Revealing Hidden Vulnerabilities in Autoencoders through Gradient Signal Restoration

Chethan Krishnamurthy Ramanaik
Arjun Roy
Tobias Callies
Eirini Ntoutsi
University of the Bundeswehr Munich
Germany
{chethan.krishnamurthy,arjun.roy,tobias.callies,eirini.ntoutsi}@unibw.de

Abstract

Adversarial robustness of deep autoencoders (AEs) has received less attention than that of discriminative models, although their compressed latent representations induce ill-conditioned mappings that can amplify small input perturbations and destabilize reconstructions. Existing white-box attacks for AEs, which optimize norm-bounded adversarial perturbations to maximize reconstruction damage, often converge to suboptimal perturbations, thereby potentially overstating AE robustness. We show that this limitation is linked to vanishing adversarial loss gradients during backpropagation through ill-conditioned layers, associated with near-zero singular values in their intermediate weight matrices. To address this, we propose GRILL (Gradient Signal Restoration in Ill-Conditioned Layers), a framework designed to mitigate gradient degradation and improve the reliability of adversarial robustness evaluation in encoder–decoder architectures. GRILL is designed to mitigate adversarial gradient degradation during optimization, enabling attacks to better approximate high-distortion perturbations under fixed norm constraints. Through extensive experiments across multiple AE architectures, under both sample-specific and universal attacks, as well as standard and adaptive attack settings, we show that GRILL significantly increases attack effectiveness, thereby exposing vulnerabilities hidden by existing attack limitations. Beyond AEs, we provide preliminary evidence that modern multimodal encoder–decoder architectures exhibit similar vulnerabilities.

Code: https://github.com/ChethanKodase/illcond

Keywords

adversarial robustness, autoencoders, encoder–decoder architectures, ill-conditioning, adversarial attacks

1 Introduction

Autoencoders (AEs) [17] are deployed in high-stakes applications, including image compression, reconstruction, denoising, anomaly detection, and generative modeling. By learning to approximately invert an encoder through a decoder, AEs inherently pose an inverse problem and exhibit structural non-invertibility due to dimensionality reduction. Despite this, adversarial vulnerabilities of AEs remain less studied than those of classification models [45]. Rigorous robustness evaluation requires strong and well-understood white-box attack methodologies, which are a standard tool for robustness evaluation [2], as they are essential for identifying model vulnerabilities and designing effective defense [45, 53]. AEs are a prominent class of encoder–decoder architectures, motivating investigation of similar optimization challenges in broader encoder–decoder systems.

Existing white-box attacks for AEs [8, 9, 16] optimize norm-bounded perturbations to maximize reconstruction damage. However, encoder–decoder architectures often exhibit ill-conditioning due to dimensionality reduction and approximate inversion during gradient propagation. As a result, adversarial optimization may yield suboptimal perturbations, potentially overstating robustness.

Ill-Conditioning as a Source of Adversarial Vulnerability. Ill-conditioning in neural networks can arise from highly imbalanced singular value spectra of the Jacobians of nonlinear network mappings, resulting in large condition numbers, as well as from structurally rank-deficient transformations induced by dimensionality reduction. Both effects are common in encoder–decoder architectures and can increase sensitivity to input perturbations. In practice, conditioning behavior is often approximated through the singular value spectra of intermediate layer weight matrices, which strongly influence gradient propagation [38, 43, 58]. Prior studies [44, 55] have shown that ill-conditioned networks can amplify small perturbations and increase adversarial vulnerability. Research on instabilities in deep inverse problems [3, 4] and the accuracy–stability trade-off [18] further demonstrates that deep learning pipelines, including AEs, can exhibit severe ill-conditioning and increased sensitivity to perturbations. At the same time, near-zero singular values in ill-conditioned networks can suppress adversarial gradient propagation, degrading attack optimization.

Why Small Singular Values Matter. Prior work on improving robustness has mainly focused on controlling large singular values through Lipschitz regularization and spectral normalization [6, 14]. In particular, several regularization methods [11, 19, 36, 58] explicitly constrain spectral norms, i.e., the largest singular values of weight matrices [54], to limit perturbation amplification and improve stability. In addition, implicit self-regularization effects arising from optimization dynamics [35] have been observed to promote similar spectral control during training. In contrast, considerably less attention has been paid to near-zero singular values, which can induce rank-deficiency and instability [29], suppress gradient propagation, and impair adversarial optimization, potentially preventing attacks from finding worst-case perturbations.

Ill-Conditioning Behind the Illusion of Adversarial Robustness. Prior work on dynamical isometry [38] suggests that near-zero singular values break dynamical isometry, leading to ill conditioned Jacobians causing poor gradient flow and slow or stalled

optimization. Similar effects were observed in quantized classification models during attack optimization [20], creating the illusion of adversarial robustness by simply impeding adversarial optimization rather than improving true robustness. However, the impact of ill-conditioning on gradient propagation during white-box adversarial optimization against encoder–decoder architectures remains largely unexplored. Our contributions are summarized as follows:
(i) We identify a previously unexplored failure mode in AE adversarial optimization, where near-zero singular values suppress adversarial gradients and yield suboptimal attacks.
(ii) We propose GRILL, a technique that is designed to mitigate gradient degradation during adversarial optimization in ill-conditioned encoder–decoder architectures.
(iii) We show that GRILL consistently improves attack effectiveness and exposes vulnerabilities hidden by optimization limitations.
(iv) We provide preliminary evidence that modern multimodal encoder–decoder architectures may exhibit similar vulnerabilities.

Paper organization: Section 2 reviews related work. Section 3 introduces preliminaries and attack formulations. We first introduce LGR in Section 4 for simplified decoder-side ill-conditioning settings, before presenting GRILL in Section 5 for general ill-conditioned encoder–decoder architectures. Section 6 reports experimental results and ablation studies, and Section 7 concludes the paper.

2 Related Work

Related work spans four main areas: AE architectures, white-box adversarial attacks on autoencoders, adversarial attacks on broader encoder–decoder architectures, and optimization challenges arising from ill-conditioned networks.

Architectures. Modern AEs extend the vanilla AE [25] through structural and regularization enhancements. β-VAE [24] and TC-VAE [10] encourage disentangled latent representations via KL regularization and total correlation penalties, respectively. NVAE [51] introduces hierarchical latent variables for multi-scale representations, while DiffAE [39] integrates diffusion-based decoding to improve fidelity and robustness. Masked autoencoders (MAE) [22] employ self-supervised reconstruction via random masking without explicit latent regularization. While discrete-latent models exist [52], we focus on state-of-the-art AEs with continuous latent spaces. To evaluate the broader applicability of GRILL beyond classical AEs, we additionally consider large-scale vision–language encoder–decoder models, including Gemma 3 [48] and Qwen 2.5 [26].

White-box Attacks on Autoencoders. White-box adversarial attacks optimize norm-bounded perturbations with full access to model gradients and may be targeted or untargeted, we focus on the latter. Unlike targeted attacks such as C&W [8], which induce misclassification, untargeted attacks on AEs maximize latent or output distortion under a fixed perturbation budget using metrics such as L2 distance [16] or Wasserstein distance [9]. Cosine similarity is commonly used in attacks on diffusion-based models [40, 59, 61]. Symmetric KL divergence [7] has also been explored for probabilistic latent outputs, but is inapplicable to non-probabilistic intermediate and final AE representations and is therefore excluded. While classical white-box attacks probe intrinsic model robustness, adaptive attacks [49] evaluate defenses (e.g., Hamiltonian Monte Carlo-based defense [30]) with full knowledge of the defense.

White-box Attacks on Encoder-Decoder Architectures. Adversarial attacks on multimodal encoder–decoder architectures mainly focus on disrupting intermediate representations or maximizing prediction uncertainty. Dispersion Reduction Attack (DRA) [34], Feature disruptive attack (FDA) [15], and Self Supervised Perturbation Attack (SSPA) [37] evaluate multimodal model robustness by perturbing inputs to distort latent feature spaces. Blockwise Similarity Attack (BSA) [57] further extends this idea to vision-language models through block-wise feature disruption across image and transformer encoders. More recently, Entropy Guided Attacks (EGA) [23] showed that high-entropy tokens can act as critical failure points in vision-language models, leading to adversarial outputs. We use these methods as baselines to explore the applicability of GRILL to broader encoder-decoder architectures.

Vanishing Gradients from Ill-Conditioned Weights. Dynamical isometry via orthogonal initialization improves optimization stability in deep networks, highlighting the adverse impact of vanishing gradients during adversarial optimization [38]. While gradient attenuation associated with near-zero singular values has been studied in quantized classification networks [20], the proposed remedies rely on softmax outputs and are not applicable to AEs. Gradient vanishing has also been identified as a form of *gradient obfuscation*, exploited by some defenses to hinder attack optimization [5]. Similarly, ill-conditioned encoder–decoder architectures may hinder adversarial optimization through suppressed gradient propagation. While prior work has explored hidden-layer perturbations [28, 50] and training-time spectral regularization [43, 58], it does not address optimization-induced gradient degradation that can mask vulnerabilities during adversarial optimization in encoder–decoder architectures, which is our focus.

3 Preliminaries

Notation. Let each input sample be represented as $x \in \mathbb{R}^d$. An autoencoder $\mathcal{Y} : \mathbb{R}^d \to \mathbb{R}^d$ consists of an encoder ϕ, which maps the input to a latent space $\mathbb{R}^n$, and a decoder ψ, which maps back to the original space $\mathbb{R}^d$ (typically, $n < d$):

$$\mathcal{Y}(x) \;=\; \psi \circ \phi(x), \qquad \phi : \mathbb{R}^d \to \mathbb{R}^n,\ \psi : \mathbb{R}^n \to \mathbb{R}^d$$

Generally, for a function $f : \mathbb{R}^m \to \mathbb{R}^l$ and a vector $z \in \mathbb{R}^m$, we denote by $J_z^f \in \mathbb{R}^{l \times m}$ the Jacobian matrix of f at z. We assume that all partial derivatives, and hence the Jacobian, exist throughout.

The L_p-ball around x of radius c is denoted by: $B_c^p(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\|_p \leq c\}$ where c is the attack budget. Finally, a distortion measure $\Delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is used (with d replaced by n for latent space distortions).

Multidimensional Chain Rule. The chain rule for higher dimensional maps states that the Jacobian matrix of a composed function $h = g \circ f$ is given by the matrix product of the Jacobians of g and f. In particular, for an AE, the Jacobians of the encoder, and the decoder at input x are denoted by J_x^ϕ and $J_{\phi(x)}^\psi$, respectively, yielding: $J_x^{\mathcal{Y}} = J_{\phi(x)}^\psi J_x^\phi$.

Multidimensional Taylor Approximation. Assuming suitable smoothness, function $f : \mathbb{R}^m \to \mathbb{R}^l$ can be locally approximated at z_a by

$$f(z_a) \simeq f(z) + J_z^f (z_a - z) + O(||z_a - z||^2)$$

where $O(||z_a - z||^2)$ is the remainder term, which decays quadratically. The approximation $f(z_a) \simeq f(z) + J_z^f(z_a - z)$ is commonly referred to as *first-order approximation.*

Adversarial Attack Objectives

Unlike targeted attacks such as C&W [8], which jointly optimize for misclassification and minimal perturbation magnitude, untargeted attacks on AEs are typically formulated as maximizing either output-space or latent-space distortion under a fixed perturbation budget [7, 16, 46].

Output–space Maximization (OA) *Output-space attacks* seek an adversarial input $x_a^* \in B_c^p(x)$ that maximizes reconstruction distortion between the original and adversarial inputs:

$$x_a^\star = \arg \max_{x_a \in B_c^p(x)} \Delta\big(\mathcal{Y}(x_a), \mathcal{Y}(x)\big). \tag{1}$$

Optimization: At iteration t, the adversarial input x_a^t is updated using gradient *ascent* with step size η, followed by projection onto the ball $B_c^p(x)$. This approach serves as a baseline method.

Latent–space Maximization (LA) *Latent-space attacks* [46] seek an adversarial input $x_a^* \in B_c^p(x)$ that maximizes distance between encoded latent representations:

$$x_a^\star = \arg \max_{x_a \in B_c^p(x)} \Delta\big(\phi(x_a), \phi(x)\big), \tag{2}$$

Optimization: Similar to OA attacks, but gradients are computed using latent-space distortion rather than output-space distortion. This serves as a second baseline method.

Existing works employ different distance metrics for untargeted attacks depending on the model architecture and training setting, including L_2 distance [16], Wasserstein distance [9], and cosine similarity in diffusion-based models [40, 59, 61].

Vanishing Gradients and Ill-conditioning

Condition number. For a matrix $W \in \mathbb{R}^{m \times n}$, let $\sigma_1(W)$ and $\sigma_r(W)$ denote its largest and smallest non-zero singular values. Let $r(W) \leq \min\{m, n\}$ denote its rank. The condition number of W is defined as: $\kappa(W) = \infty$ if $r(W) < n$, and $\kappa(W) = \sigma_1(W)/\sigma_r(W)$ otherwise.

A matrix is called *ill-conditioned* when $\kappa \gg 1$. Large condition numbers indicate high sensitivity to perturbations along certain directions. Neural network sensitivity has been linked to condition numbers of network parameters [44].

Ill-conditioning in AEs Since we assume a smaller latent dimension $n < d$, the encoder Jacobian is structurally rank-deficient and therefore *structurally ill-conditioned*, implying the existence of directions with reduced local sensitivity. Beyond the unavoidable ill-conditioning caused by dimensionality reduction, training can further amplify ill-conditioning in model layers (*learned ill-conditioning*). In particular, training may drive singular values toward highly uneven magnitudes, with some becoming very large and others approaching zero.

Most adversarial attack objectives for AEs tend to implicitly exploit directions associated with large singular values, since under first-order approximation these directions maximize local amplification of norm-bounded perturbations and are closely related to local Lipschitz sensitivity [1, 31]. In contrast, small non-zero singular values may attenuate adversarial gradients and hinder optimization, motivating GRILL.

4 Latent Gradient Restoration under Decoder Ill-Conditioning

There exists prior work on AE adversarial attacks optimizing only output-space divergence [16] (Eq. 1), as well as attacks optimizing only latent-space divergence [30]. Recent studies [6] show that certifiable robustness of AEs depends jointly on factors including the Lipschitz properties of the *encoder* and *decoder* [21], as well as the sensitivity of latent representations to input perturbations. This suggests that adversarial vulnerability in encoder–decoder architectures depends not only on reconstruction-space behavior, but also on the stability of latent representations and the propagation of gradients through both components. However, existing adversarial objectives do not account for optimization difficulties induced by ill-conditioned layers.

We first consider a simplified setting in which the decoder contains ill-conditioned intermediate layers while the encoder remains well-conditioned. In this setting, decoder-side ill-conditioning can attenuate adversarial gradients during backpropagation. To address this issue, we introduce *Latent Gradient Restoration* (LGR), which couples latent-space and reconstruction-space distortions such that latent-space gradients compensate for degraded reconstruction-space gradients, while reconstruction-space distortion continues to guide optimization by scaling the latent-space gradient.

We define the following maximum-damage criterion:

$$x_a^* = \arg \max_{x_a \in B_c^p(x)} \mathcal{L}(x_a). \tag{3}$$

where $\mathcal{L}(x_a)$ is defined as:

$$\mathcal{L}(x_a) = \Delta\big(\phi(x_a), \phi(x)\big) \cdot \Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big). \tag{4}$$

The product structure yields the following gradient:

$$\begin{aligned} \nabla_{x_a} \mathcal{L} = \Delta(\phi(x_a), \phi(x)) \cdot \nabla_{x_a} \Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big) \\ + \Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big) \cdot \nabla_{x_a} \Delta(\phi(x_a), \phi(x)). \end{aligned} \tag{5}$$

Gradient Degradation Mitigation Mechanism. When the decoder output gradient $\nabla_{x_a} \Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big) \to 0$ due to ill-conditioned decoder layers, its direct propagation becomes weak. However, since the encoder remains well-conditioned, the latent-space gradient $\nabla_{x_a} \Delta(\phi(x_a), \phi(x))$ remains non-negligible. Consequently, the term

$$\Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big) \cdot \nabla_{x_a} \Delta(\phi(x_a), \phi(x))$$

continues to provide a non-vanishing ascent direction, while the output-space distortion $\Delta\big(\psi \circ \phi(x_a), \psi \circ \phi(x)\big)$ still influences optimization by scaling non–vanishing latent–space gradient $\nabla_{x_a} \Delta(\phi(x_a), \phi(x))$, analogous to gradient modulation strategies studied in auxiliary-loss optimization [13]. Thus, gradient degradation caused by the ill-conditioned decoder is mitigated through latent-space gradients originating from the well-conditioned encoder, resulting in non-trivial adversarial updates.

Importantly, using a product rather than a sum introduces multiplicative coupling between latent-space and reconstruction-space distortions, allowing each distortion magnitude to scale the gradient contribution of the other. This explicit cross-weighting can improve optimization dynamics relative to naive loss summation

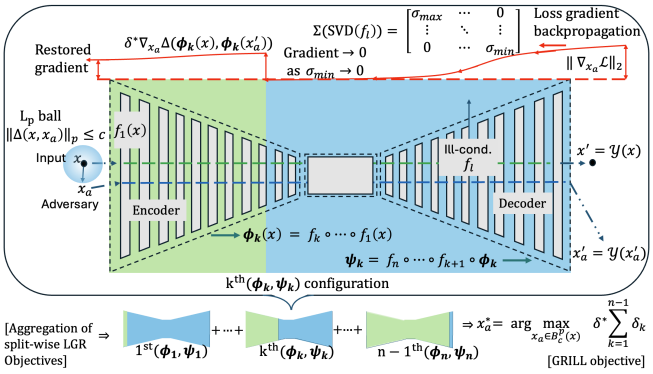


Figure 1: GRILL framework for mitigating gradient degradation in ill-conditioned encoder-decoder architectures.

(shown empirically in Section 6.6). Figure 1 illustrates the GRILL framework and the role of split-wise latent gradient restoration.

5 Gradient Signal Restoration in Ill-Conditioned Layers

LGR (Section 4) considers a simplified setting where gradient degradation originates from ill-conditioned decoder layers while the encoder remains well-conditioned. However, in practical encoder–decoder architectures, ill-conditioning can occur anywhere in the network, including intermediate encoder layers, limiting the applicability of LGR. To address this issue, we introduce GRILL (Gradient Signal Restoration in Ill-Conditioned Layers), which considers the AE $\mathcal{Y}$ as a composition of n layers $\mathcal{Y} = f_n \circ \cdots \circ f_1$ inducing $n-1$ possible *encoder-decoder* configurations (splits) indexed by $k \in \{1, \ldots, n-1\}$. Specifically, each "split" is given by:

- $\boldsymbol{\phi}_k = f_k \circ f_{k-1} \circ \cdots \circ f_1$, serving as an encoder up to layer k,
- $\boldsymbol{\psi}_k = f_n \circ f_{n-1} \circ \cdots \circ f_{k+1}$, serving as a decoder from layer $k+1$ to n.

The latent distortion after layer f_k can thus be expressed in terms of $\boldsymbol{\phi}_k$ as

$$\delta_k = \Delta(\boldsymbol{\phi}_k(x_a), \boldsymbol{\phi}_k(x)).$$

Since $\boldsymbol{\psi}_k$ takes the representations produced by $\boldsymbol{\phi}_k$ as input, the corresponding reconstruction-space distortion for the same split is

$$\delta_k^* = \Delta\big(\boldsymbol{\psi}_k \circ \boldsymbol{\phi}_k(x_a), \boldsymbol{\psi}_k \circ \boldsymbol{\phi}_k(x)\big).$$

We therefore consider the $n-1$ splits and define the GRILL loss as

$$\mathcal{L}_{\text{GRILL}}(x_a) = \sum_{k=1}^{n-1} \delta_k \delta_k^*. \tag{6}$$

Since the decoders of all splits reconstruct the same AE output, we consider $\delta_k^* = \delta^* = \Delta(\mathcal{Y}(x_a), \mathcal{Y}(x))$. The final optimization objective becomes

$$x_a{}^* = \arg \max_{x_a \in B_c^p(x)} \delta^* \sum_{k=1}^{n-1} \delta_k. \tag{7}$$

This objective is optimized using a gradient-based update scheme under an L_p-norm perturbation constraint.

5.1 How GRILL Differs from Layer-wise Loss Summation

A Layer-wise Loss Summation (LLS)-based attack typically aggregates intermediate distortions additively:

$$\mathcal{L}_{\text{LLS}}(x_a) = \sum_{k=1}^{n-1} \delta_k(x_a), \tag{8}$$

yielding

$$\nabla_{x_a} \mathcal{L}_{\text{LLS}} = \sum_{k=1}^{n-1} \nabla_{x_a} \delta_k. \tag{9}$$

Under severe ill-conditioning, gradients propagated through deeper transformations may become strongly attenuated, causing the corresponding terms $\nabla_{x_a} \delta_k$ to approach zero. Consequently, additive aggregation cannot preserve the influence of reconstruction-space distortion once gradient degradation occurs. In contrast, GRILL introduces multiplicative coupling:

$$\mathcal{L}_{\text{GRILL}}(x_a) = \delta^*(x_a) \sum_{k=1}^{n-1} \delta_k(x_a), \tag{10}$$

with gradient

$$\nabla_{x_a} \mathcal{L}_{\text{GRILL}} = \left(\sum_{k=1}^{n-1} \delta_k \right) \nabla_{x_a} \delta^* + \delta^* \sum_{k=1}^{n-1} \nabla_{x_a} \delta_k. \tag{11}$$

Unlike additive aggregation, GRILL scales intermediate-layer gradients by the reconstruction distortion δ^*. As a result, reconstruction-space distortion continues to influence attack optimization by scaling non-vanishing gradients from less ill-conditioned representations, helping mitigate optimization stagnation under degraded gradient propagation. Section 6.6 empirically confirms that removing the reconstruction distortion term δ^* from Eq.7 reduces the formulation to simple layer-loss summation (LLS), which generates weaker attacks than GRILL.

Algorithm 1 summarizes the optimization procedure of GRILL under an L_p-norm perturbation constraint. By aggregating LGR across all *encoder-decoder* splits (line 9), GRILL alleviates gradient attenuation caused by ill-conditioned encoder layers, since different splits can still contribute non-vanishing adversarial gradients as described in Section 4 and shown in Figure 1.

6 Experiments

Section 6.1 presents the experimental setup and evaluation protocol. Section 6.2 analyzes the conditioning behavior of the evaluated models. Sections 6.3 and 6.4 evaluate universal and sample-specific attacks on AEs, respectively, while Section 6.5 studies attacks on VLMs. Finally, Section 6.6 presents ablation studies, convergence and efficiency results.

6.1 Experimental Setup

We evaluate GRILL across five AEs and two vision-language models (Gemma 3 (4.3B parameters) [48] and Qwen 2.5 (8.3B parameters) [26]) under multiple adversarial attack settings, perturbation budgets, and distance metrics (Section 6.1). Table 1 summarizes the experimental setup. While introduced for AEs, GRILL targets gradient propagation through deep nonlinear encoder–decoder transformations, motivating exploratory evaluation on modern

Algorithm 1 Universal adversarial attack with L_∞ bound on perturbation using GRILL

Input: Test set $\mathcal{D} = \{x_i\}_{i=1}^N$, trained AE $\mathcal{Y} = \psi \circ \phi = f_n \circ \cdots \circ f_1$, for a configuration of $(\boldsymbol{\psi}_k, \boldsymbol{\phi}_k)$, $\boldsymbol{\phi}_k = f_k \circ f_{k-1} \circ \cdots \circ f_1$, serves as an encoder up to layer k, $\boldsymbol{\psi}_k = f_n \circ f_{n-1} \circ \cdots \circ f_{k+1}$, serves as a decoder from layer $k+1$ to n, step size η, number of steps till convergence T, batch size B
Parameter: Perturbation budget c is the L_∞ bound
Output: Universal adversarial perturbation ρ

Initialize perturbation ρ as a near-zero tensor with small random noise:
  $\rho \sim \mathcal{U}(-\xi, \xi)$, where $\xi \ll c$
for $t = 1$ to T **do**
  for each batch $\{x_j\}_{j=1}^B$ in $\mathcal{D}$ **do**
    Compute adversarial batch: $x_j^{\text{adv}} \leftarrow x_j + \rho$
    Compute GRILL loss:
      $\delta_k \leftarrow \sum_{j=1}^B \Delta(\boldsymbol{\phi}_k(x_j^{\text{adv}}), \boldsymbol{\phi}_k(x_j))$
      $\delta^* \leftarrow \sum_{j=1}^B \Delta(\mathcal{Y}(x_j^{\text{adv}}), \mathcal{Y}(x_j))$
      $\mathcal{L}_{\text{adv}} \leftarrow \delta^* \cdot \sum_{k=1}^{n-1} \delta_k$
    Compute gradient: $g \leftarrow \nabla_\rho \mathcal{L}_{\text{adv}}$
    Update perturbation using Adam optimizer:
      $\rho \leftarrow \text{AdamStep}(\rho, \nabla_\rho \mathcal{L}_{\text{adv}})$
    Project onto L_∞ ball:
      $\rho \leftarrow \text{clip}(\rho, -c, c)$
  end for
end for
return ρ

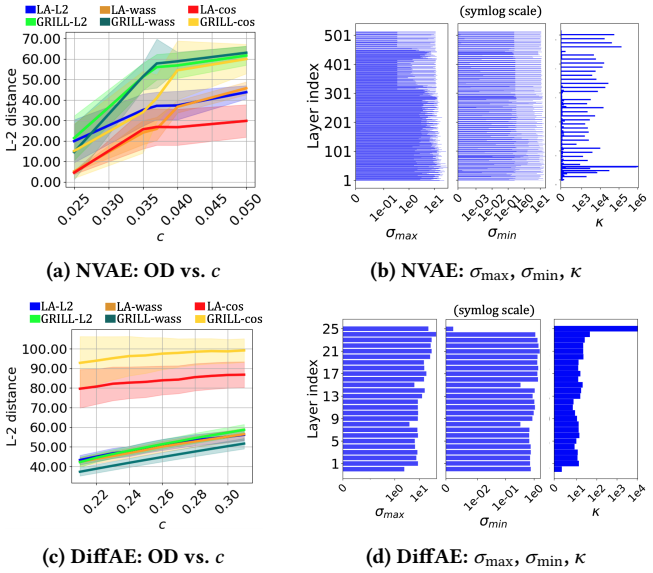
Model	Dataset	Training	L_∞ Radius c
β-VAE	CelebA[33]	From scratch	[0.04, 0.07]
TC-VAE	CelebA	From scratch	[0.04, 0.07]
NVAE[1]	CelebA	Pretrained	[0.025, 0.05]
DiffAE[2]	FFHQ[27]	Pretrained	[0.21, 0.30]
MAE[3]	ImageNet [12]	Pretrained	[0.05, 0.09]
Gemma 3[4]	ImageNet	Pretrained	[0.0005, 0.0009]
Qwen 2.5[5]	ImageNet	Pretrained	[0.001, 0.005]

Table 1: Models, datasets, training setup, and perturbations.

VLMs [32, 47]. The datasets span both face-centric data (CelebA, FFHQ) and large-scale natural image data (ImageNet), aligned with the respective training or pretraining setups. Perturbation magnitudes were selected via grid search to balance attack effectiveness and imperceptibility.

Attack Methods: We unify the existing AE attack formulations in [9, 16, 40, 59, 61] into the following AE attack baseline methods for comparison with GRILL:

- **OA** (Output-space attack): maximize output-space distortion(Eq. 1).
- **LA** (Latent-space attack): maximize latent-space distortion (Eq. 2); for NVAE, all hierarchical latent layers are attacked jointly [56]



(a) NVAE: OD vs. c (b) NVAE: $\sigma_{\max}$, $\sigma_{\min}$, κ

(c) DiffAE: OD vs. c (d) DiffAE: $\sigma_{\max}$, $\sigma_{\min}$, κ

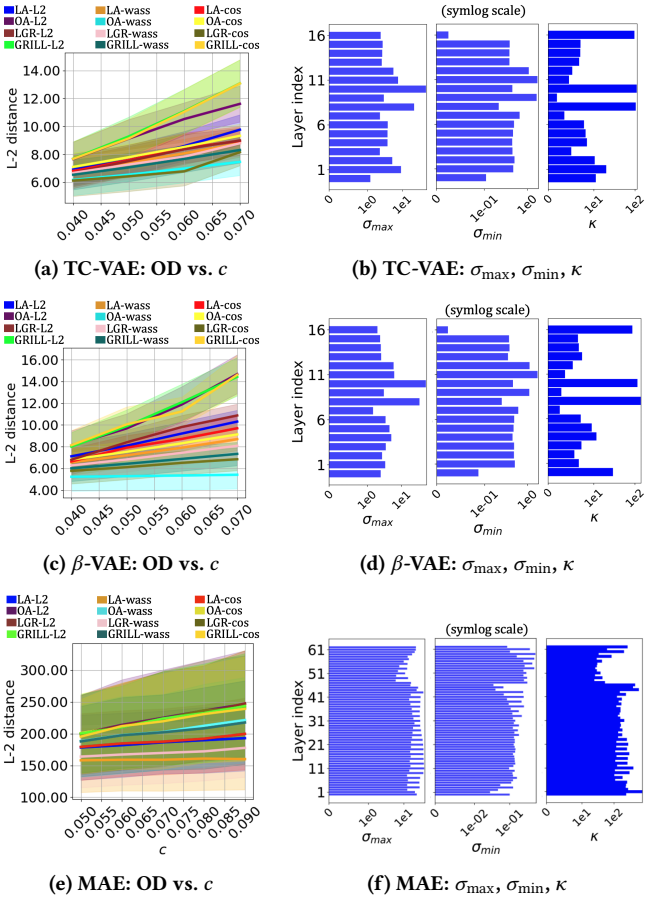
Figure 2: Universal attack performance in terms of output distortion (OD) (left) and corresponding layer-wise conditioning profiles (right) for highly ill-conditioned models.

-Latent Gradient Restoration (LGR): improves adversarial gradient propagation in decoder-only ill-conditioned settings (Eq. 3).
-Gradient Signal Restoration in Ill-Conditioned Layers (GRILL): improves adversarial gradient propagation across intermediate encoder–decoder splits.(Eq. 7).

Distance Functions. As discussed in Section 2, no single distance is universally effective, so we evaluate three common distance functions for adversarial loss: L_2 [16], Wasserstein (wass), and cosine similarity (cos) [40, 59, 61]. Attack configurations are defined by combining the attack objective (OA, LA, LGR, GRILL) and distance function (L2, wass, cos). For β-VAE, TC-VAE, and MAE, we evaluate all attack configurations. For NVAE and DiffAE, where decoding is computationally expensive, we limit evaluation to LA and GRILL, both operating solely on the encoder.

VLM Attack Strategies. For VLMs, we apply GRILL using intermediate vision and language representations within the model, and compare against representative multimodal attack baselines including **DRA** [34], **FDA** [15], **SSPA** [37], **BSA** [57], and **EGA** [23].

Evaluation measures. Following standard practice [41, 42, 56], AE attacks are evaluated using L_2 output distortion over 2,000 samples per attack configuration. For VLMs, we report qualitative results for sample-specific attacks and additionally perform quantitative evaluation on 300 samples using BERT Precision and BERT Recall [60], which measure semantic similarity between clean and adversarial outputs using contextual transformer embeddings. BERT Precision measures how well adversarial outputs preserve the semantic content of clean outputs, while BERT Recall measures how much semantic information from the clean outputs is retained after perturbation. We report both metrics under increasing perturbation budgets, where lower values indicate stronger adversarial degradation.



(a) TC-VAE: OD vs. c

(b) TC-VAE: $\sigma_{\max}$, $\sigma_{\min}$, κ

(c) β-VAE: OD vs. c

(d) β-VAE: $\sigma_{\max}$, $\sigma_{\min}$, κ

(e) MAE: OD vs. c

(f) MAE: $\sigma_{\max}$, $\sigma_{\min}$, κ

Figure 3: Universal attack performance in terms of output distortion (OD) (left) and corresponding layer-wise conditioning profiles (right) for mildly ill-conditioned models.

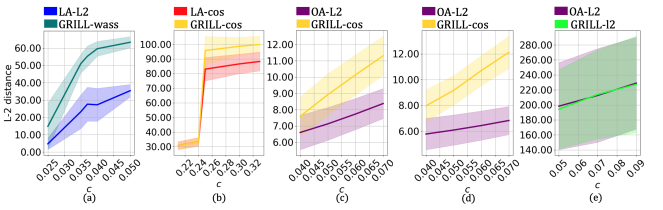
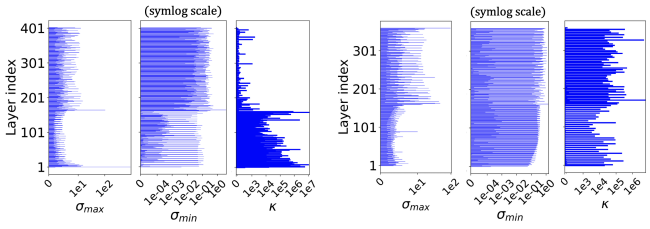


Figure 4: Universal adaptive attack performance in terms of output distortion (OD) under varying perturbation radii c for NVAE (a), DiffAE (b), TC-VAE (c), β-VAE (d), and MAE (e).

Optimization details. Universal attacks optimize a single perturbation per model and attack configuration under an L_∞-bounded perturbation radius c, whereas sample-specific attacks optimize individual perturbations for each input. All attack objectives are optimized using Adam, which has been shown to achieve fast convergence with performance comparable to L-BFGS, under an L_∞-projection constraint [8]. Based on hyperparameter tuning, we use learning rates of 10^{-4} for DiffAE, NVAE, TC-VAE, and β-VAE, 10^{-2} for MAE, and 10^{-3} for Gemma 3 and Qwen 2.5.



Gemma 3 **Qwen 2.5**

Figure 5: Layer-wise conditioning profiles ($\sigma_{\max}$, $\sigma_{\min}$ and κ)

6.2 Model conditioning behavior

To understand the effectiveness of GRILL, we first analyze the conditioning behavior of the evaluated models through the singular value spectra of intermediate transformations, using the maximum singular value $\sigma_{\max}$, minimum singular value $\sigma_{\min}$, and condition number κ. Following [58], we compute approximate singular values and condition numbers of individual intermediate layer weights for each model. We exclude intermediate operations for which the condition number κ is not well defined. Large κ values and near-zero $\sigma_{\min}$ indicate stronger ill-conditioning and an increased risk of gradient degradation.

β-VAE, TC-VAE, and MAE exhibit relatively mild ill-conditioning, as indicated by the absence of extremely small $\sigma_{\min}$ values in intermediate layers (Figures 3b, 3d, and 3f, respectively). NVAE (Figure 2b) and the VLMs Gemma 3 and Qwen 2.5 (Figure 5) exhibit widespread ill-conditioning characterized by near-zero singular values and large condition numbers. DiffAE exhibits localized severe ill-conditioning concentrated in the final encoder layer (Figure 2d).

Overall, the evaluated models span mild (β-VAE, TC-VAE, and MAE), localized (DiffAE), and severe ill-conditioning (NVAE, Gemma-3, Qwen 2.5) regimes, enabling systematic evaluation of GRILL under varying levels of gradient degradation.

6.3 Universal Attacks on AEs

We first evaluate GRILL in a *classical universal attack setup*, where AE models operate without defenses. We then consider a stronger *adaptive attack setup* in which a state-of-the-art defense mechanism is integrated into each AE model. In the classical setup, we compare GRILL against the baselines OA and LA under all attack configurations. In the adaptive setup, we evaluate the strongest baseline and GRILL configurations identified in the classical setup.

Classical Universal Attacks. For each AE model, we evaluate output distortion (OD) under increasing perturbation radii c and relate attack effectiveness to the conditioning behavior (Section 6.2).

β-VAE, TC-VAE, and MAE are *mildly ill-conditioned* (Figures 3b, 3d and 3f). Consistent with this, GRILL-based attacks perform comparably to the strongest baselines as seen in Figures 3c, 3e for β-VAE and MAE, respectively. Interestingly, despite the absence of very high κ in TC-VAE, GRILL-L_2 and GRILL-cos outperform the strongest baseline OA-L_2 attack by up to **12.66%**. This suggests that GRILL provides benefits beyond alleviating gradient degradation.

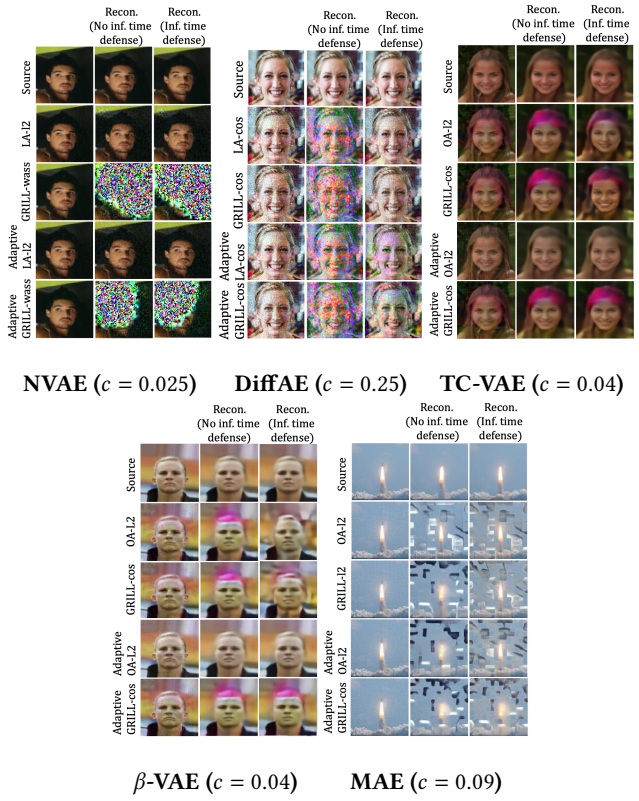


Figure 6: Universal attacks under HMC-based defense. Reconstructions are shown for baseline and GRILL attacks with and without defense.

For NVAE which exhibits *severe ill-conditioning* (Figure 2b), all GRILL-based attacks substantially outperform the baselines (Figure 2a), increasing output distortion from **38.11%** to **56.66%**.

DiffAE exhibits *localized severe ill-conditioning* in its final encoder layer (Figure 2d). Correspondingly, GRILL-cos consistently outperforms all baselines across the entire perturbation range (Figure 2c). Cosine similarity emerges as the most effective attack metric for DiffAE, with GRILL-cos exceeding the strongest baseline attack by **13.89%** to **16.31%** in output distortion.

Universal Adaptive Attacks. We next evaluate GRILL in a universal adaptive attack setting, where the adversary explicitly accounts for the deployed defense mechanism during the attack generation. A defended AE is modeled as $\mathcal{D}(x) = g(\mathcal{Y}(x))$, where $\mathcal{Y}$ is the AE and g is a latent-space defense mechanism. In our setup, g corresponds to a Hamiltonian Monte Carlo (HMC)-based latent refinement defense [30], which iteratively adjusts latent representations toward high-likelihood regions at inference time to recover stable reconstructions from adversarial perturbations while largely preserving clean inputs. Accordingly, the attack objective becomes

$$x_a^\star = \arg \max_{x_a \in B_c^p(x)} \Delta\big(\mathcal{D}(x_a), \mathcal{D}(x)\big) \tag{12}$$

We select the strongest baseline and best GRILL configuration from the classical setup and optimize attacks under active defense across varying L_∞ constraints. As shown in Figure 4, GRILL consistently achieves higher distortion scores than the strongest baselines, with gains up to **101.99%** for NVAE, **15.30%** for DiffAE, **34.96%** for TC-VAE, and **77.59%** for β-VAE (Figure 4). The gains of GRILL become more pronounced for TC-VAE and β-VAE under adaptive attacks, despite comparatively smaller gains in the classical setting, because the HMC defense suppresses baseline attacks more strongly by operating directly in latent space. Although adaptive attacks produce lower overall distortion, GRILL remains effective because, even when latent gradients vanish, the loss continues to influence adversarial updates by scaling non-vanishing gradients from earlier intermediate layers carrying adversarial signal. In contrast, no gains are observed for MAE because the HMC defense is less effective for its deterministic latent space.

Qualitative Analysis. Figure 6 shows that GRILL induces stronger output distortions than the strongest baseline attacks for NVAE and DiffAE, particularly under HMC-based defense. This effect is most pronounced for NVAE, consistent with its severe and widespread ill-conditioning. In contrast, TC-VAE, β-VAE, and MAE exhibit qualitatively similar behavior between baseline and GRILL-based attacks, consistent with their milder conditioning profiles.

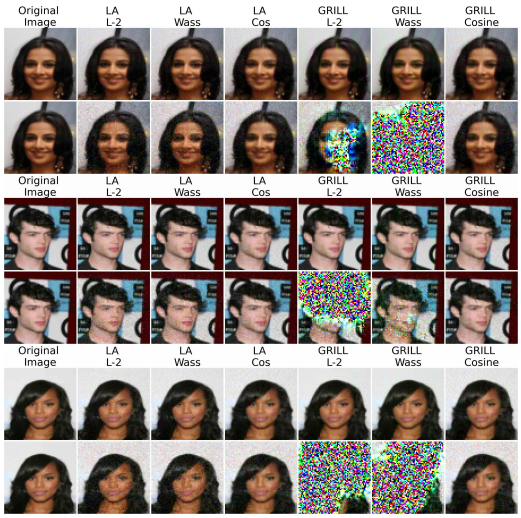
Inference-time defenses substantially suppress baseline attack distortions across all models, whereas GRILL-based attacks remain comparatively effective (Figure 6: NVAE, DiffAE, TC-VAE, β-VAE). Although GRILL does not outperform baseline attacks on MAE on average (Figure 3e), Figure 6(MAE) shows that it can still induce stronger distortions for certain samples, suggesting that GRILL remains complementary even in comparatively well-conditioned settings.

6.4 Sample-specific Attacks on AEs

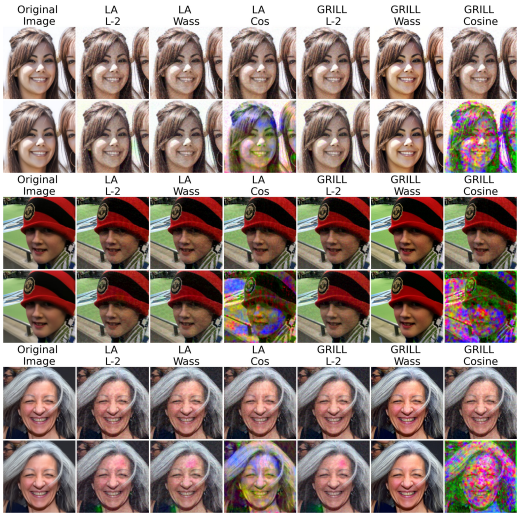
Figures 7(a) and 7(b) present qualitative sample-specific attacks on the highly ill-conditioned models NVAE and DiffAE. Under imperceptible perturbations, GRILL induces substantially stronger reconstruction distortions than the baseline LA attacks across multiple distance functions. Baseline attacks often fail to produce noticeable output degradation. These qualitative results support the conditioning analysis in Section 6.2 and further demonstrate the effectiveness of GRILL in ill-conditioned AEs.

6.5 Beyond AEs: Adversarial Attacks on Modern Encoder–Decoder Architectures

We evaluate GRILL on modern multimodal encoder–decoder architectures, Gemma 3 and Qwen 2.5. In this setting, the output logits of the VLM are treated as the attack output space, while intermediate hidden states serve as latent feature representations. Modern VLMs similarly employ layer-wise latent representations and encoder–decoder-style multimodal transformations [32, 47], motivating investigation of GRILL beyond classical AEs. Among the evaluated GRILL distance configurations, GRILL-cos consistently achieved the strongest attack performance; therefore, we report only GRILL-cos results for VLM experiments. We compare GRILL-cos against several representative white-box attack baselines, including DRA, FDA, SSPA, BSA, and EGA (see Section 2). Quantitative evaluation is reported for Gemma 3, while for the computationally more expensive Qwen 2.5 model, we report qualitative results



(a) NVAE ($c = 0.03$)



(b) DiffAE ($c = 0.08$)

Figure 7: Sample-specific attacks on highly ill-conditioned models. Clean/adversarial inputs are shown on the top, with corresponding reconstructions on the bottom.

only, as obtaining statistically significant quantitative results would require expensive evaluation.

Qualitative results: Figure 8 shows that under small perturbation budgets ($c \leq 0.002$ for Gemma 3 and $c \leq 0.0006$ for Qwen 2.5), baseline attacks induce only minor semantic variations in the generated response, indicating limited attack effectiveness. In contrast, GRILL-cos produces substantially stronger semantic degradation, including contextual distortion and loss of object fidelity.

This behavior is consistent with the conditioning analysis (Figure 5), where both models exhibit widespread intermediate-layer ill-conditioning characterized by near-zero singular values. The results suggest that GRILL remains effective on modern multimodal architectures even under highly constrained perturbation budgets.

Quantitative results: Figure 9 compares attack performance under increasing perturbation budgets using BERT Precision and Recall. Across all evaluation metrics, GRILL-cos consistently achieves lower scores than BSA, DRA, FDA, SSPA, and EGA, indicating stronger semantic degradation and more effective attacks.

Note: All qualitative results are shown for smaller perturbation budgets, where gradient degradation is prevalent and GRILL clearly outperforms baseline attacks. At higher budgets, output distortions from baseline attacks become comparable.

6.6 Ablation study, Convergence & Efficiency

We further analyze the behavior of GRILL through ablation, convergence, and efficiency studies.

GRILL vs Layer-wise loss Summation (LLS): Figure 10 compares GRILL against a variant obtained by removing the distortion term δ^* from Eq. 7, reducing the objective to simple layer-loss summation (LLS). We evaluate both formulations under the base perturbation budgets ($c = 0.025$ for NVAE and $c = 0.22$ for DiffAE, see Table 1), where gradient degradation effects are most pronounced.

GRILL consistently outperforms LLS, with particularly large gains on NVAE (Figure 10) due to its widespread ill-conditioning (Figure 2b) across multiple intermediate layers. Despite exhibiting only localized severe ill-conditioning (Figure 2d), DiffAE also shows a substantial performance gap (Figure 10: DiffAE). These results indicate that simple layer-wise loss aggregation is insufficient to mitigate gradient degradation effectively and further support the hypothesis. Additionally, Figure 9 shows that GRILL outperforms BSA, which is based on layer-wise and token-wise cosine similarity aggregation, further indicating the superiority of GRILL over conventional aggregation-based attack formulations that lack the δ^* term from Eq. 7.

Effect of the Fraction of Encoder–Decoder Splits Used in GRILL: To analyze the role of partial split aggregation, we vary the fraction of encoder–decoder splits considered in Equation 7 from 30% to 100% using GRILL-wass on NVAE, which serves as an ideal testbed due to its severe ill-conditioning. As shown in Figure 11, increasing the proportion of intermediate splits consistently improves attack strength, leading to larger output distortions.

Gradient degradation and recovery: To validate our hypothesis that weak attacks arise from gradient degradation, we compare histograms of backpropagated adversarial loss gradients for baseline attacks and GRILL during optimization. As shown in Figure 12, baseline gradients remain sharply concentrated near zero throughout optimization, indicating severe gradient degradation. In contrast, GRILL produces substantially broader, lower-kurtosis gradient distributions with larger magnitudes, indicating improved adversarial gradient propagation during optimization. The behavior is consistently observed for both NVAE and DiffAE.

Weighting strategy: We investigate conditioning-aware weighting strategies for GRILL's intermediate split losses. Specifically, we compare equal, random, and condition-number-based (κ-based) weighting schemes, in which more ill-conditioned split interfaces (larger κ) are assigned lower weights, whereas better-conditioned

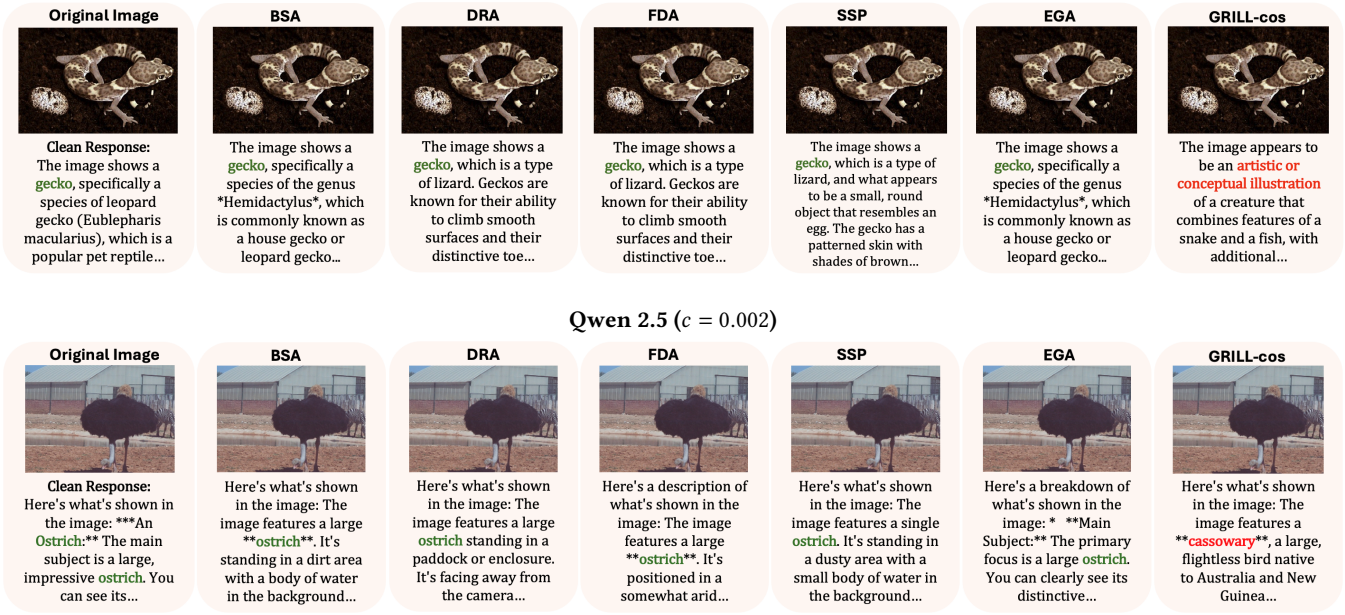


Figure 8: Adversarially perturbed images evaluated using the prompt "What is shown in this image?" under small perturbation budgets c. Ellipses (...) indicate omitted text without loss of context.

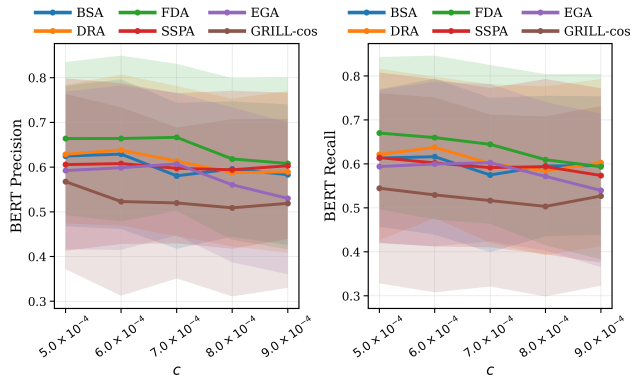


Figure 9: Gemma 3 quantitative results of sample–specific attacks

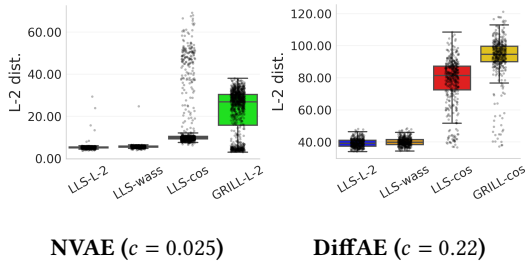


Figure 10: Ablation: GRILL vs layer-wise loss aggregation (LLS)

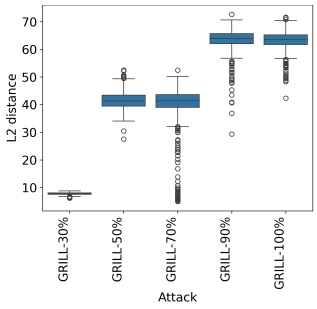
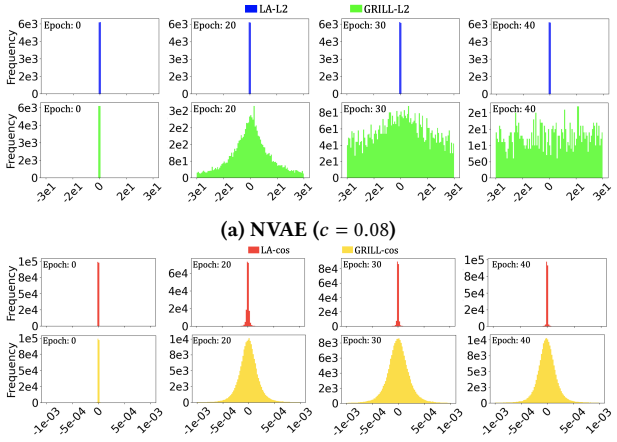


Figure 11: Output distortion distributions for split-wise GRILL ablations on NVAE under varying fractions of encoder–decoder splits.

split interfaces receive higher weights. As shown in Figures 13(a–d), κ-based weights noticeably impact optimization, particularly in ill-conditioned models like NVAE (Figure 13c). In particular, κ-based weighting improves optimization behavior for highly ill-conditioned models, suggesting that conditioning-aware weighting can further facilitate adversarial gradient propagation.

Convergence behavior: Figure 14 compares the optimization trajectories of GRILL and the strongest baseline attacks across all AE architectures. In highly ill-conditioned models such as NVAE and DiffAE, GRILL converges to substantially higher distortion levels, whereas baseline attacks plateau earlier due to degraded gradients.



(a) NVAE ($c = 0.08$)

(b) DiffAE ($c = 0.3$)

Figure 12: Adversarial loss gradient histograms for baseline and GRILL-based universal attacks.

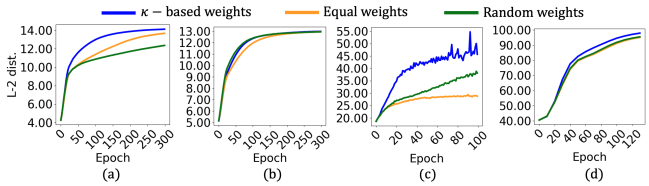
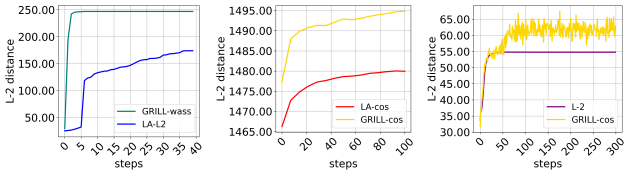
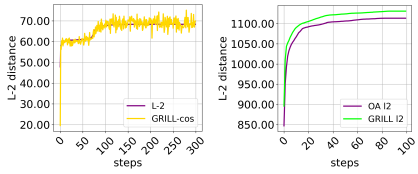


Figure 13: Equal, random, and κ-based weighting strategies for GRILL across (a) β-VAE, (b) TC-VAE, (c) NVAE (d) DiffAE.



(a) NVAE (b) DiffAE (c) TC-VAE



(d) β-VAE (e) MAE

Figure 14: Optimization trajectories of GRILL and the strongest baseline attacks across AEs.

In comparatively well-conditioned models such as MAE and β-VAE, both methods exhibit similar convergence behavior.

Efficiency: Table 2 reports the floating-point operations (FLOPs) per attack step for AE and VLM attacks. By reusing intermediate representations from the forward/backward pass, GRILL remains computationally comparable to existing latent-space attacks on AEs and competitive with state-of-the-art VLM attack baselines.

Table 2: FLOPs per attack step for AEs and VLMs

AEs		
Method	**NVAE**	**DiffAE**
LA-(L2, wass, cos)	1.73×10^{11}	1.40×10^{11}
GRILL-(L2, wass, cos)	1.73×10^{11}	1.40×10^{11}
VLMs		
Method	**Qwen 2.5**	**Gemma 3**
BSA	2.94×10^{13}	2.59×10^{13}
GRILL-cos	1.70×10^{13}	2.59×10^{13}
DRA	1.42×10^{13}	1.01×10^{13}
EGA	2.68×10^{13}	1.95×10^{13}
FDA	6.14×10^{12}	1.01×10^{13}
SSPA	6.14×10^{12}	1.56×10^{13}

7 Conclusions and Outlook

In this work, we show that ill-conditioned layers with near-zero singular values can suppress adversarial gradient propagation in encoder–decoder architectures, leading to weak norm-bounded attacks and overstated robustness. We propose GRILL, a gradient-guided adversarial optimization framework designed to mitigate gradient degradation and consistently improve attack effectiveness across diverse AE architectures. We further present results on vision–language models, providing preliminary evidence that related optimization phenomena may also arise in broader encoder–decoder architectures. These findings highlight the importance of conditioning-aware adversarial optimization for reliable robustness evaluation and motivate future research on computationally efficient adversarial evaluation methods for generative and multimodal encoder-decoder systems.

Limitations. GRILL is limited to white-box robustness evaluation in encoder–decoder architectures with continuous latent spaces, and does not directly extend to discrete latent spaces (e.g., VQ-VAE).

References

[1] Abulikemu Abuduweili and Changliu Liu. 2024. Estimating Neural Network Robustness via Lipschitz Constant and Architecture Sensitivity. *arXiv* (2024).

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* 9 (2021), 155161–155196.

[3] Vegard Antun, Nina M Gottschling, Anders C Hansen, and Ben Adcock. 2021. Deep learning in scientific computing: Understanding the instability mystery. *SIAM NEWS MARCH* 2021 (2021).

[4] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. 2020. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30088–30095.

[5] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*. PMLR, 274–283.

[6] Ben Barrett, Alexander Camuto, Matthew Willetts, and Tom Rainforth. 2022. Certifiably robust variational autoencoders. In *AISTATS*. PMLR, 3663–3683.

[7] Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, and Tom Rainforth. 2021. Towards a theoretical understanding of the robustness of variational autoencoders. In *AISTATS*. PMLR, 3565–3573.

[8] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE S&P*. Ieee, 39–57.

[9] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dj Dvijotham, and Pushmeet Kohli. 2020. Adversarially robust representations with smooth encoders. In *ICLR*.

[10] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. *NeurIPS* 31 (2018).

[11] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *ICML*. PMLR, 854–863.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

[13] Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. 2018. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224* (2018).

[14] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. *NeurIPS* 32 (2019).

[15] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. 2019. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8069–8079.

[16] George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. 2018. Adversarial attacks on variational autoencoders. *arXiv* (2018).

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[18] Nina M Gottschling, Vegard Antun, Anders C Hansen, and Ben Adcock. 2025. The Troublesome Kernel: On Hallucinations, No Free Lunches, and the Accuracy-Stability Tradeoff in Inverse Problems. *SIAM Rev.* 67, 1 (2025), 73–104.

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *NeurIPS* 30 (2017).

[20] Kartik Gupta and Thalaiyasingam Ajanthan. 2022. Improved gradient-based adversarial attacks for quantized networks. In *AAAI*, Vol. 36. 6810–6818.

[21] William W Hager. 1979. Lipschitz continuity for constrained processes. *SIAM J. Control Optim.* 17, 3 (1979), 321–338.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[23] Mengqi He, Xinyu Tian, Xin Shen, Jinhong Ni, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Few Tokens Matter: Entropy Guided Attacks on Vision-Language Models. *arXiv preprint arXiv:2512.21815* (2025).

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)* 3 (2017).

[25] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science* 313, 5786 (2006), 504–507.

[26] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186* (2024).

[27] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

[28] Valentin Khrulkov and Ivan Oseledets. 2018. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8562–8570.

[29] Andreas Kirsch et al. 2011. *An introduction to the mathematical theory of inverse problems*. Vol. 120. Springer. Theorem A.58 and Theorem 3.10 in Chapter 3 show instability due to small singular values.

[30] Anna Kuzina, Max Welling, and Jakub M Tomczak. 2022. Alleviating adversarial attacks on variational autoencoders with mcmc. *Advances in Neural Information Processing Systems* 35 (2022), 8811–8823.

[31] Anna Kuzina, Max Welling, and Jakub M Tomczak. 2021. Diagnosing vulnerability of variational auto-encoders to adversarial attacks. *arXiv* (2021).

[32] Junnan Li et al. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *ICML* (2022).

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

[34] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. 2020. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 940–949.

[35] Charles H Martin and Michael W Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *JMLR* 22, 165 (2021), 1–73.

[36] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv* (2018).

[37] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2020. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 262–271.

[38] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. 2017. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *NeurIPS* 30 (2017).

[39] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*. 10619–10629.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.

[41] Chethan Krishnamurthy Ramanaik, Arjun Roy, and Eirini Ntoutsi. 2024. Adversarial Robustness of VAEs across Intersectional Subgroups. *arXiv* (2024).

[42] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *MLSDA*. 4–11.

[43] Hanie Sedghi, Vineet Gupta, and Philip M Long. 2018. The singular values of convolutional layers. *arXiv* (2018).

[44] Abhishek Sinha, Mayank Singh, and Balaji Krishnamurthy. 2018. Neural networks in an adversarial setting and ill-conditioned weight space. In *ECML PKDD*. Springer, 177–190.

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv* (2013).

[46] P. Tabacof, J. Tavares, and E. Valle. 2016. Adversarial Images for Variational Autoencoders. In *NeurIPS Workshop*. Workshop paper.

[47] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv preprint arXiv:1908.07490* (2019).

[48] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv* (2025).

[49] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *NeurIPS* 33 (2020), 1633–1645.

[50] Olga Tsymboi, Danil Malaev, Andrei Petrovskii, and Ivan Oseledets. 2023. Layer-wise universal adversarial attack on NLP models. In *Findings of ACL*. 129–143.

[51] Arash Vahdat and Jan Kautz. 2020. NVAE: A deep hierarchical variational autoencoder. *NeurIPS* 33 (2020), 19667–19679.

[52] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS* 30 (2017).

[53] Max van Spengler, Jan Zahálka, and Pascal Mettes. 2024. Adversarial attacks on hyperbolic networks. In *ECCV*. Springer, 363–381.

[54] Aladin Virmaux and Kevin Scaman. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *NeurIPS* 31 (2018).

[55] Xian Wei, Yangyu Xu, Yanhui Huang, Hairong Lv, Hai Lan, Mingsong Chen, and Xuan Tang. 2022. Learning extremely lightweight and robust model with differentiable constraints on sparsity and condition number. In *ECCV*. Springer, 690–707.

[56] Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. 2019. Improving VAEs' robustness to adversarial attack. *arXiv* (2019).

[57] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems* 36 (2023), 52936–52956.

[58] Yuichi Yoshida and Takeru Miyato. 2017. Spectral norm regularization for improving the generalizability of deep learning. *arXiv* (2017).

[59] Yaopei Zeng, Yuanpu Cao, Bochuan Cao, Yurui Chang, Jinghui Chen, and Lu Lin. 2024. AdvI2I: Adversarial Image Attack on Image-to-Image Diffusion models. *arXiv* (2024).

[60] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[61] Haomin Zhuang, Yihua Zhang, and Sijia Liu. 2023. A pilot study of query-free adversarial attack against stable diffusion. In *CVPR*. 2385–2392.