Can GRQO based fine-tuning speed up the inference stage of the denoising pipeline by reducing the reliance on the TTA?

Reinforcement Learning Based Learning for Seismic Denoising

Andrei Oprescu[1]
Supervisors: Dr. Jing Sun[1]**, Dr. Tiexing Wang**[3]**, Jiahua Zhao**[2,4]**, Dr. Eric Verschuur**[2]

[1]EEMCS, Delft University of Technology, The Netherlands
[2]Faculty of Civil Engineering and Geosciences, Delft University of Technology, The Netherlands
[3]R&D Department, Shearwater GeoServices, UK
[4]The Cyprus Institute, Cyprus

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Andrei Oprescu[1]
Final project course: CSE3000 Research Project
Thesis committee: Dr. Jing Sun[1], Dr. Tiexing Wang[3], Jiahua Zhao[2,4], Dr. Eric Verschuur[2]
Examiner: Dr. Petr Kellnhofer[1]

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Seismic denoising is essential for subsurface imaging. Deterministic methods such as Radon transforms exist but struggle to separate signal from noise. Recent solutions like vision foundation models (VFMs) offer better signal separation, but generalize poorly across geological domains. Parameter-efficient fine-tuning with LoRA and kurtosis-guided test-time adaptation (TTA) improves cross-domain generalization, yet TTA performs weight updates during inference and collapses throughput from 114.29 to 1.09 patches per second, making real-time deployment impractical. This paper investigates whether group relative query optimization (GRQO), a reinforcement-learning fine-tuning method applied before deployment, can reduce reliance on TTA. A LoRA-adapted DINOv3 backbone with a U-Net decoder is fine-tuned with a GRQO objective that combines a multi-head reward, KL-divergence regularization against a frozen reference, and a head-diversity penalty. On a 4,000-patch unseen synthetic data set, SFT+GRQO alone improves signal retention (MS-SSIM-R 0.638→0.604) while preserving a similar inference speed. GRQO does not fully replace TTA, but it lowers the optimal TTA budget for signal retention: the best configuration for MS-SSIM-R pairs GRQO with only 25 TTA epochs (MS-SSIM 0.862, MS-SSIM-R 0.452), and longer adaptation degrades performance and decreases throughput. However, GRQO alone minimally deteriorates performance when evaluated on real seismic data (LS 0.373→0.381). These results indicate that pre-deployment reinforcement learning can shorten the costly real-time adaptation and increase signal retention, but may perform worse on unseen seismic domains.

1 Introduction

Subsurface imaging heavily relies on seismic monitoring for applications such as crude oil exploration, carbon storage and hazard management (Hudson et al., 2024). Both active and passive seismic surveys are used to image the subsurface (Zhao et al., 2026), but the reliability of raw seismic data is greatly compromised by complex noise environments, such as ambient noise and coupling noise in distributed acoustic sensing (DAS) systems (Taha et al., 2025). Existing noise mitigation methods, such as wavelet and Radon transforms, struggle to separate signal from noise and require expert parameter tuning for each new domain, making them labor-intensive and difficult to scale (Zhao et al., 2026).

In recent years, vision foundation models (VFMs) have emerged as a promising alternative, as convolutional neural networks and transformers can learn non-linear patterns from data with less expert intervention. However, VFMs are highly sensitive to domain shift (Hou & Messud, 2021), and current training strategies that aggregate multiple seismic datasets still generalize poorly to unseen geological domains (Maguire et al., 2024; Sheng et al., 2025). Supervised fine-tuning further requires extensive data and compute, and risks catastrophic forgetting.

To reduce the impact of the rarity of datasets, previous research has tried using synthetic datasets generated through physics-based modeling (Merrifield et al., 2022). However, while synthetic datasets offer controlled conditions, they lack the noise complexity of real-world datasets, resulting in less meaningful data for training the foundation model.

A recent foundation model based approach introduces parameter efficient fine-tuning (PEFT) to address some of these issues, specifically using LoRA (Zhao et al., 2026). This removes the need for large computational resources for fine-tuning the VFM, and mitigates catastrophic forgetting by freezing the pre-trained backbone's weights and adding smaller

weight matrices trained in parallel. Instead of fine-tuning the LoRA weight matrices on previous data, Zhao et al. (2026) opt for kurtosis-guided test-time adaptation (TTA). By exploiting the statistical properties of seismic data, the LoRA trainable parameters are updated through unsupervised training during the TTA phase on high-kurtosis patches from real-time data. Although this drastically improves the ability of the model to generalize to new domains, it also drastically reduces inference speed, making it difficult to adopt in real-world deployments.

This tradeoff between domain generalization and inference speed motivates the need to shorten the demanding TTA process. As an alternative to unsupervised backbone adaptation at inference time, one option is to use reinforcement learning based fine-tuning, such as group relative query optimization (GRQO). This training technique addresses the TTA bottleneck by shifting the adaptation burden from real-time weight updates to a pre-deployment optimization phase. Since GRQO optimizes based on relative reward signals without requiring ground-truth labels, it waives the need for large labeled datasets and improves generalization in vision prompting scenarios Pan et al., 2025.

Since the TTA phase updates LoRA weights to identify and denoise high-kurtosis patches, a similar approach is used to train the model with GRQO. High-kurtosis patches are incentivized by the GRQO reward structure, and patches that achieve a higher signal-to-noise ratio after denoising are rewarded. Through the reinforcement learning algorithm, the LoRA matrices are updated to fit better to the training data while generalizing across multiple domains.

The main research question of this paper is: can GRQO replace the TTA step completely, enabling efficient real-time seismic denoising? In addition, it is also valuable to examine how GRQO performs in conjunction with a reduced TTA phase, where GRQO provides some generalization across domains but also uses TTA to a lesser extent by training on less epochs.

2 Related Work

2.1 Seismic Denoising: From Traditional Methods to Deep Learning

Subsurface imaging data is heavily compromised by complex noise. Traditional mitigation techniques, like Radon and wavelet transforms, struggle with dynamic signal separation and require expert, per-domain tuning (Zhao et al., 2026). Data-driven deep learning, particularly convolutional neural networks (CNNs), improved automatic signal isolation without heavy manual parameterization (Yu et al., 2019). However, these supervised models are highly sensitive to domain shift, failing to generalize to unseen geological settings with different noise profiles (Maguire et al., 2024). Supervised fine-tuning therefore demands extensive labeled data for every new target, risking catastrophic forgetting.

2.2 Vision Foundation Models and Parameter-Efficient Fine-Tuning

To avoid training domain-specific networks from scratch, recent approaches use Vision Foundation Models (VFMs). Models like DINOv3 (Siméoni et al., 2025) and the domain-specific Seismic Foundation Model (SFM) (Sheng et al., 2025) extract generalized features for complex seismic interpretation. However, fully fine-tuning these massive models per

site is computationally prohibitive. Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA) (Hu et al., 2022), addresses this bottleneck. By freezing the backbone and optimizing only low-rank matrices, LoRA calibrates models to site-specific noise without the overhead of full-network retraining (Zhao et al., 2026).

2.3 The Inference Bottleneck: TTA vs. Reinforcement Learning

While LoRA reduces parameters, handling domain shift during active deployment remains a bottleneck. Test-Time Adaptation (TTA) mitigates distribution shifts by dynamically updating parameters on unlabeled test samples (Sun et al., 2020). In seismic denoising, TTA updates LoRA matrices during inference using high-kurtosis patches (Zhao et al., 2026). However, real-time backpropagation drastically reduces throughput, hindering real-time applications. Conversely, Group Relative Query Optimization (GRQO) shifts adaptation to the pre-deployment phase (Pan et al., 2025). By structurally optimizing the network for generalization via reinforcement learning, GRQO may reduce or eliminate the need for inference-heavy TTA, which is the focus of this paper.

3 Data Preparation

In this section, the data collection and preparation are described. Firstly, the dataset is presented, followed by the image patching strategy. The process that determines which patches the model trains on is then described.

3.1 Dataset Acquisition and Description

Two datasets are taken from the open-source seismic denoising dataset provided by Sheng et al. (2025). The first set comprises 2000 synthetic patches with accompanying denoised ground-truth images, created using geological forward modeling. When running the experiments, 1600 patches will be used as the training data and 400 will be used as the test set.

The second dataset is a subset derived from a large compilation of global seismic surveys. This set contains 4000 patches without ground-truth labels. These patches are comprised of real-world data collected from seismic surveys.

A third dataset will also be used to assess how well the model generalizes across different domains. This is the ThinkOnward Image Impeccable dataset (ThinkOnward, 2024), a synthetic dataset. It is comprised of 3 volumes of data, which are then split into 19,800 patches. Since this dataset will be used with TTA for evaluating generalization, the same number of randomly selected patches (4000) will be used as the real world dataset.

All patches are of size 224×224 pixels and are created in the same manner as in Sheng et al. (2025): 224×224 snippets are extracted with overlap, keeping 1 in 5 images both crossline and inline. This results in patches that capture more seismic information while preventing overfitting from overly similar samples.

3.2 Data Pre-processing Pipeline

Once the patches are loaded for training, they are converted into a format the backbone can process. Seismic amplitudes exhibit large dynamic ranges, which destabilizes deep neural

network training. Therefore, patch pixel values are normalized using z-score normalization.

Because DINOv3 Siméoni et al., 2025 only accepts 3D inputs while the patches contain 2D data, the patch values are copied across all three color channels. Geometric transformations such as rotation and flipping are intentionally omitted, as they would not preserve the signal integrity and kinematic properties of seismic traces.

3.3 High-Kurtosis Patch Identification

The SFT and GRQO training loops use all patches described above. However, during TTA it is preferable to avoid training on patches that are dominated by noise, as the model would otherwise fit random noise rather than seismic signals.

Seismic reflection events exhibit sparsity and high non-Gaussianity relative to background noise, yielding higher kurtosis values. The kurtosis of a patch is:

$$\kappa(x_{\text{patch}}) = \frac{\mathbb{E}[(x_{\text{patch}} - \mu)^4]}{\sigma^4}$$

where x_{patch} contains the patch pixel values. The bottom 5% of patches by kurtosis are discarded, keeping most informative data while reducing the risk of training on background noise.

4 Methodology

This section introduces the GRQO training architecture for the seismic denoising pipeline of Zhao et al. (2026). It describes the backbone VFM selection, the LoRA-based PEFT pipeline, the baseline TTA method, and the GRQO fine-tuning adaptation.

4.1 Pipeline Overview

To create the proposed framework, a suitable vision transformer is chosen for the seismic denoising task. The DINOv3 VFM Siméoni et al., 2025 is selected for its ability to extract global and local features from image patch inputs, and its capability to understand semantic features such as edges and textures relevant to seismic trace interpretation.

4.2 Vision Foundation Model & LoRA Integration

To extract robust semantic features from the seismic input, this framework uses DINOv3 Siméoni et al., 2025 as the backbone VFM. DINOv3 relies on a Vision Transformer (ViT) design that divides input data into non-overlapping patches, processing them through a stack of multi-head self-attention layers to capture global and local dependencies. The core self-attention operation projects seismic embeddings into a feature space using dense projection weight matrices for the Query (W^Q), Key (W^K), and Value (W^V).

Adapting these high-dimensional projection matrices entirely to a new seismic domain is computationally prohibitive and prone to overfitting, often leading to catastrophic forgetting. To achieve parameter-efficient adaptation, Low-Rank Adaptation (LoRA) Hu et al., 2022 is injected into the transformer blocks following the approach of Zhao et al. (2026). LoRA freezes the pre-trained weight matrix W_0 and introduces trainable low-rank matrices

$A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ to run in parallel, where d is the original embedding dimension and r is the low-rank bottleneck. The forward pass of a LoRA-modified layer is:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

A is initialized with random Gaussian noise and B with zeros, so the initial update ΔW is zero. LoRA parameters are applied to the Query, Key, Value, and Output projection matrices.

To enable self-calibration to site-specific noise without labeled data, the LoRA parameters are refined unsupervised. The framework retains the top 95% of high-kurtosis patches (see Section 3.3) for self-training, discarding the bottom 5% to prevent fitting background noise.

4.3 The Baseline: Test-Time Adaptation

While LoRA significantly reduces the trainable parameter count, addressing domain shift during deployment introduces a new computational bottleneck. To handle unseen field conditions, the baseline framework of Zhao et al. (2026) employs a Test-Time Adaptation (TTA) module. This module leverages the non-Gaussian statistical characteristics of seismic data (Section 3.3) to select high-information patches during inference.

During the TTA phase, the algorithm isolates patches with the highest kurtosis scores and uses them for unsupervised learning, requiring backpropagation to update the LoRA matrices A and B *during inference*. This real-time weight updating severely degrades throughput. While a LoRA-adapted DINOv3 model without TTA achieves 114.29 patches per second, activating the multi-epoch TTA process drops throughput to just 1.09 patches per second. This latency is a critical bottleneck for real-time monitoring applications such as microseismic event detection, motivating the shift to the pre-deployment GRQO framework.

4.4 Group Relative Query Optimization Algorithm

Due to the high compute cost of TTA, GRQO is used to shift the adaptation burden from inference to the pre-deployment fine-tuning phase Pan et al., 2025. With this, the framework of Zhao et al. (2026) becomes more effective in real-time denoising scenarios. Moreover, GRQO is a reinforcement-learning fine-tuning framework, so TTA can still be used alongside it for further cross-domain generalization.

The model trained with GRQO is a DINOv3 backbone pre-trained with 20 epochs of SFT (Figure 1). Keeping the LoRA matrices allows TTA to optionally be applied after GRQO, leveraging the domain knowledge acquired during the GRQO phase for more robust generalization. In the original GRQO implementation, the task is one-shot object detection, which naturally produces a distribution of predictions per query. Since our model produces one denoised patch per query, K decoder heads are implemented after the DINOv3 encoder to produce K denoised candidates per input. Let x be the noisy input patch, y the clean target, and $\hat{y}_k$ the prediction of decoder head $k \in \{0, \ldots, K-1\}$. Batches are indexed by $b \in \{1, \ldots, B\}$ and pixels by $i \in \{1, \ldots, HW\}$. The residual of head k is $r_k = x - \hat{y}_k$.

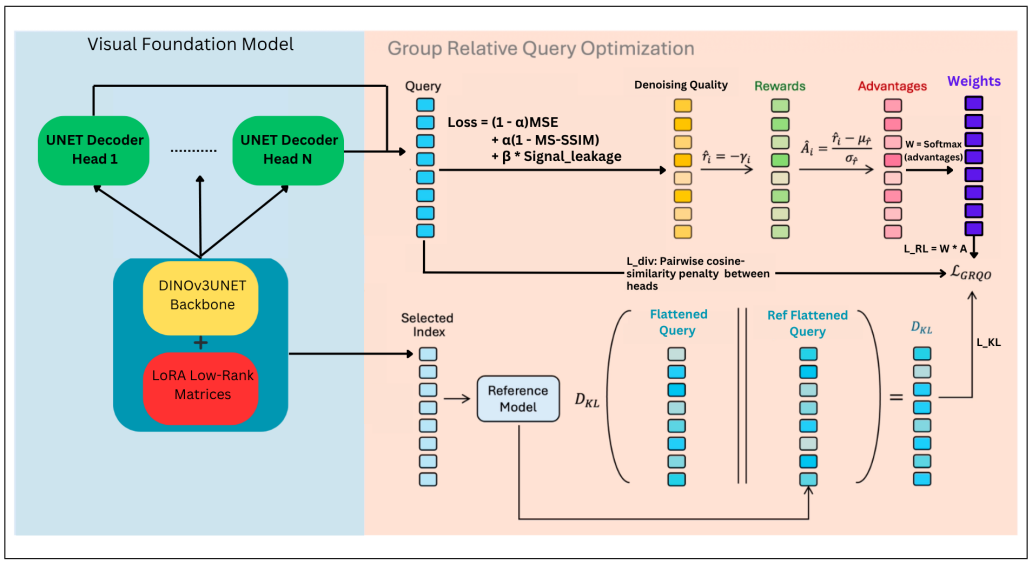

Figure 1: The DINOv3 backbone with the GRQO fine-tuning method. A noisy patch is passed through a single encoder and decoded by multiple non-identical decoder heads. The upper path runs the RL algorithm; the lower path applies KL-divergence regularization against a frozen reference model. Both combine to form the GRQO loss.

To induce diversity in the decoder heads, the head weights are perturbed at initialization:

$$\begin{aligned} \boldsymbol{\theta}_0 &= \boldsymbol{\theta}^{\mathrm{SFT}}, \\ \boldsymbol{\theta}_k &= \boldsymbol{\theta}^{\mathrm{SFT}} + \rho\,\mathrm{RMS}\big(\boldsymbol{\theta}^{\mathrm{SFT}}\big)\,\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ k \geq 1, \end{aligned} \tag{2}$$

$$\mathrm{RMS}(\boldsymbol{\theta}) = \frac{\|\boldsymbol{\theta}\|_2}{\sqrt{N} + \epsilon}, \quad N = \dim(\boldsymbol{\theta}), \tag{3}$$

where ρ is the perturbation fraction. Head 0 retains the exact SFT weights.

Head Diversity Penalty. To prevent the K heads from collapsing to identical solutions despite the initial perturbation, a pairwise cosine-similarity penalty is added to the training objective. For each head's flattened, unit-normalised output $\hat{y}_k^\flat$, the diversity loss is the mean squared cosine similarity over all $\binom{K}{2}$ head pairs:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{\binom{K}{2}} \sum_{i<j} \left(\frac{\hat{y}_i^\flat \cdot \hat{y}_j^\flat}{\|\hat{y}_i^\flat\|_2 \|\hat{y}_j^\flat\|_2} \right)^2. \tag{4}$$

Squaring the cosine term penalises both correlated ($\cos \approx +1$) and anti-correlated ($\cos \approx -1$) pairs equally, driving heads toward orthogonal denoising strategies.

Reinforcement Learning Pipeline. Each query is denoised by the K heads, and an intermediate loss is computed that combines MSE, MS-SSIM Wang et al., 2003, and a

signal-leakage penalty:

$$\mathcal{L}_{b,k}^{\mathrm{MSE}} = \frac{1}{HW} \sum_{i} \big(\hat{y}_{k,b,i} - y_{b,i}\big)^2, \tag{5}$$

$$\mathcal{L}_{b,k}^{\mathrm{SSIM}} = 1 - \text{MS-SSIM}\big(\hat{y}_{k,b}, y_b\big), \tag{6}$$

$$\mathcal{L}_{b,k}^{\mathrm{rec}} = \alpha\, \mathcal{L}_{b,k}^{\mathrm{SSIM}} + (1-\alpha)\, \mathcal{L}_{b,k}^{\mathrm{MSE}}. \tag{7}$$

$$S_{b,k} = \big|\mathrm{corr}(r_{k,b}, y_b)\big| = \frac{\big|\sum_i (r_{k,b,i} - \bar{r}_{k,b})(y_{b,i} - \bar{y}_b)\big|}{\|r_{k,b} - \bar{r}_{k,b}\|_2 \|y_b - \bar{y}_b\|_2 + \epsilon}. \tag{8}$$

$$\mathcal{L}_{b,k}^{\mathrm{aug}} = \mathcal{L}_{b,k}^{\mathrm{rec}} + \beta S_{b,k}. \tag{9}$$

$\alpha = 0.5$ balances MS-SSIM and MSE; $\beta = 0.1$ scales the signal-leakage penalty. The reward for each head is:

$$R_{b,k} = -\mathcal{L}_{b,k}^{\mathrm{aug}} \tag{10}$$

Advantages are normalized within each batch to remove the arbitrary scale of raw rewards:

$$\begin{aligned} A_{b,k} &= \mathrm{clip}\left(\frac{R_{b,k} - \frac{1}{K}\sum_j R_{b,j}}{\sigma_b + \epsilon},\, -2,\, +2\right), \\ \sigma_b &= \sqrt{\frac{1}{K}\sum_j \big(R_{b,j} - \frac{1}{K}\sum_m R_{b,m}\big)^2}. \end{aligned} \tag{11}$$

The clip to $[-2, +2]$ prevents a single head from dominating the softmax weighting. Softmax weights derived from the advantages are used to scale the per-head losses:

$$\begin{aligned} w_{b,k} &= \frac{\exp(A_{b,k}/\tau)}{\sum_j \exp(A_{b,j}/\tau)}, \\ \mathcal{L}_{\mathrm{RL}} &= \frac{1}{B} \sum_b \sum_k w_{b,k}\, \mathcal{L}_{b,k}^{\mathrm{aug}}, \end{aligned} \tag{12}$$

where τ is the temperature.

Kullback–Leibler Divergence Pipeline. To stabilise training under the high-variance RL gradient signals and to prevent distributional shift between the training model and the reference model, KL-divergence regularization is applied Pan et al., 2025. The reference model shares the same DINOv3 architecture and is pre-trained with 20 epochs of SFT but its weights are frozen throughout GRQO training. The KL divergence is computed between the softmax-normalised pixel-level distributions of the reference and training model outputs:

$$\begin{aligned} \tilde{u}_{b,i} &= \mathrm{clip}(u_{b,i}, -50, 50), \\ \mathrm{softmax}_i(\tilde{u}_b) &= \frac{\exp(\tilde{u}_{b,i})}{\sum_j \exp(\tilde{u}_{b,j})}, \end{aligned} \tag{13}$$

$$\begin{aligned} P_{b,i} &= \mathrm{softmax}_i\big(\tilde{y}_b^{\mathrm{ref}}\big), \\ Q_{b,i} &= \mathrm{softmax}_i\big(\tilde{\hat{y}}_b\big), \quad \bar{y}_b = \frac{1}{K}\sum_k \hat{y}_{k,b}, \end{aligned} \tag{14}$$

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{B} \sum_{b} \mathrm{KL}\big(P_b \,\|\, Q_b\big). \tag{15}$$

Finally, all loss components are combined to yield the complete GRQO objective:

$$\begin{aligned} \mathcal{L}_{\mathrm{all}} &= \frac{1}{BK} \sum_{b} \sum_{k} \mathcal{L}_{b,k}^{\mathrm{aug}}, \\ \mathcal{L}_{\mathrm{GRQO}} &= \lambda_{\mathrm{all}}\, \mathcal{L}_{\mathrm{all}} + \lambda_{\mathrm{rl}}\, \mathcal{L}_{\mathrm{RL}} + \lambda_{\mathrm{kl}}\, \mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{div}}\, \mathcal{L}_{\mathrm{div}}. \end{aligned} \tag{16}$$

Each term in $\mathcal{L}_{\mathrm{GRQO}}$ encodes a property that good seismic denoising must satisfy before deployment, so the model depends less on test-time adaptation to learn it. The reconstruction term $\mathcal{L}_{b,k}^{\mathrm{rec}}$ defines a clean patch by pairing MSE for amplitude fidelity with MS-SSIM for the multi-scale structure of reflectors. The signal-leakage penalty $S_{b,k}$ is the seismic-specific term: by penalizing correlation between the removed residual and the clean target, it suppresses noise without erasing weak coherent events – the signal-retention property that MS-SSIM-R measures.

The remaining terms make this learnable without labels. $\mathcal{L}_{\mathrm{RL}}$ has the K heads compete on the same patch, normalizes their advantages within the group, and softmax-weights the better candidates, so the model prefers cleaner reconstructions through the same unsupervised objective as TTA but shifted to pre-deployment; $\mathcal{L}_{\mathrm{div}}$ stops the heads collapsing, since this group-relative comparison is only informative when the candidates differ. Finally, $\mathcal{L}_{\mathrm{KL}}$ anchors the model to the frozen SFT reference, stabilising the high-variance RL gradients and keeping the backbone from drifting from its generalizable features, which lets GRQO embed domain structure in the LoRA matrices and lower the TTA budget.

5 Experiments

This section describes the experimental setup, including the model architecture, pre-trained baselines, fine-tuning hyperparameters, and evaluation metrics.

5.1 Experimental Setup

The pipeline is implemented in PyTorch due to its compatibility with DINOv3 and flexible training loop configuration. The DINOv3 model is initialized via the Hugging Face library, and LoRA is integrated using the PEFT library in Python. Training and evaluation are accelerated using an NVIDIA RTX 4090 GPU. The seismic patches fed into the GRQO training loop are normalized using z-score normalization to stabilize training.

5.2 Model, Pre-Trained Baselines & GRQO Fine-tuning Architecture

The backbone chosen is DINOv3-ViT-S/16 Siméoni et al., 2025 with a U-Net decoder Ronneberger et al., 2015. Since GRQO requires a strong reference model, the backbone is first pre-trained with 20 epochs of SFT. Table 6 in Appendix A summarizes the SFT configuration. LoRA is injected into the `qkv` and `proj` attention layers with r=16 and α=64, matching the DINOv3 configuration. After SFT pre-training, the model serves as both the reference model and the starting point for GRQO training. The GRQO hyperparameters are listed in Table 1.

Table 1: GRQO-v5 fine-tuning hyperparameters.

Setting	**Value**
Optimizer	AdamW, wd = 0.01
LR (LoRA + decoder)	5×10^{-6}
LR schedule	cosine
Epochs / batch	10 / 16
Decoder heads K	8, perturbed by 5% RMS
α (MS-SSIM / MSE mix)	0.5
β (leakage penalty)	0.1
Temperature τ	1.0
λ_{all} / λ_{rl} / λ_{KL}	scheduled 1.0→0.2 / scheduled 0.2→1.0 / 10^{-3}
λ_{div} (head diversity)	0.05
KL reference	Frozen 20-ep SFT
Runs	3
Seeds	0, 1, 2

5.3 Evaluation Metrics

The model is evaluated using 5 metrics:

- **MS-SSIM** Wang et al., 2003: structural preservation between the denoised and clean reference patch (higher is better).
- **MS-SSIM-R**: structural similarity of the removed residual relative to the clean reference patch (lower is better; a lower value indicates that the removed component does not resemble the original signal).
- **PSNR**: peak signal-to-noise ratio of the denoised patch relative to the clean reference patch (higher is better).
- **Throughput**: inference speed in files per second, measuring whether GRQO with optional TTA is more computationally efficient than full TTA alone (Higher is better).
- **LS** (Local Similarity): the locally-windowed normalized cross-correlation between the denoised patch and the removed residual (lower is better; a low value indicates low local correlation with the retained signal)

5.4 Experimental Configurations & Baselines

The core question is whether GRQO can account for the domain generalization provided by TTA. Multiple configurations are evaluated after GRQO, TTA, or just the base model, as shown in Table 2. The SFT-50 model serves as an upper-bound reference, representing 50 supervised training epochs. The SFT-20 model serves as a model that has not yet converged and still can converge using GRQO to generalize better.

Table 2: DINOv3 experimental comparison matrix. All variants share the DINOv3-ViT-S/16 backbone and U-Net decoder. The base SFT-20 and SFT-50 models will be trained once more with GRQO-10. All of them will be adapted on the test data with TTA-25, and TTA-50 for the SFT-20 models to compare with the SFT-50 versions.

Variant	**GRQO**	**TTA epochs**
SFT-20	×	25, 50
SFT-50	×	25
SFT-20 + GRQO-10	✓	25, 50
SFT-50 + GRQO-10	✓	25

Table 3: DINOv3 quantitative results on the 400-sample synthetic test set. ↑ higher is better; ↓ lower is better. Values are mean ± standard deviation over 3 seeds. Throughput is measured as files per second during inference; for TTA variants it also includes the unsupervised adaptation time, dropping substantially due to real-time weight updates.

Variant	**MS-SSIM ↑**	**MS-SSIM-R ↓**	**PSNR (dB) ↑**	**Throughput (files/s)**
SFT-20	0.942 ± 0.017	0.540 ± 0.001	29.40 ± 1.36	114.29
SFT-50	0.956 ± 0.022	0.532 ± 0.001	**30.53** ± 1.78	114.29
SFT-20 + GRQO-10	0.944 ± 0.011	0.511 ± 0.002	29.26 ± 0.55	78.43
SFT-50 + GRQO-10	**0.957** ± 0.018	**0.510** ± 0.004	30.47 ± 1.04	78.43
SFT-20 + TTA-25	0.788 ± 0.015	0.532 ± 0.004	23.41 ± 0.90	2.14
SFT-50 + TTA-25	0.787 ± 0.001	0.547 ± 0.008	22.30 ± 1.14	2.16
SFT-20 + GRQO-10 + TTA-25	0.795 ± 0.010	0.533 ± 0.006	20.77 ± 1.34	1.43
SFT-50 + GRQO-10 + TTA-25	0.786 ± 0.019	0.532 ± 0.005	20.86 ± 1.36	1.44
SFT-20 + TTA-50	0.780 ± 0.005	0.532 ± 0.008	20.16 ± 0.11	1.09
SFT-20 + GRQO-10 + TTA-50	0.788 ± 0.005	0.545 ± 0.003	20.04 ± 0.05	0.72

6 Results & Discussion

Table 3 presents quantitative results across the 400-sample synthetic test set. The SFT-20 and SFT-50 variants are evaluated with their GRQO-trained and TTA variants.

Effect of GRQO alone. GRQO pre-conditioning alone yields a meaningful improvement in signal retention, as seen in Table 3: MS-SSIM-R drops from 0.540 (SFT-20 baseline) to 0.511, indicating that less seismic signal is removed during denoising. Simultaneously, MS-SSIM and PSNR both stay close to the baseline (0.944 vs. 0.942; 29.26 dB vs. 29.40 dB), approaching the SFT-50 upper bound despite starting from only 20 supervised epochs. Inference throughput remains high (≈78 files/s), confirming that GRQO does not introduce deployment latency.

Interaction with TTA. After applying TTA on the test set, the model's performance on denoising deteriorates. MS-SSIM drops from 0.942 to 0.788 on the SFT-20 model, and it drops from 0.944 to 0.795 for the SFT-20 GRQO-10 variant. Since the metrics drop for both SFT-only and the GRQO variant, this points to an issue in the TTA stage rather than GRQO. The likely culprit is the fact that the test set is made up of 400 images, meaning that the unsupervised adaptation does not have enough patches to train on for it to gain a good understanding of the noise model. Therefore, to test this hypothesis, a larger test

Table 4: DINOv3 quantitative results on the 4000-sample Image-Impeccable test set. ↑ higher is better; ↓ lower is better. Values are mean ± standard deviation over 3 seeds. Throughput is measured as files per second during inference; for TTA variants it also includes the unsupervised adaptation time, dropping substantially due to real-time weight updates.

Variant	**MS-SSIM ↑**	**MS-SSIM-R ↓**	**PSNR (dB) ↑**	**Throughput (files/s)**
SFT-20	0.862 ± 0.001	0.638 ± 0.002	23.69 ± 0.07	113.96
SFT-50	0.868 ± 0.001	0.618 ± 0.003	24.06 ± 0.08	114.29
SFT-20 + GRQO-10	0.862 ± 0.001	0.604 ± 0.006	23.31 ± 0.11	79.21
SFT-50 + GRQO-10	0.868 ± 0.001	0.599 ± 0.006	23.74 ± 0.09	79.21
SFT-20 + TTA-25	0.873 ± 0.001	0.496 ± 0.009	25.92 ± 0.70	2.20
SFT-50 + TTA-25	0.863 ± 0.005	0.521 ± 0.006	26.23 ± 0.52	2.24
SFT-20 + GRQO-10 + TTA-25	0.862 ± 0.001	**0.452** ± 0.019	24.88 ± 0.19	1.40
SFT-50 + GRQO-10 + TTA-25	0.881 ± 0.021	0.483 ± 0.016	26.15 ± 1.25	1.40
SFT-20 + TTA-50	0.858 ± 0.000	0.563 ± 0.000	**27.18** ± 0.00	1.06
SFT-20 + GRQO-10 + TTA-50	**0.886** ± 0.000	0.475 ± 0.000	27.06 ± 0.00	0.71

set can be used to evaluate the model, such as the Image Impeccable dataset, which is explored below.

Effects on Generalization. To see if training with GRQO on a dataset helps the model better generalize to other unseen datasets from different domains, an evaluation on another synthetic dataset can be performed. For this, the Image Impeccable dataset was used. GRQO displays good denoising performance on an unseen dataset. Despite MS-SSIM marginally changing, the MS-SSIM-R drops by 0.034 for the SFT-20 baseline, displaying a capability to remove less signal, as seen in Table 4. Moreover, the TTA variants of the models further push the model to generalize on unseen data, supporting the hypothesis that TTA on the SFM synthetic data decreases performance due to a small amount of data. SFT-20 GRQO-10 with TTA-25 improves upon the SFT-20 TTA-25 model by 0.044 on MS-SSIM-R, outperforming the SFT-50 TTA-10 baselines as well.

However, synthetic datasets may not be representative of the real world as it is based on assumptions and approximations of the earth's crust. Therefore, evaluating the GRQO model against a real world seismic survey is necessary to evaluate whether the model can help the model generalize. The data used are 4000 noisy patches from a seismic survey in the SFM dataset.

From this experiment, the GRQO is seen to perform similar to the SFT-only configuration at denoising real world domains it has not seen before. The SFT-20 + TTA-25 configuration performs best on this dataset.

Denoised Patches Visualization To examine how the patches are denoised visually by the different models, Figure 2 displays an example of a denoised patch from the synthetic SFM test set by them with their performance metrics. From the figure, you can see that the residual contains less traces of signal in the GRQO trained model. To observe the performance of GRQO on denoising real-wrold patches, Figure 3 shows the patches denoised by the SFT-20 and the SFT-20 GRQO-10 models. From the image, it can be observed that GRQO will denoise seismic signals with varying performance compared to the SFT-only

Table 5: Local similarity (LS, ↓ lower is better) of each DINOv3 variant on the 4000-sample real-world test set. Real data has no clean reference, so the reference-free LS metric is reported. Values are mean ± standard deviation over 3 seeds.

	SFT-20	SFT-50	SFT-20 + GRQO-10	SFT-50 + GRQO-10	SFT-20 + TTA-25	SFT-50 + TTA-25	SFT-20 + GRQO-10 + TTA-25	SFT-50 + GRQO-10 + TTA-25
LS ↓	0.383 ±0.000	0.373 ±0.000	0.398 ±0.004	0.381 ±0.003	**0.339** ±0.008	0.343 ±0.019	0.362 ±0.011	0.347 ±0.008

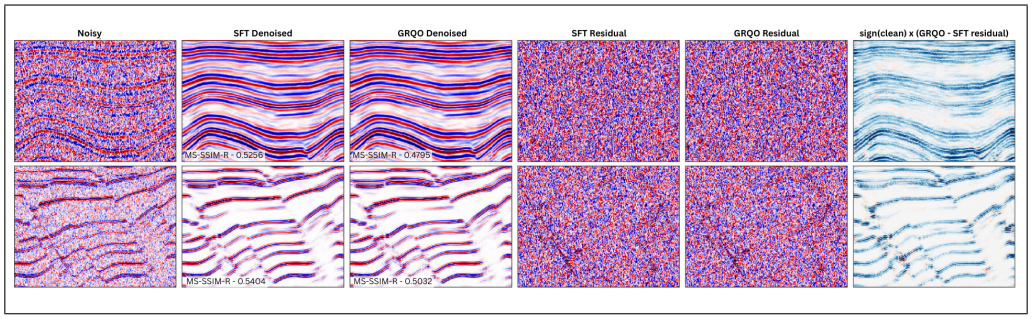

Figure 2: The DINOv3 backbone with SFT-20 against SFT-20 GRQO-10 on the SFM synthetic dataset. The right-most plots compare the signal leakage difference between the two configurations. Blue = More signal leakage by the SFT-20 model. Red = More signal leakage by the SFT-20 GRQO-10 model. The top row represents the best patch denoised by the GRQO trained model, while the bottom row represents the worst patch denoised by the GRQO model.

model in the same patch, while in ideal cases only removing noise as much as the SFT-only model.

Answering the research question. GRQO cannot fully replace TTA: the SFT + GRQO variants do not reach the denoising quality of SFT + TTA in any of the training configurations. However, GRQO substantially reduces the optimal TTA budget: the SFT-20 GRQO-10 TTA-25 configuration, outperforms SFT-50 TTA-50, and performs similar to its TTA-50 counterpart, effectively doubling its throughput. Even though MS-SSIM improves only marginally, the primary benefit of GRQO is improved signal retention (lower MS-SSIM-R) at near-zero deployment overhead. An example of the signal retention improvement GRQO provides can be seen in Figure 2.

However, GRQO only shows positive results only on domains that are similar to the ones it was trained on. When evaluated on real seismic surveys, the GRQO model's performance marginally degenerates with just the SFT+GRQO-10 variant: a 0.015 increase in LS for the SFT-20 variant and a 0.008 increase for the SFT-50 variant. After the TTA phase, the gap becomes smaller: the SFT-20 variant LS increases by 0.023 while the SFT-50 variant increases by 0.004. The greatest change in the LS metric is equivalent to 6.78% of the SFT-only variant's LS. Even though the change is minimal, it might still mean that some faint signals might get removed.

Discussing the real data performance. The reason GRQO does not replace TTA is explained by the encoder's latent geometry. DINOv3 embeddings (CLS and mean-pooled

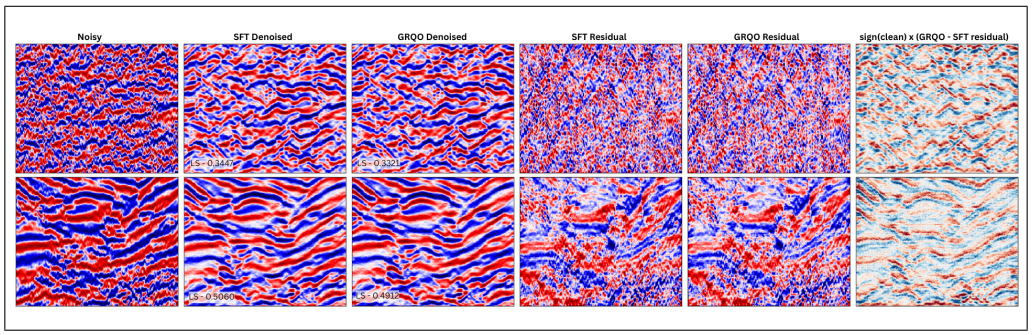

Figure 3: The DINOv3 backbone with SFT-20 against SFT-20 GRQO-10 on the SFM real-world dataset. The rightmost plots compare the signal leakage difference between the two configurations. Blue = More signal leakage by the SFT-20 model. Red = More signal leakage by the SFT-20 GRQO-10 model. The top row represents the best patch denoised by the GRQO trained model; the bottom row represents the worst patch denoised by the GRQO model (relative to LS). It does not seem like a model consistently performs worse, but they each perform better in some scenarios

patch tokens), projected with UMAP (Figure 4), place the synthetic training and test patches on a single compact island, confirming intact in-distribution generalization. The two out-of-domain datasets are distributed differently: the real-world SFM data forms an isolated cluster, and the ThinkOnward dataset is fragmented across thin strands, not in a singular island. An in-distribution versus out-of-distribution probe separates the two with AUC = 1.000, and a kurtosis axis organizes the domains: ThinkOnward spans low and high kurtosis, the real-world cluster is confined to low kurtosis, and the training patches show almost no kurtosis spread.

GRQO does not make the clusters more similar however. With the encoder frozen, the reward and decoder heads optimized only on in-distribution patches, the clusters and the AUC probe are indistinguishable from the pre-GRQO baseline. Separability, however, does not imply poor denoising: ThinkOnward is cleanly separable yet denoises well, because as synthetic data it preserves the high-kurtosis, non-Gaussian signal-and-noise structure the heads were trained on. The real-world SFM data, confined to low kurtosis and differing in its noise statistics, lacks that shared structure — and this is where GRQO alone falls short.

Therefore if an out-of-distribution domain still shares the training signal structure, as ThinkOnward does, GRQO is able to generalize and improve the results; where it does not, as with real-world data, closing the gap requires in-domain adaptation. TTA supplies this by fitting the LoRA weights to the target domain's own high-kurtosis patches at inference. GRQO can pre-condition those weights and shorten adaptation, but genuine cross-domain generalization still requires exposure to in-domain data.

7 Conclusion

This paper investigated whether GRQO-based pre-deployment fine-tuning can reduce the reliance on TTA for seismic denoising with LoRA-adapted vision foundation models. The

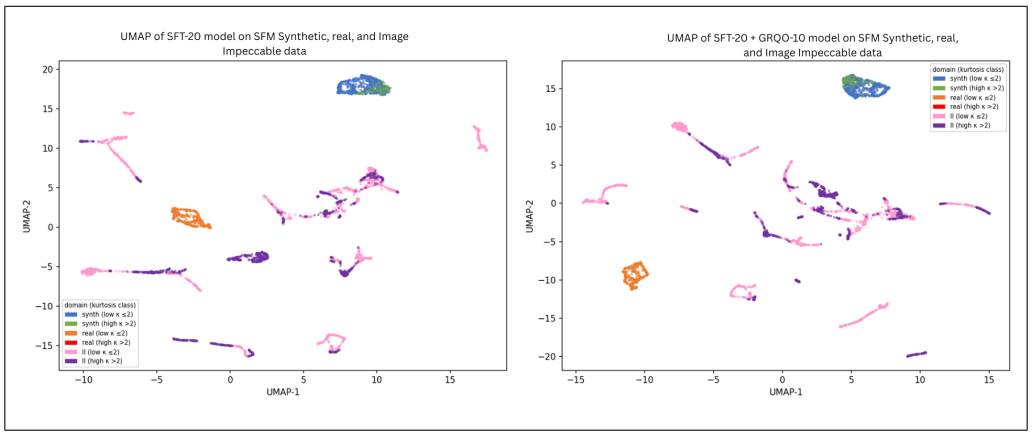

Figure 4: UMAP latent space representations of the DINOv3 embeddings. Left: SFT-20 model. Right: SFT-20 + GRQO-10 model.

results show that GRQO alone meaningfully improves signal retention on familiar domains (MS-SSIM-R: 0.511 vs. 0.540 for the SFT-20 baseline) and exceeds the performance of the SFT-50 upper bound, without any additional TTA at inference time. However it does not improve the model's ability to generalize to real-world domains from synthetic data.

Therefore, GRQO cannot fully eliminate the need for TTA. The best results are achieved by combining GRQO with a reduced 25-epoch TTA phase, reaching MS-SSIM of 0.862 and MS-SSIM-R of 0.452 on the Image Impeccable dataset, while having slightly negative effect on the real-world dataset. Notably, extending TTA beyond 25 epochs after GRQO does not significantly improve performance, suggesting that GRQO pre-conditioning accelerates TTA convergence and that over-adaptation becomes harmful. The key practical finding is therefore that GRQO reduces the optimal TTA budget, enabling a shorter real-time adaptation phase.

Several limitations of this study warrant acknowledgement. First, the GRQO objective combines four loss terms — $\mathcal{L}_{\text{RL}}$, $\mathcal{L}_{\text{all}}$, $\mathcal{L}_{\text{KL}}$, and $\mathcal{L}_{\text{div}}$ — yet no ablation was performed to isolate which components drive the observed improvements most in signal retention; it is therefore unclear to what extent each component contributes to the training loop. Additionally, all experiments train from a single dataset of 1,600 training patches covering one domain; this limits the diversity of noise types and geological settings seen during GRQO training, making it necessary for future research to assess how GRQO performs when trained on diverse synthetic datasets.

Moreover, a more robust model such as the SFM (trained on many different seismic data domains) could provide better results as it is already trained on multiple domains, making the exploration of GRQO on SFM worthwhile. Additionally, the DINOv3 backbone could also be evaluated on other datasets, such as the F3 or Penobscot datasets from different domains to conclusively determine whether GRQO can improve on unseen real-world domains.

8 Responsible Research

This improvement over the training pipeline of VFMs for seismic denoising also carries some ethical considerations with it that must be addressed.

Because this pipeline makes crude oil exploration more efficient, it has the potential to accelerate fossil fuel extraction and worsen climate change and global warming. It must also be highlighted that over-reliance on automated models such as the ones trained with the GRQO pipeline can create safety-risks if, for example, the denoising model suppresses critical earthquake warnings if it hallucinates. It must also be noted that training such large vision models also creates a carbon footprint. Even though a model must be trained only once before put in production, this training pipeline still utilizes TTA, which trains the model even when being used in production. However, the reduced optimal TTA budget for the GRQO models reduces the amount of training required during inference, reducing its carbon footprint.

Moreover, the reproducibility of the experiment is also taken into consideration. Every configuration in my comparison matrix of the results shares the same backbone - DINOv3-ViT-S/16 + U-Net decoder - and the same LoRA placement (q/k/v/o proj, r=16, α=64). Since the optimizer and scheduling remains the same across experiments, the only independent variables are the GRQO and TTA configurations. To ensure reproducibility and correctness, the experiments were repeated across 3 different seeds, and the train/test splits of the data were kept the same. It is also important to note that the displayed metrics are not solely based off of any plotted patches or graphics, but of the entire dataset the configurations are tested on. The training pipeline was adapted from Zhao et al. (2026) and their work, and can be accessed at https://github.com/AndreiOprescu/RP-project.

AI tools, namely Claude Code and Gemini, were also used for producing the code. Specifically, AI was used for code navigation, implementation, debugging support, suggesting ideas, and critiquing plans, cleaning up code and documentation, with thorough human supervision. Additionally, it was also used for checking grammatical errors, creating a template and resolving formatting issues in the written report. Running the experiments, reviewing the code and the numerical results were done by the author, which takes responsibility for the correctness, originality and integrity of the submitted work.

References

Hou, S., & Messud, J. (2021). Machine learning for seismic processing: The path to fulfilling promises. *First International Meeting for Applied Geoscience & Energy*, 3204–3208.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Hudson, T. S., Kettlety, T., Kendall, J.-M., O'Toole, T., Jupe, A., Shail, R. K., & Grand, A. (2024). Seismic node arrays for enhanced understanding and monitoring of geothermal systems. *The Seismic Record*, *4*(3), 161–171. https://doi.org/10.1785/0320240019

Maguire, R., Schmandt, B., Wang, B., Kong, Q., & Sanchez, P. (2024). Generalization of deep-learning models for classification of local distance earthquakes and explosions across various geologic settings. *Seismological Research Letters*, *95*, 2229–2238.

Merrifield, T. P., Griffith, D. P., Zamanian, S. A., Gesbert, S., Sen, S., De La Torre Guzman, J., Potter, R. D., & Kuehl, H. (2022). Synthetic seismic data for training deep learning networks. *Interpretation*, *10*(3), SE31–SE39.

Pan, C., He, W., Tu, Z., & Ren, L. (2025). DINO-R1: Incentivizing reasoning capability in vision foundation models.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Sheng, H., Wu, X., Si, X., Li, J., Zhang, S., & Duan, X. (2025). Seismic foundation model (SFM): A new generation deep-learning model in geophysics. *Geophysics*, *90*, IM59–IM79.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., & Massa, F. (2025). DINOv3.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., & Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. *International Conference on Machine Learning (ICML)*, 9229–9248.

Taha, B. A., Addie, A. J., Haider, A. J., Osman, S. A., Ramli, M. Z., & Arsad, N. (2025). A review of seismic detection using fiber optic distributed acoustic sensing: From telecommunication cables to earthquake sensors. *Natural Hazards*, *121*(12), 13927–13959. https://doi.org/10.1007/s11069-025-07370-5

ThinkOnward. (2024). *Image impeccable*. GitHub. Retrieved June 20, 2026, from https://github.com/thinkonward/challenges/tree/main/geoscience/image-impeccable

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *Asilomar Conference on Signals, Systems and Computers*, 1398–1402.

Yu, S., Ma, J., & Wang, W. (2019). Deep learning for denoising. *Geophysics*, *84*(6), V333–V350. https://doi.org/10.1190/geo2018-0668.1

Zhao, J., bin Waheed, U., Sun, J., Cui, Y., Savva, N., & Verschuur, E. (2026). Parameter-efficient adaptation of pre-trained vision foundation models for active and passive seismic data denoising.

A Supplementary Tables

Table 6: DINOv3-UNet SFT training configuration.

Setting	**Value**
Architecture	
Backbone	DINOv3-ViT-S/16 (frozen, 21.6M params)
LoRA rank / alpha	r=16, α=64 (scaling = 4)
LoRA modules	`q, k, v, o_proj`
Decoder	4-stage U-Net Ronneberger et al., 2015, $(192, 96, 48, 24)$
Trainable params	1.86M / 23.46M (7.95%)
Training	
Epochs / batch size	20 / 24
Optimizer	AdamW, lr $= 6 \times 10^{-4}$, wd = 0.01
LR schedule	3-epoch linear warmup + cosine decay
Loss	$(1-\alpha)\,\mathcal{L}_{\text{MSE}} + \alpha\,(1-\text{MS-SSIM})$, α=0.5
MS-SSIM	5 scales, 11×11 Gaussian window
Input norm.	Per-sample z-score (μ, σ) of noisy patch

B Supplementary Figures

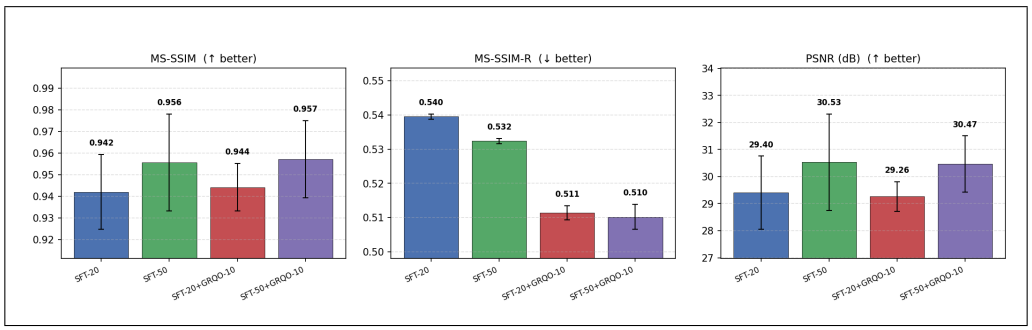

Figure 5: Evaluation metrics on the SFM synthetic dataset.

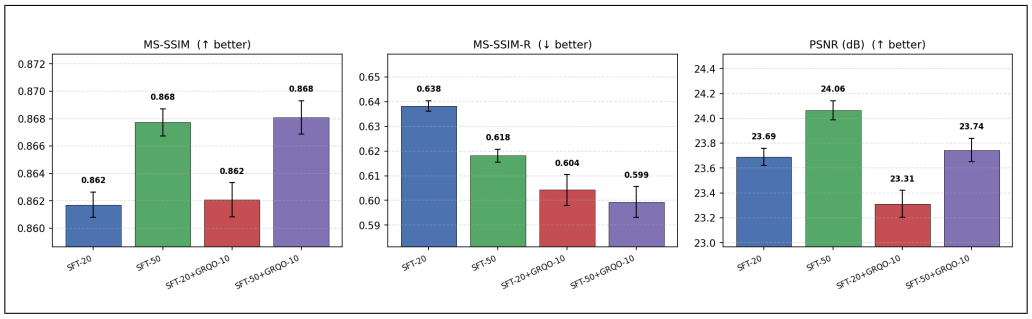

Figure 6: Evaluation metrics on the Image Impeccable dataset.

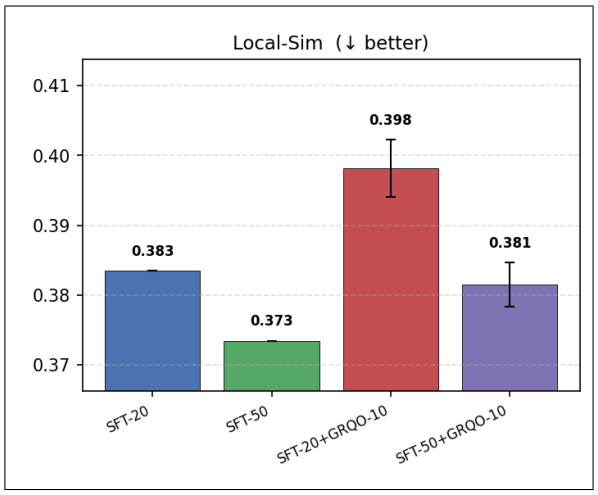

Figure 7: Evaluation metrics on the SFM real world dataset.

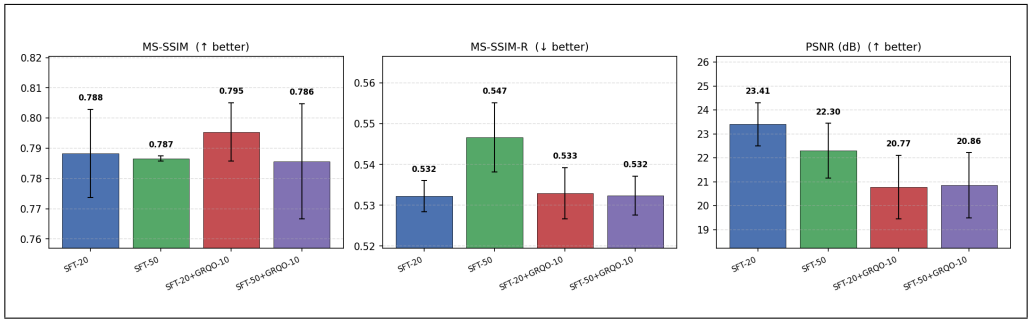

Figure 8: Evaluation metrics on the SFM synthetic dataset, with TTA-25.

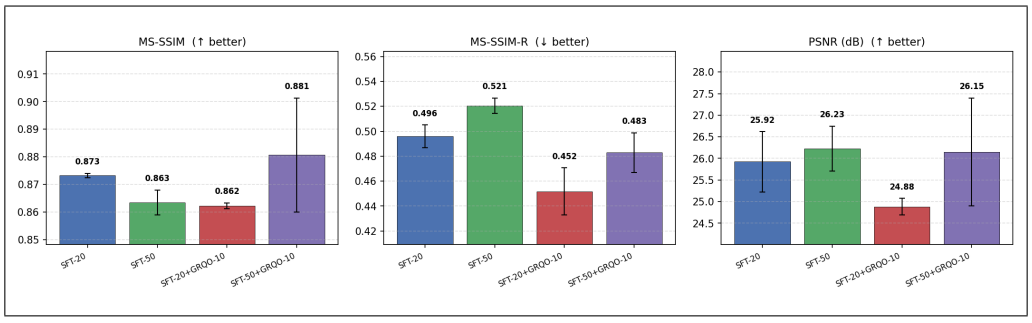

Figure 9: Evaluation metrics on the Image Impeccable dataset, with TTA-25.

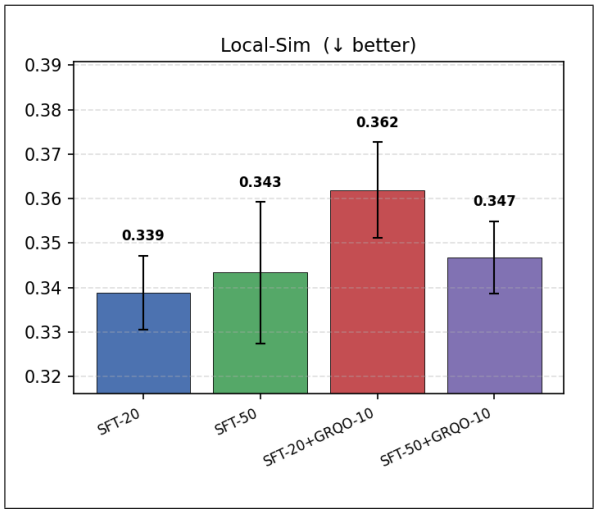

Figure 10: Evaluation metrics on the SFM real world dataset, with TTA-25.

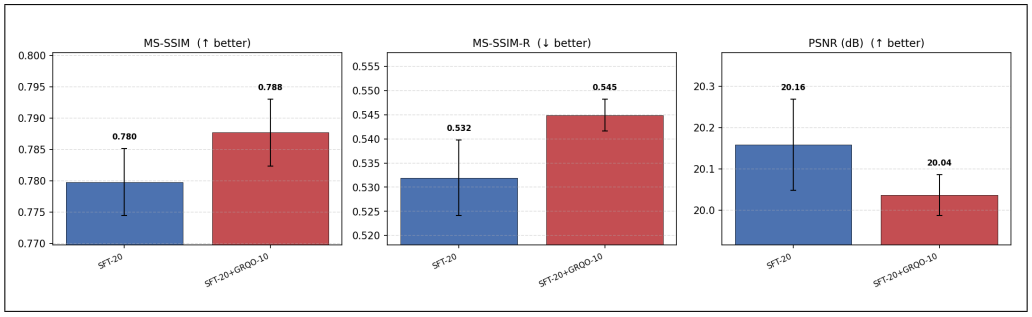

Figure 11: Evaluation metrics on the SFM synthetic dataset, with TTA-50.

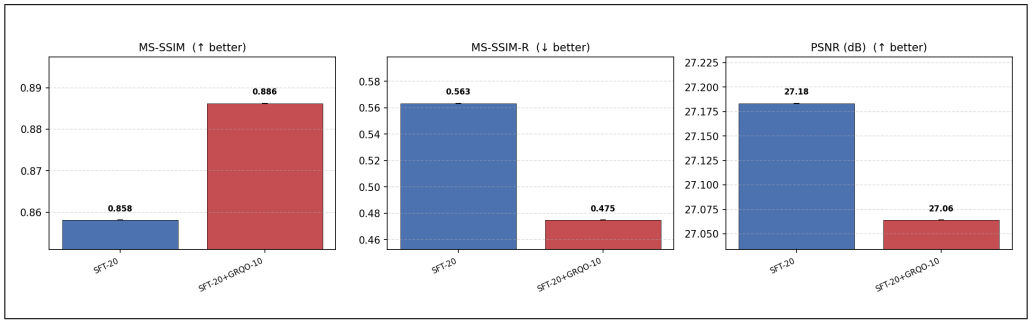

Figure 12: Evaluation metrics on the Image Impeccable synthetic dataset, with TTA-50.

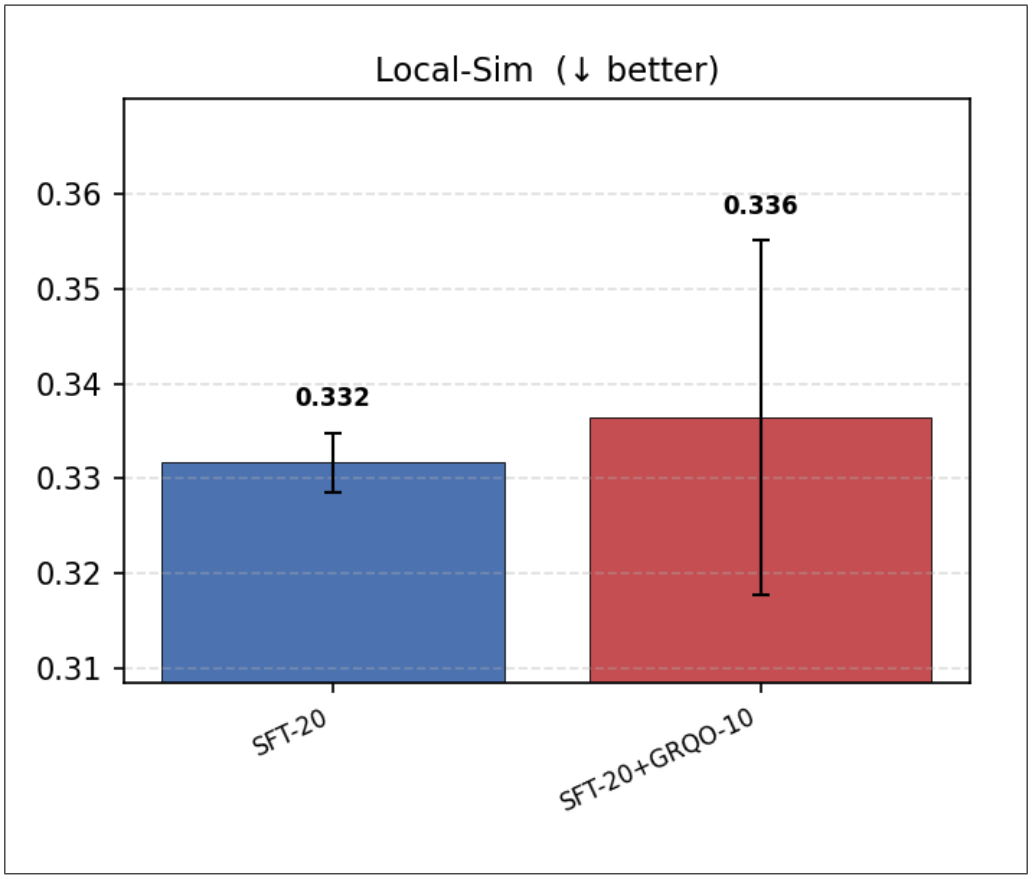

Figure 13: Evaluation metrics on the SFM real world dataset, with TTA-50.