = 0$ . Further, suppose  $p_{f(Z,G)|Z} \neq p_{f(-Z,G)|Z}$ . Then we can choose the content/style encodings to be  $\hat{Z} = Z \text{sgn}(G)$ ,  $\hat{G} = G$  and the decoder  $\hat{f}(\hat{Z}, \hat{G}) = f(\hat{Z} \text{sgn}(\hat{G}), \hat{G})$ , where  $\text{sgn}(x)$  denotes the sign of  $x$ . Then  $(\hat{Z}, \hat{G})$  are  $(0,0)$ -disentangled but not  $\epsilon$ -editable for some  $\epsilon > 0$ .

Intuitively, the encoder flips the sign of the content variable depending on the style. This transformation preserves both independence and informativeness, but causes ambiguity when recombining  $\hat{Z}$  with a new style  $\hat{G}'$ , since the decoder cannot distinguish whether the original sign should be restored. As a result, samples generated from new combinations do not match the original content distribution.

**Weakly supervised disentanglement.** When neither  $Z$  nor  $G$  is observed, disentanglement is in general impossible for nonlinear mixing functions: the observable pdf  $p_X$  alone does not identify whether the data is generated from a disentangled or entangled LVM [26]. We consider two practical settings where *side information* is available to help resolve this ambiguity.

**Definition 3.3** (Content disentanglement). Assume Eq. 6, and suppose a known style function  $g: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_G}$  is given such that  $g(f(z, G)) = G$  for all  $z$ . The goal of content disentanglement is to learn encodings  $(\hat{Z}, \hat{G})$  from  $X$  that are  $(\epsilon, \nu)$ -disentangled and  $\epsilon'$ -editable.

This setting appears in applications such as image editing and voice conversion [49, 38, 50]. For example, in voice conversion,  $g(X)$  is a speaker embedding extracted from a pre-trained speaker recognition model (e.g. speaker embeddings from ECAPA-TDNN, which achieves almost perfect speaker verification error rate [51]). This assumption mirrors common practice in the voice conversion literature and is not unique to our framework. In image editing,  $g(\cdot)$  could represent learned text embeddings of editing instructions.

**Definition 3.4** (Multi-view disentanglement). Assume there are multiple views  $X^1, \dots, X^{n_V}$ , where each  $X^i$  is generated as  $X^i = \sqrt{1 - \delta_i^2} f_i(Z, G^i) + \delta_i N^i$ ,  $1 \leq i \leq n_V$ , with i.i.d view-specific styles  $G^i$ 's, i.i.d standard Gaussian noise  $N^i$ 's, view-specific noise levels  $\delta_i$ 's and invertible view-specific mixing functions  $f_i: \mathbb{R}^{d_{G_i}} \times \mathbb{R}^{d_{Z_i}} \mapsto \mathbb{R}^{d_{X_i}}$ . Then the task of multi-view disentanglement is to learn encodings  $(\hat{Z}, \hat{G}^i)$  for all  $i$  such that  $(\hat{Z}, \hat{G}^i)$  are  $(\epsilon, \nu)$ -disentangled and  $\epsilon'$ -editable.

This setting is prevalent in SSL (e.g., [52–58]). When  $f_i \equiv f_1$  (e.g. multiple camera views), each  $X^i$  may correspond to a different augmentation. When  $f_i$  differ (e.g., across sensory modalities), the views may represent distinct but semantically aligned representations. For clarity, we focus on the unimodal two-view case ( $n_V = 2$ ), which readily generalizes to multimodal scenarios with  $n_V > 2$ . To facilitate our theoretical analysis of DM-based disentanglement, we adopt the following mild assumptions, which are common in the analysis of DMs [59, 42, 41].

**Assumption 3.5.** The sample  $X$  is sub-gaussian with second moment  $\sigma_X^2 d_X$ .

**Assumption 3.6.** The score function of the sample pdf  $p_X$  is  $\lambda_s$ -Lipschitz.

Assumption 3.5 ensures that sample values do not exhibit heavy tails, which could destabilize the diffusion process. Assumption 3.6 ensures that the score function does not change too abruptly, preventing discontinuities that could hinder accurate recovery of content and style during denoising.

## 4 Theory: diffusion model-based disentanglement

In this section, we first present our theoretical results for DM-based disentanglement, and then discuss multi-view disentanglement, and finally provide results for disentanglement of independent subspaces.

### 4.1 Content disentanglement with diffusion models

**Overview.** This section analyzes the ability of DMs to achieve approximate disentanglement in the content disentanglement task defined in Definition 3.3. First, we introduce a conditional DM trained with a regularized score matching objective tailored for this task. Second, we show that this model can learn content and style encodings  $(\hat{Z}, \hat{G})$  that are  $(\epsilon, \nu)$ -disentangled with arbitrarily small  $\epsilon, \nu$  as the noise level  $\delta \rightarrow 0$ . Finally, we discuss why such encodings may still fail to be  $\epsilon'$ -editable for some  $\epsilon' > 0$ . Formal proofs are provided in Appendix B.

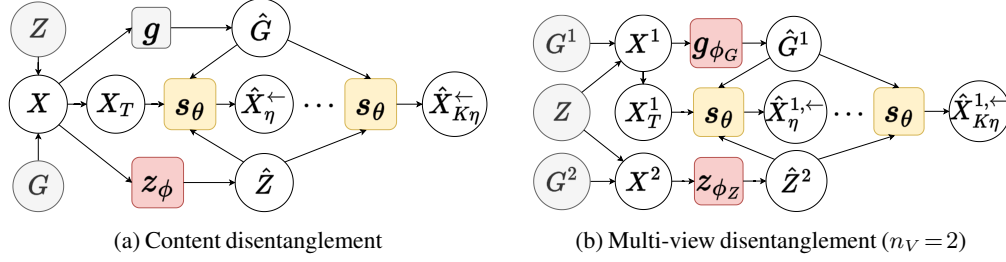


Figure 1: **Graphical model of diffusion-based disentanglement under different types of weak supervision.** Shaded nodes denote latent variables; clear nodes denote observed variables. Learnable components are marked in red and orange, while the frozen component is shown in gray. **Left: Content disentanglement.** The style encoder  $g(\cdot)$  is known and fixed. The model learns a content encoder  $z_\phi$  and a score estimator  $s_\theta$  to estimate the score of  $p_{X|\hat{Z},\hat{G}}$ , thereby disentangle content  $Z$  from style. **Right: Multi-view disentanglement.** Given paired views  $(X^1, X^2)$  sharing content  $Z$ , the model learns to recover  $Z$  and style  $G^1$  of view 1 by estimating the score of  $p_{X|\hat{Z}^2,\hat{G}^1}$  using  $s_\theta$ .

The model architecture is illustrated in Figure 1a. Style is represented by a fixed encoder  $\hat{G} := g(X_{t_0})$ , where  $g(\cdot)$  is assumed known. Content is learned through a trainable encoder yielding  $\hat{Z} = z_\phi(X_{t_0})$ . To enable controllable generation, both  $\hat{Z}$  and  $G$  are fed as inputs to a conditional score estimator  $s_\theta: \mathbb{R}^{d_X} \times \mathbb{R}^{d_Z} \times \mathbb{R}^{d_G} \times [0, T] \mapsto \mathbb{R}^{d_X}$ , which estimates the score of the *conditional* pdf  $p_{X|\hat{Z},\hat{G}}$ . We train this model using the following *regularized* score matching objective: hyperparameters  $\gamma, \rho > 0$ :

$$L_c^{\gamma,\rho}(\theta, \phi) := \underbrace{\mathbb{E}_{t, p_0 p_{t_0|0} p_{t|t_0}} \left\| s_\theta(X_t, \hat{Z}, \hat{G}, t) + \frac{N_t}{\sigma(t)} \right\|^2}_{\text{Conditional score matching loss}} + \underbrace{\gamma(I(z_\phi(X_{t_0}); X) - \rho)_+}_{\text{Content information regularizer}}, \quad (7)$$

where  $(x)_+ := \max\{x, 0\}$ . The **score matching loss** encourages the model to match the conditional score, while the **regularizer** limits the information that the content encoder can extract from the input, preventing overfitting via direct copying. In practice, mutual information is often approximated using tractable variational bounds [60–62], which serve as surrogates for the **content information regularizer**. Using the above setup, we prove the following theorem.

**Theorem 4.1.** Suppose Assumption 3.5-3.6 hold, and  $(\theta^*, \phi^*)$  be a minimizer of  $L_c^{\gamma,\rho}$  defined in Eq. 7 with  $\rho = I(Z; X) + C_1 \delta, \gamma \geq C_2/T$  for some  $C_1, C_2 > 0$ . Set  $t_0 = -\log(1 - \delta^2)^{1/2}$  and  $\hat{Z} := z_{\phi^*}(X)$ . Then for any  $\delta < \min\left\{\frac{1}{2}, \frac{1}{\sqrt{d_X}}\right\}$ , the encodings  $(\hat{Z}, \hat{G}) = (z_{\phi^*}(X_{t_0}), g(X_{t_0}))$  are  $(\epsilon, \nu)$ -disentangled with  $\epsilon = C_3 \lambda_s \sigma_X^2 d_X \delta, \nu = C_4 \sigma_X^2 \delta^2$  for some constants  $C_3, C_4 > 0$ .

**Intuition.** Theorem 4.1 shows that under sub-gaussian tail assumptions and Lipschitz-continuous scores, DM can achieve  $(\epsilon, \nu)$ -disentanglement with  $(\epsilon, \nu) \rightarrow 0$  as the noise level  $\delta \rightarrow 0$ . The rate at which disentanglement improves depends inversely on the Lipschitz constant — higher sensitivity in the score function slows disentanglement by amplifying noise-induced variations. The regularized objective mitigates this trading off between **predictive power** and the amount of **content information** retained in the content encoding  $\hat{Z}$ .

Theorem 4.1 extends to cases where content and style are only approximately independent, i.e.,  $I(Z; G) \leq \epsilon_1$  for some  $\epsilon_1 > 0$ , by treating the dependency as a perturbation (Appendix B.7). Assumption 3.5 can also be relaxed to bounded variances, as the proof relies on moment control. However, it is important to note that approximate disentanglement in the MI sense does not imply editability. As illustrated in Example 3.1, the model may still leak style information into the content encoding. This limitation, which we refer to as *content distortion*, prevents achieving vanishing  $\epsilon'$ -editability even as  $\delta \rightarrow 0$ .

## 4.2 Multi-view disentanglement with diffusion models

**Overview.** This section analyzes the ability of DMs to achieve editability in the multi-view disentanglement setting defined in Definition 3.4. We introduce a DM trained with a modified score matching objective, analogous to the content disentanglement case. We then show that the learned

encodings  $(\hat{Z}, \hat{G})$  are  $\epsilon$ -editable with  $\epsilon \rightarrow 0$  as  $\delta \rightarrow 0$ . Finally, we discuss how the result generalizes to more than two views and non i.i.d styles, and what it reveals about the role of different types of weak supervision. Full proofs are in Appendix C.

The model architecture is illustrated in Figure 1b. We define the encoders  $\hat{Z}^i := z_{\phi_Z}(X_{t_0}^i)$  and  $\hat{G}^i := g_{\phi_G}(X_{t_0}^i)$  for each view  $i$ . The score estimator  $s_\theta(X_t, \hat{Z}^2, \hat{G}^1, t)$  is trained with the following loss:

$$L_m(\theta, \phi) := \underbrace{\mathbb{E}_{t, p_0 p_{t_0|0} p_{t|0}} \left\| s_\theta(X_t^1, \hat{Z}^2, \hat{G}^1, t) + \frac{N_t^1}{\sigma(t)} \right\|^2}_{\text{Conditional score matching loss}} + \underbrace{I(\hat{G}^1; X^2)}_{\text{Style information regularizer}}, \quad (8)$$

where  $\phi = [\phi_Z, \phi_G]$ . The **regularizer** encourages  $\hat{G}^1$  to encode only the style of  $X^1$ . At inference time, we sample from the estimated reverse process Eq. 5 but use the *conditional* score estimator to guide generation. Under this setup, We prove the following theorem.

**Theorem 4.2.** *Under Assumption 3.5-3.6 and additional regularity conditions (Assumption C.2-C.4). Consider the two-view, unimodal setting with  $\delta_1 =: \delta$  and  $\delta_2 = 0$ . Let  $(\theta^*, \phi_Z^*, \phi_G^*)$  be a minimizer of  $L_m$  and  $\hat{Z}^i = z_{\phi_Z^*}(X^i)$ ,  $\hat{G}^i = g_{\phi_G^*}(X^i)$  for each view  $i$ . Set  $t_0 = -\log(1 - \delta^2)^{1/2}$  and the diffusion step size  $\eta := C_5 \frac{\delta^3}{\lambda_X^2 T}$  for some constant  $C_5 > 0$ . Then for some constants  $C_6, C_7 > 0$ : (i)  $I(\hat{Z}^2; \hat{G}^1) \leq C_6 \lambda_s \sigma_X d_x \delta$ ; (ii) the content encoding  $\hat{Z}^2$  and the style encoding  $\hat{G}^1$  are  $C_7 \sqrt{\lambda_s \sigma_X^2 d_X} \delta$ -editable.*

**Intuition.** Theorem 4.2 establishes that in the multi-view setting, DMs can achieve  $(\epsilon, \nu)$ -disentanglement and  $\epsilon$ -editability with vanishingly small  $\epsilon$ , even though  $\nu$  cannot be made arbitrarily small. In other words, perfect separation of content and style is theoretically achievable in terms of independence and editability, even if some information loss is unavoidable. The editability error depends on the Lipschitz constant of the decoder, the intrinsic data variance, and the latent dimensionality—factors that all amplify sensitivity to noise. To mitigate this, we introduce a **style information regularizer**, which suppresses content leakage into the style encoder. This prevents entanglement and enables reliable "mix-and-match" generation. Without this regularization, residual content in the style embedding can undermine both disentanglement and editing performance.

While Theorem 4.2 is stated for i.i.d. styles across views, it readily extends to nearly independent styles, i.e., those satisfying  $I(G^1; G^2 | Z) \leq \epsilon_I$  for some small  $\epsilon_I$  (see Appendix C.4). Our framework also handles more than two views and heterogeneous view-specific mixing functions. A key challenge arises when content and style are highly correlated or when the sample size is small. In these settings, noise can break delicate statistical dependencies, leading to spurious entanglement, especially under limited data, which can mask the true latent structure further due to overfitting. These issues are not covered by prior identifiability results [48, 29], which assume noiseless generation and invertible mixing functions.

### 4.3 Disentanglement of independent subspaces with diffusion models

**Overview.** This section considers a *special case* of content disentanglement – when the underlying distribution follows an *independent subspace model* (ISM) [45–47, 63]. ISMs have served as a foundation in classical disentanglement literature and are also supported empirically in recent self-supervised learning studies [64], with applications in both discriminative and generative disentanglement tasks [65, 66, 39, 50, 67]. Under the ISM setting, we prove that a carefully constructed DM trained with gradient descent can simultaneously achieve  $(\epsilon, \nu)$ -disentanglement and  $\epsilon'$ -editability even in the finite-sample setting – offering stronger theoretical guarantees than those in the general content disentanglement case (Section 4.1). We also outline how these results can be extended to the multi-view case in the discussion. Full proofs are provided in Appendix D. To begin, we formally define an ISM below.

**Definition 4.3.** *An independent subspace model (ISM) is an LVM defined by*

$$Z \sim p_Z, G \sim p_G, X = A_Z Z + A_G G, \quad (9)$$

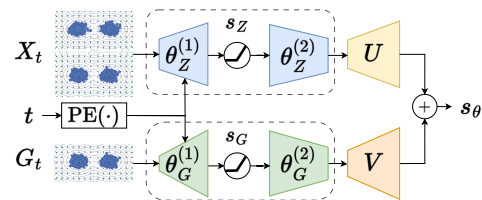


Figure 2: **Dual-encoder score network** for content-style disentanglement of ISM.

Under the ISM setting, we prove that a carefully constructed DM trained with gradient descent can simultaneously achieve  $(\epsilon, \nu)$ -disentanglement and  $\epsilon'$ -editability even in the finite-sample setting – offering stronger theoretical guarantees than those in the general content disentanglement case (Section 4.1). We also outline how these results can be extended to the multi-view case in the discussion. Full proofs are provided in Appendix D. To begin, we formally define an ISM below.

where  $A_Z \in \mathbb{R}^{d_X \times d_Z}$ ,  $A_G \in \mathbb{R}^{d_X \times d_G}$  are orthogonal matrices such that their column spaces are orthogonal and span  $\mathbb{R}^{d_X}$ . That is,  $R(A_Z)^\perp = R(A_G)$  with  $d_Z + d_G = d_X$ , where  $R(A)$  is the column space of matrix  $A$ . Let  $X_t$  denote the noisy version of  $X$  at time  $t$  in the diffusion process. Then we define  $Z_t := A_Z^\top X_t =: z(X_t)$ ,  $G_t := A_G^\top X_t =: g(X_t)$ .

This model generalizes the setting in [41], which assumes Gaussian noise for  $G_t$ . A key property of ISMs is that the score function of the marginal  $p_t(X_t)$  is *decomposable* as shown below.

**Lemma 4.4.** For any  $t \geq 0$ , the score of the pdf  $p_t(X_t)$  under the ISM satisfies

$$s^*(X_t, t) := \nabla_x \log p_t(X_t) = A_Z \nabla_z \log p_{Z_t}(z(X_t)) + A_G \nabla_g \log p_{G_t}(g(X_t)).$$

Lemma 4.4 shows that for ISM, the score function is a linear combination of the score functions of content and style. Motivated by Lemma 4.4, we propose a *dual encoder network* for learning  $s^*(x, t)$ , as shown in Figure 2:

$$s_\theta(x, t) := U s_Z^{\theta_Z}(x, t) + V s_G^{\theta_G}(g(x), t) := \text{NN}([x, \text{PE}(t)]) + \text{NN}([g(x), \text{PE}(t)]). \quad (10)$$

where  $\text{NN}(\cdot)$  denotes a two-layer ReLU neural net and  $\text{PE}(t)$  is a time position encoding. The first branch of the dual network computes the content score  $s_Z(X_t, t)$  of the noisy sample  $X_t$ , while the second branch computes the style score  $s_G(G_t, t)$  from the noisy style  $G_t$ . The two scores are combined to produce the final score function  $s_\theta(X_t, t)$ . Unlike earlier sections,  $s_\theta$  here is *unconditional*. To train this network, we use a regularized score matching loss:

$$L_n^{\lambda_r}(\theta) = \underbrace{2\mathbb{E}_{t, \hat{p}_t^n(x)} \|s_\theta(x, t) - s^*(x, t)\|_2^2}_{L_{0,n}: \text{score matching loss}} + \underbrace{2\lambda_r \mathbb{E}_{t, \hat{p}_t^n(x)} \|V s_G(g(x), t) - s^*(x, t)\|_2^2}_{L_{r,n}: \text{style guidance regularizer}} + \underbrace{\frac{1}{2} L_{b,n}(\theta)}_{\text{Balancing loss}}, \quad (11)$$

where  $\lambda_r > 0$  is the **style guidance weight** and  $\hat{p}_t^n$  is the empirical pdf. The **style guidance regularizer** encourages style separation by **forcing the score network to explain part of the data using only style**, thereby implicitly **discouraging it from encoding style information redundantly in the content branch**; The **balancing loss**, defined in Appendix D, helps prevent poor local minima. For general latent variable models, this is a heuristic regularization rather than a formal MI minimization, and we will add a clearer explanation and acknowledge. The following theorem analyzes training dynamics by studying the *gradient flow*:

$$[\dot{U}, \dot{V}] = [-\nabla_U L_n^{\lambda_r}(\theta), -\nabla_V L_n^{\lambda_r}(\theta)], \quad [\dot{\theta}_Z, \dot{\theta}_G] = [-\nabla_{\theta_Z} L_n^{\lambda_r}(\theta), -\nabla_{\theta_G} L_n^{\lambda_r}(\theta)]. \quad (12)$$

**Theorem 4.5.** Under Assumption 3.5-3.6, for some  $t_0$  dependent on  $n$  and let  $\min\{d_T, d_H\} \rightarrow \infty$ , and let the positional encoding  $\text{PE}(\cdot)$  be bounded and linearly independent over  $t \in [t_0, T]$ . Define  $P_M$  to be the projection matrix onto  $R(M)$ , and  $\sigma_i(s)$  to be the  $i$ -th largest singular value of the operator  $s$ . Then for some  $\lambda_r$ , Eq. 12 converges to a critical point  $\hat{\theta} := [\hat{\theta}_Z, \hat{\theta}_G, \hat{U}, \hat{V}]$  such that with probability at least  $1 - O(\frac{1}{n})$ : 1) Encodings  $\hat{Z} := P_{\hat{U}} X$  and  $\hat{G} := g(X)$  are  $\left(O\left(\frac{d_X^{5/4} \log^{3/4} n}{\sigma_{d_Z}(s_Z^*) n^{1/4}}\right), 0\right)$ -disentangled; 2) Let  $\hat{Z}_{t_0} := \hat{Z} + \sigma(t_0) N_{t_0}$ . Then  $(\hat{Z}_{t_0}, \hat{G})$  are  $O\left(\frac{d_X^{7/4} \log^{9/16} n}{T \min\{\sigma_{d_Z}^{1/2}(s_Z^*), \sigma_{d_G}^{1/2}(s_G^*)\} n^{1/16}}\right)$ -editable.

**Intuition.** Theorem 4.5 shows that for ISMs, a DM trained with gradient descent can recover the content and style subspaces as the number of samples  $n \rightarrow \infty$ . This enables both approximate  $(\epsilon, \nu)$ -disentanglement and  $\epsilon'$ -editability. A key component is the **style guidance regularizer**, which encourages the model to separate content and style by encouraging the model to use the style information during **score matching**. Importantly, the speed and quality of subspace recovery depend on the strength of the signal in the score function. In particular, recovery is faster when the content and style score functions have larger minimum singular value. This highlights a practical consideration: well-separated or high-contrast latent factors lead to more reliable disentanglement, while near-degenerate cases may require additional regularization or supervision.

Theorem 4.5 extends to the *multi-view* setting using the dual encoder network  $s_{\theta, \phi}(X_t^1, t) = U s_Z(X_t^2, t) + V s_G(X_t^1, t)$ , where  $s_Z$  extracts content from the second view. In this case, we use a *content guidance loss*  $L'_{r,n}$  applied to  $(U, s_Z)$  instead of  $(V, s_G)$ , to encourage reliance on content information from the second view and suppress residual content in the style encoder  $s_G$ . Our bound depends on the data dimension  $d_X$ , but extensions to lower-dimensional latent representations are

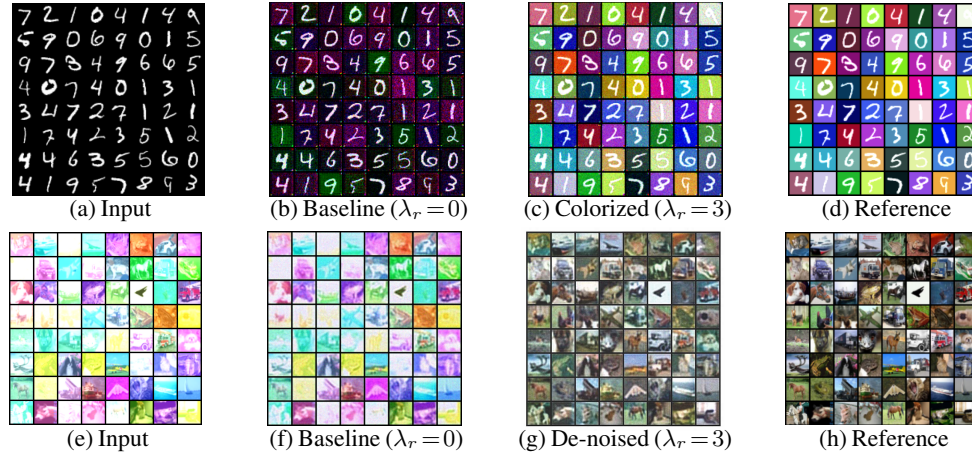


Figure 3: **Disentanglement results on MNIST and CIFAR-10.** Top: **Disentanglement results on MNIST.** The content is the gray-scale digit image, and the style is the background color. Bottom: **Disentanglement results on CIFAR10.** The content is the clean image and the style is the corruption on the image. Both 3b-3c and 3f-3g suggest disentanglement is achieved with the style guidance loss.

Table 1: **Quantitative results for image denoising and face editing on CelebA test sets for various models.** Conditional (cond.) InfoGAN and cond.  $\beta$ -VAE stand for InfoGAN [69] and  $\beta$ -VAE [70] adapted to the content disentanglement setting by conditioning on the observed style variable. DDPM stands for denoising diffusion probabilistic model (DDPM) [33].

|                                 | Denoising             |                    | Face Editing             |                       |
|---------------------------------|-----------------------|--------------------|--------------------------|-----------------------|
|                                 | LPIPS( $\downarrow$ ) | SSIM( $\uparrow$ ) | CLIP Score( $\uparrow$ ) | LPIPS( $\downarrow$ ) |
| cond. InfoGAN                   | 0.83                  | 0.16               | 0.24                     | 0.39                  |
| cond. $\beta$ -VAE              | 0.57                  | 0.17               | 0.25                     | 0.37                  |
| DDPM (ours), $\lambda_r = 0$    | 0.53                  | 0.18               | 0.25                     | 0.35                  |
| DDPM (ours), $\lambda_r = 0.03$ | 0.47                  | 0.19               | 0.26                     | <b>0.30</b>           |
| DDPM (ours), $\lambda_r = 0.3$  | 0.35                  | 0.23               | 0.26                     | 0.32                  |
| DDPM (ours), $\lambda_r = 3$    | <b>0.33</b>           | <b>0.24</b>        | <b>0.26</b>              | 0.38                  |

possible via residual connections in the score network in Figure 2. While our focus is on unconditional score matching, our results extend to conditional settings with appropriate corrections for cross terms. Lastly, although our analysis assumes infinite-width two-layer ReLU networks, similar convergence behavior may hold for deeper or finite-width networks [68].

## 5 Experiments

This section presents empirical evaluation of our theoretical framework by testing whether DMs can achieve approximate disentanglement under settings in Section 4.1-4.3. We begin with synthetic datasets generated from Gaussian mixture models (GMM), which instantiate the ISM framework from Section 4.3 to validate guarantees in Theorem 4.5. We then move to more realistic settings using standard image datasets. Specifically, we apply our DM-based disentanglement method to two tasks: image colorization on MNIST [71] and image denoising on CIFAR10 [72]. Finally, we validate Theorem 4.1-4.2 on a real-world speech task, voice conversion (VC) adaptation, by adopting the DM-based VC framework from [38].

**Implementation details.** For the GMM dataset, we use a two-layer ReLU network consistent with Theorem 4.5. For MNIST and CIFAR, we use a U-Net [73] architecture, following common DM design [74]. We optimize a regularized score-matching objective, inspired by Eq. 11:

$$L_c^{\lambda_r}(\theta, \phi) := \mathbb{E}_{t, p_0 p_{t_0} | p_t | t_0} \left\| s_{\theta}(X_t, \hat{Z}, G, t) + \frac{N_t}{\sigma(t)} \right\|^2 + \lambda_r \mathbb{E}_{t, p_t} \left\| s_{\theta}(X_t, \mathbf{0}_{d_Z}, G, t) + \frac{N_t}{\sigma(t)} \right\|^2, \quad (13)$$

where  $\lambda_r$  controls the **style guidance weight**. Note that we omit the balancing term for empirical simplicity, and style guidance is applied by zeroing out content, equivalent to the style regularizer in Eq. 11 when the input data has ISM structure.

**GMM disentanglement results.** First, we conduct subspace recovery experiments on *latent subspace GMMs* (LS-GMM), a class of LSMs where each subspace follows a GMM. Figure 4 shows the subspace recovery error as a function of the style guidance weight  $\lambda_r$  and the sample size  $n$ . Consistent with the predictions of Theorem 4.5, the LSGMM achieves the smallest subspace reconstruction error when the style guidance weight is sufficiently large, and the result is consistent across different noise schedulers. Moreover, since all score networks are wide, two-layer MLPs trained using gradient-based methods, these results provide further support for Theorem 4.5. Figure 4 also reveals a sublinear decay rate of the subspace recovery error as the sample size increases, aligning with Theorem 4.5.

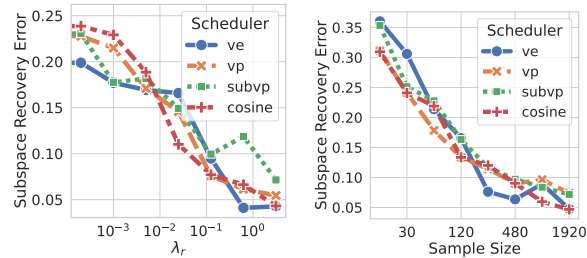


Figure 4: **GMM disentanglement results with the score estimator in Eq. 10.** The subspace recovery error (defined in Appendix E.1) is normalized between  $[0,1]$ . In four random trials, DM consistently recovers (error  $< 0.1$ ) the correct content subspace and achieves disentanglement with sufficiently large style guidance weight  $\lambda_r$  and sample size.

**Image disentanglement results.** Next, we validate our theoretical findings on image data from MNIST and CIFAR-10. For MNIST, we perform *colorization*, treating digit shape as content and background color as style. The results are visualized in Figure 3. Without regularization, the model fails to disentangle these factors, simply copying the input (Figure 3b). With the proposed regularizer, it successfully disentangles color from shape (Figure 3c). On CIFAR10, we test image *denoising* where the content is the clean image and the style is an independent noise. In this setup, we introduce a random color shift as the noise, though our approach can, in principle, be extended to other types of independent noise. The results show similar improvements with regularization, consistent with Theorem 4.5. We have further verified our theory on two more disentanglement tasks on the CelebA image dataset. First, we perform denoising with the same setting as CIFAR 10 and compared our approach with two standard baselines, infoGAN [69] and beta-VAE [70] adapted to the content disentanglement setting by conditioning on the observed style variable. As shown in the first two columns of Table 1, we observe the same trend as our previous experiments. Further, we perform face editing with *gender* as the style variable. As shown in the last two columns of Table 1, our method is superior to the conditional GAN baseline in terms of style transfer (CLIP score) and content preservation (LPIPS [75] score). Additional examples and quantitative results using different regularization weights for multiple metrics are provided in Appendix E.2.

**Speech disentanglement results.** Lastly, we apply DM-based disentanglement to a real-world application: *voice conversion adaptation* (VCA). In this setup, the style corresponds to speaker identity, while the content captures attributes like emotional state or health condition. Our goal is to learn representations that disentangle these factors, enabling robust classification under speaker shifts. We use speech emotion recognition on the IEMOCAP dataset [76] as a testbed, where generalization across unseen speakers serves as a proxy for successful disentanglement. This task instantiates *multi-view* disentanglement (Definition 3.4): different speakers provide style views, and the shared emotion across recordings is the content. We are also actively developing additional vision-domain multi-view experiments.

As shown in Figure 5, our DM-based approach achieves higher classification accuracy than

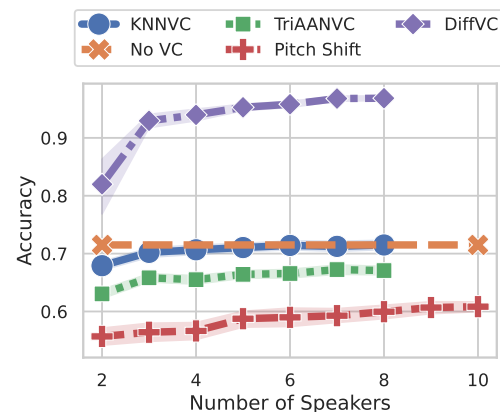


Figure 5: **Emotion recognition results on IEMOCAP as a probing task for speech disentanglement.** DM-based disentanglement between emotion (content) and speaker (style) outperforms other methods. Data augmentation using multiple speakers further improves disentanglement.

several baselines, including no conversion, pitch shifting, and recent VCA models [65, 77]. Performance improves as more target speakers are used during training, consistent with our theoretical prediction of Theorem 4.1 that DMs can indeed achieve approximate disentanglement for speech data. It also validates Theorem 4.2 by demonstrating that multi-style data enhances content-style disentanglement. Further, more speech metrics and additional experiments beyond emotion classification (e.g., Alzheimer detection [78], Amyotrophic Lateral Sclerosis severity [79]) are detailed in Appendix E.3. Further, we also note that our current experiments focus on controlled settings aligned with the theory. Extensions to large-scale experimentation are left as an exciting future avenue of research.

## 6 Related works

**Diffusion model theory.** Early theoretical works on DMs analyzed their ability to learn data distributions, under different assumptions like log-Sobolev inequality [59], bounded moments [80, 42] and score approximation in  $L^\infty$  [81] and  $L^2$  [59, 42] norms. Later works compared DMs with likelihood-based models [82], and studied their capability to recover Gaussian mixtures [83–86], Ising models [87], low-dimensional subspaces [41, 66, 88, 89] and manifold structures [90]. Recently, [91, 92] analyzed the convergence of conditional DMs and the role of classifier-free guidance. Others have analyzed the training [93–95] and sampling [95–97] dynamics of DMs.

**Disentangled representation learning.** Disentanglement is defined via factorized representation [98, 21–23], group equivariance [99] or approximate independence [100]. It underpins many advances in self-supervised learning [52, 54, 101, 56], multimodal learning [102–104] and controllable generation [49, 38, 35, 36, 105]. Theoretically, disentanglement traces back to independent component analysis (ICA) [106], later extended to correlated factors [48, 29] and studied through the lens of modern SSL techniques such as data augmentation, contrastive loss [107] and self-distillation [108]. A key result [26] shows that unsupervised disentanglement is impossible without additional inductive biases, leading to weakly supervised methods using multiple views [29, 30], auxiliary labels [27, 109, 28], temporal cues [110, 111] and isometric constraints [112]. Other works analyze the disentangling capacity of (variational) auto-encoders [49, 113–116, 100, 117, 118].

## 7 Conclusion

We presented the first learning-theoretic framework for DM-based disentanglement, addressing unsolved challenges of prior works that focused primarily on deterministic, invertible mixing processes. Our framework introduces two notions of approximate disentanglement that generalize classical formulations to stochastic, non-invertible settings and shows how DMs can achieve them with partial labels or multi-view inputs. Moreover, in the special case of ISMs, we derive stronger guarantees, including finite-sample global convergence for gradient-based training. Our experiments across several domains, spanning Gaussian mixture recovery, image colorization, denoising, and voice conversion, show that theory-guided training methods, such as style-guidance regularization, lead to improved disentanglement and better downstream performance. Our framework lays the foundation for principled disentanglement with DMs and future works on more complex latent structures and modalities.

## Acknowledgments and Disclosure of Funding

This research was supported by Takeda Development Center Americas, INC. (successor in interest to Millennium Pharmaceuticals, INC.) This research could not be possible without the immense contributions from people living with ALS who have participated for months and years in the ALS Research Collaborative Study. We thank the ALS patient and caregiver community, in particular Augie’s Quest for a Cure, for financially supporting ALS TDI’s data collection work.

## References

- [1] Y. A. Farha, A. Richard, and J. Gall, “When will you do what? anticipating temporal occurrences of activities,” in *CVPR*, pp. 5343–5352, 2018.
- [2] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *ECCV*, pp. 649–665, 2018.

- [3] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in *ECCV*, September 2018.
- [4] M. Hamilton, A. Zisserman, J. R. Hershey, and W. T. Freeman, “Separating the “chirp” from the “chat”: Self-supervised visual grounding of sound and language,” in *CVPR*, pp. 13117–13127, 2024.
- [5] E. Araujo, A. Rouditchenko, Y. Gong, S. Bhati, S. Thomas, B. Kingsbury, L. Karlinsky, R. Feris, J. R. Glass, and H. Kuehne, “CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment,” in *CVPR*, (Nashville, USA), 2025.
- [6] C. Wunsch, *The Ocean Circulation Inverse Problem*. Cambridge University Press, 1996.
- [7] R. Vigário, V. Jousmäki, M. Hämläinen, R. Hari, and E. Oja, “Independent component analysis for identification of artifacts in magnetoencephalographic recordings,” in *Neural Information Processing System*, vol. 10, 1997.
- [8] G. D. Brown, S. Yamada, and T. J. Sejnowski, “Independent component analysis at the neural cocktail party,” *Trends in Neurosciences*, vol. 24, no. 1, pp. 54–63, 2001.
- [9] M. S. Zhdanov, *Geophysical Inverse Theory and Regularization Problems*. Elsevier, 2002.
- [10] W. Choi, C. Fang-Yen, K. Badizadegan, S. Oh, N. Lue, R. R. Dasari, and M. S. Feld, “Tomographic phase microscopy,” *Nature Methods*, vol. 4, pp. 717–719, Sept. 2007.
- [11] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse mri: The application of compressed sensing for rapid mr imaging,” *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, Dec. 2007.
- [12] J. Virieux and S. Operto, “An overview of full-waveform inversion in exploration geophysics,” *Geophysics*, vol. 74, no. 6, pp. WCC1–WCC26, 2009.
- [13] K. Chadon and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*. Springer Science & Business Media, 1 ed., 2012.
- [14] M. A. Iglesias, K. J. Law, and A. M. Stuart, “Ensemble kalman methods for inverse problems,” *Inverse Problems*, vol. 29, no. 4, p. 045001, 2013.
- [15] A. Chael, K. Bouman, M. Johnson, M. Wielgus, L. Blackburn, C.-K. Chan, J. R. Farah, D. Palumbo, and D. Pesce, “eht-imaging: v1.1.0: Imaging interferometric data with regularized maximum likelihood.” <https://doi.org/10.5281/zenodo.2562874>, 2019. Software.
- [16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.
- [17] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, and W. Manassra, “Improving image generation with better captions.” <https://cdn.openai.com/papers/dall-e-3.pdf>, 2023. Accessed: 2025-05-09.
- [18] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, “Audiobox: Unified audio generation with natural language prompts,” 2023.
- [19] OpenAI, “Sora: Creating video from text,” 2024. Accessed: 2025-05-09.
- [20] B. Seed Team, “Seed-TTS: A family of high-quality versatile speech generation models,” *arXiv preprint arXiv:2406.02430*, 2024.
- [21] P. Comon, “Independent component analysis, a new concept?,” *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [22] A. Hyvärinen and P. Pajunen, “Nonlinear independent component analysis: Existence and uniqueness results,” *Neural Networks*, vol. 12, pp. 429–439, April 1999.

- [23] Y. Bengio and Y. LeCun, “Scaling learning algorithms towards AI,” in *Large Scale Kernel Machines*, MIT Press, 2007.
- [24] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning basic visual concepts with a constrained variational framework,” in *International Conference on Learning Representations*, 2017.
- [25] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, “Toward causal representation learning,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
- [26] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, PMLR, Jun 2019.
- [27] J. Chen and K. Batmanghelich, “Weakly supervised disentanglement by pairwise similarities,” in *AAAI*, vol. 34, pp. 3495–3502, 2020.
- [28] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, “Weakly supervised disentanglement with guarantees,” in *International Conference on Learning Representations*, 2020.
- [29] I. Daunhawer, A. Bizeul, E. Palumbo, A. Marx, and J. E. Vogt, “Identifiability results for multimodal contrastive learning,” in *International Conference on Learning Representations*, 2023.
- [30] D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello, “Multi-view causal representation learning with partial observability,” in *International Conference on Learning Representations*, 2024.
- [31] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2256–2265, PMLR, 7 2015.
- [32] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” in *Neural Information Processing System*, 2019.
- [33] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Neural Information Processing System*, vol. 33, pp. 6840–6851, 2020.
- [34] R. A. Fisher, “The detection of linkage with ‘dominant’ abnormalities,” *Annals of Eugenics*, vol. 6, no. 2, pp. 187–201, 1935.
- [35] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, “Uncovering the disentanglement capability in text-to-image diffusion models,” in *CVPR*, pp. 1900–1910, 2023.
- [36] T. Yang, Y. Wang, Y. Lu, and N. Zheng, “Disdiff: Unsupervised disentanglement of diffusion probabilistic models,” in *Neural Information Processing System*, 2023.
- [37] D. A. Hudson, D. Zoran, M. Malinowski, A. K. Lampinen, A. Jaegle, J. L. McClelland, L. Matthey, F. Hill, and A. Lerchner, “Soda: Bottleneck diffusion models for representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. to appear, June 2024.
- [38] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, “Diffusion-based voice conversion with fast maximum likelihood sampling scheme,” in *International Conference on Learning Representations*, 2022.
- [39] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, “Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation,” in *Proc. INTERSPEECH 2023*, pp. 2283–2287, 2023.

- [40] A. Hyvärinen and P. Pajunen, “Nonlinear independent component analysis: Existence and uniqueness results,” *Neural Networks*, vol. 12, pp. 429–439, Apr. 1999.
- [41] M. Chen, K. Huang, T. Zhao, and M. Wang, “Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data,” in *International Conference on Machine Learning*, 2023.
- [42] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. Zhang, “Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions,” in *International Conference on Learning Representations*, 2023.
- [43] B. D. Anderson, “Reverse-time diffusion equation models,” *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.
- [44] C. Eastwood and C. K. I. Williams, “A framework for the quantitative evaluation of disentangled representations,” in *International Conference on Learning Representations*, 2018.
- [45] M. A. Casey and A. Westner, “Separation of mixed audio sources by independent subspace analysis,” in *Proceedings of the International Computer Music Conference (ICMC)*, pp. 154–161, 2000.
- [46] A. Hyvärinen and P. Hoyer, “Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces,” *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.
- [47] F. Theis, “Towards a general independent subspace analysis,” in *Neural Information Processing System*, vol. 19, pp. 1361–1368, 2006.
- [48] J. V. Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, “Self-supervised learning with data augmentations provably isolates content from style,” in *Neural Information Processing System* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.
- [49] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “AutoVC: Zero-shot voice style transfer with only autoencoder loss,” in *International Conference on Machine Learning*, pp. 5210–5219, 2019.
- [50] F. Huang, K. Zeng, and W. Zhu, “DiffVC+: Improving diffusion-based voice conversion for speaker anonymization,” in *Interspeech*, 2024.
- [51] B. Desplanques, J. Thienpondt, and K. Demuyne, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” in *Interspeech*, 2020.
- [52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR, 13–18 Jul 2020.
- [53] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent: A new approach to self-supervised learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 21271–21284, 2020.
- [54] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.
- [55] X. Chen and K. He, “Exploring simple siamese representation learning,” in *CVPR*, (Nashville, TN, USA), pp. 15745–15753, 2021.
- [56] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

- [57] L. Logeswaran and H. Lee, “An efficient framework for learning sentence representations,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [58] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pretraining for speech recognition,” in *Proceedings of Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pp. 3465–3469, 2019.
- [59] H. Lee, J. Lu, and Y. Tan, “Convergence for score-based generative modeling with polynomial complexity,” in *Neural Information Processing System* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.
- [60] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, “Mutual information neural estimation,” in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 531–540, PMLR, 2018.
- [61] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [62] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, “Club: A contrastive log-ratio upper bound of mutual information,” in *International Conference on Machine Learning*, pp. 1779–1788, PMLR, 2020.
- [63] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, IEEE, 2011.
- [64] O. D. Liu, H. Tang, and S. Goldwater, “Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces,” in *Interspeech*, 2023.
- [65] M. Baas, B. van Niekerc, and H. Kamper, “Voice conversion with just nearest neighbors,” in *Interspeech*, 2023.
- [66] L. Rout, N. Raoof, G. Daras, C. Caramanis, A. Dimakis, and S. Shakkottai, “Solving linear inverse problems provably via posterior sampling with latent diffusion models,” in *Neural Information Processing System*, 2023.
- [67] H. Kamper, B. van Niekerc, J. Zaïdi, and M.-A. Carbonneau, “Linearvc: Linear transformations of self-supervised features through the lens of voice conversion,” in *Interspeech*, 2025.
- [68] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, “Gradient descent finds global minima of deep neural networks,” in *International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685, PMLR, 09–15 Jun 2019.
- [69] X. Chen, Y. Duan, R. Houthoof, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Neural Information Processing System, NIPS’16*, (Red Hook, NY, USA), pp. 2180–2188, Curran Associates Inc., 2016.
- [70] H. Kim and A. Mnih, “Disentangling by factorising,” in *International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, PMLR, 10–15 Jul 2018.
- [71] L. Deng, “The mnist database of handwritten digit images for machine learning research,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
- [72] A. Krizhevsky, V. Nair, and G. Hinton, “Learning multiple layers of features from tiny images.” <http://www.cs.toronto.edu/~kriz/cifar.html>, 2009. Accessed: 2025-01-25.
- [73] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

- [74] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *International Conference on Learning Representations*, 2021.
- [75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018.
- [76] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.-N. Chang, S. Lee, and S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” in *Language Resources and Evaluation (LREC)*, pp. 335–339, 2008.
- [77] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, “Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion,” in *ICASSP*, 2023.
- [78] S. Luz, F. Haider, S. de la Fuente, and B. M. Davida Fromm, “Alzheimer’s dementia recognition through spontaneous speech: The adress challenge,” in *Proceedings of Interspeech 2020*, pp. 2172–2176, 2020.
- [79] F. Vieira, S. Venugopalan, A. Premasiri, *et al.*, “A machine-learning based objective measure for ALS disease severity,” *NPJ Digital Medicine*, vol. 5, p. 45, 2022.
- [80] A. Block, Y. Mroueh, and A. Rakhlin, “Generative modeling with denoising auto-encoders and langevin sampling,” *arXiv preprint arXiv:2002.00107*, 2020.
- [81] V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet, “Diffusion schrödinger bridge with applications to score-based generative modeling,” in *Neural Information Processing System* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 17695–17709, Curran Associates, Inc., 2021.
- [82] C. Pabbaraju, D. Rohatgi, A. Sevekari, H. Lee, A. Moitra, and A. Risteski, “Provable benefits of score matching,” in *Neural Information Processing System*, 2023.
- [83] K. Shah, S. Chen, and A. Klivans, “Learning mixtures of gaussians using the ddpm objective,” in *Neural Information Processing System*, 2023.
- [84] Y. Wu, M. Chen, Z. Li, M. Wang, and Y. Wei, “Theoretical insights for diffusion guidance: A case study for gaussian mixture models,” *arXiv preprint arXiv:2403.01639*, 2024.
- [85] S. Chen, V. Kontonis, and K. Shah, “Learning general gaussian mixtures with efficient score matching,” *arXiv preprint arXiv:2404.18893*, 2024.
- [86] G. Li, C. Cai, and Y. Wei, “Dimension-free convergence of diffusion models for approximate gaussian mixtures,” *arXiv preprint arXiv:2504.05300*, 2025.
- [87] S. Mei and Y. Wu, “Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models,” *arXiv preprint arXiv:2309.11420*, 2023.
- [88] J. Y.-C. Hu, W. Wu, Z. Li, S. Pi, Z. Song, and H. Liu, “On statistical rates and provably efficient criteria of latent diffusion transformers (dits),” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [89] P. Wang, H. Zhang, Z. Zhang, S. Chen, Y. Ma, and Q. Qu, “Diffusion models learn low-dimensional distributions via subspace clustering,” *arXiv preprint arXiv:2409.02426*, 2024.
- [90] V. D. Bortoli, A. Thornton, J. H. Hutchinson, J. A. Mathieu, Y. Teh, and A. Doucet, “Convergence of denoising diffusion models under the manifold hypothesis,” *Transactions on Machine Learning Research (TMLR)*, 2022.
- [91] M. Chidambaram, K. Gatmiry, S. Chen, H. Lee, and J. Lu, “What does guidance do? a fine-grained analysis in a simple setting,” in *Neural Information Processing System*, 2024.
- [92] H. Fu, Z. Yang, M. Wang, and M. Chen, “Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory,” *ArXiv*, vol. abs/2403.11968, 2024.

- [93] P. Li, Z. Li, H. Zhang, and J. Bian, “On the generalization properties of diffusion models,” in *Neural Information Processing System*, vol. 36, pp. 2097–2127, 2023.
- [94] A. Han, W. Huang, Y. Cao, and D. Zou, “On the feature learning in diffusion models,” in *International Conference on Learning Representations*, 2025.
- [95] Y. Wang, Y. He, and M. Tao, “Evaluating the design space of diffusion-based generative models,” *arXiv preprint arXiv:2406.12839*, 2024.
- [96] G. Li, Y. Wei, Y. Chi, and Y. Chen, “A sharp convergence theory for the probability flow odes of diffusion models,” *arXiv preprint arXiv:2408.02320*, 2024.
- [97] M. Li and S. Chen, “Critical windows: Non-asymptotic theory for feature emergence in diffusion models,” in *International Conference on Machine Learning*, pp. 27474–27498, 2024.
- [98] J. Schmidhuber, “Learning factorial codes by predictability minimization,” *Neural Computation*, vol. 4, no. 6, pp. 863–879, 1992.
- [99] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, “Towards a definition of disentangled representations,” *arXiv preprint arXiv:1812.02230*, 2018.
- [100] Z. Pan, L. Niu, J. Zhang, and L. Zhang, “Disentangled information bottleneck,” in *AAAI*, 2021.
- [101] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139, pp. 12310–12320, PMLR, 2021.
- [102] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multi-modal transformer for unaligned multimodal language sequences,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 6558–6569, Association for Computational Linguistics, July 2019.
- [103] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic, “Geometric multimodal contrastive representation learning,” in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 17782–17800, PMLR, 17–23 Jul 2022.
- [104] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, “Multimae: Multi-modal multi-task masked autoencoders,” in *Computer Vision – ECCV 2022*, pp. 348–367, Springer Nature Switzerland, 2022.
- [105] J. Hahm, J. Lee, S. Kim, and J. Lee, “Isometric representation learning for disentangled latent space of diffusion models,” in *International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, pp. 17224–17245, PMLR, 2024.
- [106] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.
- [107] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel, “Contrastive learning inverts the data generating process,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990, PMLR, 2021.
- [108] Y. Tian, X. Chen, and S. Ganguli, “Understanding self-supervised learning dynamics without contrastive pairs,” in *International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139, pp. 10268–10278, PMLR, 2021.
- [109] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, “Weakly-supervised disentanglement without compromises,” in *International Conference on Machine Learning*, vol. 119, pp. 6348–6359, PMLR, 2020.
- [110] A. Hyvärinen and H. Morioka, “Unsupervised feature extraction by time-contrastive learning and nonlinear ICA,” in *Neural Information Processing System*, pp. 3765–3773, 2016.

- [111] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen, “ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 10235–10246, 2020.
- [112] D. Horan, E. Richardson, and Y. Weiss, “When is unsupervised disentanglement possible?,” in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.
- [113] M. Rolinek, D. Zietlow, and G. Martius, “Variational autoencoders pursue pca directions (by accident),” in *CVPR*, pp. 12406–12415, 2019.
- [114] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, “Disentangling disentanglement in variational autoencoders,” in *International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 4402–4412, PMLR, 2019.
- [115] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen, “Variational autoencoders and nonlinear ICA: A unifying framework,” in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2207–2217, PMLR, 2020.
- [116] X. Bao, J. Lucas, S. Sachdeva, and R. Grosse, “Regularized linear autoencoders recover the principal components, eventually,” in *Neural Information Processing System*, pp. 6971–6981, 2020.
- [117] P. Bhowal, A. Soni, and S. Rambhatla, “Why do variational autoencoders really promote disentanglement?,” in *Forty-first International Conference on Machine Learning*, 2024.
- [118] T. Uscidda, L. Eyring, K. Roth, F. J. Theis, Z. Akata, and marco cuturi, “Disentangled representation learning with the gromov-monge gap,” in *International Conference on Learning Representations*, 2025.
- [119] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *International Conference on Learning Representations*, 2018.
- [120] C. Anil, J. Lucas, and R. Grosse, “Sorting out lipschitz function approximation,” in *International Conference on Machine Learning*, pp. 291–301, PMLR, 2019.
- [121] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, “Regularisation of neural networks by enforcing lipschitz continuity,” *Machine Learning*, vol. 110, no. 2, pp. 393–416, 2021.
- [122] J.-F. Le Gall, *Brownian Motion, Martingales, and Stochastic Calculus*, vol. 274 of *Graduate Texts in Mathematics*. Springer, 2016.
- [123] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Neural Information Processing System*, 2018.
- [124] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.
- [125] R. Ge, C. Jin, and Y. Zheng, “No spurious local minima in nonconvex low rank problems: A unified geometric analysis,” in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1233–1242, PMLR, 06–11 Aug 2017.
- [126] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, “Gradient descent only converges to minimizers,” in *Conference on Learning Theory*, 2016.
- [127] B. Hajek and M. Raginsky, “ECE 543: Statistical Learning Theory.” <http://maxim.ece.illinois.edu/teaching/SLT/>, 2021. Accessed: 2025-01-25.
- [128] R. Vershynin, *High-Dimensional Probability—An Introduction with Applications in Data Science*. Cambridge University Press, Sept. 2018.
- [129] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*, 2023.

- [130] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024.
- [131] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
- [132] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, 2021.
- [133] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Neural Information Processing System*, 2020.
- [134] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, T. Gregory, Z. Lin, N. Gimselshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.
- [135] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperriere, M. Rouvier, R. D. Mori, and Y. Esteve, "Open-source conversational ai with SpeechBrain 1.0," 2024.
- [136] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Neural Information Processing System*, 2020.
- [137] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.
- [138] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Interspeech*, 2024.
- [139] L. Wang, Y. Gong, N. Dawalatabad, M. Vilela, K. Placek, B. Tracey, Y. Gong, A. Premasiri, F. Vieira, and J. Glass, "Automatic prediction of amyotrophic lateral sclerosis progression using longitudinal speech transformer," in *Interspeech*, 2024.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Yes, the main claims in the abstract and introduction are an accurate reflection of the paper's contributions and scope. The stated goals and innovations are clearly substantiated in the theoretical sections (Sec. 4.1-4.3) and experimental sections (Sec. 5), and there is no overstatement or misalignment between the framing and the actual content. The abstract provides a faithful summary, and the introduction sets realistic expectations for what the paper achieves.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss limitations in the discussion of each section throughout the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full set of assumptions and a complete and correct proof for each theorem are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All information needed to produce the main experimental results are provided in the paper and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and code will be released upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the training and test details are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We average results over multiple runs and report error bars or other statistical significance metrics for the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The details about computation resources are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research is ethical according to the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper presents a foundational research not tied to particular applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are fully licensed.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: All assets are well-documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper has no such risk.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not used for non-standard purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## Contents

|                                                            |           |
|------------------------------------------------------------|-----------|
| <b>A Proof of Example 3.1</b>                              | <b>27</b> |
| <b>B Proof of Theorem 4.1</b>                              | <b>27</b> |
| B.1 Main proof . . . . .                                   | 27        |
| B.2 Proof of Lemma B.2 . . . . .                           | 29        |
| B.3 Proof of Lemma B.3 . . . . .                           | 31        |
| B.4 Proof of Lemma B.4 . . . . .                           | 32        |
| B.5 Proof of Lemma B.5 . . . . .                           | 32        |
| B.6 Proof of Lemma B.6 . . . . .                           | 32        |
| B.7 Extension to correlated content and style . . . . .    | 33        |
| <b>C Proof of Theorem 4.2</b>                              | <b>33</b> |
| C.1 Main proof . . . . .                                   | 33        |
| C.2 Proof of Lemma C.5 . . . . .                           | 34        |
| C.3 Proof of Lemma C.6 . . . . .                           | 36        |
| C.4 Extension to correlated view-specific styles . . . . . | 36        |
| <b>D Proof of Theorem 4.5</b>                              | <b>36</b> |
| D.1 Main proof . . . . .                                   | 37        |
| D.2 Proof of Lemma 4.4 . . . . .                           | 41        |
| D.3 Proof of Theorem D.2 . . . . .                         | 41        |
| D.4 Proof of Theorem D.3 . . . . .                         | 42        |
| D.5 Proof of Lemma D.5 . . . . .                           | 44        |
| D.6 Proof of Lemma D.6 . . . . .                           | 46        |
| D.7 Proof of Lemma D.7 . . . . .                           | 47        |
| D.8 Proof of Lemma D.8 . . . . .                           | 47        |
| D.9 Proof of Lemma D.4 . . . . .                           | 48        |
| D.10 Proof of Lemma D.9 . . . . .                          | 52        |
| D.11 Proof of Lemma D.10 . . . . .                         | 53        |
| <b>E Experiment details</b>                                | <b>53</b> |
| E.1 Latent subspace GMM disentanglement . . . . .          | 53        |
| E.2 Image disentanglement . . . . .                        | 54        |
| E.3 Speech disentanglement . . . . .                       | 56        |

## A Proof of Example 3.1

**Example A.1.** (Example 3.1, restated) Suppose content  $Z \sim \mathcal{N}(0,1)$  and  $G \sim \mathcal{N}(0,1)$  are independent standard Gaussian r.v.'s and consider the noiseless setting with  $\delta=0$ . Further, suppose  $p_{f(Z,G)|Z} \neq p_{f(-Z,G)|Z}$ . Then we can choose the content/style encodings to be  $\hat{Z} = Z\text{sgn}(G)$ ,  $\hat{G} = G$  and the decoder  $\hat{f}(\hat{Z}, \hat{G}) = f(\hat{Z}\text{sgn}(\hat{G}), \hat{G})$ , where  $\text{sgn}(x)$  denotes the sign of  $x$ . Then  $(\hat{Z}, \hat{G})$  are  $(0,0)$ -disentangled but not  $\epsilon$ -editable for some  $\epsilon > 0$ .

*Proof.* First of all, notice that  $\hat{Z} \sim \mathcal{N}(0,1)$  regardless of the value of  $\hat{G}$  owing to the even symmetry of the standard Gaussian distribution, which implies  $\hat{Z} \perp\!\!\!\perp \hat{G}$ , or equivalently  $I(\hat{Z}; \hat{G}) = 0$ . Second, by construction,  $\hat{f}(\hat{Z}, \hat{G}) = f(Z\text{sgn}^2(G), G) = f(Z, G) = X$ , and thus  $I(\hat{Z}, \hat{G}; X) - I(Z, G; X) = 0$ . Therefore by definition,  $(\hat{Z}, \hat{G})$  are  $(0,0)$ -disentangled. However, for an i.i.d sample  $\hat{G}' \sim p_G$ , the conditional distribution  $p_{\hat{f}(\hat{Z}, \hat{G}')|Z} = \frac{1}{2}p_{f(-Z, \hat{G}')|Z} + \frac{1}{2}p_{f(Z, \hat{G}')|Z} \neq p_{f(Z, G)|Z}$  due to the fact that  $\text{sgn}(\hat{G})\text{sgn}(\hat{G}')$  is a symmetric Bernoulli variable and  $p_{f(Z, G)|Z} \neq p_{f(-Z, G)|Z}$ . Therefore,  $(\hat{Z}, \hat{G})$  is not  $\epsilon$ -editable for  $\epsilon = d_{\text{TV}}(p_{\hat{f}(\hat{Z}, \hat{G}')|Z}, p_{X|Z}) > 0$ .  $\square$

## B Proof of Theorem 4.1

For clarity, we restate the theorem below.

**Theorem B.1.** Suppose Assumption 3.5-3.6 hold, and  $(\theta^*, \phi^*)$  be a minimizer of  $L_c^{\gamma, \rho}$  defined in Eq. 7 with  $\rho = I(Z; X) + C_1\delta, \gamma \geq C_2/T$  for some  $C_1, C_2 > 0$ . Set  $t_0 = -\log(1-\delta^2)^{1/2}$  and  $\hat{Z} := z_{\phi^*}(X)$ . Then for any  $\delta < \min\left\{\frac{1}{2}, \frac{1}{\sqrt{d_X}}\right\}$ , the encodings  $(\hat{Z}, \hat{G}) = (z_{\phi^*}(X_{t_0}), g(X_{t_0}))$  are  $(\epsilon, \nu)$ -disentangled with  $\epsilon = C_3\lambda_s\sigma_X^2d_X\delta, \nu = C_4\sigma_X^2\delta^2$  for some constants  $C_3, C_4 > 0$ .

### B.1 Main proof

Our result relies crucially on the following lemma.

**Lemma B.2.** There exists  $(\theta_1, \phi_1)$  such that for

$$\delta < \min\left\{\frac{1}{2}, \frac{1}{\sqrt{d_X}}\right\}, t_0 = -\log(1-\delta^2)^{1/2}, t_1 = -\log(1-\delta)^{1/2},$$

the followings hold:

1. The followings hold for the MIs  $I(z_{\phi_1}(X_{t_0}); X), I(g(X_{t_0}); X)$  and  $I(z_{\phi_1}(X_{t_0}); g(X_{t_0})|X)$ :

$$\begin{aligned} I(z_{\phi_1}(X_{t_0}); X) &= I(Z; X) + O(\lambda_s\sigma_X d_X\delta), \\ I(g(X_{t_0}); X) &= I(G; X) + O(\lambda_s\sigma_X d_X\delta), \\ I(z_{\phi_1}(X_{t_0}); g(X_{t_0})|X) &= I(Z; G|X) + O(\lambda_s\sigma_X d_X\delta); \end{aligned}$$

2. The conditional score matching loss satisfies

$$L_c(\theta_1, \phi_1) \leq \frac{(1+\sigma_X^2\delta^2)\delta^2 d_X (e^{2T} - e^{2t_1})}{2(T-t_1)(e^{2t_1} - 1)(e^{2T} - 1)} = O\left(\frac{\delta d_X}{T}\right).$$

First, we assume Lemma B.2 to be true and defer its proof to Section B.2. Therefore, by the property of  $(\theta_1, \phi_1)$  and the optimality of  $(\theta^*, \phi^*)$ ,

$$\begin{aligned} L_c^{\gamma, \rho}(\theta^*, \phi^*) &= L_c(\theta^*, \phi^*) + \gamma(I(z_{\phi^*}(X_{t_0}); X) - \rho)_+ \\ &\leq L_c(\theta_1, \phi_1) + \gamma(I(z_{\phi_1}(X_{t_0}); X) - I(Z; X)) + \gamma(I(Z; X) - \rho)_+ \\ &\leq \frac{(1+\sigma_X^2\delta^2)\delta d_X}{2(T-t_1)} + C_1\gamma\delta + \gamma(I(Z; X) - \rho)_+, \end{aligned}$$

for some  $C_1 = O(\lambda_s \sigma_X d_X)$  and  $C_2 := 1 + \sigma_X^2 \delta^2$ . Since both  $L_c$  and  $I(Z; X)$  are nonnegative, this implies

$$L_c(\theta^*, \phi^*) \leq \frac{(1 + \sigma_X^2 \delta^2) \delta d_X}{2(T - t_1)} + C_1 \gamma \delta + \gamma(I(Z; X) - \rho)_+,$$

$$I(z_{\phi^*}(X_{t_0}); X) \leq I(Z; X) + \frac{(1 + \sigma_X^2 \delta^2) \delta d_X}{2\gamma(T - t_1)} + C_1 \delta.$$

Choose  $\gamma = \frac{1 + \sigma_X^2 \delta^2}{2(T - t_1)}$  and  $\rho = I(Z; X) + O(\delta)$ , then we have

$$I(z_{\phi^*}(X_{t_0}); X) \leq I(Z; X) + C'_1 \delta,$$

where  $C'_1 = O(\lambda_s \sigma_X d_X)$ .

Let  $h(X) := -\int_{\mathcal{X}} p(x) \log p(x) dx$  denote the differential entropy of continuous random variable  $X$ . Then by definition, we have

$$\begin{aligned} I(Z, G; X) &= h(Z, G) - h(Z, G|X) \\ &= h(Z) - h(Z|X) + h(G) - h(G|X) + I(Z; G|X) \\ &= I(Z; X) + I(G; X) + I(Z; G|X), \\ I(X_T, \hat{Z}, \hat{G}; X) &= I(\hat{Z}, \hat{G}; X) + I(X_T; X|\hat{Z}, \hat{G}) \\ &= h(\hat{Z}, \hat{G}) - h(\hat{Z}, \hat{G}|X) + I(X_T; X|\hat{Z}, \hat{G}) \\ &= I(\hat{Z}; X) + I(\hat{G}; X) + I(\hat{Z}; \hat{G}|X) - I(\hat{Z}; \hat{G}) + I(X_T; X|\hat{Z}, \hat{G}). \end{aligned}$$

As a result,

$$\begin{aligned} I(\hat{Z}; \hat{G}) &= I(\hat{Z}; X) + I(\hat{G}; X) + I(\hat{Z}, \hat{G}|X) + I(X_T; X|\hat{Z}, \hat{G}) - I(X_T, \hat{Z}, \hat{G}; X) \\ &= \underbrace{I(\hat{Z}; X) - I(Z; X)}_{(i)} + \underbrace{I(\hat{G}; X) - I(G; X)}_{(ii)} + \underbrace{I(\hat{Z}; \hat{G}|X) - I(Z; G|X)}_{(iii)} + \\ &\quad \underbrace{I(Z, G; X) - I(X_T, \hat{Z}, \hat{G}; X)}_{(iv)} + \underbrace{I(X_T; X|\hat{Z}, \hat{G})}_{(v)} \end{aligned} \quad (14)$$

where terms (i)(ii)(iii) can be upper bounded by item 1 as  $3C_1 \delta$ . To bound the term (iv), use the definition  $I(Z, G; X)$ :

$$I(Z, G; X) = h(X) - h(N) = h(X) - \frac{1}{2} \log(2\pi e \delta^2 d_X),$$

and apply the maximum entropy inequality on  $I(X_T, z(X_{t_0}), G; X)$ :

$$\begin{aligned} I(X_T, z(X_{t_0}), \hat{G}; X) &\geq h(X) - \frac{1}{2} \log 2\pi e \mathbb{E} \left\| (e^T - e^{-T}) s_{\theta^*}(X_T, \hat{Z}, \hat{G}, T) + e^T X_T - X \right\|^2 \\ &\geq h(X) - \frac{1}{2} \log 2\pi e^{2T+1} (1 - e^{-2T})^2 \lim_{t_1 \rightarrow T} L_c(\theta^*, \phi^*) \\ &\geq h(X) - \frac{1}{2} \log 2\pi e C_2 \delta^2 d_X = I(Z, G; X) - \frac{1}{2} \log C_2. \end{aligned}$$

The last inequality uses the optimality of  $(\theta^*, \phi^*)$  and therefore  $s_{\theta}(x, z, g, t)$ 's needs to achieve minimal loss at any  $t \in [t_1, T]$ , which is upper-bounded by item 2 of Lemma B.2 as

$$\lim_{t_1 \rightarrow T} L_c(\theta_1, \phi_1) \leq \lim_{t_1 \rightarrow T} \frac{C_2 \delta^2 d_X (e^{2T} - e^{2t_1})}{2(T - t_1)(e^{2t_1} - 1)(e^{2T} - 1)} = \frac{C_2 \delta^2 d_X}{e^{2T}(1 - e^{-2T})^2}.$$

To bound term (v) of the RHS of Eq. 14, notice that  $(\hat{Z}, \hat{G})$  and  $X_{t_0}$  are invertible, and thus

$$I(X_T; X|\hat{Z}, \hat{G}) = I(X_T; X|X_{t_0}) = 0.$$

Combining the bounds and choose  $T := \Omega(\log \frac{1}{\delta})$ , we conclude that

$$I(z_{\phi^*}(X_{t_0}); \hat{G}) \leq \frac{1}{2} \log C_2 + 3C_1 \delta = O(\sigma_X^2 \delta^2 + \lambda_s \sigma_X d_X \delta) = O(\lambda_s \sigma_X^2 d_X \delta).$$

## B.2 Proof of Lemma B.2

To prove the lemma, we need the following technical lemmas, whose proofs are deferred to B.3, B.4 and B.5 respectively.

**Lemma B.3.** Suppose random variable  $Y = \alpha X + N \sim p_Y$ , for independent random variables  $X \sim p_X$  and  $N \sim \mathcal{N}(0, \sigma^2 I_d)$ , where  $\alpha := \sqrt{1 - \sigma^2}$ . Then, for any distribution  $q(x|z)$  with an  $L$ -Lipschitz score function in  $x$  for any  $z \in \mathcal{X}$ , then the following inequality holds:

$$\mathbb{E}_{p_{XY}Z(x,y,z)} \log \frac{q(y|z)}{q(x|z)} \leq CL[\sigma^2(1 + \sigma^2)\mathbb{E}\|X\|^2 + \sigma^2 d + \sigma \sqrt{\mathbb{E}\|X\|^2 d}],$$

for some constant  $C > 0$  independent of  $\sigma$ ,  $\mathbb{E}\|X\|^2$  and  $d$ .

**Lemma B.4.** For  $\delta < 1/2$  and  $\alpha := \sqrt{1 - \delta^2}$ , the KL divergence between two Gaussian distributions  $\mathcal{N}(\alpha\mu, \delta^2 I_d)$  and  $\mathcal{N}(\frac{1}{\alpha}\mu, \frac{\delta^2}{\alpha^2} I_d)$  is upper-bounded as

$$D_{\text{KL}}\left(\mathcal{N}(\alpha\mu, \delta^2 I_d) \left\| \mathcal{N}\left(\frac{1}{\alpha}\mu, \frac{\delta^2}{\alpha^2} I_d\right)\right.\right) \leq \left(\frac{\|\mu\|^2}{2} + \frac{d}{6}\right)\delta^2.$$

**Lemma B.5.** Suppose random variable  $Y = \alpha X + N \sim p_Y$ , for independent random variables  $X \sim p_X$  and  $N \sim \mathcal{N}(0, \sigma^2 I_d)$ , where  $\alpha := \sqrt{1 - \sigma^2}$ . Further, suppose  $p_X(x) > 0$  for any  $x \in \mathbb{R}^d$  and the score function of  $p_X$  is  $L$ -Lipschitz, then the following holds

$$\max\{D_{\text{KL}}(p_X \| p_Y), D_{\text{KL}}(p_Y \| p_X)\} \leq CL(\mathbb{E}\|X\|^2)^{1/2} \delta d,$$

for some constant  $C > 0$ .

Assuming Lemma B.3-B.5, define  $\bar{X} := f(Z, G)$  and choose  $t_0 := \frac{1}{2} \log \frac{1}{1 - \delta^2}$ . Then by the property of the OU process,

$$X_{t_0} = \alpha X + \delta N_{t_0},$$

where  $\alpha := \sqrt{1 - \delta^2}$ . Next, since the mixing function  $f$  is invertible, we can write its inverse  $f^{-1} : \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_Z} \times \mathbb{R}^{d_G}$  as  $f^{-1}(x) =: [z(x), g(x)]$ , where we call  $z : \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_Z}$  the *true content encoder* and  $g : \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_G}$  the *true style encoder* such that  $z(\bar{X}) = Z, g(\bar{X}) = G$ . By Definition 3.3, we have access to the true style encoder  $g(\cdot)$ . Set  $\phi_1$  such that  $z_{\phi_1}(x) = z(x)$ . Further, define

$$q(z|x) := p_{z(X_{t_0})|X}(z|x),$$

$$q(z) := p_{z(X_{t_0})}(z),$$

$$p(z|x) := p_{Z|X}(z|x),$$

$$p(z) := p_Z(z),$$

then by definition,

$$\begin{aligned} & I(z(X_{t_0}); X) - I(Z; X) \\ &= \mathbb{E}_{p_X} [D_{\text{KL}}(q(Z|X) \| q(Z)) - D_{\text{KL}}(p(Z|X) \| p(Z))] \\ &= \mathbb{E}_{q_X} [D_{\text{KL}}(q(Z|X) \| p(Z)) - D_{\text{KL}}(q(Z) \| p(Z)) - D_{\text{KL}}(p(Z|X) \| p(Z))] \\ &\leq \mathbb{E}_{p_X} [D_{\text{KL}}(q(Z|X) \| p(Z)) - D_{\text{KL}}(p(Z|X) \| p(Z))] \\ &\leq \mathbb{E}_{p_X} D_{\text{KL}}(q(Z|X) \| p(Z|X)), \end{aligned}$$

where the first inequality uses the non-negativity of  $D_{\text{KL}}$  and the second inequality uses the triangle inequality of  $D_{\text{KL}}$ :  $D_{\text{KL}}(p \| q) \leq D_{\text{KL}}(p \| r) + D_{\text{KL}}(r \| q)$  for any pdfs  $p, q, r$ . Further, by the data processing inequality,

$$\begin{aligned} & \mathbb{E}_{p_X} D_{\text{KL}}(q(Z|X) \| p(Z|X)) \\ &= \mathbb{E}_{p_X} D_{\text{KL}}(p_{z(X_{t_0})|X} \| p_{z(X)|X}) \\ &\leq \mathbb{E}_{p_X} D_{\text{KL}}(p_{X_{t_0}|X} \| p_{\bar{X}|X}) \\ &= \mathbb{E}_{p_{XX_{t_0}}(x,y)} \log \frac{\mathcal{N}(y | \alpha t_0, x, \delta^2 I_{d_X})}{p_{\bar{X}}(y) \mathcal{N}(x | \alpha y, \delta^2 I_{d_X}) / p_X(x)} \\ &= \mathbb{E}_{p_{XX_{t_0}}(x,y)} \log \frac{p_X(x)}{p_{\bar{X}}(y)} + O((d_X + \mathbb{E}\|X\|)\delta^2) \\ &= D_{\text{KL}}(p_X \| p_{\bar{X}}) + \mathbb{E}_{p_{XX_{t_0}}(x,y)} \log \frac{p_{\bar{X}}(x)}{p_{\bar{X}}(y)} + O((d_X + \mathbb{E}\|X\|)\delta^2), \end{aligned}$$

where the second-to-last equality uses Lemma B.4:

$$\begin{aligned} & \mathbb{E}_{p_{X|X_{t_0}}(x,y)} \log \frac{\mathcal{N}(y|\alpha x, \delta^2 I_{d_X})}{\mathcal{N}(x|\alpha y, \delta^2 I_{d_X})} \\ &= \mathbb{E}_{p_{X|X_{t_0}}(x,y)} \log \frac{\mathcal{N}(y|\alpha_{t_0} x, \delta_{t_0}^2 I_{d_X})}{\mathcal{N}(y|\frac{1}{\alpha} x, \frac{\delta^2}{\alpha^2} I_{d_X})} + d_X \log(1/\alpha) \\ &= O((d_X + \mathbb{E}\|X\|)\delta^2). \end{aligned} \quad (15)$$

To proceed, notice further that

$$\begin{aligned} & D_{\text{KL}}(p_X \| p_{\bar{X}}) + \mathbb{E}_{p_{X|X_{t_0}}(x,y)} \log \frac{p_{\bar{X}}(x)}{p_{\bar{X}}(y)} \\ &= D_{\text{KL}}(p_X \| p_{\bar{X}}) + O(\lambda_s \sigma_X d_X \delta) \\ &= O(\lambda_s d_X \sigma_X \delta), \end{aligned} \quad (16)$$

where the first equality uses Lemma B.3 with  $\delta < 1/\sqrt{d_X}$ , and the second inequality uses Lemma B.5. Combining this with Eq. 15 yields

$$I(z(X_{t_0}); X) - I(Z; X) = O(\lambda_s \sigma_X d_X \delta) = C_1 \delta,$$

where  $C_1 := O(\lambda_s \sigma_X d_X)$ . Using a similar strategy, we can prove that

$$I(\hat{G}; X) - I(G; X) = O(\lambda_s \sigma_X d_X \delta) = C_1 \delta.$$

For the last MI  $I(\hat{Z}; \hat{G}|X)$ , define

$$q(z, g|x) := p_{\hat{Z}|\hat{G}|X}(z, g|x), p(z, g|x) := p_{ZG|X}(z, g|x),$$

and notice that

$$\begin{aligned} & I(\hat{Z}; \hat{G}|X) - I(Z; G|X) \\ &= \mathbb{E}_{p_X} [D_{\text{KL}}(q(Z, G|X) \| q(Z|X)q(G|X)) - D_{\text{KL}}(p(Z, G|X) \| p(Z|X)p(G|X))] \\ &= \mathbb{E}_{p_X} [D_{\text{KL}}(q(Z, G|X) \| p(Z|X)p(G|X)) - D_{\text{KL}}(p(Z, G|X) \| p(Z|X)p(G|X))] - \\ & \quad \mathbb{E}_{p_X} [D_{\text{KL}}(p(Z|X) \| q(Z|X)) + D_{\text{KL}}(p(G|X) \| q(G|X))] \\ &\leq \mathbb{E}_{p_X} [D_{\text{KL}}(q(Z, G|X) \| p(Z|X)p(G|X)) - D_{\text{KL}}(p(Z, G|X) \| p(Z|X)p(G|X))] \\ &\leq \mathbb{E}_{p_X} D_{\text{KL}}(q(Z, G|X) \| p(Z, G|X)) \\ &\leq \mathbb{E}_{p_X} D_{\text{KL}}(p_{X_{t_0}|X} \| p_{\bar{X}|X}) = O(\lambda_s d_X \sigma_X \delta) = C_1 \delta, \end{aligned}$$

where we again apply non-negativity of  $D_{\text{KL}}$  on the first inequality and triangle inequality of  $D_{\text{KL}}$  on the second inequality. For the last equality, we combine Eq. 15 and Eq. 16.

To prove item 2, set  $\theta_1$  so that the score function

$$s_{\theta_1}(X_t, \hat{Z}, \hat{G}, t) = \frac{\exp(-t)f(\hat{Z}, \hat{G}) - X_t}{1 - \exp(-2t)}.$$

Then by setting  $t_1 := \frac{1}{2} \log \frac{1}{1-\delta}$ , the loss  $L_c$  becomes

$$\begin{aligned} L_c(\theta_1, \phi_1) &= \frac{1}{T-t_1} \int_{t_1}^T \frac{\exp(-2t) \mathbb{E} \|f(\hat{Z}, \hat{G}) - X\|^2}{(1 - \exp(-2t))^2} dt \\ &=: \frac{\tilde{L}_c(\theta_1, \phi_1)}{T-t_1} \int_{e^{t_1}}^{e^T} \frac{d\tau}{2(\tau-1)^2} \\ &= \frac{\tilde{L}_c(\theta_1, \phi_1)}{2(T-t_1)} \left( \frac{1}{e^{2t_1}-1} - \frac{1}{e^{2T}-1} \right) \leq \frac{\tilde{L}_c(\theta_1, \phi_1)(e^{2T}-e^{2t_1})}{2(T-t_1)(e^{2t_1}-1)(e^{2T}-1)}, \end{aligned}$$

Further, notice that

$$\begin{aligned} \tilde{L}_c(\theta_1, \phi_1) &:= \mathbb{E} \|f(\hat{Z}, \hat{G}) - X\|^2 \\ &= \mathbb{E} \|f(z(X_{t_0}), g(X_{t_0})) - f(z(X), g(X))\|^2 \\ &= \mathbb{E} \|X_{t_0} - X\|^2 \\ &\leq C_2 \delta^2 d_X, \end{aligned}$$

where  $C_2 := 1 + \delta^2 \sigma_X^2$ . The second equality uses the definition of inverse functions so that  $f(z(x), g(x)) = x, \forall x \in \mathbb{R}^{d_X}$ . The last inequality uses the fact that

$$\begin{aligned}\mathbb{E}\|X_{t_0} - X\|^2 &= \mathbb{E}\|X - \bar{X}\|^2 \\ &= \mathbb{E}\|(\alpha - 1)X + N_t\|^2 \\ &= (1 - \sqrt{1 - \delta^2})^2 \mathbb{E}\|X\|^2 + \mathbb{E}\|N_{t_0}\|^2 \\ &\leq \delta^4 \sigma_X^2 d_X + \delta^2 d_X.\end{aligned}$$

As a result, item 2 follows from

$$L_c(\theta_1, \phi_1) \leq \frac{C_2 \delta^2 d_X (e^{2T} - e^{2t_1})}{2(T - t_1)(e^{2t_1} - 1)(e^{2T} - 1)} \leq \frac{C_2 \delta d_X}{2(T - \log(1 - \delta)^{1/2})} = O\left(\frac{\delta d_X}{T}\right)$$

with the choice of  $t_1$  and  $\delta$  and the fact that  $0 < \delta < \min\{T, 1\}$ .

### B.3 Proof of Lemma B.3

To begin, we make use the following lemma proved in Section B.6.

**Lemma B.6.** *Suppose the score function  $s_q(x|z) := \nabla_x \log q(x|z)$  of the probability density  $q(x|z)$  is  $L$ -Lipschitz as a function of  $x$ , then the following inequality holds:*

$$\log \frac{q(y|z)}{q(x|z)} \leq (L\|x\| + \|s_q(\mathbf{0}_d|z)\|)\|y - x\| + \frac{L\|y - x\|^2}{2}.$$

Set  $s_q(x|z) := \nabla_x \log q(x|z)$ , then by Lemma B.6,

$$\begin{aligned}\mathbb{E}_{p_{XYZ}(x,y,z)} \log \frac{q(y|z)}{q(x|z)} \\ \leq \mathbb{E}_{p_{XYZ}(x,y,z)} (L\|x\| + \|s_q(\mathbf{0}_d|z)\|)\|y - x\| + \frac{L}{2} \mathbb{E}_{p_{XY}(x,y)} \|y - x\|^2.\end{aligned}$$

For the first term of the RHS, notice that

$$\begin{aligned}\mathbb{E}_{p_{XYZ}(x,y,z)} (L\|x\| + \|s_q(\mathbf{0}_d|z)\|)\|y - x\| \\ \leq \sqrt{\mathbb{E}(L\|X\| + \|s_q(\mathbf{0}_d|Z)\|)^2 \mathbb{E}\|Y - X\|^2} \\ = \sqrt{\mathbb{E}(L\|X\| + \|s_q(\mathbf{0}_d|Z)\|)^2} \sqrt{(1 - \alpha)^2 \mathbb{E}\|X\|^2 + \sigma^2 d} \\ \leq \sqrt{2L^2 \mathbb{E}\|X\|^2 + C_1} \sqrt{(1 - \alpha)^2 \mathbb{E}\|X\|^2 + \sigma^2 d} \\ \leq C_2 L (\sigma^2 \mathbb{E}\|X\|^2 + \sigma \sqrt{\mathbb{E}\|X\|^2 d}),\end{aligned}$$

where  $C_1 := 2 \sup_z \mathbb{E}\|s_q(\mathbf{0}_d|z)\|^2$  and  $C_2$  large enough. To bound the second term of the RHS, notice that

$$\begin{aligned}\frac{L}{2} \mathbb{E}_{p_{XY}(x,y)} \|y - x\|^2 &= \frac{L}{2} [(1 - \alpha)^2 \mathbb{E}\|X\|^2 + \mathbb{E}\|N\|^2] \\ &\leq \frac{L\sigma^2}{2} (\sigma^2 \mathbb{E}\|X\|^2 + d).\end{aligned}$$

Combining the two terms yields

$$\mathbb{E}_{p_{XYZ}(x,y,z)} \log \frac{q(y|z)}{q(x|z)} \leq CL[\sigma^2(1 + \sigma^2) \mathbb{E}\|X\|^2 + \sigma^2 d + \sigma \sqrt{\mathbb{E}\|X\|^2 d}],$$

for some  $C > 0$  large enough.

## B.4 Proof of Lemma B.4

Use the formula for the KL divergence between Gaussians:

$$\begin{aligned}
 D_{\text{KL}}(\mathcal{N}(\alpha\mu, \delta^2 I_d) \parallel \mathcal{N}(\mu/\alpha, (\delta/\alpha)^2 I_d)) \\
 &= \left( \frac{\delta^2}{(\delta/\alpha)^2} - 1 \right) d/2 + \frac{\|\alpha\mu - \mu/\alpha\|^2}{2(\delta/\alpha)^2} + d \log \frac{1}{\alpha} \\
 &= -\frac{d}{2} \delta^2 + \frac{\delta^2 \|\mu\|^2}{2} + d \log \frac{1}{\alpha} \\
 &= \frac{(\|\mu\|^2 - d) \delta^2}{2} + \frac{d}{2} \log \left( 1 + \frac{\delta^2}{1 - \delta^2} \right) \\
 &\leq \frac{(\|\mu\|^2 - d) \delta^2}{2} + \frac{2d\delta^2}{3} = \left( \frac{\|\mu\|^2}{2} + \frac{d}{6} \right) \delta^2.
 \end{aligned}$$

## B.5 Proof of Lemma B.5

By Jensen's inequality,

$$\begin{aligned}
 D_{\text{KL}}(p_Y \parallel p_X) \\
 &= \mathbb{E}_{\int p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d) dx} \log \frac{\int p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d) dx}{p_X(y)} \\
 &\leq \mathbb{E}_{p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d)} \log \frac{p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d)}{p_X(y) \mathcal{N}(x|y/\alpha, (\sigma/\alpha)^2 I_d)} \\
 &= \mathbb{E}_{p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d)} \log \frac{p_X(x)}{p_X(y)} + d \log \alpha \\
 &\leq \mathbb{E}_{p_X(x) \mathcal{N}(y|\alpha x, \sigma^2 I_d)} \log \frac{p_X(x)}{p_X(y)} - \frac{\sigma^2 d}{2} \\
 &\leq CL(\sigma^2 \mathbb{E}\|X\|^2 + \sigma^2 d + \sigma \sqrt{d} \mathbb{E}\|X\|),
 \end{aligned}$$

for some  $C > 0$ , where the last inequality uses Lemma B.3.

Similarly, apply Jensen's inequality and Lemma B.3,

$$\begin{aligned}
 D_{\text{KL}}(p_X \parallel p_Y) &= \mathbb{E}_{p_X(x)} \log \frac{p_X(x)}{\int p_X(y) \mathcal{N}(x|\alpha y, \sigma^2 I_d) dy} \\
 &\leq \mathbb{E}_{p_X(x) \mathcal{N}(y|x/\alpha, (\sigma/\alpha)^2 I_d)} \log \frac{p_X(x) \mathcal{N}(y|x/\alpha, (\sigma/\alpha)^2 I_d)}{p_X(y) \mathcal{N}(x|\alpha y, \sigma^2 I_d)} \\
 &= \mathbb{E}_{p_X(x) \mathcal{N}(y|x/\alpha, (\sigma/\alpha)^2 I_d)} \log \frac{p_X(x)}{p_X(y)} + d \log \alpha \\
 &\leq CL(\sigma^2 \mathbb{E}\|X\|^2 + \sigma^2 d + \sigma \sqrt{d} \mathbb{E}\|X\|).
 \end{aligned}$$

## B.6 Proof of Lemma B.6

by the Lipschitz property of the score function of  $s_q$ , for any  $(x, y) \in \mathcal{X}^2$ ,

$$\begin{aligned}
 \log \frac{q(y|z)}{q(x|z)} &\leq |\langle s_q(x|z), y - x \rangle| + \frac{L\|y - x\|^2}{2} \\
 &\leq \|s_q(x|z)\| \|y - x\| + \frac{L\|y - x\|^2}{2}.
 \end{aligned}$$

Apply the Lipschitz property of  $s_q$  again,

$$\log \frac{q(y|z)}{q(x|z)} \leq (L\|x\| + \|s_q(\mathbf{0}_d|z)\|) \|y - x\| + \frac{L\|y - x\|^2}{2}. \quad (17)$$

## B.7 Extension to correlated content and style

To extend Theorem 4.1 to correlated content and style, we can add an additional term to the right-hand side of Eq. 14 Theorem 4.2 as

$$I(\hat{Z};\hat{G}) = (i) + (ii) + (iii) + (iv) + (v) + I(Z;G) \leq (i) + (ii) + (iii) + (iv) + (v) + \epsilon_1, \quad (18)$$

where (i)-(v) are defined and bounded as in Appendix B.1.

## C Proof of Theorem 4.2

For clarity, we restate the theorem below.

**Theorem C.1.** *Under Assumption 3.5-3.6 and additional regularity conditions (Assumption C.2-C.4). Consider the two-view, unimodal setting with  $\delta_1 =: \delta$  and  $\delta_2 = 0$ . Let  $(\theta^*, \phi_Z^*, \phi_G^*)$  be a minimizer of  $L_m$  and  $\hat{Z}^i = z_{\phi_Z^*}(X^i)$ ,  $\hat{G}^i = g_{\phi_G^*}(X^i)$  for each view  $i$ . Set  $t_0 = -\log(1 - \delta^2)^{1/2}$  and the diffusion step size  $\eta := C_5 \frac{\delta^3}{\lambda_s^2 T}$  for some constant  $C_5 > 0$ . Then for some constants  $C_6, C_7 > 0$ : (i)  $I(\hat{Z}^2; \hat{G}^1) \leq C_6 \lambda_s \sigma_X d_X \delta$ ; (ii) the content encoding  $\hat{Z}^2$  and the style encoding  $\hat{G}^1$  are  $C_7 \sqrt{\lambda_s \sigma_X^2 d_X \delta}$ -editable.*

### C.1 Main proof

Let  $\bar{X}^i := f_i(Z, G^i) = f(Z, G^i)$ , where the last equality assumes the unimodal setting. By the invertibility of the mixing function, there exist functions  $z: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_Z}$  such that  $z(\bar{X}^i) = Z$  and  $g: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_G}$  such that  $g(\bar{X}^i) = G^i$ ,  $i \in \{1, 2\}$ . To prove this theorem, we need additional assumptions below.

**Assumption C.2.** *For any  $\theta \in \Theta, \phi \in \Phi$ , the estimated score function  $s_\theta(x, z, g, t)$  is  $\frac{\lambda_\Theta}{\sigma(t)^2}$ -Lipschitz in all arguments.*

**Assumption C.3.** *The style encoder  $g_{\phi_G}: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_G}$  is  $\lambda_g$ -Lipschitz in  $x \in \mathbb{X}$  and the content encoder  $z_{\phi_Z}: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_Z}$  is  $\lambda_z$ -Lipschitz. Further, there exists  $\phi_{1,G} \in \Phi_G$  such that  $z_{\phi_{1,Z}}(\cdot) = z(\cdot)$  and  $g_{\phi_{1,G}}(\cdot) = g(\cdot)$ .*

**Assumption C.4.** *The mixing function  $f$  is  $\lambda_f$ -Lipschitz.*

Assumption C.2 ensures the function class for the bottleneck and the score function are Lipschitz, analogous to the setting in [41]. Assumption C.3 assumes that the content and style encoders are both Lipschitz and the true content and style functions are realizable by the function classes  $\Phi_Z$  and  $\Phi_G$  respectively. This is a mild assumption since the style and content encoders are assumed to be NNs and common choices of neural architectures are either already Lipschitz or constrained to be so through regularizations [119–121]. For Assumption C.4, note that the invertibility of  $f_\theta$  is not strictly necessary and can be relaxed to injectivity by partitioning the domain.

Again, we start by introducing a couple helpful lemmas and postponing their proofs to Section C.2 and C.3.

**Lemma C.5.** *There exists  $(\theta_1, \phi_{Z,1}, \phi_{G,1})$  such that for*

$$\delta < \min \left\{ \frac{1}{2}, \frac{1}{\sqrt{d_X}} \right\}, \quad t_0 = -\log(1 - \delta^2)^{1/2}, \quad t_1 = -\log(1 - \delta)^{1/2},$$

*the followings hold:*

1. *The followings hold for MIs  $I(\hat{G}^1; X^2)$  and  $I(\hat{G}^1; Z)$ :*

$$I(\hat{G}^2; X) = O(\lambda_s \sigma_X d_X \delta),$$

$$I(\hat{G}^1; Z) = O(\lambda_s \sigma_X d_X \delta);$$

2. *The regularized score matching loss for the multi-view disentanglement in Eq. 8 satisfies*

$$L_m(\theta_1, \phi_1) \leq \frac{\lambda_f \lambda_g (\sigma_X^2 \delta^2 + 2) \delta^2 d_X (e^{2T} - e^{2t_1})}{2(T - t_1)(e^{2t_1} - 1)(e^{2T} - 1)} = O\left(\frac{\lambda_f \lambda_g d_X \delta}{T}\right),$$

*where  $\phi_1 := [\phi_{Z,1}, \phi_{G,1}]$ .*

**Lemma C.6.** (Novikov's condition) *The following bound holds:*

$$\mathbb{E}_{X^1, (X_t^{1,\leftarrow})_t} \exp \left( \frac{1}{2} \int_0^{T-t} \|s_\theta(X_t^{1,\leftarrow}, \hat{Z}^2, \hat{G}^1, T-t) - \nabla_x \log p_{T-t_0}(X_t^{1,\leftarrow} | X^1)\|^2 dt \right) < \infty$$

Let  $z: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_Z}$  and  $g: \mathbb{R}^{d_X} \mapsto \mathbb{R}^{d_G}$  be the true content and style encoders defined in Section B.1, and define  $\bar{X}^i := f(Z^i, G^i)$ ,  $i \in \{1, 2, 3\}$ . By Definition 3.4, we have  $X^2 = \bar{X}^2$ . Further, define  $\hat{X}^{21} := \hat{X}_{T-t_1}^{21,\leftarrow}$  to be a sample from the following estimated backward process:

$$d\hat{X}_t^{21,\leftarrow} = [\hat{X}_t^{21,\leftarrow} + 2s_{\theta^*}(\hat{X}_{k\eta}^{21,\leftarrow}, \hat{Z}^2, \hat{G}^1, T-k\eta)]dt + \sqrt{2}dB_t^{\leftarrow}, \hat{X}_0^{21,\leftarrow} \sim p_T, t \in [k\eta, (k+1)\eta]. \quad (19)$$

We begin by proving item (i) of the theorem. To this end, we apply the first part of Lemma C.5 and data processing inequality:

$$I(\hat{G}^1; \hat{Z}^2) \leq I(\hat{G}^1; X^2) = O(\lambda_s \sigma_X d_X \delta). \quad (20)$$

Next, we prove item (ii). Then apply triangle inequality:

$$\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | Z}, p_{\hat{X}^{23} | Z}) \leq \underbrace{\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | Z}, p_{\hat{X}^{21} | Z})}_{(a)} + \underbrace{\mathbb{E}_{p_Z} d_{\text{TV}}(p_{\hat{X}^{21} | Z}, p_{\hat{X}^{23} | Z})}_{(b)}. \quad (21)$$

To bound term (a), we check that Novikov's condition holds by Lemma C.6, and apply Girsanov's theorem on the estimated backward process Eq. 19 with  $t_1 := -\log(1-\delta)^{1/2}$  followed by data processing inequality:

$$\begin{aligned} \mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | Z}, p_{\hat{X}^{21} | Z}) &\leq \mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | X^1}, p_{\hat{X}^{21} | X^1}) \\ &= O\left((\sqrt{d_X \eta} + \sigma_X^2 \eta) \lambda_\Theta \sqrt{T} / \delta + \sqrt{L_c(\theta^*, \phi_Z^*, \phi_G^*) T}\right) \\ &= O\left((\sqrt{d_X \eta} + \sigma_X^2 \eta) \lambda_\Theta \sqrt{T} / \delta + \sqrt{\lambda_f \lambda_g \delta d_X}\right), \end{aligned}$$

where the last equality applies Lemma C.5. Set  $\eta := \frac{\delta^3}{\lambda_\Theta^2 T}$ , we obtain

$$\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | Z}, p_{\hat{X}^{21} | Z}) = O\left(\sqrt{\lambda_f \lambda_g \sigma_X^2 d_X \delta}\right).$$

To bound the second term (b), applying data processing inequality, Pinsker's inequality, triangle inequality and data processing inequality again in that order,

$$\begin{aligned} &d_{\text{TV}}(p_{\hat{X}^{11} | Z}, p_{\hat{X}^{13} | Z}) \\ &\leq d_{\text{TV}}(p_{\hat{Z}^2 \hat{G}^1 | Z}, p_{\hat{Z}^2 \hat{G}^3 | Z}) \leq d_{\text{TV}}(p_{\hat{Z}^2 \hat{G}^1 | Z}, p_{\hat{Z}^2 | Z} p_{\hat{G}^1 | Z}) + d_{\text{TV}}(p_{\hat{Z}^2 \hat{G}^2 | Z}, p_{\hat{Z}^2 \hat{G}^3 | Z}) \\ &= d_{\text{TV}}(p_{\hat{Z}^2 \hat{G}^1 | Z}, p_{\hat{Z}^2 | Z} p_{\hat{G}^1 | Z}) + d_{\text{TV}}(p_{\hat{Z}^2 | Z} p_{\hat{G}^1 | Z}, p_{\hat{Z}^2 | Z} p_{\hat{G}^3 | Z}) \\ &\leq \sqrt{\frac{I(\hat{Z}^2; \hat{G}^1 | Z)}{2}} + \sqrt{\frac{I(\hat{G}^1; Z)}{2}} \leq \sqrt{\frac{I(X^2; X^1 | Z)}{2}} + O(\sqrt{\lambda_s \sigma_X d_X \delta}) = O(\sqrt{\lambda_s \sigma_X d_X \delta}), \quad (22) \end{aligned}$$

where the last inequality uses Lemma C.5 and the last equality using the conditional independence  $X^1 \perp\!\!\!\perp X^2 | Z$ .

Combining the bounds on (i)(ii), we obtain

$$\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}^1 | Z}, p_{\hat{X}^{23} | Z}) = O\left(\sqrt{\sigma_X^2 d_X \delta} + \sqrt{\lambda_s \sigma_X d_X \delta}\right) = O\left(\sqrt{(\lambda_s + \lambda_f \lambda_g) \sigma_X^2 d_X \delta}\right).$$

## C.2 Proof of Lemma C.5

We start by proving item 1. By Definition 3.4, the styles of different views are independent, i.e.  $G^1 \perp\!\!\!\perp X^2$  and thus  $I(G^1; X^2) = 0$ . Choose parameters  $(\theta_1, \phi_{Z,1}, \phi_{G,1})$  such that

$$\hat{f}_{\theta_1}(x) = f(x), \quad z_{\phi_{Z,1}}(x) = z(x), \quad g_{\phi_{G,1}}(x) = g(x).$$

Further, define

$$\begin{aligned} q(g|x^2) &:= p_{\hat{G}^1|X^2}(g|x^2), \quad p(g|x^2) := p_{G^1|X^2}(g|x^2), \\ q(g|x^1) &:= p_{\hat{G}^1|X^1}(g|x^1), \quad p(g|x^1) := p_{G^1|X^1}(g|x^1), \\ p(x^1|x^2) &:= p_{X^1|X^2}(x^1|x^2), \end{aligned}$$

and by a similar argument as in the proof of Lemma B.2 in Section B.2,

$$I(\hat{G}^1; X^2) = I(\hat{G}^1; X^2) - I(G^1; X^2) \leq \mathbb{E}_{X^2} D_{\text{KL}}(q(g|x^2) || p(g|x^2)).$$

Further, by data processing inequality,

$$\begin{aligned} \mathbb{E}_{p_{X^2}} D_{\text{KL}}(q(g|x^2) || p(g|x^2)) &\leq \mathbb{E}_{p_{X^2}} D_{\text{KL}}(p_{\hat{G}^1, X^1|X^2} || p_{G^1, X^1|X^2}) \\ &= \mathbb{E}_{X^2} D_{\text{KL}}(p(x^1|x^2)q(g|x^1) || p(x^1|x^2)p(g|x^1)) \\ &= \mathbb{E}_{p_{X^1}} D_{\text{KL}}(q(g|x^1) || p(g|x^1)), \end{aligned} \quad (23)$$

where the first equality uses the fact that  $(\hat{G}^1, G^1) \perp\!\!\!\perp X^2 | X^1$  and thus

$$p_{\hat{G}^1|X^1, X^2}(g|x^1, x^2) = q(g|x^1), \quad p_{G^1|X^1, X^2}(g|x^1, x^2) = p(g|x^1).$$

Again by a similar argument as in Section B.2,

$$\begin{aligned} \mathbb{E}_{p_{X^1}} D_{\text{KL}}(q(g|x^1) || p(g|x^1)) &= O(\lambda_s \sigma_X d_X \delta_1) \\ \implies I(\hat{G}^1; X^2) &= O(\lambda_s \sigma_X d_X \delta_1). \end{aligned}$$

As a result, by the chain rule of MI,

$$\begin{aligned} I(\hat{G}^1; Z, X^2) &= I(\hat{G}^1; Z) = I(\hat{G}^1; X^2) + I(\hat{G}^1; Z | X^2) \\ &= I(\hat{G}^1; Z | X^2) + O(\lambda_s \sigma_X d_X \delta) \\ &= O(\lambda_s \sigma_X d_X \delta), \end{aligned} \quad (24)$$

where the first inequality uses the conditional independence  $\hat{G}^1 \perp\!\!\!\perp X^2 | Z$  and the last equality uses the conditional independence  $\hat{G}^1 \perp\!\!\!\perp Z | \bar{X}^2 = X^2$ . This concludes the proof of item 1.

To prove item 2, consider the score function of the form

$$s_{\theta_1}(X_t^1, \hat{Z}^2, \hat{G}^1, t) = \frac{\exp(-t)f(\hat{Z}^2, \hat{G}^1)}{1 - \exp(-2t)} = \frac{\exp(-t)f(Z, \hat{G}^1)}{1 - \exp(-2t)},$$

and set  $t_1 := \frac{1}{2} \log \frac{1}{1-\delta}$ , the loss  $L_c$  becomes

$$\begin{aligned} L_m(\theta_1, \phi_1) &= \frac{1}{T-t_1} \int_{t_1}^T \frac{\exp(-2t) \mathbb{E} \|f(Z, \hat{G}^1) - X^1\|^2}{(1 - \exp(-2t))^2} dt \\ &= \frac{\tilde{L}_m(\theta_1, \phi_1)}{2(T-t_1)} \left( \frac{1}{e^{2t_1} - 1} - \frac{1}{e^{2T} - 1} \right) \leq \frac{\tilde{L}_m(\theta_1, \phi_1)(e^{2T} - e^{2t_1})}{2(T-t_1)(e^{2t_1} - 1)(e^{2T} - 1)}. \end{aligned}$$

By Assumption C.3,

$$\begin{aligned} \tilde{L}_m(\theta_1, \phi_1) &:= \mathbb{E} \|f(Z, \hat{G}^1) - X^1\|^2 = \mathbb{E} \|f(Z, \hat{G}^1) - f(Z, G^1)\|^2 \\ &\leq \lambda_f \lambda_g \|X_{t_0}^1 - \bar{X}^1\|^2 \leq \lambda_f \lambda_g (\sigma_X^2 \delta^2 + 2) \delta^2 d_X, \end{aligned}$$

where the inequality uses the fact that

$$\begin{aligned} \mathbb{E} \|X_{t_0}^1 - \bar{X}^1\|^2 &= \mathbb{E} \|\alpha^2 \bar{X}^1 + \alpha N^1 + \delta N_{t_0}^1 - \bar{X}^1\|^2 \\ &= \delta^4 \mathbb{E} \|\bar{X}\|^2 + (1 - \delta^2) \delta^2 d_X + \delta^2 d_X \\ &= ((\sigma_X^2 - 1) \delta^2 + 2) \delta^2 d_X. \end{aligned}$$

Plugging this into Eq. 24 yields

$$\begin{aligned} L_m(\theta_1, \phi_1) &\leq \frac{\lambda_f \lambda_g (\sigma_X^2 \delta^2 + 2) \delta^2 d_X (e^{2T} - e^{2t_1})}{2(T-t_1)(e^{2t_1} - 1)(e^{2T} - 1)} \\ &= O\left(\frac{\lambda_f \lambda_g \delta d_X}{T}\right). \end{aligned}$$

This proves item 2 of the lemma.

### C.3 Proof of Lemma C.6

First, by the equivalence of the forward and reverse process,

$$\begin{aligned} & \mathbb{E}_{X^1, (X_t^1, \leftarrow)_t} \exp \left( \frac{1}{2} \int_0^{T-t_1} \|s_\theta(X_t^1, \leftarrow, \hat{Z}^2, \hat{G}^1, T-t) - \nabla_x \log p_{T-t}(X_t^1, \leftarrow | X^1)\|^2 dt \right) \\ &= \mathbb{E}_{p_{X^1}} \mathbb{E}_{t, p_{t|0}(\cdot | X^1)} \exp \left( \frac{1}{2} \int_{t_1}^T \|s_\theta(X_t^1, \hat{Z}^2, \hat{G}^1, t) - \nabla_x \log p_t(X_t^1 | X^1)\|^2 dt \right) \end{aligned} \quad (25)$$

$$= \mathbb{E}_{p_{X^1}} \mathbb{E}_{t, p_{t|0}(\cdot | X^1)} \exp \left( \int_{t_1}^T \frac{1}{2\sigma(t)^4} \left\| \sigma(t)^2 s_\theta(X_t^1, \hat{Z}^2, \hat{G}^1, t) + \sigma(t) N_t^1 \right\|^2 dt \right), \quad (26)$$

where the second-to-last equality uses the closed-form formula of the conditional score function of an OU process and  $\sigma(t) := \sqrt{1 - e^{-2t}}$ . Define  $\lambda := 2\lambda_\Theta(1 + \lambda_z + \lambda_g)$ , and by Assumption C.2 again, we have

$$\begin{aligned} & \mathbb{E} \exp \left( \frac{1}{2\sigma(t)^4} \int_{t_1}^T \left\| \sigma(t)^2 s_\theta(X_t^1, \hat{Z}^2, \hat{G}^1, t) + \sigma(t) N_t^1 \right\|^2 dt \right) \\ & \lesssim \mathbb{E} \exp \left( \int_{t_1}^T \frac{\lambda^2 (\|X_t^1\|^2 + \|X_{t_0}^2\|^2 + \|X_{t_0}^1\|^2 + t^2) + \sigma(t)^2 \|N_t^1\|^2}{\sigma(t)^4} dt \right) \\ & \lesssim \mathbb{E} \exp \left( 2\lambda^2 (2\|\bar{X}^1\|^2 + \|\bar{X}^2\|^2) \int_{t_1}^T \frac{dt}{\sigma(t)^4} + 2\lambda^2 \int_{t_1}^T \frac{\|N_{t_0}^1\|^2 + \|N_{t_0}^2\|^2 + \frac{\lambda^2 + \sigma(t)^2}{\lambda^2} \|N_t^1\|^2}{\sigma(t)^4} dt + \int_{t_1}^T \frac{\lambda^2 t^2 dt}{2\sigma(t)^4} \right) \\ & \lesssim \mathbb{E} \exp \left( \frac{2\lambda^2 T (2\|\bar{X}^1\|^2 + \|\bar{X}^2\|^2 + (3+\delta)d_X \|\sup_{t_1 \leq t \leq T} N_{t,i}^1\|^2)}{\delta^2} + \frac{\lambda^2 T^3}{2\delta^2} \right) < \infty, \end{aligned}$$

where the last inequality uses the fact that  $\bar{X}$  is sub-gaussian while  $\sup_{t_1 \leq t \leq T} N_{t,i}, \forall 1 \leq i \leq d_X$  is sub-gaussian [122, 41] and independent of  $\bar{X}$ , and the second-to-last inequality uses the bounds

$$\begin{aligned} \int_{t_1}^T \frac{dt}{\sigma(t)^4} &\leq \frac{T}{\sigma(t_1)^4} = \frac{T}{(1 - e^{-2t_1})^2} = \frac{T}{\delta^2} \\ \int_{t_1}^T \frac{dt}{\sigma(t)^2} &\leq \frac{T}{\sigma(t_1)^2} = \frac{T}{1 - e^{-2t_1}} = \frac{T}{\delta}. \end{aligned}$$

### C.4 Extension to correlated view-specific styles

To extend Theorem 4.2 to correlated view-specific styles with  $I(G^1; G^2 | Z) \leq \epsilon_I$ , we can replace every step in the proof of Theorem 4.2 that involves the conditional independence  $X^1 \perp\!\!\!\perp X^2 | Z$  and thus  $I(X^1; X^2 | Z) = 0$  (e.g., Eq. 22 and Eq. 24) with

$$I(X^1; X^2 | Z) \leq I(G^1; G^2 | Z) \leq \epsilon_I,$$

by data processing inequality. The rest of the proofs follow as before.

## D Proof of Theorem 4.5

For clarity, we restate the theorem below.

**Theorem D.1.** *Under Assumption 3.5-3.6, for some  $t_0$  dependent on  $n$  and let  $\min\{d_T, d_H\} \rightarrow \infty$ , and let the positional encoding  $PE(\cdot)$  be bounded and linearly independent over  $t \in [t_0, T]$ . Define  $P_M$  to be the projection matrix onto  $R(M)$ , and  $\sigma_i(s)$  to be the  $i$ -th largest singular value of the operator  $s$ . Then for some  $\lambda_r$ , Eq. 12 converges to a critical point  $\hat{\theta} := [\hat{\theta}_Z, \hat{\theta}_G, \hat{U}, \hat{V}]$  such that with probability at least  $1 - O(\frac{1}{n})$ : 1) Encodings  $\hat{Z} := P_{\hat{U}} X$  and  $\hat{G} := g(X)$  are  $\left( O\left( \frac{d_X^{5/4} \log^{3/4} n}{\sigma_{d_Z}(s_Z^*) n^{1/4}} \right), 0 \right)$ -disentangled; 2)*

*Let  $\hat{Z}_{t_0} := \hat{Z} + \sigma(t_0) N_{t_0}$ . Then  $(\hat{Z}_{t_0}, \hat{G})$  are  $O\left( \frac{d_X^{7/4} \log^{9/16} n}{T \min\{\sigma_{d_Z}^{1/2}(s_Z^*), \sigma_{d_G}^{1/2}(s_G^*)\} n^{1/16}} \right)$ -editable.*

## D.1 Main proof

To prove the theorem, we need Lemma 4.4 proved in Section D.2 and the following theorems, whose proofs are postponed to Section D.3-D.4. Note that in the proofs, we drop the subscript  $n$  and use the population versions of the losses when  $n \rightarrow \infty$ . In subsequent analysis, we use the following balancing loss:

$$L_{b,n} := \mathbb{E}_t \|\tilde{U}^\top \tilde{U} - \mathbb{E}_{\hat{p}_t^n(x)} \tilde{s}(x,t)^\top \tilde{s}(x,t)\|^2 + \lambda_r \mathbb{E}_t \|V^\top V - \mathbb{E}_{\hat{p}_t^n(x)} s_G(g(x),t) s_G(g(x),t)^\top\|^2,$$

where  $\tilde{U} := [U, V]$  and  $\tilde{s}(x,t) := [s_Z(x,t)^\top, s_G(x,t)^\top]^\top$

**Theorem D.2.** *For the linear subspace model in Definition 4.3 and the objective in Eq. 11 with  $n \rightarrow \infty$ , then any minimizer  $(U^*, V^*)$  of Eq. 11 satisfy  $R(U^*) = R(A_Z)$  and  $R(V^*) = R(A_G)$ .*

**Theorem D.3.** *Suppose  $\min\{n, d_T, d_H\} \rightarrow \infty$ , and the neural network weights are initialized by standard Gaussians. Further, choose  $PE(\cdot)$  such that  $PE(t)$ 's are bounded and linearly independent for all  $t \in [t_0, T]$ . Then for  $\lambda_r = 3$ , the system of gradient flow equations in Eq. 12 converges to a critical point  $(\hat{U}, \hat{V})$  such that  $R(\hat{U}) = R(A_Z)$ ,  $R(\hat{V}) = R(A_G)$ .*

Recall that the two branches of the dual encoder network  $s_Z^{\theta_Z}$  and  $s_G^{\theta_G}$  are defined as

$$s_Z^{\theta_Z}(x,t) =: s_Z(x,t) := \frac{1}{\sqrt{d_H}} \sum_{j=1}^{d_H} \theta_Z^{(2),j} (\theta_Z^{(1),j^\top} [x^\top, PE(t)]^\top)_+, \quad (27)$$

$$s_G^{\theta_G}(g,t) =: s_G(g,t) := \frac{1}{\sqrt{d_H}} \sum_{j=1}^{d_H} \theta_G^{(2),j} (\theta_G^{(1),j^\top} [g^\top, PE(t)]^\top)_+. \quad (28)$$

Further, define the *neural tangent kernels* (NTKs) [123] for the score functions  $s_Z$  and  $s_G$  as

$$\begin{aligned} K_Z(x,t,x',t') &:= J_{\text{vec}(\theta_Z)} s_Z(x,t)^\top J_{\text{vec}(\theta_Z)} s_Z(x',t'), \\ K_G(x,t,x',t') &:= J_{\text{vec}(\theta_G)} s_G(g(x),t)^\top J_{\text{vec}(\theta_G)} s_G(g(x'),t'), \forall (x,t,x',t') \in \mathcal{X} \times [t_0, T] \times \mathcal{X} \times [t_0, T], \end{aligned}$$

where  $\text{vec}(\theta)$  denotes the flattened version of the parameter  $\theta$ .

For i.i.d samples  $[x^1, \dots, x^n]$ , let the Hilbert space spanned by the NTKs  $K_Z(x,t, \cdot, \cdot)$ 's and  $K_G(x,t, \cdot, \cdot)$ 's be  $\mathcal{H}_{K_Z}$  and  $\mathcal{H}_{K_G}$  respectively. Further, for any  $f(x,t) := \int_{t_0}^T \int_{\mathcal{X}} K(x,t,x',t') c(x',t') dx' dt' \in \mathcal{H}_{K_Z}, c(x,t) \in \mathbb{R}^d, \forall x,t$ , define the *NTK norm* as

$$\|f\|_K := \langle f, f \rangle_K := \sqrt{\mathbb{E}_{t,t',p_t(x)p_t(x')} c(x,t)^\top K(x,t,x',t') c(x',t)}. \quad (29)$$

Further, define the *subspace score matching losses* as

$$\begin{aligned} L_Z(U, s_Z) &:= \mathbb{E}_{t,p_t(x)} \|U s_Z(x,t) - A_Z s_Z^*(x,t)\|^2 =: \mathbb{E}_{t,p_t(x)} \ell_Z(x,t; U, s_Z) \\ L_G(V, s_G) &:= \mathbb{E}_{t,p_t(x)} \|V s_G(g(x),t) - A_G s_G^*(x,t)\|^2 =: \mathbb{E}_{t,p_t(x)} \ell_G(x,t; V, s_G), \end{aligned}$$

and their empirical versions as

$$\begin{aligned} \hat{L}_Z(U, s_Z) &:= \frac{1}{n} \sum_{i=1}^n \ell_Z(\tilde{x}^i; U, s_Z) \\ \hat{L}_G(V, s_G) &:= \frac{1}{n} \sum_{i=1}^n \ell_G(\tilde{x}^i; V, s_G). \end{aligned}$$

We also need the following lemma proved in Section D.9.

**Lemma D.4.** (*generalization error bound*) *Let  $\min\{d_T, d_H\} \rightarrow \infty$ , and for i.i.d samples  $[x^1, \dots, x^n]$ , let the reproducing kernel Hilbert space (RKHS) spanned by the NTKs  $K_Z(x,t, \cdot, \cdot)$ 's and  $K_G(x,t, \cdot, \cdot)$ 's be  $\mathcal{H}_{K_Z}$  and  $\mathcal{H}_{K_G}$  respectively, and denote  $K_{Z,(x,t)} := K_Z(x,t, \cdot, \cdot)$ . Further, define the function classes*

$\mathcal{S}_Z, \mathcal{U}, \mathcal{S}_G, \mathcal{V}$  as

$$\begin{aligned}\mathcal{S}_Z &:= \left\{ f = \sum_{i=1}^N c_i K_{Z, \tilde{x}_i} \mid \forall N \in \mathbb{N}, \|f\|_K \leq C_Z \sqrt{\frac{\sigma_1(s_Z^*) d_Z^{1/2}}{\lambda_{\min}(K_Z^*)}}, \forall [\tilde{x}_1, \dots, \tilde{x}_N] \in (\mathcal{X} \times [t_0, T])^N \right\}, \\ \mathcal{U} &:= \left\{ U : \|U\|_F \leq C_Z \sqrt{\sigma_1(s_Z^*) d_Z^{1/2}} \right\}, \\ \mathcal{S}_G &:= \left\{ f = \sum_{i=1}^N c_i K_{G, \tilde{x}_i} \mid \forall N \in \mathbb{N}, \|f\|_K \leq C_G \sqrt{\frac{\sigma_1(s_G^*) d_G^{1/2}}{\lambda_{\min}(K_G^*)}}, \forall [\tilde{x}_1, \dots, \tilde{x}_N] \in (\mathcal{X} \times [t_0, T])^N \right\}, \\ \mathcal{V} &:= \left\{ V : \|V\|_F \leq C_G \sqrt{\sigma_1(s_G^*) d_G^{1/2}} \right\},\end{aligned}$$

where  $\sigma_i(A)$  is the  $i$ -th largest singular value of the operator  $A$ . Then with probability at least  $1 - O(\frac{1}{n})$ ,

$$\begin{aligned}L_Z(\hat{U}, \hat{s}_Z) &\leq \min_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} L_Z(U, s_Z) + O\left(\sqrt{\frac{d_X^5 \log^3 n}{n}}\right), \\ L_G(\hat{V}, \hat{s}_G) &\leq \min_{(V, s_G) \in \mathcal{V} \times \mathcal{S}_G} L_G(V, s_G) + O\left(\sqrt{\frac{d_X^5 \log^3 n}{n}}\right).\end{aligned}$$

Now we are ready to prove the Theorem 4.5. To this end, we first prove the following statements:

1. The population score matching loss satisfies:  $L_0(\hat{\theta}) = O\left(\sqrt{\frac{d_X^5 \log^3 n}{n}}\right)$ ;
2. The subspace recovery errors are bounded by:

$$\|P_{\hat{U}} - A_Z A_Z^\top\|_F^2 = O\left(\sqrt{\frac{d_X^5 \log^3 n}{\sigma_{d_Z}^4(s_Z^*) n}}\right), \quad \|P_{\hat{V}} - A_G A_G^\top\|_F^2 = O\left(\sqrt{\frac{d_X^5 \log^3 n}{\sigma_{d_G}^4(s_G^*) n}}\right).$$

First, by the universal approximation theorem of neural networks (e.g. [124]) and the Lipschitzness of the true score functions, for  $C_Z, C_G$  large enough,  $s_Z^* \in \mathcal{S}_Z$  and  $s_G^* \in \mathcal{S}_G$  and therefore

$$\begin{aligned}\min_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} L_Z(U, s_Z) &= L_Z(A_Z, s_Z^*) = 0, \\ \min_{(V, s_G) \in \mathcal{V} \times \mathcal{S}_G} L_G(V, s_G) &= L_G(A_G, s_G^*) = 0.\end{aligned}$$

As a result, we have with probability at least  $1 - O(\frac{1}{n})$ ,

$$L_Z(\hat{U}, \hat{s}_Z) \leq \epsilon(n), \quad L_G(\hat{V}, \hat{s}_G) \leq \epsilon(n),$$

where  $\epsilon(n) = O\left(\sqrt{\frac{d_X^5 \log^3 n}{n}}\right)$ . Therefore, we prove item 1 by noticing

$$L_0(\hat{U}, \hat{V}, \hat{s}_Z, \hat{s}_G) \leq 2L_Z(\hat{U}, \hat{s}_Z) + 2L_G(\hat{V}, \hat{s}_G) \leq 2\epsilon(n) = O\left(\sqrt{\frac{d_X^5 \log^3 n}{n}}\right).$$

We proceed to prove item 2. Let  $Q_Z \in \mathbb{R}^{d_X \times d_Z}, Q_G \in \mathbb{R}^{d_X \times d_G}$  be orthogonal matrices such that

$$P_{R(\hat{U})} = Q_Z Q_Z^\top, P_{R(\hat{V})} = Q_G Q_G^\top,$$

then we have  $\hat{U} \hat{s}_Z := Q_Z Q_Z^\top \hat{U} \hat{s}_Z$  and  $\hat{V} \hat{s}_G := Q_G Q_G^\top \hat{V} \hat{s}_G$ ,

Then applying Lemma 7 of [41] yields with probability at least  $1 - 4\delta$ ,

$$\begin{aligned}\|P_{R(\hat{U})} - A_Z A_Z^\top\|^2 &\leq \frac{\epsilon(n)}{\sigma_{d_Z}(s_Z^*)^2} \\ \|P_{R(\hat{V})} - A_G A_G^\top\|^2 &\leq \frac{\epsilon(n)}{\sigma_{d_G}(s_G^*)^2}.\end{aligned}$$

To prove  $(\epsilon, \nu)$ -disentanglement, define  $Z_{\parallel} := f_1(Z) := P_U A_Z Z$  and  $Z_{\perp} := f_2(G) := P_U A_G G$ . Then by definition,

$$\hat{Z} = Z_{\parallel} + Z_{\perp} = f_1(Z) + f_2(G).$$

Since  $Z_{\parallel}$  is a function of  $Z$  and  $\hat{G} = G$ , we have the independence relations  $Z_{\parallel} \perp\!\!\!\perp \hat{G}$  and  $\hat{Z} \perp\!\!\!\perp \hat{G} | Z_{\perp}$ , and therefore by data processing inequality,

$$\begin{aligned} I(\hat{Z}; \hat{G}) &\leq I(\hat{Z}; Z_{\perp}) \\ &= \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,y)} \log \frac{p_{\hat{Z}Z_{\perp}}(z,y)}{p_{\hat{Z}}(z)p_{Z_{\perp}}(y)} \\ &= \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{\hat{Z}}(z)} \\ &= \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,y)} \log \frac{p_{Z_{\parallel}}(z)}{p_{\hat{Z}}(z)} + \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{Z_{\parallel}}(z)} \\ &= \mathbb{E}_{p_{\hat{Z}}(z)} \log \frac{p_{Z_{\parallel}}(z)}{p_{\hat{Z}}(z)} + \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{Z_{\parallel}}(z)} \\ &= -D_{\text{KL}}(p_{\hat{Z}} \| p_{Z_{\parallel}}) + \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,z-y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{Z_{\parallel}}(z)} \\ &\leq \mathbb{E}_{p_{\hat{Z}Z_{\perp}}(z,z-y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{Z_{\parallel}}(z)} = \mathbb{E}_{p_{Z_{\perp}}(y)p_{Z_{\parallel}}(z-y)} \log \frac{p_{Z_{\parallel}}(z-y)}{p_{Z_{\parallel}}(z)} \\ &= O\left(\mathbb{E}_{p_{Z_{\parallel}}(z-y)p_{Z_{\perp}}(y)} \|z-y\| \|y\| + \mathbb{E}_{p_{Z_{\perp}}(y)} \|y\|^2\right) = O(\|P_U A_G\| + \|P_U A_G\|^2), \end{aligned}$$

where the last inequality uses the nonnegativity of KL divergence and the second-to-last equality combines Lemma B.6 with the fact that  $Z_{\parallel} = P_U A_Z Z$  is a 1-Lipschitz function of  $Z$  and therefore its score function along the column space of  $P_U A_Z$  is  $\lambda_s$ -Lipschitz. To upper bound  $\|P_U A_G\|^2$ , we utilize item 2 by noticing that

$$\|P_U A_G\|^2 = \frac{1}{2} (\|P_U - A_Z A_Z^{\top}\|^2 + \text{rank}(U) - d_Z) \leq \frac{\epsilon(n)}{2\sigma_{d_Z}(s_Z^*)^2},$$

where the last inequality uses the fact that  $\text{rank}(U) \leq d_Z$ . As a result,

$$I(\hat{Z}; \hat{G}) = O\left(\frac{\sqrt{\epsilon(n)}}{\sigma_{d_Z}(s_Z^*)}\right) = O\left(\frac{d_X^{5/4} \log^{3/4} n}{\sigma_{d_Z}(s_Z^*) n^{1/4}}\right).$$

Further, we would like to bound the mutual information gap  $I(\hat{Z}, \hat{G}; X) - I(Z, G; X)$  by finding an estimator of the sample  $X$  from  $(\hat{Z}, \hat{G})$ . To this end, we show that  $[P_U, P_V]$  is invertible with high probability. This is the case since by item 2,

$$\|P_{\hat{U}} + P_{\hat{V}} - I_{d_X}\| \leq \|P_{\hat{U}} - A_Z A_Z^{\top}\| + \|P_{\hat{V}} - A_G A_G^{\top}\| = O\left(\frac{\log^{3/4} n}{n^{1/4}}\right)$$

with high probability. Therefore, by matrix perturbation inequality,

$$\sigma_{d_X}([P_{\hat{U}}, P_{\hat{V}}]) \geq 1 - O\left(\frac{\log^{3/4} n}{n^{1/4}}\right) > 0$$

for sufficiently large  $n$ . As a result, we conclude there exists an  $\hat{f}(P_{\hat{U}} X, P_{\hat{V}} X) := X$  that can reconstruct  $X$  and therefore preserve its information perfectly. Therefore,  $(\hat{Z}, \hat{G})$  are  $\left(O\left(\frac{d_X^{5/4} \log^{3/4} n}{\sigma_{d_Z}(s_Z^*) n^{1/4}}\right), 0\right)$ -disentangled.

To prove  $\epsilon$ -editability, choose  $t_0 := \frac{1}{2} \log \frac{1}{1-\delta^2}$  and  $t_1 := \frac{1}{2} \log \frac{1}{1-\delta}$  for some  $\delta > 0$  and let  $q(\cdot | \hat{Z}_{t_0}, \hat{G})$  be the pdf of the following estimated reverse diffusion process at time  $T - t_1$ :

$$d\hat{X}_t^{\leftarrow} = [\hat{X}_t^{\leftarrow} + 2s_{\hat{\theta}}(\hat{X}_{k\eta}^{\leftarrow}, \hat{Z}_{t_0}, \hat{G}, T - k\eta)]dt + \sqrt{2}dB_t^{\leftarrow}, \hat{X}_0^{\leftarrow} \sim p_T, t \in [k\eta, (k+1)\eta],$$

where

$$s_{\hat{\theta}}(x, z, g, t) := \frac{e^{-t}x - P_{\hat{U}}z - P_{\hat{V}}g}{\sigma(t)^2}. \quad (30)$$

For generated samples  $\hat{X} \sim q(\cdot | \hat{Z}_{t_0}, \hat{G})$  and  $\hat{X}' \sim q(\cdot | \hat{Z}_{t_0}, \hat{G}')$ , where  $\hat{G}' \sim p_G$  is an i.i.d sample of  $\hat{G}$ , we have

$$\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}|Z}, p_{\hat{X}'|Z}) \leq \underbrace{\mathbb{E}_{p_Z} d_{\text{TV}}(p_{X_{t_1}|Z}, p_{\hat{X}|Z})}_{(i)} + \underbrace{\mathbb{E}_{p_Z} d_{\text{TV}}(p_{\hat{X}|Z}, p_{\hat{X}'|Z})}_{(ii)}.$$

To bound (i), notice that the score estimation error of  $s_{\hat{\theta}}$  is

$$\begin{aligned} L_c &:= \mathbb{E}_{t, \hat{Z}_{t_0}, \hat{G}, X_t} \left\| s_{\hat{\theta}}(X_t, \hat{Z}_{t_0}, \hat{G}, t) + \frac{N_t}{\sigma(t)^2} \right\|^2 \\ &= \frac{1}{T-t_1} \int_{t_1}^T \frac{e^{-2t} \mathbb{E} \left\| \hat{Z}_{t_0} + \hat{G} - X \right\|^2}{\sigma(t)^4} dt \\ &= \frac{(e^{2T} - e^{2t_1})}{2(T-t_1)(e^{2t_1} - 1)(e^{2T} - 1)} \mathbb{E} \left\| P_U \hat{Z}_{t_0} + P_V \hat{G} - X \right\|^2 \\ &= O \left( \frac{d_X}{\delta T} \mathbb{E} \left\| P_U \hat{Z}_{t_0} + P_V \hat{G} - X \right\|^2 \right) \\ &= O \left( \frac{d_X^{7/2} \log^{9/8} n}{T \min\{\sigma_{d_Z}^2(s_Z^*), \sigma_{d_G}^2(s_G^*)\} n^{7/16}} + \frac{d_X}{T n^{1/16}} \right), \end{aligned}$$

where the last equality uses  $\delta := \frac{\log^{3/8} n}{n^{1/16}}$  and the fact that

$$\begin{aligned} &\mathbb{E} \left\| P_U \hat{Z}_{t_0} + P_V \hat{G} - X \right\|^2 \\ &\leq 2\mathbb{E} \left\| (P_U A_Z A_Z^\top - A_Z A_Z^\top) X \right\|^2 + 2\mathbb{E} \left\| (P_V A_G A_G^\top - A_G A_G^\top) X \right\|^2 + 2\sigma(t_0)^2 \\ &= O \left( \frac{d_X^{5/2} \log^{3/2} n}{\min\{\sigma_{d_Z}^2(s_Z^*), \sigma_{d_G}^2(s_G^*)\} n^{1/2}} + \frac{\log^{3/8} n}{n^{1/8}} \right). \end{aligned}$$

Further, using an argument similar to Section C.3, we can check that the Novikov's condition holds for  $s_{\hat{\theta}}$ , and using an argument similar to Section C.2, we can apply Girsanov's theorem to show that

$$(i) = O \left( \frac{d_X^{7/4} \log^{9/16} n}{T \min\{\sigma_{d_Z}(s_Z^*), \sigma_{d_G}(s_G^*)\} n^{7/32}} + \frac{d_X}{T n^{1/32}} \right).$$

Further, notice that

$$\begin{aligned} &I(\hat{Z}_{t_0}; \hat{G} | Z) \\ &= \mathbb{E}_{p_{\hat{Z}_{t_0}|G|Z}} \log \frac{p_{\hat{Z}_{t_0}|G,Z}(\hat{z}|g,z)}{p_{\hat{Z}_{t_0}|Z}(\hat{z}|z)} \\ &= \mathbb{E}_{p_{\hat{Z}_{t_0}|G|Z}} \log \frac{p_{\delta N_{t_0}}(\hat{z} - f_1(z) - f_2(g))}{p_{f_2(g) + \delta N_{t_0}}(\hat{z} - f_1(z))} \\ &= -D_{\text{KL}}(p_{\delta N_{t_0}} \| p_{f_2(g) + \delta N_{t_0}}) + \mathbb{E}_{p_{\hat{Z}_{t_0}|Z}} \log \frac{p_{\delta N_{t_0}}(\hat{z} - f_1(z) - f_2(g))}{p_{\delta N_{t_0}}(\hat{z} - f_1(z))} \\ &\leq \frac{1}{\delta^2} O \left( \mathbb{E}_{p_{\hat{Z}_{t_0}|Z}} \|\hat{z} - f_1(z)\| \|f_2(g)\| + \mathbb{E}_{p_{G|Z}} \|f_2(g)\|^2 \right) \\ &= O \left( \frac{1}{\delta^2} \|P_U A_G\| \right) = O \left( \frac{d_X^{5/4} \log^{3/4} n}{\delta^2 \sigma_{d_Z}(s_Z^*) n^{1/4}} \right) = O \left( \frac{d_X^{5/4}}{\sigma_{d_Z}(s_Z^*) n^{1/8}} \right), \end{aligned}$$

where the inequality uses Lemma B.6. By Pinsker's inequality and data processing inequality similar to the proof of Theorem 4.2, we have

$$d_{\text{TV}}(p_{X_{t_1}|Z}, p_{\hat{X}'|Z}) = O\left(\frac{d_X^{5/8}}{\sigma_{d_Z}(s_Z^*)^{1/2} n^{1/16}}\right).$$

Combining the bounds on (i)(ii), we have  $\hat{Z}_{t_0}$  and  $\hat{G}$  are  $O\left(\frac{d_X^{7/4} \log^{9/16} n}{T \min\{\sigma_{d_Z}^{1/2}(s_Z^*), \sigma_{d_G}^{1/2}(s_G^*)\} n^{1/16}}\right)$ -editable.

## D.2 Proof of Lemma 4.4

Let  $J_x f(x) \in \mathbb{R}^{d_1 \times d_2}$  denote the Jacobian matrix of vector function  $f: \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$  with respect to vector  $x \in \mathbb{R}^{d_1}$ . Use the change of variables formula for pdf on  $X_t = A_Z Z_t + A_G G_t$ , we have

$$\begin{aligned} p_t(x) &= p_{Z_t G_t}(z(x), g(x)) |\det(J_x[z(x)^\top, g(x)^\top]^\top)| \\ &= p_{Z_t G_t}(z(x), g(x)) |\det(A)| = p_{Z_t G_t}(z(x), g(x)), \end{aligned}$$

where the last equality uses the fact that the matrix  $A = [A_Z, A_G]$  is orthogonal and thus  $|\det(A)| = 1$ . Further, by Definition 4.3,  $Z_t \perp G_t, \forall t \geq 0$  since  $Z \perp G$  and  $A_Z^\top N_t \perp A_G^\top N_t$  due to the orthogonality of  $R(A_Z)$  and  $R(A_G)$ . Therefore,

$$p_t(x) = p_{Z_t G_t}(z(x), g(x)) = p_{Z_t}(z(x)) p_{G_t}(g(x)).$$

Plugging this into the score of  $p_t(X)$  and applying the chain rule yields

$$\begin{aligned} \nabla_x \log p_t(x) &= J_x z(x) \nabla_z \log p_Z(z(x)) + J_x g(x) \nabla_g \log p_G(g(x)) \\ &= A_Z \nabla_z \log p_Z(z(x)) + A_G \nabla_g \log p_G(g(x)). \end{aligned}$$

## D.3 Proof of Theorem D.2

Let  $s_Z^*(x) := \nabla_z \log p_{Z_t}(z(x))$  and  $s_G^*(g) := \nabla_g \log p_{G_t}(g)$  and  $P_A$  be the projection matrix onto  $R(A)$ , then for any  $(s_Z, s_G, U, V)$ ,

$$\begin{aligned} &L_0(s_Z, s_G, U, V) \\ &:= \mathbb{E}_{t, p_t(x)} \|(P_{A_Z}(U s_Z(x, t) + V s_G(g(x), t)) - A_Z s_Z^*(x, t)) + (P_{A_G}(U s_Z(x, t) + V s_G(g(x), t)) - A_G s_G^*(g(x), t))\|^2 \\ &= \mathbb{E}_{t, p_t(x, t)} [\|P_{A_Z}(U s_Z(x, t) + V s_G(g(x), t)) - A_Z s_Z^*(x, t)\|^2 + \\ &\quad \|P_{A_G}(U s_Z(x, t) + V s_G(g(x), t)) - A_G s_G^*(g(x), t)\|^2] \geq 0 = L_0(s_Z^*, s_G^*, A_Z, A_G). \quad (31) \end{aligned}$$

To analyze the equality condition, notice by the fact that minimizing over a larger set leads to smaller loss and the independence between  $Z_t$  and  $G_t$ ,

$$\begin{aligned} &\mathbb{E}_{t, p_t(x)} \|P_{A_Z}(U s_Z(x, t) + V s_G(g(x), t)) - A_Z s_Z^*(x, t)\|^2 \\ &\geq \mathbb{E}_{t, p_t(x)} \|\mathbb{E}_{p_t(x)} [A_Z s_Z^*(x, t) | U s_Z(x, t) + V s_G(g(x), t)] - A_Z s_Z^*(x, t)\|^2 \\ &\geq \mathbb{E}_{t, p_t(x)} \|\mathbb{E}_{p_t(x)} [A_Z s_Z^*(x, t) | U s_Z(x, t), V s_G(g(x), t)] - A_Z s_Z^*(x, t)\|^2 \\ &= \mathbb{E}_{t, p_t(x)} \|\mathbb{E}_{p_t(x)} [A_Z s_Z^*(x, t) | U s_Z(x, t)] - A_Z s_Z^*(x, t)\|^2, \quad (32) \end{aligned}$$

with equality if and only if

$$\begin{aligned} &\mathbb{E}_{p_t(x)} (s_Z^*(x, t) - P_{A_Z} V s_G(g(x), t)) s_G(g(x), t)^\top = 0, \forall t \in [0, T] \\ &\Rightarrow P_{A_Z} V s_G(g(x), t) = 0, \text{a.s.}, \forall x, t. \end{aligned}$$

by the orthogonality principle. As a result, the equality of Eq. 32 is achieved if and only if

$$P_{A_Z} U s_Z(x, t) = A_Z s_Z^*(x, t), \quad (33)$$

$$P_{A_Z} V s_G(g(x), t) = 0, \quad (34)$$

$$P_{A_G} U s_Z(x, t) + V s_G(g(x), t) = A_G s_G^*(g(x), t), \text{a.s.}, \forall x, t. \quad (35)$$

Now, we turn our attention to the regularizer  $L_r$  and notice that for any  $(s_Z, U)$  that satisfies Eq. 33-35,

$$\begin{aligned} L_r(s_G, V) &\geq \mathbb{E}_{t, p_t(x)} \|V s_G(g(x), t) - \nabla_x \log p_t(x)\|^2 = \mathbb{E}_{t, p_t(x)} \|U s_Z(x, t)\|^2 \\ &\geq \mathbb{E}_{t, p_t(x)} \|P_{A_Z} U s_Z(x, t)\|^2 = \mathbb{E}_{t, p_t(x)} \|A_Z s_Z^*(x, t)\|^2, \end{aligned}$$

where both equalities are achieved if and only if

$$\begin{aligned} V^\top V &= \mathbb{E} s_G(g, t) s_G(g, t)^\top, \\ \|P_{A_G} U s_Z(x, t)\| &= 0, \text{ a.s., } \forall x, t. \end{aligned}$$

Finally, we shall show that for any  $(A_Z, A_G, s_Z^*, s_G^*)$ , there exists some optimal solution  $(U, V, s_Z, s_G)$  such that

$$U s_Z = A_Z s_Z^*, V s_G = A_G s_G^*, \quad (36)$$

$$V^\top V = \mathbb{E} s_G(g, t) s_G(g, t)^\top, L_b(s_Z, s_G, U, V) = 0. \quad (37)$$

To this end, consider the SVD of the operator  $A_Z s_Z^*(x, t) + A_G s_G^*(g(x), t)$  as

$$\begin{aligned} \forall x, t, \quad A_Z s_Z^*(x, t) + A_G s_G^*(g(x), t) &= [\Phi_Z, \Phi_G] \begin{bmatrix} \Sigma_Z & \mathbf{0}_{d_Z \times d_G} \\ \mathbf{0}_{d_G \times d_Z} & \Sigma_G \end{bmatrix} \begin{bmatrix} f_Z(x, t) \\ f_G(g(x), t) \end{bmatrix} \\ \text{s.t.} \quad \Sigma_Z &= \text{diag}(\sigma_1(s_Z^*), \dots, \sigma_{d_Z}(s_Z^*)), \Sigma_G = \text{diag}(\sigma_1(s_G^*), \dots, \sigma_{d_G}(s_G^*)), \\ &\begin{bmatrix} \Phi_Z^\top \\ \Phi_G^\top \end{bmatrix} [\Phi_Z, \Phi_G] = \mathbb{E} \begin{bmatrix} f_Z(x, t) \\ f_G(g(x), t) \end{bmatrix} [f_Z(x, t)^\top, f_G(g(x), t)^\top] = I_{d_X} \end{aligned}$$

Set  $U := \Phi_Z \Sigma_Z^{1/2}$ ,  $s_Z^*(x, t) = \Sigma_Z^{1/2} f_Z(x, t)$ ,  $V := \Phi_G \Sigma_G^{1/2}$ ,  $s_G^*(x, t) = \Sigma_G^{1/2} f_G(g(x), t)$ , we have  $(U, V, s_Z, s_G)$  satisfies Eq. 36. Further, notice that

$$\begin{aligned} U^\top U &= \Sigma_Z^{1/2} \Phi_Z^\top \Phi_Z \Sigma_Z^{1/2} = \Sigma_Z = \Sigma_Z^{1/2} \mathbb{E} f_Z(x, t) f_Z(x, t)^\top \Sigma_Z^{1/2} = \mathbb{E} s_Z(x, t) s_Z(x, t)^\top, \\ U^\top V &= \Sigma_Z^{1/2} \Phi_Z^\top \Phi_G \Sigma_G^{1/2} = 0 = \Sigma_Z^{1/2} \mathbb{E} f_Z(x, t) f_G(g(x), t)^\top \Sigma_G^{1/2} = \mathbb{E} s_Z(x, t) s_G(g(x), t)^\top, \\ V^\top V &= \Sigma_G^{1/2} \Phi_G^\top \Phi_G \Sigma_G^{1/2} = \Sigma_G = \Sigma_G^{1/2} \mathbb{E} f_G(g, t) f_G(g, t)^\top \Sigma_G^{1/2} = \mathbb{E} s_G(g, t) s_G(g, t)^\top. \end{aligned}$$

Therefore,  $(U, V, s_Z, s_G)$  also satisfies Eq. 37.

#### D.4 Proof of Theorem D.3

Define matrices

$$\begin{aligned} W(x, t) &:= \begin{bmatrix} U & V \\ s_Z(x, t)^\top & s_G(g(x), t)^\top \end{bmatrix} =: \begin{bmatrix} \tilde{U} \\ s(x, t)^\top \end{bmatrix}, \\ \tilde{W}(x, t) &:= \begin{bmatrix} W(x, t) \\ \mathbf{0}_{(d_X+1) \times d_Z}, \sqrt{\frac{\lambda_r}{3}} W_G(x, t) \end{bmatrix}, \\ W^*(x, t) &:= \begin{bmatrix} A_Z & A_G \\ s_Z^*(x, t)^\top & s_G^*(g(x), t)^\top \end{bmatrix} =: \begin{bmatrix} A \\ s^*(x, t)^\top \end{bmatrix}, \\ \tilde{W}^*(x, t) &:= \begin{bmatrix} W^*(x, t) \\ \mathbf{0}_{(d_X+1) \times d_Z}, \sqrt{\frac{\lambda_r}{3}} W_G^*(x, t) \end{bmatrix}, \\ N(x, t) &:= W(x, t) W(x, t)^\top, N^*(x, t) := W^*(x, t) W^*(x, t)^\top, \\ \tilde{N}(x, t) &:= \tilde{W}(x, t) \tilde{W}(x, t)^\top, \tilde{N}^*(x, t) := \tilde{W}^*(x, t) \tilde{W}^*(x, t)^\top. \end{aligned}$$

Further, define the *direction of improvement* of  $W(x, t)$ 's and  $\tilde{W}(x, t)$ 's respectively as

$$\begin{aligned} \Delta(x, t) &:= W(x, t) - W^*(x, t) R, \\ \tilde{\Delta}(x, t) &:= \tilde{W}(x, t) - \tilde{W}^*(x, t) R, \\ R &:= \underset{R: R^\top R = R R^\top = I_{d_X}}{\operatorname{argmin}} \mathbb{E}_{t, p_t(x)} [\|\tilde{W}(x, t) - \tilde{W}^*(x, t) R\|^2] \\ &:= \underset{R: R^\top R = R R^\top = I_{d_X}}{\operatorname{argmin}} \mathbb{E}_{t, p_t(x)} \left[ \|W(x, t) - W^*(x, t) R\|^2 + \frac{\lambda_r}{3} \|W_G(x, t) - W_G^*(x, t) R\|^2 \right], \end{aligned}$$

where for any matrices  $M$ , define  $M_Z$  to be its first  $d_Z$  columns and  $M_G$  to be its  $(d_Z + 1)$ -th through  $d_X$ -th columns.

For any set of matrices  $\{C(y, t)\}_{y \in \mathcal{Y}, t \in [0, T]}$  and probability measures  $q_t(y)$ 's, define random matrix  $\mathbf{C}$  as

$$\mathbf{C} = C(y, t) \quad \text{w.p.} \quad p(t)q_t(y),$$

where  $p(t)$  is some fixed distribution of the diffusion time  $t$ . Further, define a blockwise representation of  $\mathbf{C}$  as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{c}_2 \\ \mathbf{c}_1 & c_{11} \end{bmatrix},$$

where  $\mathbf{C}_0$  is  $\mathbf{C}$  deleting the last row and column.

For a pair of random matrices  $(\mathbf{C}_1, \mathbf{C}_2)$ , define

$$\mathbf{C}_1 \mathbf{C}_2 = C_1(y, t) C_2(y', t) \quad \text{w.p.} \quad p(t)q_t(y)q_t(y').$$

Next, let  $[\mathbf{A}, \mathbf{B}]_{\mathcal{K}}$  denote the *bilinear form* between random matrices  $(\mathbf{A}, \mathbf{B})$  weighted by the operator  $\mathcal{K}$ , and  $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{K}} = [\mathbf{A}, \mathbf{B}]_{\mathcal{K}}$  be an *inner product* between random matrices  $\mathbf{A}$  and  $\mathbf{B}$  if  $\mathcal{K}$  is positive definite. Then we define the following bilinear forms between random matrices  $(\mathbf{A}, \mathbf{B})$  with weight operators  $\mathcal{I}, \mathcal{G}$  and  $\mathcal{H}_0$  respectively as

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{I}} &:= \mathbb{E}_{t, q_t(y)} \langle A(y, t), B(y, t) \rangle \\ [\mathbf{A}, \mathbf{B}]_{\mathcal{G}} &:= \mathbb{E}_{t, q_t(y)} [\langle A_0(y, t), B_0(y, t) \rangle + a_{11}(y, t)b_{11}(y, t) - \\ &\quad \langle a_1(y, t), b_1(y, t) \rangle - \langle a_2(y, t), b_2(y, t) \rangle], \\ [\mathbf{A}, \mathbf{B}]_{\mathcal{H}_0} &:= \mathbb{E}_{t, q_t(y)} [\langle a_1(y, t), b_1(y, t) \rangle + \langle a_2(y, t), b_2(y, t) \rangle]. \end{aligned}$$

It can be verified that  $\langle \cdot, \cdot \rangle_{\mathcal{I}}$  is indeed an inner product satisfying properties such as conjugate symmetry, linearity in the first argument and positive definiteness, and therefore we can define the norm  $\|\mathbf{A}\|_{\mathcal{I}} := \langle \mathbf{A}, \mathbf{A} \rangle_{\mathcal{I}}$ . In addition, it can be checked that  $[\cdot, \cdot]_{\mathcal{G}}$  and  $[\cdot, \cdot]_{\mathcal{H}_0}$  are conjugate symmetric and linear in the first argument. An important relation we use repeatedly later is the fact that

$$2[\mathbf{C}, \mathbf{C}]_{\mathcal{H}_0} + [\mathbf{C}, \mathbf{C}]_{\mathcal{G}} = [\mathbf{C}, \mathbf{C}]_{2\mathcal{H}_0 + \mathcal{G}} = \|\mathbf{C}\|_{\mathcal{I}}^2, \quad (38)$$

since

$$[\mathbf{C}, \mathbf{C}]_{2\mathcal{H}_0 + \mathcal{G}} = 2(\|\mathbf{c}_1\|_{\mathcal{I}}^2 + \|\mathbf{c}_2\|_{\mathcal{I}}^2) + \|\mathbf{C}_0\|_{\mathcal{I}}^2 + \|c_{11}\|_{\mathcal{I}}^2 - \|\mathbf{c}_1\|_{\mathcal{I}}^2 - \|\mathbf{c}_2\|_{\mathcal{I}}^2 = \|\mathbf{C}\|_{\mathcal{I}}^2.$$

Under these definitions, we prove the following Lemma in Section D.5.

**Lemma D.5.** Let  $\tilde{L}^{\lambda_r}(\mathbf{W}) := L^{\lambda_r}(\theta)$  with  $\lambda_r = 3$ , and  $\mathbf{W}$  be an approximate critical point of  $\tilde{L}(\mathbf{W})$  so that  $\langle \nabla \tilde{L}(\mathbf{W}), \Delta' \rangle_{\mathcal{I}} \leq \epsilon \|\Delta'\|_{\mathcal{I}}$  for any random matrix  $\Delta'$ . Then the following holds

$$[\Delta, \Delta]_{\nabla^2 \tilde{L}^{\lambda_r}(\mathbf{W})} \leq \|\tilde{\Delta} \tilde{\Delta}^\top\|_{\mathcal{I}}^2 - 3\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 + \epsilon \|\Delta\|_{\mathcal{I}}, \quad (39)$$

where  $\nabla^2 f(\mathbf{W})$  denotes the Hessian operator of the functional  $f$  at  $\mathbf{W}$ .

To proceed, we use the following lemma for random matrices analogous to Lemma 40 and 41 in [125] and defer its proof to Section D.6.

**Lemma D.6.** If  $\mathbb{E}_{p(\mathbf{X})}[U(\mathbf{X})^\top Y(\mathbf{X})]$  is a positive semi-definite (PSD) matrix, then for independent, identically distributed random variables  $\mathbf{X}, \mathbf{X}'$ ,

$$\begin{aligned} &\mathbb{E} \|U(\mathbf{X})U(\mathbf{X}')^\top - Y(\mathbf{X})Y(\mathbf{X}')^\top\|^2 \geq \\ &\max \left\{ \frac{1}{2} \mathbb{E} \|(U(\mathbf{X}) - Y(\mathbf{X}))(U(\mathbf{X}') - Y(\mathbf{X}'))^\top\|^2, 2(\sqrt{2} - 1) \mathbb{E} \|(U(\mathbf{X}) - Y(\mathbf{X}))U(\mathbf{X}')^\top\|^2 \right\}. \end{aligned}$$

Further, we prove in Section D.7 the following lemma showing that gradient descent converges to a local optimum of the objective  $\tilde{L}$ .

**Lemma D.7.** The gradient flow equation Eq. 12 converges in probability to a solution  $(\hat{U}, \hat{V}, \hat{\theta}_Z, \hat{\theta}_G)$  such that for some  $\epsilon > 0$  and random matrix  $\Delta'(t)$ :

$$\begin{aligned} \langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta' \rangle_{\mathcal{I}} &\leq \epsilon, \\ [\Delta', \Delta']_{\nabla^2 \tilde{L}^{\lambda_r}} &\geq 0. \end{aligned}$$

To prove the theorem, first consider the singular value decomposition  $\mathbb{E}_{t,p_t(x)} \tilde{W}^*(x,t)^\top \tilde{W}(x,t) =: \Phi \Sigma f^\top$ , and by the definition of the direction of improvement,

$$\begin{aligned} R &:= \operatorname{argmin}_{RR^\top = R^\top R = I_{d_X}} \|\tilde{\mathbf{W}} - \tilde{\mathbf{W}}^* R\|_{\mathcal{I}}^2 \\ &= \operatorname{argmax}_{RR^\top = R^\top R = I_{d_X}} \langle \tilde{\mathbf{W}}, \tilde{\mathbf{W}}^* R \rangle_{\mathcal{I}} = \operatorname{argmax}_{RR^\top = R^\top R = I_{d_X}} \langle \mathbb{E}_{t,p_t(x)} \tilde{W}^*(x,t)^\top \tilde{W}(x,t), R \rangle \\ &= \operatorname{argmax}_{RR^\top = R^\top R = I_{d_X}} \langle \Sigma, \Phi^\top R f \rangle = \Phi f^\top, \end{aligned}$$

where the last equality holds since  $R' := \Phi^\top R f$  is orthogonal,  $|R'_{ii}| \leq 1$  and

$$\langle \Sigma, R' \rangle \leq \sum_{i=1}^{d_X} \Sigma_{ii},$$

with equality iff  $R' = I_{d_X}$ . As a result,

$$\mathbb{E}_{t,p_t(x)} \tilde{W}^\top(x,t) \tilde{W}^*(x,t) R = f \Sigma f^\top = \mathbb{E}_{p_t(x),t} R^\top \tilde{W}^{*\top}(x,t) \tilde{W}(x,t)$$

is PSD. Applying Lemma D.6 on  $U(\mathbf{X}) = \tilde{\mathbf{W}}^* R$  and  $Y(\mathbf{X}) = \tilde{\mathbf{W}}$  yields

$$\begin{aligned} &\|\tilde{\Delta} \tilde{\Delta}^\top\|_{\mathcal{I}}^2 - 3\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 \\ &\leq 2\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 - 3\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 = -\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 \leq -2(\sqrt{2}-1)\|\tilde{\Delta}\|_{\mathcal{I}}^2 \leq -0.8\|\tilde{\Delta}\|_{\mathcal{I}}^2. \end{aligned}$$

Combining this with Lemma D.5 yields

$$[\Delta, \Delta]_{\nabla^2 \tilde{L}^{\lambda_r}} \leq -0.8\|\tilde{\Delta}\|_{\mathcal{I}}^2,$$

Therefore, we have the LHS to be positive only if  $\|\tilde{\Delta}\|_{\mathcal{I}} = 0$ . Set  $\epsilon = 0$  and applying Theorem D.2 yields  $R(U) = R(A_Z), R(V) = R(A_G)$ .

## D.5 Proof of Lemma D.5

Define  $\tilde{L}_b(\mathbf{W}) := [\mathbf{N}, \mathbf{N}]_{\mathcal{G}}^2$ , then it can be verified that

$$\begin{aligned} L_0(\theta) &= \frac{1}{2}[\mathbf{N} - \mathbf{N}^*, \mathbf{N} - \mathbf{N}^*]_{\mathcal{H}_0}^2 =: \tilde{L}_0(\mathbf{W}), \\ L_b(\theta) &= [\mathbf{N}, \mathbf{N}]_{\mathcal{G}}^2 + \lambda_r[\tilde{\mathbf{N}}, \tilde{\mathbf{N}}]_{\mathcal{G}}^2 = \tilde{L}_b(\mathbf{W}) + \lambda_r \tilde{L}_b(\tilde{\mathbf{W}}). \end{aligned}$$

Define  $\tilde{L}_r(\mathbf{W}) := L_r(\theta) + \lambda_r \tilde{L}_b(\tilde{\mathbf{W}})$ , then we have

$$\tilde{L}^{\lambda_r}(\mathbf{W}) = 2\tilde{L}_0(\mathbf{W}) + \frac{1}{2}\tilde{L}_b(\mathbf{W}) + \lambda_r \tilde{L}_r(\mathbf{W}) = [\mathbf{N} - \mathbf{N}^*, \mathbf{N} - \mathbf{N}^*]_{\mathcal{H}_0}^2 + \frac{1}{2}[\mathbf{N}, \mathbf{N}]_{\mathcal{G}}^2 + \lambda_r \tilde{L}_r(\mathbf{W}).$$

Then, consider the Fréchet derivative of  $\tilde{L}_t(\mathbf{W})$  along  $\Delta(t)$ , it can be shown that

$$\begin{aligned} \langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta \rangle_{\mathcal{I}} &= [\mathbf{N} - \mathbf{N}^*, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{2\mathcal{H}_0} + [\mathbf{N}, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{\mathcal{G}} + \lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} \\ &= [\mathbf{N} - \mathbf{N}^*, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}} + [\mathbf{N}^*, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{\mathcal{G}} + \lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} \\ &= [\mathbf{N} - \mathbf{N}^*, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}} + 2[\mathbf{N}^*, \mathbf{N}]_{\mathcal{G}} + \lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}}, \quad (40) \end{aligned}$$

where the last equality uses the fact that  $[\mathbf{N}^*, \mathbf{W}^* \mathbf{W}^\top]_{\mathcal{G}} = [\mathbf{N}^*, \mathbf{W} \mathbf{W}^{*\top}]_{\mathcal{G}} = 0$ . To see this, notice that

$$\begin{aligned} [\mathbf{N}^*, \mathbf{W}^* \mathbf{W}^\top]_{\mathcal{G}} &= \langle A A^\top, \tilde{U} A^\top \rangle + \mathbb{E}_{t,p_t(x)p_t(x')} s^*(x,t)^\top s^*(x',t) s(x,t)^\top s^*(x',t) \\ &\quad - \mathbb{E}_{t,p_t(x)} s^*(x,t)^\top A^\top \tilde{U} s^*(x,t) - \mathbb{E}_{t,p_t(x)} s^*(x,t)^\top A^\top A s(x,t) \\ &= \operatorname{Tr}(A(A^\top A - \mathbb{E}_{t,p_t(x)}[s^*(x,t)s^*(x,t)])\tilde{U}) + \\ &\quad \mathbb{E}_{t,p_t(x)}[s^*(x,t)^\top (\mathbb{E}_{p_t(x',t)}[s^*(x',t)s^*(x',t)]^\top - A^\top A) s(x,t)] = 0, \end{aligned}$$

where the last inequality uses the fact that  $\tilde{L}_b(\mathbf{W}^*) = 0$  and therefore

$$\begin{aligned} U^{*\top} U^* &= \mathbb{E}_{t, p_t(x)} s_Z^*(x, t) s_Z^*(x, t)^\top, \\ V^{*\top} V^* &= \mathbb{E}_{t, p_t(x)} s_G^*(x, t) s_G^*(x, t)^\top. \end{aligned}$$

Similarly, we can show that  $[\mathbf{N}^*, \mathbf{W}\mathbf{W}^{*\top}]_{\mathcal{G}} = 0$ .

Now, consider the Hessian of  $\tilde{L}$  along  $\Delta$  by taking the Fréchet derivative of Eq. 40,

$$\begin{aligned} [\Delta, \Delta]_{\nabla^2 \tilde{L}^{\lambda_r}} &= [\Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}}^2 + 2[\mathbf{N} - \mathbf{N}^*, \Delta \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}} + \\ &\quad 2[\mathbf{N}^*, \mathbf{W} \Delta^\top + \Delta \mathbf{W}^\top]_{\mathcal{G}} + \lambda_r [\Delta, \Delta]_{\nabla^2 \tilde{L}_r(\mathbf{W})} \\ &= \|\Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top\|_{\mathcal{I}}^2 + 2[\mathbf{N} - \mathbf{N}^*, \Delta \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}} + 2[\mathbf{N}^*, \mathbf{W} \Delta^\top + \Delta \mathbf{W}^\top]_{\mathcal{G}} + \lambda_r [\Delta, \Delta]_{\nabla^2 \tilde{L}_r(\mathbf{W})}, \end{aligned}$$

where we use Eq. 38 in the last equality. For the first term of the right-hand side, by the choice of  $\Delta$ ,

$$\begin{aligned} \|\Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top\|_{\mathcal{I}}^2 &= \|\mathbf{N} - \mathbf{N}^* + \Delta \Delta^\top\|_{\mathcal{I}}^2 \\ &= \|\Delta \Delta^\top\|_{\mathcal{I}}^2 + 2\langle \mathbf{N} - \mathbf{N}^*, \Delta \mathbf{W}^\top + \mathbf{W} \Delta^\top \rangle_{\mathcal{I}} - \|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2 \\ &= \|\Delta \Delta^\top\|_{\mathcal{I}}^2 + 2\langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - 4\langle \mathbf{N}^*, \mathbf{N} \rangle_{\mathcal{G}} - 2\lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - \|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2. \end{aligned}$$

For the second term,

$$\begin{aligned} [\mathbf{N} - \mathbf{N}^*, \Delta \Delta^\top]_{2\mathcal{H}_0 + \mathcal{G}} &= [\mathbf{N} - \mathbf{N}^*, \mathbf{W} \Delta^\top + \Delta \mathbf{W}^\top]_{2\mathcal{H}_0 + \mathcal{G}} - [\mathbf{N} - \mathbf{N}^*, \mathbf{N} - \mathbf{N}^*]_{2\mathcal{H}_0 + \mathcal{G}} \\ &= \langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - 2\langle \mathbf{N}^*, \mathbf{N} \rangle_{\mathcal{G}} - \lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - \|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2. \quad (41) \end{aligned}$$

Combined with the fact that

$$[\mathbf{N}^*, \mathbf{W} \Delta^\top + \Delta \mathbf{W}^\top]_{\mathcal{G}} = 2[\mathbf{N}^*, \mathbf{N}]_{\mathcal{G}},$$

we obtain

$$\begin{aligned} [\Delta, \Delta]_{\nabla^2 \tilde{L}^{\lambda_r}} &= \|\Delta \Delta^\top\|_{\mathcal{I}}^2 + 4\langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - 3\|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2 \\ &\quad - 4[\mathbf{N}, \mathbf{N}^*]_{\mathcal{G}} - 4\lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} + \lambda_r [\Delta, \Delta]_{\nabla^2 \tilde{L}_r(\mathbf{W})} \\ &\leq \|\Delta \Delta^\top\|_{\mathcal{I}}^2 - 3\|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2 + 4\langle \nabla \tilde{L}^{\lambda_r}(\mathbf{W}), \Delta \rangle_{\mathcal{I}} - 4\lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} + \lambda_r [\Delta, \Delta]_{\nabla^2 \tilde{L}_r(\mathbf{W})} \\ &\leq \|\Delta \Delta^\top\|_{\mathcal{I}}^2 - 3\|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2 + 4\epsilon \|\Delta\|_{\mathcal{I}} - 4\lambda_r \langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}} + \lambda_r [\Delta, \Delta]_{\nabla^2 \tilde{L}_r(\mathbf{W})}, \quad (42) \end{aligned}$$

where the first inequality uses the fact that

$$\begin{aligned} [\mathbf{N}, \mathbf{N}^*]_{\mathcal{G}} &= \langle U U^\top, A A^\top \rangle + \mathbb{E}_{t, p_t(x) p_t(x')} s^*(x, t)^\top s^*(x', t) s(x, t)^\top x(x', t) - 2\mathbb{E}_{t, p_t(x)} s(x, t) U^\top A s^*(x', t) \\ &= \|U^\top A - \mathbb{E}_{t, p_t(x)} s(x, t) s^*(x, t)^\top\|^2 \geq 0, \end{aligned}$$

and the second inequality uses the condition that  $\mathbf{W}$  is an approximate critical point of  $\tilde{L}^{\lambda_r}(\mathbf{W})$ . It remains to bound the  $\tilde{L}_r$  related terms. Define  $\mathbf{N}^Z := \mathbf{W}_Z \mathbf{W}_Z^\top$ ,  $\mathbf{N}^G := \mathbf{W}_G \mathbf{W}_G^\top$  and similarly  $\mathbf{N}^{Z*}$  and  $\mathbf{N}^{G*}$ . Notice that

$$\begin{aligned} \tilde{L}_r(\mathbf{W}) &= [\mathbf{N}^G - \mathbf{N}^*, \mathbf{N}^G - \mathbf{N}^*]_{\mathcal{H}_0} + \frac{1}{2} [\mathbf{N}^G, \mathbf{N}^G]_{\mathcal{G}} \\ &= [\mathbf{N}^G - \mathbf{N}^{G*}, \mathbf{N}^G - \mathbf{N}^{G*}]_{\mathcal{H}_0}^2 - 2[\mathbf{N}^G - \mathbf{N}^{G*}, \mathbf{N}^{Z*}]_{\mathcal{H}_0} + [\mathbf{N}^{Z*}, \mathbf{N}^{Z*}]_{\mathcal{H}_0} + \frac{1}{2} [\mathbf{N}^G, \mathbf{N}^G]_{\mathcal{G}} \\ &= [\mathbf{N}^G - \mathbf{N}^{G*}, \mathbf{N}^G - \mathbf{N}^{G*}]_{\mathcal{H}_0} + \frac{1}{2} [\mathbf{N}^G, \mathbf{N}^G]_{\mathcal{G}} = 2\tilde{L}_0(\mathbf{W}_G) + \frac{1}{2} \tilde{L}_b(\mathbf{W}_G) = \tilde{L}^0(\mathbf{W}_G), \end{aligned}$$

where the second-to-last equality uses the fact that

$$\begin{aligned} [\mathbf{N}^G - \mathbf{N}^{G*}, \mathbf{N}^{Z*}]_{\mathcal{H}_0} &= 2\mathbb{E}_{t, p_t(x)} \langle V s_G(g(x), t) - A_G s_G^*(g(x), t), A_Z s_Z^*(z(x), t) \rangle \\ &= 2\mathbb{E}_{t, p_t(x)} \langle V s_G(g(x), t), A_Z s_Z^*(z(x), t) \rangle \\ &= 2\langle V \mathbb{E}_{t, p_t(x)} s_G(g(x), t), A_Z \mathbb{E}_{t, p_t(x)} s_Z^*(z(x), t) \rangle = 0, \end{aligned}$$

where the second-to-last equality uses the independence of the content variable  $Z$  and the style variable  $G$  and the last equality uses the property of the score function:

$$\mathbb{E}_{p_Z} s_Z^*(z) = \mathbb{E}_{p_Z} \nabla_z \log p_Z(z) = \int \nabla_z p_Z(z) dz = \nabla_z \int p_Z(z) dz = 0.$$

Therefore, we can apply Eq. 42 with  $\lambda_r = 0$  to obtain

$$[\Delta, \Delta]_{\tilde{L}_r(\mathbf{W})} = \langle \Delta_G, \Delta_G \rangle_{\tilde{L}^0(\mathbf{W}_G)} \quad (43)$$

$$\leq \|\Delta_G \Delta_G^\top\|_{\mathcal{I}}^2 - 3\|\mathbf{N}^G - \mathbf{N}^{G*}\|_{\mathcal{I}}^2 + 4\langle \nabla \tilde{L}_r(\mathbf{W}), \Delta \rangle_{\mathcal{I}}. \quad (44)$$

Plugging Eq. 43 into Eq. 42 yields

$$\begin{aligned} [\Delta, \Delta]_{\tilde{L}^{\lambda_r}(\mathbf{W})} &\leq \|\Delta \Delta^\top\|_{\mathcal{I}}^2 + \lambda_r \|\Delta_G \Delta_G^\top\|_{\mathcal{I}} - 3(\|\mathbf{N} - \mathbf{N}^*\|_{\mathcal{I}}^2 + \lambda_r \|\mathbf{N}^G - \mathbf{N}^{G*}\|_{\mathcal{I}}^2) + \epsilon \|\Delta\|_{\mathcal{I}} \\ &= \|\tilde{\Delta} \tilde{\Delta}^\top\|_{\mathcal{I}}^2 - 3\|\tilde{\mathbf{N}} - \tilde{\mathbf{N}}^*\|_{\mathcal{I}}^2 + \epsilon \|\Delta\|_{\mathcal{I}}, \end{aligned}$$

where the last equality uses the definition of  $\mathbf{N}, \mathbf{N}^*, \Delta$  and  $\Delta^*$ . For example, for  $\mathbf{N}$ , we have

$$\begin{aligned} \|\tilde{\mathbf{N}}\|_{\mathcal{I}}^2 &= \left\| \begin{bmatrix} \mathbf{W} \mathbf{W}^\top & \sqrt{\frac{\lambda_r}{3}} \mathbf{W}_G \mathbf{W}_G^\top \\ \sqrt{\frac{\lambda_r}{3}} \mathbf{W}_G \mathbf{W}_G^\top & \frac{\lambda_r}{3} \mathbf{W}_G \mathbf{W}_G^\top \end{bmatrix} \right\|_{\mathcal{I}}^2 = \|\mathbf{N}\|_{\mathcal{I}}^2 + \lambda_r \|\mathbf{N}^G\|_{\mathcal{I}}^2 + \left( \frac{\lambda_r^2}{9} - \frac{\lambda_r}{3} \right) \|\mathbf{N}^G\|_{\mathcal{I}}^2 \\ &= \|\mathbf{N}\|_{\mathcal{I}}^2 + 3\|\mathbf{N}^G\|_{\mathcal{I}}^2, \end{aligned}$$

where in the last equality we set  $\lambda_r = 3$ .

## D.6 Proof of Lemma D.6

First, let  $\Delta(\mathbf{X}) := U(\mathbf{X}) - Y(\mathbf{X})$ , by definition,

$$\mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X}) = \mathbb{E} U(\mathbf{X})^\top U(\mathbf{X}) - \mathbb{E} Y(\mathbf{X})^\top U(\mathbf{X}) = \mathbb{E} U(\mathbf{X})^\top \Delta(\mathbf{X}),$$

and

$$\mathbb{E} \|U(\mathbf{X}) U(\mathbf{X}')^\top - Y(\mathbf{X}) Y(\mathbf{X}')^\top\|^2 = \mathbb{E} \|\Delta(\mathbf{X}) U(\mathbf{X}')^\top + U(\mathbf{X}) \Delta(\mathbf{X}')^\top - \Delta(\mathbf{X}) \Delta(\mathbf{X}')^\top\|^2.$$

Expanding the square norm,

$$\begin{aligned} &\mathbb{E} \|\Delta(\mathbf{X}) U(\mathbf{X}')^\top + U(\mathbf{X}) \Delta(\mathbf{X}')^\top - \Delta(\mathbf{X}) \Delta(\mathbf{X}')^\top\|^2 \\ &= 2\mathbb{E} \|\Delta(\mathbf{X}) U(\mathbf{X}')^\top\|^2 + \mathbb{E} \|\Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 + 2\langle \mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X}), \mathbb{E} U(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle - \\ &\quad 2\mathbb{E} \langle \Delta(\mathbf{X}) U(\mathbf{X}')^\top + U(\mathbf{X}) \Delta(\mathbf{X}')^\top, \Delta(\mathbf{X}) \Delta(\mathbf{X}')^\top \rangle \\ &= \mathbb{E} \|\Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 + 2\langle \mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X}), \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle + 2\|\mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X})\|^2 - \\ &\quad 4\langle \mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X}), \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle \\ &= \frac{1}{2} \mathbb{E} \|\Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 + 2\langle \mathbb{E} U(\mathbf{X})^\top Y, \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle + \\ &\quad \|\sqrt{2} U(\mathbf{X})^\top \Delta(\mathbf{X}) - \frac{1}{\sqrt{2}} \Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 \\ &\geq \frac{1}{2} \mathbb{E} \|\Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 = \frac{1}{2} \mathbb{E} \|(U(\mathbf{X}) - Y(\mathbf{X}))(U(\mathbf{X}') - Y(\mathbf{X}'))^\top\|^2, \end{aligned}$$

where the second equality uses the symmetry of  $\mathbb{E} U(\mathbf{X})^\top \Delta(\mathbf{X})$  the last inequality uses the PSD of  $\mathbb{E} U(\mathbf{X})^\top Y(\mathbf{X})$ .

Similarly,

$$\begin{aligned} &\mathbb{E} \|U(\mathbf{X}) U(\mathbf{X}')^\top - Y(\mathbf{X}) Y(\mathbf{X}')^\top\|^2 \\ &= 2\mathbb{E} \|\Delta(\mathbf{X}) U(\mathbf{X}')^\top\|^2 + \mathbb{E} \|\Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 + 2\|\mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X})\|^2 - \\ &\quad 4\langle \mathbb{E} \Delta(\mathbf{X})^\top U(\mathbf{X}), \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle \\ &= (2\sqrt{2} - 2) \mathbb{E} \|U(\mathbf{X}) \Delta(\mathbf{X}')^\top\|^2 + (4 - 2\sqrt{2}) \langle \mathbb{E} U(\mathbf{X})^\top Y(\mathbf{X}), \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X}) \rangle + \\ &\quad \|\sqrt{2} \mathbb{E} U(\mathbf{X})^\top \Delta(\mathbf{X}) - \mathbb{E} \Delta(\mathbf{X})^\top \Delta(\mathbf{X})\|^2 \\ &\geq (2\sqrt{2} - 2) \mathbb{E} \|U(\mathbf{X}) \Delta(\mathbf{X}')^\top\|^2 = 2(\sqrt{2} - 1) \mathbb{E} \|(U(\mathbf{X}) - Y(\mathbf{X})) U(\mathbf{X}')^\top\|^2, \end{aligned}$$

where the last inequality uses the PSD of  $\mathbb{E} U(\mathbf{X})^\top Y(\mathbf{X})$ .

## D.7 Proof of Lemma D.7

The proof relies on the following lemma.

**Lemma D.8.** *Suppose the score network parameters are initialized randomly as  $\theta_Z(0), \theta_G(0) \sim \mathcal{N}(0, I)$ . Then as  $d_H \rightarrow \infty$ , the NTKs  $K_Z$  and  $K_G$  converge to some kernels  $K_Z^*$  and  $K_G^*$  fixed during training. Further, define the minimal eigenvalues of operator  $K : \mathcal{X} \times [t_0, T] \times \mathcal{X} \times [t_0, T] \mapsto \mathbb{R}^d$  as*

$$\lambda_{\min}(K) := \inf_{v: \mathbb{E}_{t,p(x)} \|v(t, x)\|^2 = 1} \mathbb{E}_{t,t',p(x)p(x')} v(x, t)^\top K(x, t, x', t') v(x', t'),$$

*then the minimal eigenvalues of  $K_Z^*$  and  $K_G^*$  satisfy  $\min\{\lambda_{\min}(K_Z^*), \lambda_{\min}(K_G^*)\} > 0$ .*

Define the random gradient of loss  $L$  with respect to random matrix  $\mathbf{Y}$  with probability density  $q_t$  as

$$\nabla_{\mathbf{Y}} L(\mathbf{Y}) = \nabla_{Y(x,t)} L(\mathbf{Y}) \quad \text{w.p.} \quad p(t)q_t(x).$$

By the definition of the gradient flow equations in Eq. 12, we have

$$\begin{aligned} \dot{\tilde{L}}(\mathbf{W}) &= \langle \nabla_{\mathbf{W}} \tilde{L}, \dot{\mathbf{W}} \rangle_{\mathcal{I}} \\ &= -\|\nabla_U \tilde{L}(\mathbf{W})\|^2 - \|\nabla_V \tilde{L}(\mathbf{W})\|^2 - \|\mathbb{E}_{t,p_t} J_{\theta_Z} s_Z \nabla_{s_Z} \tilde{L}(\mathbf{W})\|^2 - \|\mathbb{E}_{t,p_t} J_{\theta_G} s_G \nabla_{s_G} \tilde{L}(\mathbf{W})\|^2, \end{aligned}$$

where the first two terms of the RHS by the property of the gradient flow, vanishes if and only if the gradients  $\nabla_U \tilde{L}(\mathbf{W})$  and  $\nabla_V \tilde{L}(\mathbf{W})$  become 0. For the third term of the RHS, notice that

$$\begin{aligned} &\|\mathbb{E}_{t,p_t} J_{\theta_Z} s_Z \nabla_{s_Z} \tilde{L}(\mathbf{W})\|^2 \\ &= \mathbb{E}_{t,t',p_t(x)p_{t'}(x')} \nabla_{s_Z(x,t)} \tilde{L}(\mathbf{W})^\top J_{\theta_Z} s_Z(x, t)^\top J_{\theta_Z} s_Z(x', t') \nabla_{s_Z(x', t')} \tilde{L}(\mathbf{W}) \\ &= \mathbb{E}_{t,t',p_t(x)p_{t'}(x')} \nabla_{s_Z(x,t)} \tilde{L}(\mathbf{W})^\top K_Z(x, t, x', t') \nabla_{s_Z(x', t')} \tilde{L}(\mathbf{W}) \\ &\geq \lambda_{\min}(K_Z) \int \|p(t)p_t(x) \nabla_{s_Z(x,t)} \tilde{L}(\mathbf{W})\|^2 dx dt. \end{aligned}$$

By Lemma D.8,  $\lambda_{\min}(K_Z) \xrightarrow{d_H \rightarrow \infty} \lambda_{\min}(K_Z^*) > 0$  and thus the term vanishes if and only if

$$\|p(t)p_t(x) \mathbb{E}_{t,p_t(x)} \nabla_{s_Z(x,t)} \tilde{L}(\mathbf{W})\|^2 = 0, \forall x, t.$$

Similarly, we can show that the gradient flow converges only if

$$\|p(t)p_t(x) \mathbb{E}_{t,p_t(x)} \nabla_{s_G(g(x), t)} \tilde{L}(\mathbf{W})\|^2 = 0, \forall x, t.$$

For the second order condition, we use the well-known result that gradient descent (with small noise) is able to escape saddle points almost surely [126].

## D.8 Proof of Lemma D.8

Due to the symmetry in their forms, it suffices to prove the statement for  $s_Z$  and  $K_Z$  and we omit the subscript when the context is clear. Large of large number and the standard theory on neural tangent kernel [123] yields

$$K_Z \xrightarrow{d_H \rightarrow \infty} \mathbb{E}_{\text{vec}(\theta_Z) \sim \mathcal{N}(0, I)} J_{\text{vec}(\theta_Z)} s_Z(x, t)^\top J_{\text{vec}(\theta_Z)} s_Z(x', t'), \quad (45)$$

which stays fixed during training.

Define  $\tilde{x} := [x^\top, \text{PE}(t)]^\top$  and  $a(\tilde{x}) := (\theta^{(1)\top} \tilde{x})_+$ . Later, we will slightly abuse the notation to represent  $(x, t)$  when  $\tilde{x}$  appears in the true score functions. By definition,

$$\begin{aligned} K_Z^{(2)}(x, t, x', t') &= \mathbb{E}_{\mathcal{N}(0, I)} J_{\text{vec}(\theta_Z^{(2)})} s_Z(\tilde{x})^\top J_{\text{vec}(\theta_Z^{(2)})} s_Z(\tilde{x}') \\ &= \mathbb{E}_{\mathcal{N}(0, I)} J_{\text{vec}(\theta_Z^{(2)})} \theta_Z^{(2)} a(\tilde{x})^\top J_{\text{vec}(\theta_Z^{(2)})} \theta_Z^{(2)} a(\tilde{x}') \\ &= \mathbb{E}_{\mathcal{N}(0, I)} (a(\tilde{x})^\top \otimes I_{d_Z}) (a(\tilde{x}') \otimes I_{d_X}) = \mathbb{E}_{\mathcal{N}(0, I)} a(\tilde{x})^\top a(\tilde{x}') I_{d_Z}, \end{aligned}$$

where  $\otimes$  is the Kronecker product. Notice that by definition  $K_Z^{(2)} \succeq 0$ . Similarly,

$$\begin{aligned} K_Z^{(1)}(x, t, x', t') &:= \mathbb{E}_{\mathcal{N}(0, I)} J_{\text{vec}(\theta_Z^{(1)})} s_Z(\tilde{x})^\top J_{\text{vec}(\theta_Z^{(1)})} s_Z(\tilde{x}') \\ &= \mathbb{E}_{\mathcal{N}(0, I)} J_{a(\tilde{x})} s_Z(\tilde{x})^\top J_{\theta_Z^{(1)}} a(\tilde{x})^\top J_{\theta_Z^{(1)}} a(\tilde{x}') J_{a(\tilde{x}')} s_Z(\tilde{x}') \\ &=: (\tilde{x}^\top \tilde{x}') \mathbb{E}_{\mathcal{N}(0, I)} \theta_Z^{(2)} S^2(\tilde{x}, \tilde{x}') \theta_Z^{(2)\top}, \\ &= (\tilde{x}^\top \tilde{x}') \mathbb{E}_{\mathcal{N}(0, I)} S(\tilde{x}, \tilde{x}'), \end{aligned}$$

where we use the independence between  $(\theta_Z^{(1)}, \theta_Z^{(2)})$  to cancel out  $\theta_Z^{(1)}$ , and

$$S_{ij}(\tilde{x}, \tilde{x}') := \begin{cases} \mathbb{1}[\theta^{(1),j\top} \tilde{x} \geq 0, \theta^{(1),j\top} \tilde{x}' \geq 0], & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

By assumption, we have  $\theta^{(1),j\top} \tilde{x}$ 's are zero-mean Gaussians and thus

$$\Pr[\theta^{(1),j\top} \tilde{x} \geq 0, \theta^{(1),j\top} \tilde{x}' \geq 0] = \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{\tilde{x}^\top \tilde{x}'}{\|\tilde{x}\| \|\tilde{x}'\|}\right),$$

where we apply the formula for bivariate Gaussian variable  $(N_1, N_2)$ :  $\Pr[N_1 \geq 0, N_2 \geq 0] = \frac{1}{4} + \arcsin(\rho)$  where  $\rho := \frac{\text{Cov}(N_1, N_2)}{\sqrt{\text{Var}(N_1)\text{Var}(N_2)}}$ .

As a result,

$$K_Z^{(1)}(\tilde{x}, \tilde{x}') := \frac{1}{4} \tilde{x}^\top \tilde{x}' \left[ 1 + \frac{2}{\pi} \arcsin\left(\frac{\tilde{x}^\top \tilde{x}'}{\|\tilde{x}\| \|\tilde{x}'\|}\right) \right] I_{d_Z}.$$

Notice that  $K_Z^{(1)}$  is a positive definite operator since for any finite set of distinct samples  $\tilde{X} := [\tilde{x}_1, \dots, \tilde{x}_n]$  with  $t_i \neq t_j$  for all  $i, j$ , the matrix

$$K := \{K_Z(\tilde{x}_i, \tilde{x}_j)\}_{ij} \succ \frac{1}{4} \tilde{X}^\top \tilde{X} \otimes I_{d_Z} \succ 0,$$

where we use  $\|\tilde{x}\| > 0$  and  $\text{PE}(t_i)$ 's are linearly independent from the condition of the lemma, and thus their Gram matrix  $\tilde{X}^\top \tilde{X} \succ 0$ . Consequently,  $K_Z = K_Z^{(1)} + K_Z^{(2)} \succ 0$  and therefore its minimal eigenvalue is positive.

## D.9 Proof of Lemma D.4

As  $d_H, d_T \rightarrow \infty$ , we have  $K_Z = K_Z^*, K_G = K_G^*$ . First, we prove the following lemmas in Section D.10 and D.11 respectively.

**Lemma D.9.** Suppose  $n > \max\{d_Z, d_G\}$ , and  $\min\{d_T, d_H\} \rightarrow \infty$ , the following holds for the empirical risk minimizer  $(\hat{s}_Z, \hat{s}_G, \hat{U}, \hat{V})$  with probability at least  $1 - 2\exp(-n)$ :

$$\begin{aligned} \hat{L}_Z(\hat{U}, \hat{s}_U) &= \hat{L}_G(\hat{V}, \hat{s}_V) = 0, \\ \|\hat{s}_Z\|_{K_Z} &\leq C_Z \sqrt{\frac{\sigma_1(s_Z^*) d_Z^{1/2}}{\lambda_{\min}(K_Z^*)}}, \\ \|\hat{s}_G\|_{K_G} &\leq C_G \sqrt{\frac{\sigma_1(s_G^*) d_G^{1/2}}{\lambda_{\min}(K_G^*)}}, \end{aligned}$$

for some constant  $C_Z, C_G > 0$ .

**Lemma D.10.** For the NTK of the estimated content and style score functions  $K_Z$  and  $K_G$  defined in Lemma D.7 and for any  $(x, t) \in \mathcal{X} \times [t_0, T]$ , the following holds:

$$\begin{aligned} \|K(x, t, x, t)\| &\leq \frac{3}{2} (\|x\|^2 + \|\text{PE}(t)\|^2) \\ \|K_G(x, t, x, t)\| &\leq \frac{3}{2} (\lambda_g^2 \|x\|^2 + \|\text{PE}(t)\|^2). \end{aligned}$$

Then we make use of the following properties of sub-gaussian random variables from Lemma 16 of [41].

**Lemma D.11.** Consider a probability density function  $p(x) \leq \exp(-C\|x\|_2^2/2)$  for  $x \in \mathbb{R}^d$  and constant  $C > 0$ . Let  $R$  be a fixed radius. Then the following holds

$$\int_{\|x\|>R} p(x)dx \leq \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^{d-2} \exp(-CR^2/2),$$

$$\int_{\|x\|>R} \|x\|^2 p(x)dx \leq \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^d \exp(-CR^2/2).$$

Define  $\rho_Z := C_Z \sqrt{\frac{\sigma_1(s_Z^*) d_Z^{1/2}}{\lambda_{\min}(K_Z^*)}}$ , and without loss of generality, assume  $\sup_{t \in [t_0, T]} \|\text{PE}(t)\| \leq T$ . To prove lemma D.4, we first bound the Rademacher average of  $\mathcal{S}_Z$  as

$$\begin{aligned} \mathcal{R}_n(\mathcal{S}_Z) &= \mathbb{E}_{\epsilon^n} \sup_{s_Z \in \mathcal{S}_Z(\tilde{x}^{1:n})} \frac{1}{n} \sum_{i=1}^n \langle \epsilon_i, s_Z(\tilde{x}^i) \rangle \\ &= \frac{1}{n} \mathbb{E}_{\epsilon^n} \sup_{c \in \mathcal{C}_Z(\tilde{x}^{1:n})} \sum_{i=1}^n \sum_{j=1}^{d_Z} \epsilon_{ij} \langle s_{Z,j}, K_{Z,\tilde{x}^i,j} \rangle_{K_Z} = \frac{1}{n} \mathbb{E}_{\epsilon^n} \sup_{c \in \mathcal{C}_Z(\tilde{x}^{1:n})} \sum_{j=1}^{d_Z} \left\langle s_{Z,j}, \sum_{i=1}^n \epsilon_{ij} K_{Z,\tilde{x}^i,j} \right\rangle_{K_Z} \\ &\leq \frac{\rho_Z}{n} \mathbb{E}_{\epsilon^n} \sqrt{\sum_{j=1}^{d_Z} \left\| \sum_{i=1}^n \epsilon_{ij} K_{Z,\tilde{x}^i,j} \right\|_{K_Z}^2} \leq \frac{\rho_Z}{n} \sqrt{\mathbb{E}_{\epsilon^n} \sum_{j=1}^{d_Z} \left\| \sum_{i=1}^n \epsilon_{ij} K_{Z,\tilde{x}^i,j} \right\|_{K_Z}^2} \\ &= \frac{\rho_Z}{n} \sqrt{\sum_{i=1}^n p(t^i) p_{t^i}(x^i) \mathbb{E}_{\epsilon_i} \epsilon_i^\top K_Z(\tilde{x}^i, \tilde{x}^i) \epsilon_i} \leq \frac{\rho_Z}{n} \sqrt{\sum_{i=1}^n p(t^i) p_{t^i}(x^i) \text{Tr}(K_Z(\tilde{x}^i, \tilde{x}^i))}. \end{aligned}$$

Averaging over  $\tilde{x}^n$  and applying Cauchy-Schwarz inequality, we can bound the Rademacher complexity of the data-dependent function class  $\mathcal{S}_Z$  as

$$\begin{aligned} \mathbb{E}_{\tilde{x}^n} \mathcal{R}_n(\mathcal{S}_Z) &= \mathbb{E}_{\tilde{x}^n} \frac{\rho_Z}{n} \sqrt{\sum_{i=1}^n p(t^i) p_{t^i}(x^i) \text{Tr}(K_Z(\tilde{x}^i, \tilde{x}^i))} \\ &\leq \frac{\rho_Z}{\sqrt{n}} \sqrt{\mathbb{E}_{t,t,p_t(x)p_t(x)} \text{Tr}(K_Z(\tilde{x}, \tilde{x}))} =: \frac{\rho_Z C_{K_Z}}{\sqrt{n}}, \end{aligned}$$

where  $C_{K_Z} := \sqrt{\mathbb{E}_{t,t,p_t(x)p_t(x)} \text{Tr}(K_Z(\tilde{x}, \tilde{x}))} = O(d_X)$ .

To proceed, let the content subspace matrix and score function class learned by solving Eq. 12 given training data  $[\tilde{x}^1, \dots, \tilde{x}^n]$  be  $\mathcal{U}(\tilde{x}^{1:n})$  and  $\mathcal{S}_Z(\tilde{x}^{1:n})$  respectively, then by Lemma D.6, with probability at least  $1 - 2\exp(-\Omega(n))$ , the event

$$\mathcal{E} := \mathcal{S}_Z(\tilde{x}^{1:n}) \subseteq \mathcal{S}_Z \quad (46)$$

will occur, which implies with the same probability bound, the empirical risk minimizer over  $\mathcal{S}_Z(\tilde{x}^{1:n})$  satisfies

$$\hat{L}_Z(\hat{U}, \hat{s}_Z) = \min_{(U, s_Z) \in \mathcal{U}(\tilde{x}^{1:n}) \times \mathcal{S}_Z(\tilde{x}^{1:n})} \frac{1}{n} \sum_{i=1}^n \ell_Z(\tilde{x}^i; U, s_Z) = \min_{U, s_Z \in \mathcal{S}_Z} \frac{1}{n} \sum_{i=1}^n \ell_Z(\tilde{x}^i; U, s_Z) = 0.$$

Therefore, let an empirical risk minimizer of  $L_Z$  over  $\mathcal{S}_Z$  be  $\hat{U}', \hat{s}'_Z$ , then for any  $\epsilon > 0$ , we can bound the generalization error probability as

$$\begin{aligned} &\Pr \left[ L_Z(\hat{U}, \hat{s}_Z) \geq \min_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} L_Z(U, s_Z) + \epsilon \right] \\ &\leq \Pr(\mathcal{E}) + \Pr \left[ L_Z(\hat{U}', \hat{s}'_Z) \geq \min_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} L_Z(U, s_Z) + \epsilon \right] \\ &\leq \Pr \left[ L_Z(\hat{U}', \hat{s}'_Z) \geq \min_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} L_Z(U, s_Z) + \epsilon \right] + 2\exp(-\Omega(n)). \end{aligned} \quad (47)$$

Next, since the squared loss  $\ell_Z$  is not Lipschitz with respect to  $x$ , we apply a truncation argument on  $L_Z$ . To this end, define the truncated version of  $L_Z$  as

$$L_Z^{\text{trunc}}(U, s_Z) := \mathbb{E}_{t, p_t} \ell(x, t; U, s_Z) \mathbb{1}_{\|x\| \leq R}$$

for some radius  $R > 0$ . Similarly we can define its empirical version as  $L_Z^{\text{trunc}}$ . Then  $L_Z$  admits the following decomposition:

$$\begin{aligned} & L_Z(\hat{U}', \hat{s}'_Z) - L_Z(A_Z, s_Z^*) \\ &= L_Z^{\text{trunc}}(\hat{U}', \hat{s}'_Z) - L_Z^{\text{trunc}}(A_Z, s_Z^*) + L_Z(\hat{U}, \hat{s}_Z) - L_Z^{\text{trunc}}(\hat{U}', \hat{s}'_Z) + L_Z^{\text{trunc}}(A_Z, s_Z^*) - L_Z(A_Z, s_Z^*) \\ &\leq \underbrace{L_Z^{\text{trunc}}(\hat{U}', \hat{s}'_Z) - L_Z^{\text{trunc}}(A_Z, s_Z^*)}_{(i)} + \underbrace{L_Z(\hat{U}, \hat{s}_Z) - L_Z^{\text{trunc}}(\hat{U}', \hat{s}'_Z)}_{(ii)} \end{aligned}$$

To bound (i), notice that for  $s_Z \in \mathcal{S}_Z$ , with balanced weight  $U \in \mathcal{U}_Z$ ,

$$\begin{aligned} & |\ell(x', t'; U, s_Z) \mathbb{1}_{\|x'\| \leq R} - \ell(x, t; U, s_Z) \mathbb{1}_{\|x\| \leq R}| \\ &\leq \sup_{x \in \mathcal{X}, t \in [t_0, T]} \sup_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} |\ell(x, t; U, s_Z) \mathbb{1}_{\|x\| \leq R}| \\ &\leq \sup_{x \in \mathcal{X}, t \in [t_0, T]} \sup_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} 2(\|Us_Z(x, t)\|^2 + \|s_Z^*(x, t)\|^2) \mathbb{1}_{\|x\| \leq R} \\ &= \sup_{x \in \mathcal{X}, t \in [t_0, T]} \sup_{(U, s_Z) \in \mathcal{U} \times \mathcal{S}_Z} 2\|Us_Z(x, t)\|^2 \mathbb{1}_{\|x\| \leq R} + 2\lambda_s^2(R^2 + T^2) \\ &= (3\rho_Z^4 \lambda_{\min}(K_Z^*) + 2\lambda_s^2)(R^2 + T^2) =: B, \end{aligned}$$

where the second-to-last inequality uses the Lipschitz property of  $s_Z^*$ :

$$\|s_Z^*(x, t)\|^2 \mathbb{1}_{\|x\| \leq R} \leq \lambda_s^2 \|\tilde{x}\|^2 \mathbb{1}_{\|x\| \leq R} \leq \lambda_s^2(R^2 + T^2), \quad (48)$$

where the last inequality uses Lemma D.10 as follows:

$$\begin{aligned} \|Us_Z(x, t)\| &\leq \|U\| \|s_Z(x, t)\| \\ &\leq \rho_Z \sqrt{\lambda_{\min}(K_Z)} \sqrt{\sum_{j=1}^{d_Z} \langle s_{Z,j}, K_{Z,x,t,j} \rangle_{K_Z}^2} \\ &\leq \rho_Z \sqrt{\lambda_{\min}(K_Z)} \|s_Z\|_{K_Z} \|K_{Z,x,t}\|_{K_Z} \\ &\leq \rho_Z^2 \sqrt{\lambda_{\min}(K_Z)} \|K_Z(x, t, x, t)\| \leq \rho_Z^2 \sqrt{\frac{3}{2} \lambda_{\min}(K_Z)} \|\tilde{x}\|. \end{aligned} \quad (49)$$

Using Eq. 48-49, we can show that

$$\begin{aligned} \|Us_Z(\tilde{x}^i) - A_Z s_Z^*(\tilde{x}^i)\| \mathbb{1}_{\|x^i\| \leq R} &\leq \|Us_Z(\tilde{x}^i) \mathbb{1}_{\|x^i\| \leq R}\| + \|A_Z s_Z^*(\tilde{x}^i) \mathbb{1}_{\|x^i\| \leq R}\| \\ &\leq (\rho_Z^2 \sqrt{3\lambda_{\min}(K_Z^*)/2} + \lambda_s) \sqrt{R^2 + T^2}. \end{aligned}$$

Let  $\lambda_{\text{trunc}} := 2(\rho_Z^2 \sqrt{3\lambda_{\min}(K_Z^*)/2} + \lambda_s) \sqrt{R^2 + T^2}$  and use the inequality for  $\max\{|s|, |t|\} \leq \lambda$ ,

$$|s^2 - t^2| < 2\lambda|s - t|,$$

we deduce that

$$|\ell(\tilde{x}^i; U', s'_Z) - \ell(\tilde{x}^i; U, s_Z)| \mathbb{1}_{\|x^i\| \leq R} \leq \lambda_{\text{trunc}} \|\|U' s'_Z(\tilde{x}^i) - A_Z s_Z^*(\tilde{x}^i)\| - \|Us_Z(\tilde{x}^i) - A_Z s_Z^*(\tilde{x}^i)\|\|,$$

and thus  $\ell(\tilde{x}^i; U, s_Z)$  is a  $\lambda_{\text{trunc}}$ -Lipschitz function of  $\|Us_Z(\tilde{x}^i) - A_Z s_Z^*(\tilde{x}^i)\|$ . Denote

$$\begin{aligned} \mathcal{US}_Z &:= \{Us_Z : U \in \mathcal{U}, s_Z \in \mathcal{S}_Z\}, \\ \ell \circ \mathcal{US}_Z(\tilde{x}^n) &:= \{(\ell(\tilde{x}^1; U, s_Z), \dots, \ell(\tilde{x}^n; U, s_Z)) : Us_Z \in \mathcal{US}_Z\}, \end{aligned}$$

and apply the standard generalization bound for Lipschitz function on  $L_Z^{\text{trunc}}$  (see, e.g., Theorem 9.1 of [127]) to conclude that with probability at least  $1 - \delta$ ,

$$\begin{aligned}
(i) &\leq 2\mathbb{E}_{\tilde{x}^n} \mathcal{R}_n(\ell \circ \mathcal{U} \mathcal{S}_Z(\tilde{x}^n)) + \frac{B}{2} \sqrt{\frac{8\log(1/\delta)}{n}} \\
&\leq 2\lambda_{\text{trunc}} \left( \mathbb{E} \mathcal{R}_n(\mathcal{U} \mathcal{S}_Z) + \frac{\lambda_s \sqrt{R^2 + T^2}}{\sqrt{n}} \right) + \frac{B}{2} \sqrt{\frac{8\log(1/\delta)}{n}} \\
&\leq 2\lambda_{\text{trunc}} \frac{\rho_Z^2 C_{K_Z} + \lambda_s \sqrt{R^2 + T^2}}{\sqrt{n}} + \frac{B}{2} \sqrt{\frac{8\log(1/\delta)}{n}} \\
&= O \left( \frac{\lambda_s \rho_Z^4 (R^2 + T^2)}{\sqrt{n}} + \lambda_s \rho_Z^4 (R^2 + T^2) \sqrt{\frac{\log 1/\delta}{n}} \right) = O \left( \frac{\lambda_s d_Z T^2 R^2}{\sqrt{n}} + \lambda_s d_Z T^2 R^2 \sqrt{\frac{\log 1/\delta}{n}} \right),
\end{aligned}$$

where the second inequality uses triangle inequality within the definition of Rademacher complexity followed by the bound in Eq. 48, and the third inequality uses the contraction principle of Rademacher complexity [127].

To bound (ii), notice that

$$\begin{aligned}
(ii) &= \mathbb{E}_{t, p_t(x)} \ell_Z(x, t; \hat{U}', \hat{S}'_Z) \mathbb{1}_{\|x\| > R} \\
&\leq 2\mathbb{E}_{t, p_t(x)} (\|\hat{U}' \hat{S}'_Z(x, t)\|^2 + \|s_Z^*(x, t)\|^2) \mathbb{1}_{\|x\| > R} \\
&\leq 2\mathbb{E}_{t, p_t(x)} \left[ \|\hat{U}' \hat{S}'_Z(x, t)\|^2 + \lambda_s^2 (\lambda_z^2 \|x\|^2 + t^2) \right] \mathbb{1}_{\|x\| > R} \\
&\leq 2\mathbb{E}_{t, p_t(x)} \left[ \frac{3}{2} \rho_Z^4 \lambda_{\min}(K_Z) d_Z (\|x\|^2 + \|\text{PE}(t)\|^2) + \lambda_s^2 (\lambda_z^2 \|x\|^2 + t^2) \right] \mathbb{1}_{\|x\| > R} \\
&= (3\rho_Z^4 \lambda_{\min}(K_Z) d_Z + 2\lambda_s^2 \lambda_z^2) \mathbb{E}_{t, p_t(x)} \|x\|^2 \mathbb{1}_{\|x\| > R} + \mathbb{E}_{t, p_t(x)} \left( \lambda_s^2 t^2 + \frac{3}{2} \|\text{PE}(t)\|^2 \right) \mathbb{1}_{\|x\| > R} \\
&= (3\rho_Z^4 \lambda_{\min}(K_Z) d_Z + 2\lambda_s^2 \lambda_z^2) \frac{\sigma_X^2 d_X 2^{-d_X/2+1} R^{d_X}}{\Gamma(d_X/2+1)} \exp(-R^2/2\sigma_X^2) + \\
&\quad \left( \frac{3}{2} \rho_Z^4 \lambda_{\min}(K_Z) d_Z \mathbb{E}_t \|\text{PE}(t)\|^2 + \frac{\lambda_s^2 (T^3 - t_0^3)}{3(T - t_0)} \right) \frac{\sigma_X^2 d_X 2^{-d_X/2+1}}{\Gamma(d_X/2+1)} R^{d_X-2} \exp(-R^2/2\sigma_X^2) \\
&= O \left( \frac{\lambda_s^2 T^2 \sigma_X^2 d_X^3 2^{-d_X/2+1}}{\Gamma(d_X/2+1)} R^{d_X} \exp(-R^2/2\sigma_X^2) \right).
\end{aligned}$$

where the second-to-last inequality uses Eq. 48 and the last inequality applies Eq. 49.

Take  $R := O(\sqrt{d_X \log d_X + \log n})$  such that (ii)  $\leq \frac{\lambda_s^2 d_X T^2}{n}$ , and then combining (i) and (ii) yields with probability at least  $1 - \delta$ ,

$$L_Z(\hat{U}', \hat{S}'_Z) - L_Z(A_Z, s_Z^*) \leq O \left( \frac{1 + \sqrt{\log 1/\delta}}{\sqrt{n}} (d_X^{5/2} + \log n) \right). \quad (50)$$

Setting  $\delta := \frac{1}{n}$  and combining Eq. 47 and Eq. 50 yields with probability at least  $1 - O(\frac{1}{n})$ ,

$$L_Z(\hat{U}, \hat{S}_Z) - L_Z(A_Z, s_Z^*) \leq O \left( \frac{1 + \sqrt{\log n}}{\sqrt{n}} (d_X^{5/2} + \log n) \right) = O \left( \sqrt{\frac{d_X^5 \log^3 n}{n}} \right).$$

Combining this bound with Eq. 47 yields the desired bound. And similarly, we can prove the bound for  $\hat{s}_G$ .

## D.10 Proof of Lemma D.9

As  $d_H, d_T \rightarrow \infty$ , we have  $K_Z = K_Z^*, K_G = K_G^*$ . By the optimality of  $\hat{s}_Z$  and  $\hat{s}_G$  and Theorem D.3, we have

$$\begin{aligned}\hat{U} \hat{s}_Z(x^i, t^i) &= A_Z s_Z^*(x^i, t^i), \\ \hat{V} \hat{s}_G(g(x^i), t^i) &= A_G s_G^*(g(x^i), t^i), \\ \frac{1}{n} \sum_{i=1}^n \hat{s}_Z(x^i, t^i) \hat{s}_Z(x^i, t^i) &= \hat{U}^\top \hat{U}, \\ \frac{1}{n} \sum_{i=1}^n \hat{s}_G(g(x^i), t^i) \hat{s}_G(g(x^i), t^i) &= \hat{V}^\top \hat{V}, \forall i = 1, \dots, n.\end{aligned}\quad (51)$$

The first two equalities immediately yield

$$\hat{L}_Z(\hat{U}, \hat{s}_U) = \hat{L}_G(\hat{V}, \hat{s}_V) = 0.$$

To prove the last two inequalities in the lemma, let  $\mathbb{E}_{\hat{p}_n}[f(x, t)] := \frac{1}{n} \sum_{i=1}^n f(x^i, t^i)$ . Then Eq. 51 implies that

$$\begin{aligned}& \mathbb{E}_{\hat{p}_n} \|A_Z s_Z^*(x, t)\|^2 \\ &= \mathbb{E}_{\hat{p}_n} \text{Tr}(\hat{s}_Z^\top(x, t) \hat{U}^\top \hat{U} \hat{s}_Z(x, t)) = \text{Tr}(\mathbb{E}_{\hat{p}_n} \hat{s}_Z(x, t) \hat{s}_Z^\top(x, t) \hat{U}^\top \hat{U}) \\ &= \text{Tr}(\mathbb{E}_{\hat{p}_n} \hat{s}_Z(x, t) \hat{s}_Z^\top(x, t) \mathbb{E}_{\hat{p}_n} \hat{s}_Z(x, t) \hat{s}_Z^\top(x, t)) = \|\mathbb{E}_{\hat{p}_n} \hat{s}_Z(x, t) \hat{s}_Z^\top(x, t)\|^2 \geq (\mathbb{E}_{\hat{p}_n} \|\hat{s}_Z(x, t)\|^2)^2,\end{aligned}\quad (52)$$

Let  $\Sigma_n(s) := \mathbb{E}_{\hat{p}_n} s(x, t) s(x, t)^\top$  and  $\lambda_i(M)$  be the  $i$ -th eigenvalue of the matrix  $M$ , then we have

$$\begin{aligned}\mathbb{E}_{\hat{p}_n} \|\hat{s}_Z(x, t)\|^2 &\leq \sqrt{\mathbb{E}_{\hat{p}_n} \|A_Z s_Z^*(x, t)\|^2} = \sqrt{\mathbb{E}_{\hat{p}_n} \|s_Z^*(x, t)\|^2} \\ &= \sqrt{\sum_{i=1}^{d_Z} \lambda_i(\Sigma_n(s_Z^*))} \leq \sqrt{\|\Sigma_n(s_Z^*)\|_{\text{op}} d_Z}.\end{aligned}\quad (53)$$

Further, suppose  $\hat{s}_Z(\tilde{x}) = \frac{1}{n} \sum_{i=1}^n K_Z(\tilde{x}, \tilde{x}^i) c_Z(\tilde{x}^i)$  and define

$$\bar{K}_Z := \begin{bmatrix} K_Z(\tilde{x}^1, \tilde{x}^1) & \cdots & K_Z(\tilde{x}^1, \tilde{x}^n) \\ \vdots & \ddots & \vdots \\ K_Z(\tilde{x}^n, \tilde{x}^1) & \cdots & K_Z(\tilde{x}^n, \tilde{x}^n) \end{bmatrix}, \bar{c}_Z := [c_Z(\tilde{x}^1)^\top, \dots, c_Z(\tilde{x}^n)^\top]^\top,$$

then we have

$$\begin{aligned}\|\hat{s}_Z\|_{K_Z}^2 &= \frac{1}{n} \bar{c}_Z^\top \bar{K}_Z \bar{c}_Z \leq \frac{\bar{c}_Z^\top \bar{K}_Z^2 \bar{c}_Z}{n \lambda_n(\bar{K}_Z)} = \frac{1}{n \lambda_n(\bar{K}_Z)} \sum_{j=1}^n \left\| \sum_{i=1}^n \bar{K}_Z(\tilde{x}^i, \tilde{x}^j) c_Z(\tilde{x}^j) \right\|^2 \\ &= \frac{1}{\lambda_n(\bar{K}_Z)} \mathbb{E}_{\hat{p}_n} \|\hat{s}_Z(\tilde{x})\|^2 \leq \frac{1}{\lambda_{\min}(K_Z^*)} \mathbb{E}_{\hat{p}_n} \|\hat{s}_Z(\tilde{x})\|^2 \leq \frac{\sqrt{\|\Sigma_n(s_Z^*)\|_{\text{op}} d_Z}}{\lambda_{\min}(K_Z^*)}.\end{aligned}$$

It remains to prove the concentration of the operator norm  $\|\Sigma_n(s_Z^*)\|_{\text{op}}$  around  $\|\Sigma_\infty(s_Z^*)\|_{\text{op}} =: \|\Sigma(s_Z^*)\|_{\text{op}}$ . To this end, we use the assumptions that  $s_Z^*(x, t)$  is  $\lambda_s$ -Lipschitz and  $(z, t)$  is sub-gaussian with sub-gaussian norm at most  $\sigma_Z^2 d_Z + T^2$  to conclude that  $s_Z^*(x, t)$  is sub-gaussian with sub-gaussian norm at most  $\lambda_s(\sigma_Z^2 d_Z + T^2)$ . Then Theorem 4.7.1 of [128] yields with probability at least  $1 - 2\exp(-u)$ ,

$$\|\Sigma_n(s_Z^*) - \Sigma(s_Z^*)\|_{\text{op}} \leq C \left( \sqrt{\frac{r+u}{n}} + \frac{r+u}{n} \right) \|\Sigma(s_Z^*)\|_{\text{op}}, \quad (54)$$

where  $r := \frac{\text{Tr}(\Sigma(s_Z^*))}{\|\Sigma(s_Z^*)\|_{\text{op}}}$  is the stable rank of  $\Sigma(s_Z^*)$ . As a result, with probability at least  $1 - 2\exp(-n)$ ,

$$\|\hat{s}_Z\|_{K_Z} \leq \frac{C_Z \|\Sigma(s_Z^*)\|_{\text{op}}^{1/4} d_Z^{1/4}}{\lambda_{\min}(K_Z^*)^{1/2}} = C_Z \sqrt{\frac{\sigma_1(s_Z^*) d_Z^{1/2}}{\lambda_{\min}(K_Z^*)}},$$

where  $C_Z := 1 + C^{1/2}(\sqrt{r+1} + r + 1)^{1/2}$  suffices.

Finally, using a similar argument we can show that

$$\|\hat{s}_G\|_{K_G} \leq C_G \sqrt{\frac{\sigma_1(s_G^*) d_G^{1/2}}{\lambda_{\min}(K_G^*)}}$$

with probability at least  $1 - 2\exp(-n)$ .

### D.11 Proof of Lemma D.10

As  $d_H, d_T \rightarrow \infty$ , we have  $K_Z = K_Z^*, K_G = K_G^*$ . By definition of the NTKs,

$$\begin{aligned} \|K_Z(x, t, x, t)\| &= \left\| \mathbb{E}_{\mathcal{N}(0, I)} \|a(x, t)\|^2 I_{d_Z} + \frac{1}{2} \|\tilde{x}\|^2 I_{d_Z} \right\| \\ &= \left\| \mathbb{E}_{\mathcal{N}(0, I)} \|a(x, t)\|^2 + \frac{1}{2} \|\tilde{x}\|^2 \right\| d_Z \\ &= \left\| \lim_{d_H \rightarrow \infty} \frac{1}{d_H} \sum_{j=1}^{d_H} \mathbb{E}_{\mathcal{N}(0, I)} (\tilde{x}^\top \Theta_j^{(1)})_+ (\Theta_j^{(1)\top} \tilde{x})_+ + \frac{1}{2} \|\tilde{x}\|^2 \right\| d_Z \\ &\leq \left\| \lim_{d_H \rightarrow \infty} \frac{1}{d_H} \sum_{j=1}^{d_H} \mathbb{E}_{\mathcal{N}(0, I)} (\tilde{x}^\top \Theta_j^{(1)} \Theta_j^{(1)\top} \tilde{x})_+ + \frac{1}{2} \|\tilde{x}\|^2 \right\| d_Z = \left\| \|\tilde{x}\|^2 + \frac{1}{2} \|\tilde{x}\|^2 \right\| = \frac{3}{2} \|\tilde{x}\|^2. \end{aligned}$$

Similarly, we can prove the bound for  $K_G$ .

Table 2: **Quantitative results for image disentanglement on MNIST and CIFAR10 test sets.** MSE, PSNR and SSIM stands for mean squared error, peak signal-to-noise ratio and structural similarity respectively between the generated and target samples. LPIPS [75] is a perceptual metric based on features from deep image classifiers. The results are averaged over two random trials.

|                    | MNIST                  |                       |                        |                       | CIFAR10                |                      |                        |                        |
|--------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|----------------------|------------------------|------------------------|
|                    | MSE( $\downarrow$ )    | PSNR( $\uparrow$ )    | SSIM( $\uparrow$ )     | LPIPS( $\downarrow$ ) | MSE( $\downarrow$ )    | PSNR( $\uparrow$ )   | SSIM( $\uparrow$ )     | LPIPS( $\downarrow$ )  |
| $\lambda_r = 0$    | 0.19 $\pm$ 0.01        | 5.2 $\pm$ 0.2         | 0.42 $\pm$ 0.04        | 0.30 $\pm$ 0.02       | 0.44 $\pm$ 0.03        | 3.5 $\pm$ 0.2        | 0.05 $\pm$ 0.00        | 0.62 $\pm$ 0.01        |
| $\lambda_r = 0.03$ | 0.07 $\pm$ 0.04        | 9.6 $\pm$ 2.3         | 0.53 $\pm$ 0.06        | 0.18 $\pm$ 0.04       | 0.35 $\pm$ 0.02        | 4.4 $\pm$ 0.1        | 0.05 $\pm$ 0.00        | 0.61 $\pm$ 0.02        |
| $\lambda_r = 0.3$  | 0.01 $\pm$ 0.00        | 16.0 $\pm$ 0.3        | 0.66 $\pm$ 0.00        | 0.1 $\pm$ 0.00        | 0.18 $\pm$ 0.06        | 7.3 $\pm$ 0.4        | 0.06 $\pm$ 0.00        | 0.53 $\pm$ 0.02        |
| $\lambda_r = 3$    | <b>0.01</b> $\pm$ 0.00 | <b>17.3</b> $\pm$ 0.3 | <b>0.66</b> $\pm$ 0.01 | <b>0.1</b> $\pm$ 0.00 | <b>0.11</b> $\pm$ 0.01 | <b>9.3</b> $\pm$ 0.2 | <b>0.07</b> $\pm$ 0.00 | <b>0.49</b> $\pm$ 0.00 |

## E Experiment details

### E.1 Latent subspace GMM disentanglement

For the synthetic disentanglement experiment, we choose the subspace dimension to be  $d_Z = d_G = 5$  and sample the content variable via  $Z \sim \frac{1}{2} \mathcal{N}(\mu_1^Z, \sigma_0^2 \mathbf{I}_{d_Z}) + \frac{1}{2} \mathcal{N}(\mu_2^Z, \sigma_0^2 \mathbf{I}_{d_Z})$  and the style variable via  $G \sim \frac{1}{2} \mathcal{N}(\mu_1^G, \sigma_0^2 \mathbf{I}_{d_G}) + \frac{1}{2} \mathcal{N}(\mu_2^G, \sigma_0^2 \mathbf{I}_{d_G})$ , where  $\sigma_0 = 0.1$ . In this way, we generate i.i.d 4000 samples for training.

We follow the network architecture shown in Figure 2 with  $d_H = 512$ . For the time embedding  $\text{PE}(\cdot)$ , we opt for a *Gaussian Fourier projection layer* to encode temporal information between  $(0, 1]$  defined as:

$$\text{PE}(t) := \begin{bmatrix} \sin(2\pi\Omega t) \\ \cos(2\pi\Omega t) \end{bmatrix},$$

where  $\Omega \sim \mathcal{N}(\mathbf{0}_{512}, 9000 \mathbf{I}_{512})$ . We train the models for 10,000 steps with an Adam [131] optimizer with learning rate  $10^{-5}$  and batch size equal to the entire training set.

Table 10: **The default noise schedule hyperparameters** for the synthetic data experiments. Continuous time ( $t \in [10^{-5}, 1]$ ) is used in the expression.

| Name        | $\alpha(t)$                                          | $\sigma^2(t)$                                  |
|-------------|------------------------------------------------------|------------------------------------------------|
| VE          | 1                                                    | $\frac{25^{2t}-1}{2\log 25}$                   |
| VP          | $e^{-0.05t-4.975t^2}$                                | $1 - e^{-0.1t-9.95t^2}$                        |
| sub-VP      | $e^{-0.05t-4.975t^2}$                                | $(1 - e^{-0.1t-9.95t^2})^2$                    |
| VP (cosine) | $e^{-\frac{t}{2} - \frac{1}{\pi} \sin(\frac{t}{2})}$ | $1 - e^{-t - \frac{2}{\pi} \sin(\frac{t}{2})}$ |

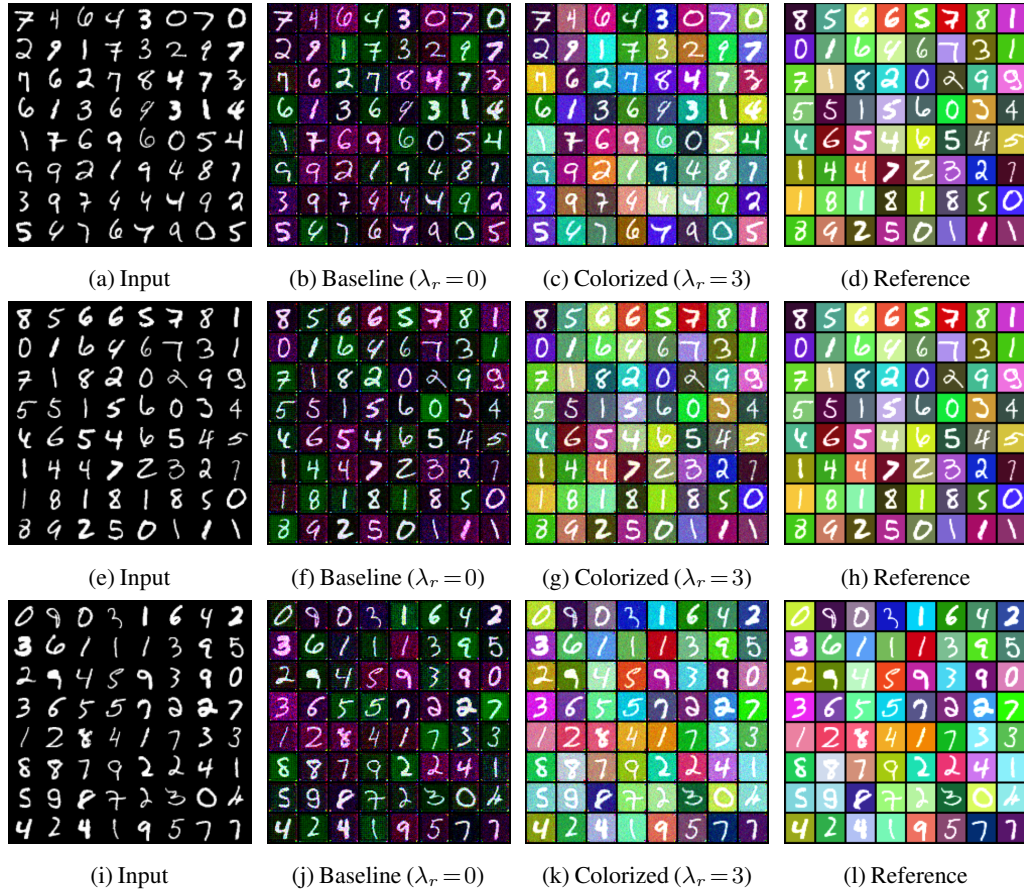


Figure 6: More image colorization results on MNIST

To ensure convergence, we pretrained the speaker score network  $\hat{s}^G$ . We experiment with various noise schedulers, including the variance exploding (VE), vanilla variance preserving (VP) [33], sub-VP [74] and cosine VP [132]. The detailed schedule hyperparameters are listed in Table 10 and are chosen based on rules of thumbs in [133, 74]. To evaluate the subspace recovery, we use the subspace recovery error defined as:

$$d(U, V, A_Z, A_G) := \frac{1}{2d_X} (\|P_U - P_{A_Z}\|_F^2 + \|P_V - P_{A_G}\|_F^2), \quad (55)$$

whose range is  $[0, 1]$ .

## E.2 Image disentanglement

For all experiments, we use a Gaussian Fourier projection layer to encode temporal information between  $(0, 1]$ . Further, the U-Net architecture used as the score function for MNIST and CIFAR10 are shown in Table 3 and Table 4 respectively. To capture the content information of the image, we use the  $16 \times 16$  feature map from a pretrained vit-small-patch16-224-dino variant of the DINO model [56]. The feature map is then resampled to the same size as the image.

For both datasets, we train the DM using an Adam optimizer [131] with a batch size of 128 and a learning rate  $10^{-4}$  for 50 epochs. A VE schedule is used during conditional score matching. During inference, we use probability flow [74] with 500 steps to perform sampling. For the CIFAR10 denoising experiment, we feed a all-zero matrix as the noise map. All models are implemented in Pytorch [134] on two A5000 GPUs. The training time is approximately an hour for both datasets and the inference is approximately 10 seconds for 64 samples.



Figure 7: More image denoising results on CIFAR10

Full quantitative results on MNIST and CIFAR10 are shown in Table 2. And more examples are shown in Figure 6 for MNIST and Figure 7 for CIFAR10.

Table 3: **U-Net score network used in the MNIST colorization experiment.** The input is  $9 \times 28 \times 28$  created by stacking the image, the projected and resized DINO feature map to  $3 \times 28 \times 28$  and the background color vector broadcasted to  $3 \times 28 \times 28$ .  $a + b$  denotes that component  $a$  accepts hidden embedding from the previous layer and component  $b$  accepts the time embedding, and the outputs of  $a$  and  $b$  are added with proper broadcasting. “Conv2d” refers to 2-D convolutional layer, “GroupNorm” stands for group normalization layer, and “ConvTrans2d” stands for 2-D transposed convolutional layer.

| MNIST U-Net                                                                             |
|-----------------------------------------------------------------------------------------|
| $384 \times 3 \times 1 \times 1$ Conv2d with stride 1 (DINO projection layer)           |
| $9 \times 32 \times 3 \times 3$ Conv2d with stride 2 + $1024 \times 32$ Linear          |
| GroupNorm with 4 groups<br>Swish activation                                             |
| $32 \times 64 \times 3 \times 3$ Conv2d with stride 2 + $1024 \times 64$ Linear         |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $64 \times 128 \times 3 \times 3$ Conv2d with stride 2 + $1024 \times 128$ Linear       |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $128 \times 256 \times 3 \times 3$ Conv2d with stride 2 + $1024 \times 256$ Linear      |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $256 \times 128 \times 4 \times 4$ ConvTrans2d with stride 2 + $1024 \times 128$ Linear |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $256 \times 64 \times 4 \times 4$ ConvTrans2d with stride 2 + $1024 \times 64$ Linear   |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $128 \times 32 \times 4 \times 4$ ConvTrans2d with stride 2 + $1024 \times 32$ Linear   |
| GroupNorm with 32 groups<br>Swish activation                                            |
| $64 \times 3 \times 3 \times 3$ ConvTrans2d with stride 1                               |

### E.3 Speech disentanglement

The dataset statistics are shown in Table 5. The overall results for realistic datasets are shown in Table 6

For the IEMOCAP dataset, we use a system available on SpeechBrain [135] that finetunes on the wav2vec 2.0 backbone [136] with a multi-layer perceptron classifier (MLP) [137]. The classifier is trained using Adam optimizer for 30 epochs with a batch size of 4 and a learning rate of  $10^{-4}$  for the MLP and the  $10^{-5}$  learning rate for wav2vec 2.0 weights. The system is then evaluated using the standard classification accuracy metric and 5-fold cross validation [76, 138]. For each fold, we use all 8 speakers from the training set as target speakers.

On the ALS and ADReSS, we use whisper-medium [129] features, as they have shown to be the most effective for speech impairment classification [139]. To avoid unfair comparison, We concatenate hidden representations over *all* layers of the whisper-medium encoder rather than selecting a particular layer and perform mean pooling over the frame-level features. For both datasets, we follow the standard splits used in previous works [79] to have no overlaps between speaker in the training and test sets. And for both datasets, we use the 15 most frequent speakers from the training set as target speakers for the VC to achieve maximize conversion quality via better speaker representation.

Table 4: **U-Net score network used in the CIFAR10 denoising experiment.** The input is  $12 \times 32 \times 32$  created by stacking the image, the projected and resized DINO feature map to  $3 \times 32 \times 32$  and the noise map broadcasted to  $3 \times 32 \times 32$ . A dual encoder with separate U-Nets for the content and the style variables is used and  $a+b+c$  denotes that component  $a$  accepts the content embedding from the previous layer, component  $b$  accepts the previous style embedding and  $c$  accepts the time embedding. Further, the outputs of  $a,c$  and the outputs of  $b,c$  are added separately with proper broadcasting.  $a+b$  here denotes that  $a$  accepts the previous content embedding,  $b$  accepts the previous style embedding and the two outputs are added.

| CIFAR10 U-Net                                                                       |  |
|-------------------------------------------------------------------------------------|--|
| 384 × 6 × 1 × 1 Conv2d with stride 1 (DINO projection layer)                        |  |
| 6 × 32 × 3 × 3 + 3 × 32 × 3 × 3 Conv2d with stride 2 + 1024 × 32 Linear             |  |
| GroupNorm with 4 groups<br>Swish activation                                         |  |
| 32 × 64 × 3 × 3 + 32 × 64 × 3 × 3 Conv2d with stride 2 + 1024 × 64 Linear           |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 64 × 128 × 3 × 3 + 64 × 128 × 3 × 3 Conv2d with stride 2 + 1024 × 128 Linear        |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 128 × 256 × 3 × 3 + 128 × 256 × 3 × 3 Conv2d with stride 2 + 1024 × 256 Linear      |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 256 × 128 × 4 × 4 + 256 × 128 × 4 × 4 ConvTrans2d with stride 2 + 1024 × 128 Linear |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 256 × 64 × 4 × 4 + 256 × 64 × 4 × 4 ConvTrans2d with stride 2 + 1024 × 64 Linear    |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 128 × 32 × 4 × 4 + 128 × 32 × 4 × 4 ConvTrans2d with stride 2 + 1024 × 32 Linear    |  |
| GroupNorm with 32 groups<br>Swish activation                                        |  |
| 64 × 3 × 3 × 3 + 64 × 3 × 3 × 3 ConvTrans2d with stride 1                           |  |

Table 5: **Datasets and VC-adapted classifiers used during realistic data experiments**

|         | $ \mathcal{V} $ | Feature          | Classifier | #Classifiers | Reference | VC        | DM-based | Reference |
|---------|-----------------|------------------|------------|--------------|-----------|-----------|----------|-----------|
| IEMOCAP | 4               | wav2vec 2.0 base | MLP        | 8            | [76]      | TriAAN-VC | No       | [77]      |
| ADReSS  | 2               | whisper-medium   | SVM        | 15           | [78]      | KNN-VC    | No       | [65]      |
| ALS-TDI | 5               | whisper-medium   | SVM        | 15           | [79]      | Diff-VC   | Yes      | [38]      |

We apply the VCs in mostly a zero-shot, plug-and-play fashion, and leave finetuning to specific datasets for future works. For the Diff-VC, we use the publicly available score network and vocoder checkpoints trained on LibriTTS and adopt the original inference hyperparameter settings for all experiments. Similarly, we use the pretrained models and for other VC models. Further, we use a maximum of 120 second speech from the target speaker to compute the target speaker embedding for all models except KNN-VC, where we use all the target speech as the pool for nearest neighbor search. We also compare VC adaptation with common data augmentation technique such as pitch shifting, where we shift the pitch of all the speech utterances to equally spaced pitch levels over the  $F0$  range of the training speech

Table 6: **Overall results on realistic datasets.** All metrics are between 0-100. A: single (average); B: single (best); MV: majority vote; SMV: soft majority vote.

| VC type        | Impairment   |             |             |             |             |             |             |             | Emotion                 |             |             |             |
|----------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------------------|-------------|-------------|-------------|
|                | ALS-TDI, F1↑ |             |             |             | ADReSS, F1↑ |             |             |             | IEMOCAP, Acc. (5-fold)↑ |             |             |             |
|                | A            | B           | MV          | SMV         | A           | B           | MV          | SMV         | A                       | B           | MV          | SMV         |
| No VC          | 54.9         | 54.9        | 54.9        | 54.9        | 70.6        | 70.6        | 70.6        | 70.6        | 71.5                    | 71.5        | 71.5        | 71.5        |
| Pitch shifting | 55.8         | 60.3        | 57.6        | <u>61.5</u> | 71.2        | 77.1        | 77.1        | 68.8        | 60.6                    | 55.1        | 61.1        | 61.1        |
| KNN-VC         | <u>55.8</u>  | <u>61.7</u> | <u>64.8</u> | <u>49.9</u> | 71.5        | <u>79.2</u> | <u>79.2</u> | <u>83.3</u> | 70.4                    | 69.3        | 71.4        | 71.5        |
| TriAAN-VC      | 55.7         | 60.7        | 61.7        | 53.3        | <u>72.4</u> | 75.0        | 77.1        | <u>83.3</u> | 65.1                    | 64.1        | 66.8        | 67.2        |
| Diff-VC        | 47.0         | 51.2        | 50.3        | 49.2        | 65.6        | 69.4        | 66.7        | 70.8        | <u>87.0</u>             | <u>94.3</u> | <u>96.5</u> | <u>97.2</u> |

Table 7: **Emotion recognition results on IEMOCAP.** 8 speakers in the training set of each fold are used as target speakers. WER stands for word error rate computed using the `whisper-large-v2` model [129] and lower is better. MOS stands for mean opinion score computed using UTMOS [130] and higher is better.

| VC type        | Voting type   | Accuracy |      |      |      |      |      | WER (↓) | MOS (↑) |
|----------------|---------------|----------|------|------|------|------|------|---------|---------|
|                |               | 1        | 2    | 3    | 4    | 5    | Avg. |         |         |
| No VC          | -             | 72.6     | 76.6 | 68.9 | 68.9 | 70.3 | 71.5 | 9.0     | 2.3     |
| Pitch shifting | single (best) | 64.0     | 65.3 | 57.4 | 57.7 | 58.6 | 60.6 | 36.4    | 1.3     |
|                | single (avg)  | 61.3     | 62.1 | 50.2 | 48.9 | 53.0 | 55.1 |         |         |
|                | majority      | 61.2     | 65.3 | 58.5 | 57.8 | 62.5 | 61.1 |         |         |
|                | soft majority | 60.8     | 65.4 | 57.8 | 57.5 | 61.5 | 61.1 |         |         |
| KNN-VC         | single (best) | 71.2     | 75.4 | 68.3 | 71.9 | 69.1 | 70.4 | 9.8     | 1.6     |
|                | single (avg)  | 69.6     | 72.6 | 67.0 | 69.9 | 67.4 | 69.3 |         |         |
|                | majority      | 70.3     | 75.6 | 68.5 | 72.8 | 70.0 | 71.4 |         |         |
|                | soft majority | 70.3     | 76.1 | 68.5 | 72.8 | 69.9 | 71.5 |         |         |
| TriAAN-VC      | single (best) | 65.5     | 66.9 | 63.0 | 67.5 | 62.6 | 65.1 | 16.6    | 1.9     |
|                | single (avg)  | 64.6     | 66.3 | 61.1 | 65.6 | 63.1 | 64.1 |         |         |
|                | majority      | 66.9     | 69.0 | 63.9 | 67.9 | 66.5 | 66.8 |         |         |
|                | soft majority | 66.6     | 69.9 | 63.6 | 68.8 | 67.0 | 67.2 |         |         |
| Diff-VC        | single (avg)  | 87.0     | 88.3 | 86.2 | 87.6 | 86.1 | 87.0 | 22.9    | 2.5     |
|                | single (best) | 94.4     | 94.8 | 94.2 | 95.2 | 92.9 | 94.3 |         |         |
|                | majority      | 97.5     | 96.7 | 95.2 | 98.1 | 94.9 | 96.5 |         |         |
|                | soft majority | 97.7     | 97.6 | 96.3 | 98.7 | 95.6 | 97.2 |         |         |

data with levels equal to the number of target speakers and train separate classifiers for each level as in the case of using VC adaptation.

For ALS severity classification as shown in Table 6, KNN-VC achieves the best performance among the VCs, reaching 65% macro-F1 with 15 target speakers and hard majority voting, compared to 54.9% when training without VC adaptation and 61.7% with pitch shifting. For cognitive impairment detection as shown in Table 6, TriAAN-VC performed the best followed by the KNN-VC method, both achieved 83.3% macro-F1 with soft majority voting, which is 12.7% better than methods without VC adaptation and 14.5% and 6.2% better than the pitch shifting adaptation using hard and soft majority voting respectively. On IEMOCAP, we found that Diff-VC performs the best, reaching an average of 97.2% accuracy, which is 25.7% better than the no-VC classifier and 36.1% than the pitch shifting adaptation. Though a phenomenon out of the scope of predictions by our theory, we hypothesized that such “specialization” of the VC methods is due to the different level of generalization ability of different VCs to latent variables *other than* the speaker identity, such as recording conditions and health conditions of the speaker. For instance, Diff-VC does not perform well on ALS compared to KNN-VC, probably due to the domain mismatch between the health conditions of its training set, which contains little pathological speech, compared to KNN-VC which uses the WavLM representation trained on much larger speech dataset with diverse speech.

Table 8: Alzheimer detection results on ADReSS

| VC type        | Voting type   | Precision | Recall | F1   | Accuracy |
|----------------|---------------|-----------|--------|------|----------|
| No VC          | -             | 71.4      | 70.8   | 70.6 | 70.8     |
| Pitch shifting | single (avg)  | 71.8      | 71.4   | 71.4 | 71.2     |
|                | single (best) | 77.1      | 77.1   | 77.1 | 77.1     |
|                | majority      | 77.1      | 77.1   | 77.1 | 77.1     |
|                | soft majority | 68.8      | 68.8   | 68.7 | 68.8     |
| KNN-VC         | single (avg)  | 71.8      | 71.5   | 71.4 | 71.5     |
|                | single (best) | 80.0      | 79.2   | 79.1 | 79.2     |
|                | majority      | 79.4      | 79.2   | 79.1 | 79.2     |
|                | soft majority | 83.6      | 83.3   | 83.3 | 83.3     |
| TriAAN-VC      | single (avg)  | 72.5      | 72.4   | 72.4 | 72.4     |
|                | single (best) | 75.2      | 75.0   | 75.0 | 75.0     |
|                | majority      | 77.5      | 77.1   | 77.0 | 77.1     |
|                | soft majority | 83.3      | 83.3   | 83.3 | 83.3     |
| Diff-VC        | single (avg)  | 65.7      | 65.4   | 65.4 | 65.6     |
|                | single (best) | 69.4      | 69.4   | 69.4 | 69.4     |
|                | majority      | 66.7      | 66.7   | 66.7 | 66.7     |
|                | soft majority | 72.2      | 70.8   | 70.4 | 70.8     |

Table 9: ALS severity classification results on ALS-TDI with a whisper-medium+SVM classifier

| VC type        | Voting type   | Precision↑ | Recall↑ | F1↑  |
|----------------|---------------|------------|---------|------|
| No VC          | -             | 59.8       | 53.7    | 54.9 |
| Pitch shifting | single (avg.) | 60.5       | 54.1    | 55.8 |
|                | single (best) | 67.4       | 57.7    | 60.3 |
|                | majority      | 73.0       | 54.9    | 57.6 |
|                | soft majority | 68.4       | 59.0    | 61.5 |
| KNN-VC         | single (avg.) | 58.5       | 54.6    | 55.8 |
|                | single (best) | 65.7       | 59.5    | 61.7 |
|                | majority      | 67.9       | 62.9    | 64.8 |
|                | soft majority | 51.1       | 49.6    | 49.9 |
| TriAAN-VC      | single (avg.) | 60.2       | 54.5    | 55.7 |
|                | single (best) | 68.0       | 58.2    | 60.7 |
|                | majority      | 69.9       | 59.1    | 61.7 |
|                | soft majority | 54.1       | 52.8    | 53.3 |
| Diff-VC        | single (avg.) | 48.2       | 47.0    | 47.0 |
|                | single (best) | 53.0       | 51.0    | 51.2 |
|                | majority      | 49.8       | 50.9    | 50.3 |
|                | soft majority | 50.1       | 48.8    | 49.2 |

As to the advantage of hard vs. soft voting, we observe different trends across different datasets and VC methods. On ALS-TDI, hard voting works better than soft voting by 8.4% and 16% for the best two methods Diff-VC and KNN-VC, though worse by 3.9% and 1.3% for pitch shifting and Diff-VC. On IEMOCAP, the gap between soft and hard voting is negligible, with soft majority voting shows a 0.1%-0.7% edge over hard majority voting across VC methods. On ADReSS, we found soft voting methods to be better than hard voting for all the VC methods by 4.1%–6.2%, while worse for the pitch shifting method by 8.3% (68.8% vs. 77.1%). Since soft voting uses a random classifier for voting, it tends to perform well when the model is “confidently” correct and “hesitantly” wrong, as it puts more weights on confident classifiers than hesitant ones. This suggests that the average confidence score estimated in terms of the classifier posteriors on incorrect examples will be high for classifier

ensembles that perform well with hard voting than soft voting. Table 7 8 9 show the complete results for the realistic dataset experiments.