**Exploring continuous and discrete embeddings from pretrained tokenizer in application to seismic data**

## Introduction

Raw prestack seismic data are expensive to acquire and remain one of the most valuable digital assets in the oil and gas industry, yet much of this data is rarely reused beyond processing and imaging cycle. Conventional workflows rely on high-performance computing to iteratively solve inverse problems of geophysics with compute-optimized numerical methods operating on datasets reaching petabyte scale. At the same time, the shear size of raw seismic data creates a structural barrier to the development of large-scale data-driven models, particularly foundation models that require repeated exposure to massive, heterogeneous datasets. While there are established methods for compression of seismic data, their objective is to optimize IO and storage requirements, making the resulting opaque binary streams unsuitable for direct use in neural networks. We explore neural embeddings and classic compression side-by-side to spot the features lost and preserved throughout the encoding-decoding cycle (Figure 1).

Recent advances in computer vision and multimodal learning have demonstrated that compact latent embeddings, rather than raw data, are the key enabler for scalable training and inference (Rombach et al., 2022). In seismic, the investigation for descriptive waveform embeddings and neural compression was studied in Lasscock et al. (2024); Palgunadi et al. (2024); Chen et al. (2025). Foundational models operate not on pixels or waveforms directly, but on representations learned by encoders that aggressively remove redundancy while preserving task-relevant structure (Cheng et al., 2025). Seismic wavefields share several structural properties with other modalities that have benefited from representation learning. Shot gathers and common-image gathers are naturally organized as 2D arrays (time–offset or time–angle) with strong local correlations, while full 3D or 4D seismic cubes resemble video volumes where spatial continuity and temporal causality constrain the evolution of amplitudes. In video models, spatiotemporal encoders explicitly preserve the ordering of frames along time, which is analogous to honoring the causal propagation of seismic events along the time axis. This analogy suggests that general-purpose visual and video tokenizers, if properly adapted, may provide useful low-dimensional embeddings for seismic data without requiring domain-specific architectural changes.

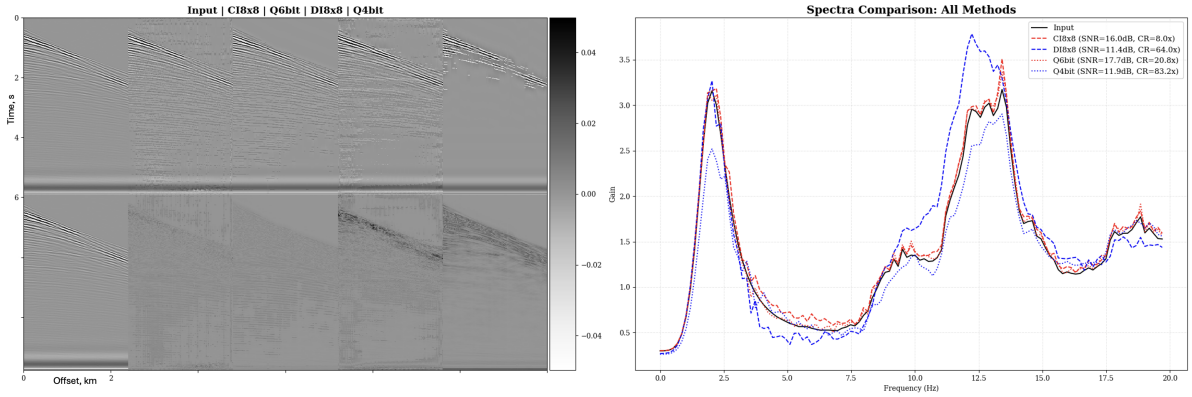


***Figure 1*** *Reconstruction errors by continuous (CI8x8) and discrete Cosmos image tokenizers (DI8x8), and NVCOMP compression on the data quantized to 4 and 6 bit (left). Respective power spectra (right).*

This work explores the capability of unmodified pretrained Cosmos Tokenizer (Agarwal et al., 2025), originally proposed for generic visual data, to produce meaningful embeddings for 2D seismic shot gathers. The tokenization in the context of foundational models implies producing vector-space embeddings suitable for either diffusion model training (continuous tokenization), or for autoregressive models with fixed-size vocabulary (discrete tokenization). We focus on understanding the mapping from time-offset domain to the structured latent space of continuous and discrete embeddings, and assessing the compression ratio trade-offs against reconstruction quality, without any seismic-specific fine-tuning. Building on this analysis, the study outlines a path toward autoencoder-style fine-tuning on large prestack archives, with the goal of reusing long-idle seismic assets to train seismic foundational models using robust, reusable embeddings.

### Latent Space Embeddings for Seismic Data

Latent embeddings can be broadly categorized into continuous and discrete representations. Continuous embeddings encode input data into dense real-valued tensors and are commonly used in diffusion and regression-based models. Discrete embeddings, by contrast, quantize latent representations into a finite vocabulary of codes (vocabulary of 64,000 for Cosmos), enabling autoregressive and token-based modeling paradigms such as training transformer models. In the context of seismic data, continuous embeddings are expected to better preserve smooth amplitude variations and relative energy content, whereas discrete embeddings offer significantly higher compression ratios at the cost of quantization artifacts. Importantly, both representations differ fundamentally from classical compression outputs: they remain structured tensors that can be consumed directly by downstream neural networks.

The Cosmos Tokenizer (Agarwal et al., 2025) employs a lightweight encoder–decoder architecture with explicit temporal causality, originally designed for joint image and video tokenization. Inputs are first transformed using a multi-level wavelet decomposition, reducing spatial and temporal redundancy while emphasizing semantically meaningful components. The encoder then applies a sequence of causal spatio-temporal convolutions and attention layers, producing a compact latent tensor with configurable spatial and temporal downsampling factors. Continuous variants output real-valued embeddings of fixed channel dimension, while discrete variants apply finite-scalar quantization, FSQ, (Mentzer et al., 2023) to map latents into a finite codebook. The decoder mirrors the encoder, reconstructing the input signal from the latent representation. Although trained on natural images and videos, the architecture's reliance on wavelets, locality, and causality makes it well aligned with the physics of seismic wave propagation. Training follows an autoencoder-style objective focused on reconstructing high-resolution images and long videos from their tokens. Unlike standard transformers where embeddings are learned "on the fly" for a single model, the Cosmos tokenizer is trained once, frozen, and reused across tasks and architectures, so representation learning is explicitly separated from downstream modeling which makes the approach suitable for offline preprocessing of large raw seismic datasets.

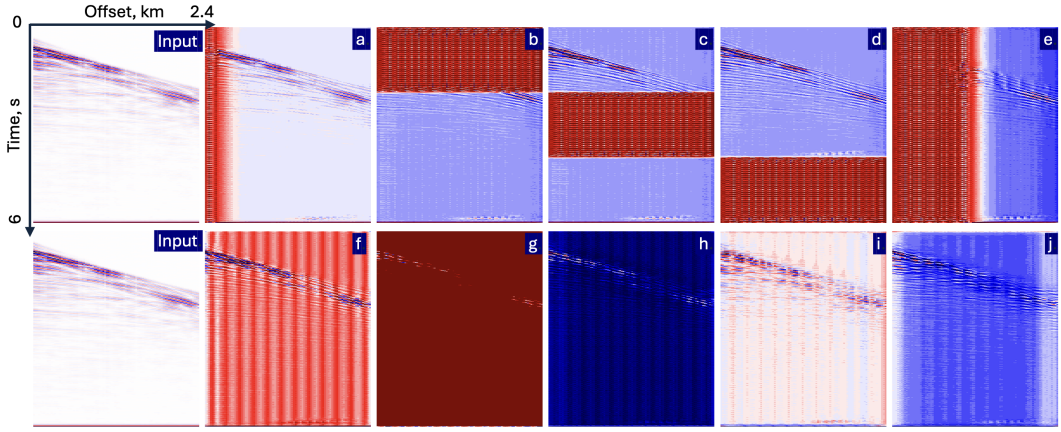


***Figure 2*** *The impact of zero-masking in the latent space of continuous embeddings. Spatial masking in the latent space directly projects into the data domain (a-e). Channel muting in impacts resolutions similarly to wavelet transformation.*

### Results

We conduct experiments on the Mobil AVO Viking Graben Line 12 open dataset (Keys and Foster, 1998). The survey consists of shot gathers acquired in the North Sea, with a sampling interval of 4 ms and 120 spatial traces with a receiver spacing of 25 m. We trim each input gather to the nearest multiple of 32 for efficient inference and low-pass filter the data below 20 Hz. The input data to encoder thus measures $(96, 1492)$ for offset and temporal dimension respectively. Each shot gather is treated as a single 2D input frame in the time–offset domain. We evaluate pretrained Cosmos image tokenizers without any seismic-specific fine-tuning, using both continuous (CI) and discrete (DI) variants with spatial strides of $8 \times 8$ and $16 \times 16$. The resulting latent representations are decoded back to the data domain and compared against the original gathers.

We start with understanding the impact of masking in the latent space of continuous tokenizer which measures $(16, 12, 184)$ encoding channels, and spatial dimensions. Figure 2 suggests direct spatial relation between the latent space of continuous tokenizer and decoded seismic shot gather. When a single-element trace is being mutted in the latent space, this results in a corrupted area roughly corresponding to the width of a perception field of convolutional kernels in the decoder. In other experiments, where the larger parts of the latent space are set to zero, the direct spatial relation with data domain remains prominent, while near-zero frequency artifacts emerge.

Since classic compression methods and neural latent-space representations are serving different purposes, we only compare them within compression ratio discussion. As baselines, we include NVCOMP lossless codecs and uniform quantization at 4–16 bits per sample. This setup allows a direct comparison between classical entropy/quantization-based compression and learned latent-space embeddings. Reconstruction quality is assessed using signal, structural, and physics-aware metrics summarized in table in Figure 3. Signal-level fidelity is measured by signal-to-noise ratio(SNR) and PSNR, while structural preservation is quantified via structural similarity(SSIM) and linear correlation. Error-based measures include NRMS and MAE. In addition, frequency-weighted amplitude and phase ($\times 1e3$) difference metrics emphasize low-frequency content, which is critical for seismic interpretation and inversion.

| Method | Type | Ratio | Size% | SNR | PSNR | Amp | Phase | SSIM | Corr | NRMS | MAE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cosmos-Tokenizer-DI16x16 | Lossy | 256.00x | 0.4% | 5.0 | 33.5 | 0.1868 | 0.1697 | 0.8506 | 0.8411 | 0.0177 | 0.00472 |
| Cosmos-Tokenizer-CI16x16 | Lossy | 32.00x | 3.1% | 9.6 | 38.1 | 0.1250 | 0.0934 | 0.9536 | 0.9643 | 0.0108 | 0.00319 |
| Cosmos-Tokenizer-DI8x8 | Lossy | 64.00x | 1.6% | 11.4 | 39.9 | 0.0718 | 0.1084 | 0.9454 | 0.9657 | 0.0087 | 0.00273 |
| NVCOMP Quant-4bit | Lossy | 83.23x | 1.2% | 11.9 | 40.4 | 0.0645 | 0.0778 | 0.9397 | 0.9681 | 0.0083 | 0.00225 |
| Cosmos-Tokenizer-CI8x8 | Lossy | 8.00x | 12.5% | 16.0 | 44.5 | 0.0193 | 0.0343 | 0.9731 | 0.9879 | 0.0052 | 0.00200 |
| NVCOMP Quant-6bit | Lossy | 20.79x | 4.8% | 17.7 | 46.2 | 0.0050 | 0.0427 | 0.9742 | 0.9918 | 0.0042 | 0.00152 |
| NVCOMP Quant-8bit | Lossy | 10.79x | 9.3% | 29.4 | 57.9 | 0.0069 | 0.0049 | 0.9987 | 0.9994 | 0.0011 | 0.00045 |
| NVCOMP Quant-10bit | Lossy | 5.95x | 16.8% | 40.9 | 69.4 | 0.0031 | 0.0014 | 0.9999 | 1.0000 | 0.0003 | 0.00013 |
| NVCOMP Quant-12bit | Lossy | 3.73x | 26.8% | 52.1 | 80.8 | 0.0008 | 0.0004 | 1.0000 | 1.0000 | 0.0001 | 0.00004 |
| NVCOMP Quant-16bit | Lossy | 1.94x | 51.6% | 64.4 | 104.9 | 0.0001 | 0.0001 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP LZ4 | Lossless | 1.00x | 100.4% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP Snappy | Lossless | 0.99x | 100.8% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP GDeflate | Lossless | 1.05x | 95.1% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP Deflate | Lossless | 1.06x | 94.8% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP ANS | Lossless | 1.01x | 99.0% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |
| NVCOMP Zstd | Lossless | 1.06x | 94.7% | 64.7 | 180.7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 | 0.00000 |

***Figure 3*** *Quantitative evaluation of seismic data compression. A comparison between pretrained neural tokenizers (Cosmos DI/CI) and standard NVCOMP compression techniques across reconstruction quality and structural similarity metrics.*

Lossless NVCOMP methods achieve perfect reconstruction, as expected, but offer limited compression ratios close to unity. Quantized NVCOMP baselines provide a smooth quality–compression trade-off: the most compact 4-bit quantization reaches a compression ratio of approximately $83\times$ but exhibits reduced SNR and SSIM. Higher-bit configurations progressively recover near-perfect fidelity at the expense of reduced compression efficiency (Figure 4).

Discrete embeddings with aggressive downsampling, DI16$\times$16, achieve compression ratios up to $256\times$, reducing storage to less than 0.5% of the original volume. Meanwhile, high-frequency components, SNR and correlation decrease substantially, yet the reconstructions retain coherent low-frequency structure. At more moderate settings, DI8$\times$8, discrete embeddings provide compression ratio of $64\times$ while maintaining SSIM above and correlation on par with classic 4-bit compression. Continuous embeddings exhibit a more favorable balance between fidelity and compression. The CI8$\times$8 configuration achieves $8\times$ compression comparable to NVCOMP 6-bit quantization.

The observed behavior reflects the fundamental design of the two embedding types. Discrete embeddings prioritize token compactness and semantic abstraction, aggressively removing high-frequency detail while preserving large-scale, low-frequency structure, which makes them well suited for extreme compression, indexing, retrieval, and autoregressive or transformer-based foundation models. Continuous embeddings retain richer amplitude and phase information at the cost of lower compression, making them more appropriate for tasks requiring waveform fidelity, such as reconstruction, conditioning, inversion, or diffusion-based generative modeling.

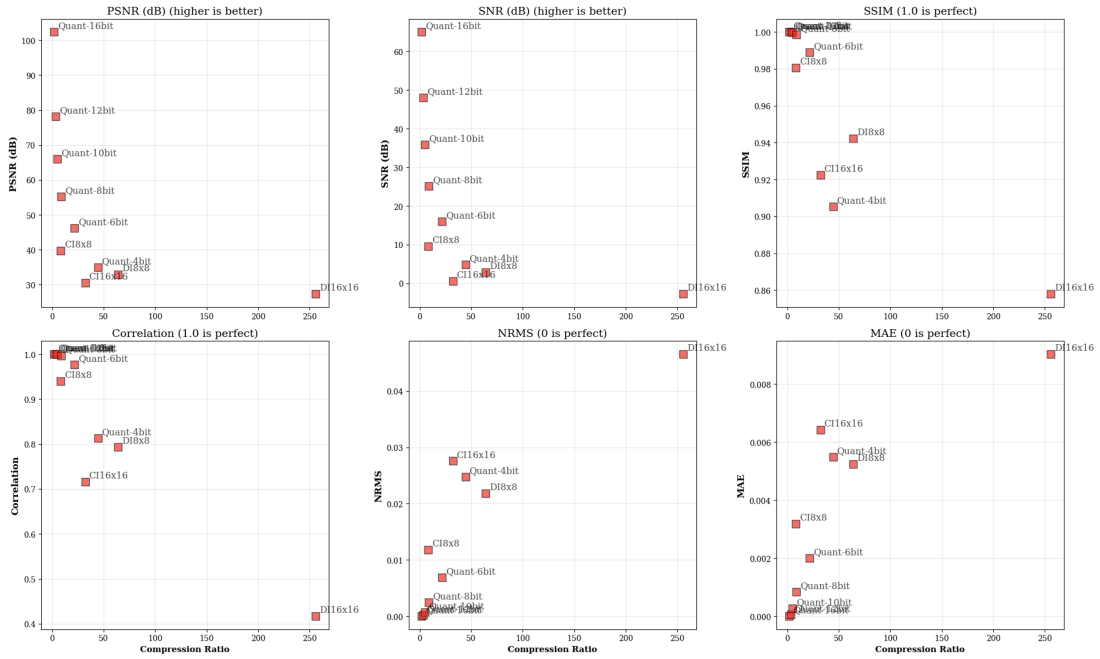


***Figure 4*** *Tradeoff between compression ratio and reconstruction accuracy across various metrics for neural representations and deterministic compression by NVComp with quantization.*

## Conclusions

The ability to compress raw prestack seismic data into compact, information-preserving embeddings opens a pathway toward scalable seismic foundational models. Despite being trained exclusively on natural images and videos, the pretrained Cosmos tokenizers outperform initial expectations when applied to seismic data, preserving coherent low-frequency structure and large-scale wavefield organization even at high compression ratios. This behavior is likely explained by strong local correlations and scale separation captured by wavelet decompositions enable the tokenizer to exploit redundancy patterns that are modality-agnostic rather than domain-specific.

## References

Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y. et al. [2025] Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.

Chen, Y., Saad, O.M., Chen, Y. and Savvaidis, A. [2025] Deep learning for seismic data compression in distributed acoustic sensing. *IEEE Transactions on Geoscience and Remote Sensing*.

Cheng, S., Harsuko, R. and Alkhalifah, T. [2025] A generative foundation model for an all-in-one seismic processing framework. *arXiv preprint arXiv:2502.01111*.

Keys, R.G. and Foster, D.J. [1998] *Comparison of seismic inversion methods on a single real data set*. Society of Exploration Geophysicists.

Lasscock, B., Sansal, A. and Valenciano, A. [2024] Encoding the subsurface in 3D with seismic. In: *Fourth International Meeting for Applied Geoscience & Energy*. Society of Exploration Geophysicists and American Association of Petroleum . . . , 617–621.

Mentzer, F., Minnen, D., Agustsson, E. and Tschannen, M. [2023] Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Palgunadi, K.H., Bergmeister, A., Bosisio, A., Ermert, L., Koroni, M., Perraudin, N., Dirmeier, S. and Meier, M.A. [2024] High resolution seismic waveform generation using denoising diffusion. *arXiv preprint arXiv:2410.19343*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. [2022] High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.