A Framework for Compressive On-Chip Action Potential Recording

Pumiao Yan, *Graduate Student Member, IEEE*, Dante G. Muratore, *Senior Member, IEEE*, E.J. Chichilnisky, Boris Murmann, *Fellow, IEEE*, and Tsachy Weissman, *Fellow, IEEE*

***Abstract*—Scaling neural recording systems to thousands of channels creates extreme bandwidth demands, posing a challenge for resource-constrained, implantable devices. This work introduces an adaptive, multi-stage compression framework for high-bandwidth neural interfaces. The system combines a Wired-OR analog-to-digital compressive readout with a digital core that adaptively requantizes, selectively samples, and encodes the neural signals. Although prior work suggests that action potential recordings can be re-quantized to approximately the signal-to-noise (SNR) number of bits without significantly degrading decoding performance, our results show that the required resolution can often be reduced even further. By matching the number of quantization levels to the electrode's maximum SNR ($\lceil \log_2 \text{SNR} \rceil$ number of bits), we retain waveform fidelity while eliminating unnecessary precision that primarily captures noise. Recorded spike samples are selected using a mutual information-based criterion to preserve both spatial and temporal discriminative waveform features. A static entropy coder completes the pipeline with low computation overhead compression optimized for neural signal statistics. Evaluated on 512-channel macaque retina *ex vivo* data, the system preserves 90% of spikes while achieving a 1098× total compression over baseline.**

***Index Terms*—Analog-to-digital compression, brain-machine interfaces, compression algorithm, neural interfaces, A/D conversion.**

I. Introduction

HIGH-DENSITY neural interfaces capable of recording from thousands of neurons at single-cell resolution are transforming both neuroscience and clinical neurotechnology [1]. These interfaces provide access to fine-grained activity across large neural populations, which offer unprecedented insight into complex interactions between neurons and their cooperative behaviors [2], [3], [4], [5], [6], [7]. Applications of neural interfaces range from basic studies of sensory processing to brain-machine interfaces for motor rehabilitation, vision restoration, and sensory augmentation [8], [9]. Meeting the demands of long-term, high-resolution *in vivo* recording requires systems that have many simultaneous recording channels and high temporal resolution, and must operate fully wirelessly for stable chronic use.

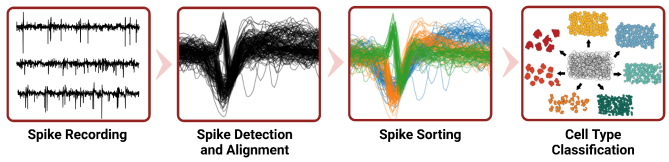

Fig. 1. Action potential signal processing pipeline.

To support these applications, wired high-density microelectrode arrays (MEAs) with thousands of electrodes have been developed in research, enabling large-scale neural recordings with fine spatial and temporal resolution [10], [11]. However, existing fully implantable systems remain limited to roughly a thousand simultaneous channels [12] and are unable to preserve spike waveform information. As channel counts scale up, power consumption, silicon area, and data transmission requirements grow proportionally. Processing and transmitting this volume of data within the power and area constraints of a wireless implant becomes increasingly difficult. This motivates the need for hardware-friendly and power-efficient compression that reduces data through the data acquisition pipeline, without discarding critical spike-level information needed for downstream analysis.

In neural recordings, most of the information relevant to downstream processing is contained in the shape and timing of action potentials [13], [14]. These features enable essential tasks such as spike detection, spike sorting, and cell-type classification (the spike processing pipeline of such tasks is shown in Fig. 1). Spike detection identifies the timing and location of events, sorting clusters them by neuronal source, and cell-type classification associates units with specific cell types based on waveform and firing statistics. While the information required for each task varies (e.g., shape, amplitude, or timing), waveform shape remains a key feature, especially for more advanced analyses such as cell-type classification [15], [16], [17]. As a result, the structure of the spike waveform directly influences which

Received 1 May 2025; revised 5 August 2025; accepted 18 September 2025. Date of publication 29 September 2025; date of current version 29 April 2026. This project was supported in part by Stanford's Wu Tsai Neurosciences Institute. The work of Pumiao Yan was supported by a Stanford Bio-X SIGF fellowship. *(Corresponding author: Pumiao Yan.)*

Pumiao Yan is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: pumiaoy@stanford.edu).

Dante G. Muratore is with the Microelectronics Department, Delft University of Technology, The Netherlands.

E.J. Chichilnisky is with the Department of Neurosurgery and Ophthalmology, Stanford University, USA, and also with the Hansen Experimental Physics Laboratory, Stanford University, USA.

Boris Murmann is with the Department of Electrical and Computer Engineering, University of Hawaii, USA.

Tsachy Weissman is with the Department of Electrical Engineering, Stanford University, USA.

Digital Object Identifier 10.1109/TBME.2025.3615514

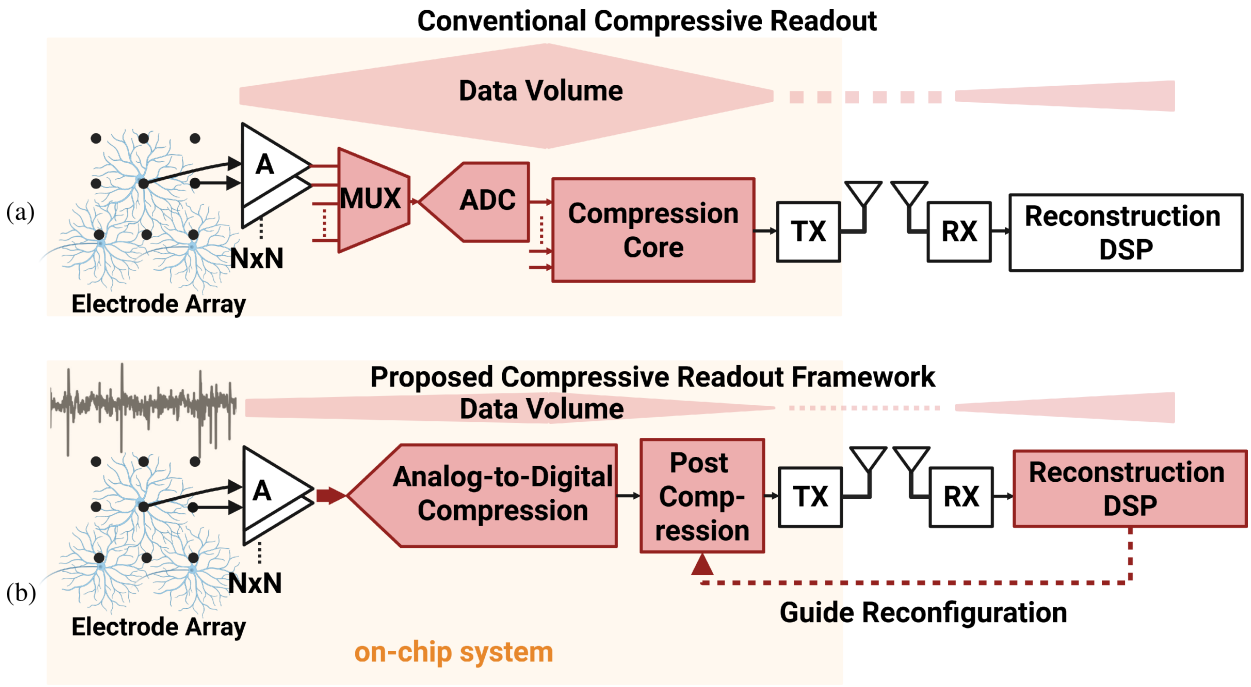

Fig. 2. (a) Conventional and (b) proposed on-chip compressive readout framework.

features are retained for downstream interpretation and which can be disregarded as noise or redundancy.

Conventional compression techniques typically rely on thresholding or temporal sub-sampling to reduce data volume [18], [19], [20], [21], [22], [23]. While thresholding can be effective for simple decoding tasks, it discards waveform details needed for sorting and classification. Setting optimal thresholds also requires channel-specific tuning, which increases system complexity and power usage. Other approaches propose on-chip spike sorting [24], [25], [26] or compression following digitization [27], [28], [29], but digitizing full-resolution signals from thousands of channels remains a major bottleneck.

To avoid overwhelming bandwidth at any stage, compression must begin as close as possible to the analog front-end [30]. Ideally, this compression would be adaptive—preserving the relevant spike information for a given task while discarding non-informative baseline activity. As shown in Fig. 2, this requires rethinking the signal processing pipeline to integrate lossy compression at the analog-to-digital (A/D) interface.

In the conventional system architecture (Fig. 2(a)), signals from a multiplexed electrode array are first digitized in full by an ADC before undergoing any compression. This design imposes a heavy bandwidth and power burden at the front-end and requires ADCs that are capable of handling continuous high-rate data, including non-informative baseline samples. Compression occurs only after digitization, requiring more processing power and bandwidth.

Recent efforts have explored embedding spike sorting in hardware to enable low-latency, on-chip processing. For example, our prior work [31] presented a 1024-channel spike-sorting chip using event-driven detection and self-organizing map clustering, while Chen et al. [32] proposed an unsupervised geometry-aware sorting architecture. These approaches show promise, but typically focus on either spatial or temporal features, not both, and remain limited by power constraints.

In this work, we present a compressive readout framework that builds on previous analog-to-digital compression architectures [33], [34] and incorporates adaptive digital processing tailored to neural signals. As shown in Fig. 2(b), the system features a multi-stage pipeline that combines a Wired-OR analog-to-digital readout with an on-chip digital core that adaptively re-quantizes, selectively samples, and encodes spike waveforms. These operations are configured by an external module using signal statistics and waveform templates. Using this calibration-driven approach, the hardware applies compression strategies optimized for both data rate and signal fidelity.

We evaluate this framework through emulation on 512-channel *ex vivo* primate retina recordings. Results show that our system can preserve 90% of spike events – an improvement of 8–10% over state-of-the-art methods [31] – while achieving a 1098× total compression ratio over baseline recordings. These findings demonstrate that signal-aware, multi-stage compression can achieve substantial data reduction without sacrificing downstream performance.

This paper extends our prior work on Wired-OR compression by introducing and analyzing the full adaptive pipeline. Specifically:

- We develop a re-quantization strategy based on electrode SNR, eliminating excess quantization levels.
- We propose a mutual information–based selective sampling method that preserves both spatial and temporal waveform features.
- We implement a lightweight entropy coding scheme for efficient final-stage compression.

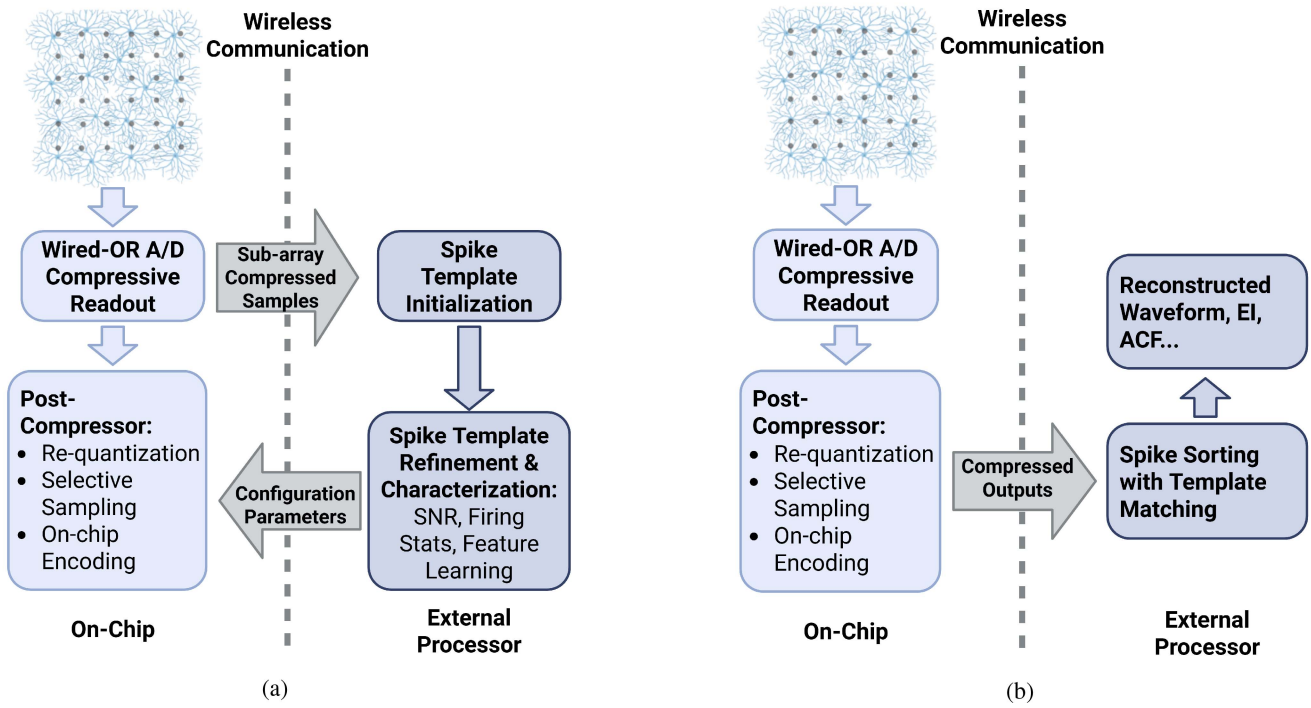

Fig. 3. On-chip compressive readout framework. (a) Calibration Phase. (b) Compression Phase.

- We evaluate the pipeline on real 512-channel recordings and analyze performance across spike detection and sorting tasks.

The rest of this paper is organized as follows. Section II introduces the multi-stage compression framework, including the Wired-OR analog-to-digital readout and the adaptive digital post compression core. Section III presents simulation and experimental results using 512-channel neural recordings. Section IV discusses hardware implications and comparisons with existing methods. Section V concludes the paper.

II. Multi-Stage Compression Framework

We now present the implementation details of the proposed compressive readout framework, focusing on how compression is distributed across analog and digital domains to minimize bandwidth and power. The proposed system operates in two main phases: calibration and compression. During the Calibration Phase (Fig. 3(a)), a representative subset of recorded neural signals is analyzed offline to extract key statistical and structural features. Characteristic spike waveforms are identified, and relevant metrics such as SNR, firing rates, and salient waveform features are estimated. These metrics guide the selection of optimal compression parameters, which are loaded into the on-chip digital compression core. Tailoring the system parameters to the underlying neural activity in this way ensures that compression decisions, such as quantization levels or sampling intervals, are application-aware and reflective of the unique signal properties observed during calibration.

Once the calibration parameters have been established, the system transitions into the Compression Phase (Fig. 3(b)). Neural signals are continuously acquired, and the first level of compression occurs immediately at the analog front-end through the Wired-OR A/D compressive readout. This mixed-signal architecture exploits the sparsity of neural activity: spike-related voltages, which deviate sharply from baseline and are less frequent, are more likely to be uniquely encoded, while baseline-level voltages, being more frequent, are prone to collision in the row and column readout and naturally suppressed. As a result, the data volume is substantially reduced while preserving nearly all spike-relevant samples, alleviating pressure on downstream processing. Following this mixed-signal compression, a digital post-compression core refines the data further through re-quantization, selective sampling of spike waveforms, and low-overhead on-chip encoding. This multi-stage strategy ensures that each subsequent compression step operates on a smaller data stream, thereby reducing both power consumption and bandwidth requirements without sacrificing essential spike information.

A. Wired-OR Analog-to-Digital Compressive Readout

The Wired-OR analog-to-digital (A/D) compressive readout [33], [34], [35] (as shown in Fig. 2(b)) serves as the first stage of the proposed multi-stage compression framework, directly reducing data at the A/D interface. This hardware-efficient and event-driven approach acquires neural data while simultaneously performing data compression and channel multiplexing. Each pixel samples the input voltage and encodes it into a pulse position based on a globally distributed ramp. Pulses from multiple pixels are merged on shared wires using Wired-OR circuitry. If only one pixel fires at a given time step, both its

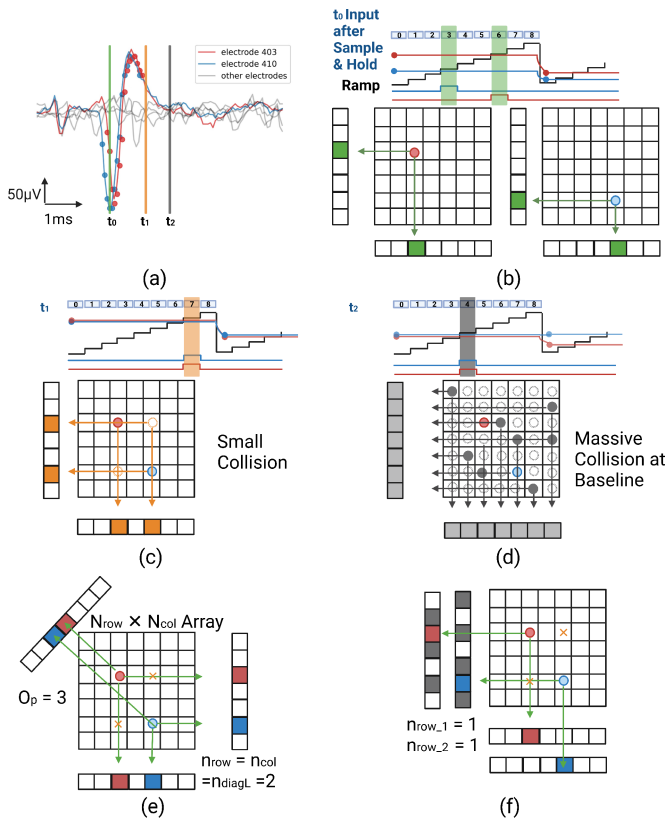

Fig. 4. Wired-OR readout concept [34]. (a) A snippet of action potential waveform seen on different electrodes. (b) Conversion of voltage to pulse position and collision-free readout of one pixel. (c) A collision between two pixels. (d) Massive collision across the array at the baseline level. (e) Diagonal wiring conceptual drawing. (f) Interleaving wiring.

location and voltage level can be uniquely decoded (as shown in Fig. 4(b)). If multiple pulses coincide—i.e., a collision occurs—the data is discarded, achieving compression by exploiting signal sparsity (see Fig. 4(c)-(d). This technique exploits the sparsity and diversity of neural signals to minimize redundant sampling and digitization while maintaining the key features required for downstream analysis.

Unlike conventional A/D architectures that sample continuously across all channels, the Wired-OR readout refrains mostly spike-related events, significantly reducing data rates and computational overhead. This spike-driven encoding process ensures that only the most relevant neural activity is retained, making the system highly scalable and power-efficient for brain-machine interface applications. Due to their temporal and spatial sparsity, spike-related samples are more likely to be unique, they are retained, while baseline-level voltages—which are more frequent and thus prone to collision—are naturally filtered out. Wiring strategies such as diagonal [34] and interleaved [33] layouts can help resolve small collisions and recover more spike information, also shown in Fig. 4(e)–(f).

By pre-compressing neural data at the analog interface, the Wired-OR readout reduces the data bandwidth burden on subsequent processing stages by $15 \sim 50\times$, depending on the wiring strategy and neural firing rate. This efficiency was validated in ex vivo experiments using a taped-out chip, where compression rates ranged from $111.2\times$ (single-wire) to $38.8\times$ (four-wire) configurations [35]. The digital compression core, detailed in Section II-B, further refines the signal through re-quantization, selective sampling, and on-chip encoding, ensuring optimal balance between compression efficiency and neural signal fidelity.

B. Digital Compression Core

The output of the Wired-OR readout stage is a sparse set of events, each represented by its row and column address along with the amplitude (encoded in ADC counts). These elements - spatial address and amplitude - form the input to the post compression core, which aims to further reduce data size by keeping critical information available for downstream analyses. The digital core first treats address and amplitude data as separate streams, optimizing compression strategies tailored to each. For the amplitude, we explore re-quantization to a lower resolution, leveraging the insight that many spikes can be accurately reconstructed at a coarser bit depth. In addition, selective sampling is applied to discard redundant or low-information spike samples. The remaining critical samples are then passed through a lightweight entropy coding stage inspired by Huffman coding. Instead of performing full encoding in hardware, we use a precomputed lookup table that approximates optimal codes based on amplitude and spatial statistics gathered during the calibration phase. This hybrid scheme achieves high compression ratios while maintaining hardware simplicity and preserving spike timing and shape characteristics essential for downstream decoding and sorting.

1) Re-Quantization: State-of-the-art neural interface systems are equipped with ADCs of 10-16 b resolution [10], [12], [36], [37], [38], [39] driven by multiple considerations such as signal drift, motion artifacts, and dynamic range. Although prior work [19] suggests that action potential data can be re-quantized to about the SNR number of bits without significantly degrading decoding performance, our results indicate that the required re-quantization resolution can often be reduced even further. In an electrode of a MEA, the measured voltage $V(z,t)$ can be modeled as [40]:

$$V(z,t) = \sum_k A_{\sigma(k)}\big(z,\, t - t_k\big) \cdot x(k) \;+\; N(t). \qquad (1)$$

Here, z denotes the electrode location, t is the continuous time, and k indexes distinct neurons or clusters. Each cluster $\sigma(k)$ has an associated waveform template $A_{\sigma(k)}$, shifted by the spike time t_k and scaled by the spike amplitude $x(k)$. The term $N(t)$ captures additive noise present in the recording.

For a spike event, the SNR is approximated by [36]:

$$SNR = \frac{V_{\textit{peak-peak amplitude}}}{V_{\sigma,channel}} \qquad (2)$$

The amplitude of the spike peak-to-peak ($V_{\textit{peak-peak amplitude}}$) is determined by identifying the electrode with the most significant peak-to-peak difference for each spike. The noise level ($V_{\sigma,channel}$) is calculated as the median absolute deviation when no action potential is detected on the electrode.

Intuitively speaking, if we look at the recording of spikes on a certain electrode, the observed voltage level is the summation of

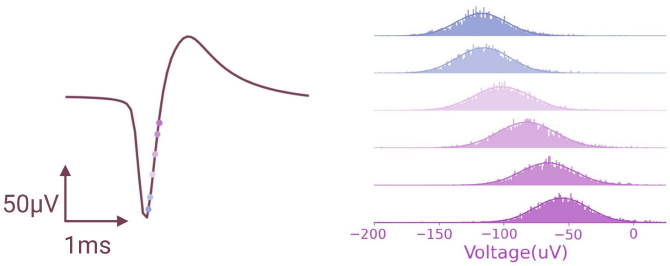

Fig. 5. Empirical demonstration of additive Gaussian noise in extracellular action potential recording. (a) Average spike waveform of a given cell-electrode pair. (b) Real, measured voltage distribution at a specific sample point across multiple spike occurrences.

the underlying noise-free signal and noise that can be modeled as a Gaussian distribution (see Fig. 5). When the electrode records spikes only up to a certain $SNR_{max,electrode}$ (following the definition of (2)), the quantizer only needs to differentiate N_{steps} quantization levels, where:

$$N_{steps} = SNR_{max,electrode} \tag{3}$$

because finer resolution would merely differentiate noise. Hence, the minimum bit depth required is:

$$B_{min} = \lceil \log_2 SNR_{max,electrode} \rceil \tag{4}$$

From a theoretical perspective, this interplay between universal filtering and lossy compression aligns with established results in information theory and signal processing. The foundational works of [41], [42], [43] have characterized noise filtering via data compression and indirect rate-distortion problems, respectively, underscoring the effectiveness of compression-based denoising. Additionally, the principles articulated in [44], [45] reveal that compressing noisy data can inherently facilitate noise reduction when the compression is optimized for a distortion tuned to the level and characteristics of the noise. Practical manifestations of this approach are extensively discussed in prior research [44], [46], reinforcing the viability of compression-driven filtering methods in real-world scenarios.

Further justification for our use of uniform scalar quantization followed by universal filtering arises from both theory and practice. It is well-established that, under squared-error distortion and in the presence of additive Gaussian noise, uniform scalar quantization achieves near-optimal performance [47]. This setting applies directly to neural signals, which are commonly modeled as clean spike waveforms corrupted by zero-mean Gaussian noise (see Fig. 6). Prior work [47] has shown that quantizing the noisy signal using a quantizer designed for the clean source yields distortion nearly matching that of the optimal scheme. Thus, uniform scalar quantization not only simplifies implementation by obviating the need for optimized quantizer thresholds but also incurs minimal performance loss.[1]

[1]Intuitively, the observed voltage at a given electrode is the superposition of the underlying spike signal and Gaussian noise. In such settings, quantization optimized for the clean signal often performs nearly as well as one optimized for the noisy observation. See [47] for theoretical discussion.

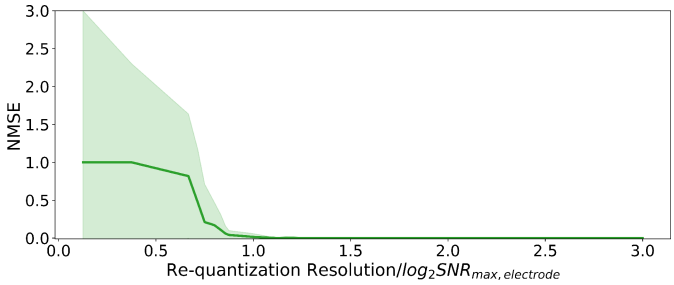

Fig. 6. Normalized mean square error (NMSE) versus re-quantization resolution, where the x-axis is expressed as the ratio of each electrode's chosen bit depth to B_{min}.

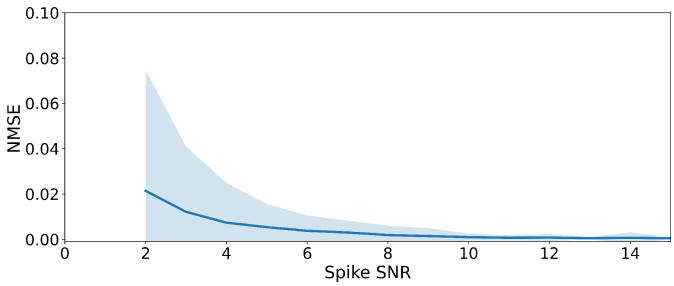

Fig. 7. Normalized Mean Square Error (NMSE) versus spike SNR for re-quantization at B_{min} bits.

An experiment of requantizing the action potential recording from an originally 12-bit quantized dataset is conducted. The requantization resolution for each electrode in three 512-channel primate *ex vivo* retina recording datasets is normalized to the proposed B_{min} defined by (4). As shown in Fig. 6, the green curve shows how NMSE[2] decreases sharply as the bit depth approaches or exceeds B_{min}, indicating that the re-quantization error can be substantially mitigated by allocating just enough bits to match the electrode's maximum SNR-based bits. The shaded region depicts the variability across different channels, illustrating the NMSE is consistently small across electrodes with a requantization resolution of B_{min}.

We also studied the average spike waveform distortion introduced by re-quantizing with B_{min} bits resolution across different cell-electrode pairs where spikes have a wide range of SNR. Fig. 7 shows that for $SNR \geq 5$, less than 1% of NMSE is introduced by re-quantization.

During the calibration phase (Fig. 3(a)), a single sweep across all channels is performed to extract signal statistics used to estimate the SNR and determine B_{min} for each channel. In this mode, full waveforms from each electrode are sampled and digitized without thresholding or compression, enabling accurate baseline noise characterization.

Because SNR can vary over time due to factors such as electrode impedance changes, tissue response, or other nonstationary effects, the system includes a recalibration mechanism. It periodically re-enters calibration mode to update noise estimates

[2]Here the mean square error is normalized to the spike peak to peak amplitude $V_{peak-peak\ amplitude}$.

and recompute B_{min}. The re-calibration interval is empirically determined based on the application, as different recording modalities (e.g., retinal vs. intracortical) exhibit different SNR dynamics. This adaptive calibration ensures that the quantizer remains well-tuned to current signal conditions, preserving both compression efficiency and signal fidelity.

2) Selective Sampling: Following the compressive Wired-OR readout and digital re-quantization, further compression gains can be realized by reducing the number of time samples retained for each spiking event. Rather than transmitting all of the Wired-OR outputs, we identify a small set of time points that carry the most discriminative information across neural units. This selective sampling process eliminates redundant or less-informative samples, preserving only those samples essential for downstream tasks such as spike sorting or neural decoding. Selective sampling is configured during the Calibration Phase, where informative samples are identified off-chip based on statistical differences across spike waveforms. The selected sample indices are then programmed into the on-chip digital core to enable compact spike representation during real-time data streaming.

Our approach is motivated by information-theoretic principles, yet implemented using lightweight operations suited to streaming. The procedure unfolds in three stages: an initial unsupervised clustering, mutual information-based sample ranking, and optional online refinement.

a) Unsupervised Clustering Initialization: We first start with performing an initial unsupervised clustering on the full Wired-OR spike waveforms. Here we use k-means clustering (with k set to a conservative number of 5, empirically chosen). This provides provisional cluster assignments without any feature selection bias.

Given spike waveforms of length 61[3] (each approximated as a true spike shape plus Gaussian noise), we need to find which time sample indices are most informative for distinguishing unknown clusters (typically 3–5 clusters). In an unsupervised scenario (no prior labels), we seek statistically sound, efficient feature-ranking methods that highlight time points contributing most to cluster separation. We focus on global importance across all clusters seen on the electrode (not per-cluster specific) and avoid heavy machine-learning models. Below, we outline the principle of our method. We also discuss their computational efficiency and how they can fit into a real-time feature selection pipeline.

b) Information-Theoretic Sample Selection: While mutual information between each sample requantized value X_t and the unit identity C is an ideal theoretical measure of importance,

$$I(C; X_t) = H(X_t) - \sum_j p_j H(X_t|C = j), \tag{5}$$

its computation is costly and ill-suited to real-time or low-power implementations. We therefore approximate this metric under the Gaussian noise model (see (1)), where mutual information simplifies to a ratio of variances:

$$\text{Score}(t) = \frac{\text{Var}[\mu_c(t)]}{\mathbb{E}_c[\sigma_c^2(t)]}, \tag{6}$$

with $\mu_c(t)$ as the mean waveform value at time t for cluster c, and $\sigma_c^2(t)$ as the within-cluster variance. This discriminability score captures the same intuition: samples with high inter-unit variability and low intra-unit noise are more informative for separating spikes from different sources.

We compute this score for each time index and retain only the top 3–5 samples with the highest discriminability, forming a compact, information-rich representation of each spike. While we do not claim this number to be theoretically optimal, it was found to be empirically sufficient across multiple datasets. This choice reflects a practical tradeoff between reducing input dimensionality and preserving clustering performance.

c) Refined Clustering and Online Update: After feature selection, clustering is rerun using the reduced feature vectors. This not only accelerates computation but can improve accuracy by suppressing noisy and less-informative samples. To support long-term use in dynamic recording environments, the discriminability scores can be updated incrementally as a subset of new spikes, with all Wired-OR spike samples, are periodically transmitted. We maintain running estimates of per-cluster means and variances at each time point, enabling online updates to feature rankings with minimal overhead. These statistics converge quickly and require only $O(k)$ computation per spike, allowing the system to adapt to changing neural conditions while maintaining compression efficiency.

3) On-Chip Encoding: To efficiently encode remaining samples, we implement a Huffman-inspired entropy coder using a precomputed lookup table. During the system's calibration phase, histograms of address and amplitude values are collected and used to build codebooks optimized for the neural signal distribution. The codebook is refreshed at each recalibration, keeping the coding efficiency near the true entropy of the evolving signal distribution.

Instead of constructing Huffman trees at runtime, which is expensive in both time and silicon area, we generate static Huffman tables offline and store them on-chip. These tables translate each sample's spatial and amplitude value into a fixed-length or variable-length code via a simple lookup. This method eliminates the need for real-time histogramming or adaptive coding logic, enabling ultra-low-power, high-speed compression.

The channel coordinates and spike amplitude exhibit skewed distributions due to structured neural activity, allowing Huffman coding to achieve substantial gains. Coupled with selective sampling and re-quantization, this final stage completes the multi-step pipeline from raw events to fully compressed and hardware-ready output.

III. Simulation and Results

To evaluate the performance of our multi-stage compression framework, we conduct experiments on 512-channel *ex vivo* primate retina datasets recorded at 20 kHz. These datasets

[3]This is an empirical number dependent on the sampling rate of the system.

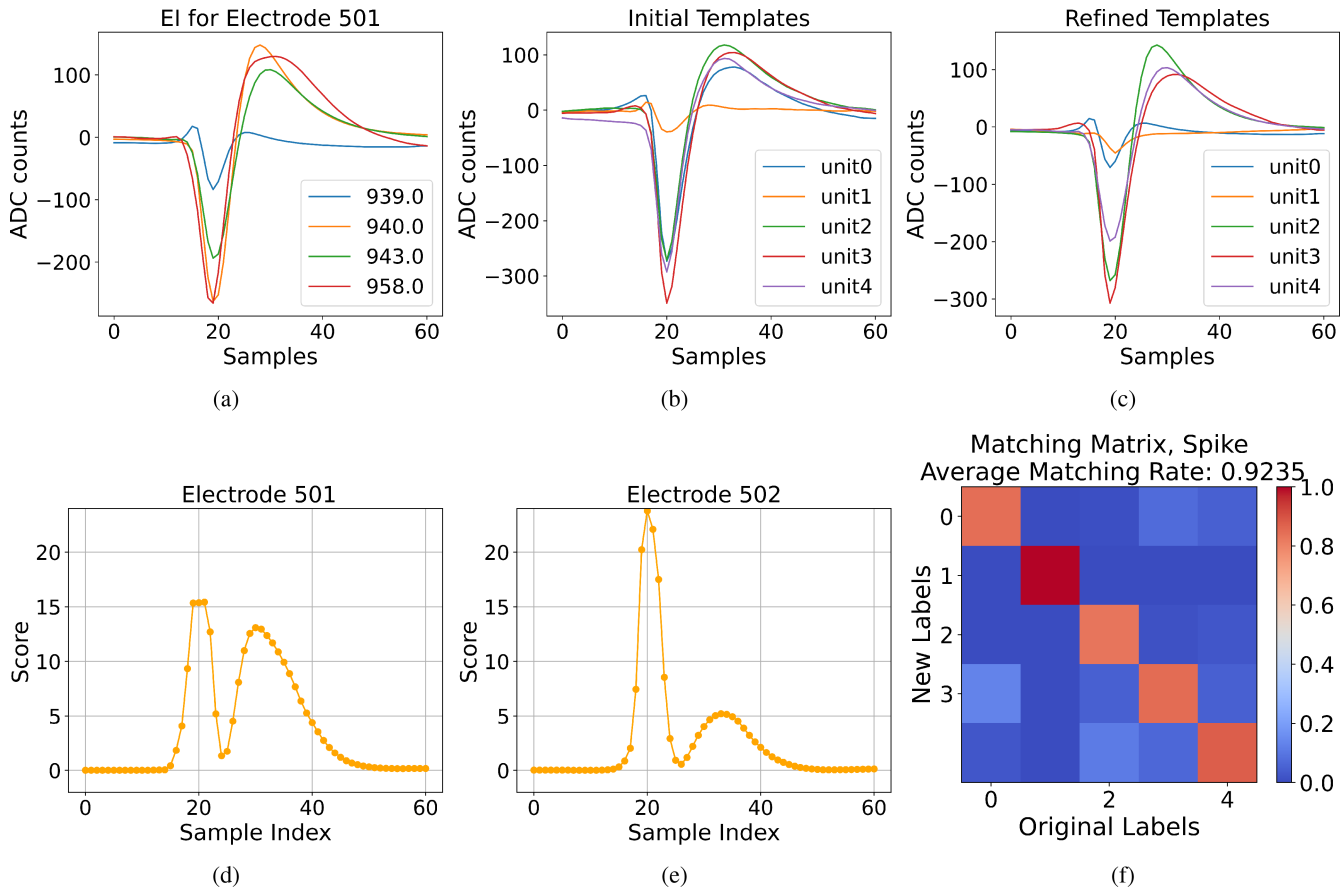

Fig. 8. An example of processing through electrode 501 in simulation. (a) Ground truth electrical image (EI) of the cells on electrode 501. (b) Initialized templates by clustering the detected and aligned events. (c) Refined and updated templates after iterating through the dataset. (d) Discriminability score of electrode 501 given the collected statistics after the calibration phase. (e) Discriminability score of electrode 502, which is near 501. (f) The matching matrix of the final clustering results and ground truth labels (different cells with the largest spike on electrode 501 and "garbage").

are processed in software to emulate the full pipeline, including Wired-OR analog-to-digital compressive readout, requantization, selective sampling, and entropy coding. Each stage is designed to reduce the total data volume while preserving critical information for downstream spike processing. Parameter recalibration was performed every 10 minutes, which was sufficient to maintain stable compression performance over time. We further benchmark our framework against state-of-the-art on-chip spike sorters [25], [31] using a publicly available multi-channel Neuropixels dataset [48].

Fig. 8 provides a detailed example of the full pipeline applied to one channel and its neighboring electrodes. The top row (see Fig. 8(a)–(c)) shows the extracted electrical images (EIs[4]) of units recorded on electrode 501 in the array, initial spike templates obtained through unsupervised k-means clustering, and the panel (c) shows the refined templates obtained after re-clustering using only the discriminative samples selected during the selective sampling phase, as described in the refined clustering and online update stage. This step enhances both accuracy and efficiency by removing noisy or redundant dimensions from the feature space. The bottom row highlights the discriminability scores computed for two adjacent electrodes: Fig. 8(d) for electrode 501 and Fig. 8(e) for electrode 502. Notably, in this example, electrode 502—despite not being the primary site of the spike—exhibits more peaked and informative samples for class separation. This underscores the value of spatially informed selective sampling, where nearby electrodes can provide higher-discriminability features than the electrode with the largest spike amplitude.

Fig. 8(f) shows a spike matching matrix between the clustering output from compressed data and ground truth unit labels. Clusters are aligned based on their dominant electrode. To address duplicate clusters across nearby electrodes, we apply a merging procedure based on the Hierarchical Adaptive Means (HAM) clustering strategy [49], which compares inter-cluster distances against a noise-informed threshold and consolidates redundant units in a hardware-efficient manner. The resulting average matching rate in this example is 92.35%.

To assess generalization beyond a single electrode, we evaluated the percentage of spikes preserved across the entire dataset. Fig. 9 summarizes spike recall as a function of signal-to-noise ratio (SNR) for different cell-electrode pair in the dataset. We

[4]Electrical images are generated by averaging spike waveforms to capture the characteristic signature of each cell–electrode pair.

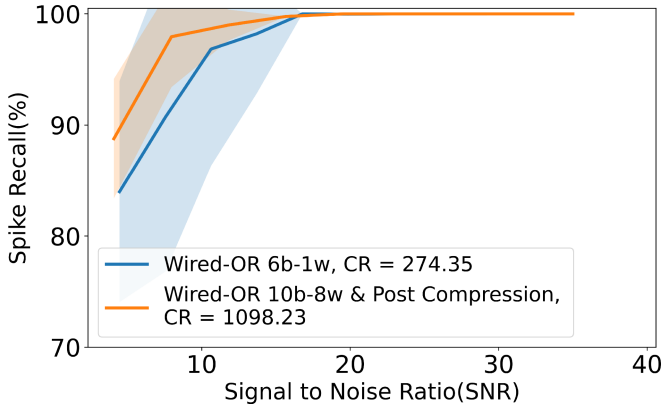

Fig. 9. Spike recall for ExVivo-1, across varying SNRs. The full pipeline achieves higher recall at higher compression.

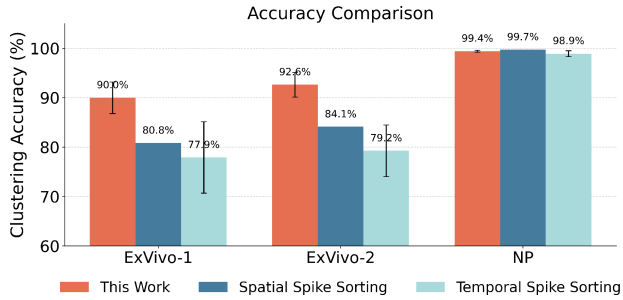

Fig. 10. Spike clustering accuracy across two retina datasets and one public Neuropixel dataset comparing three methods: our work using discriminability-driven selective sampling, the spatial spike sorting method proposed in [31], and a temporal single-electrode baseline. Our method achieves the highest average accuracy and more consistent performance across units.

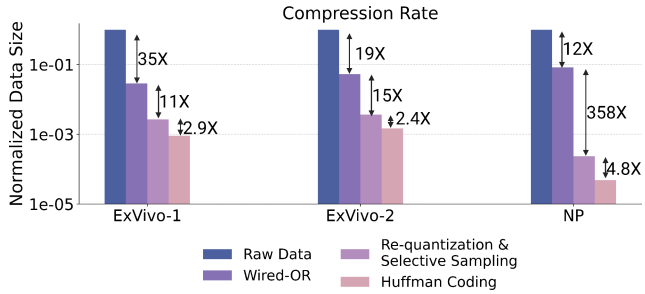

Fig. 11. Compression rate comparison across pipeline stages for three datasets. Starting from raw data, the Wired-OR stage provides the first major reduction. Re-quantization and selective sampling yield another 11×, followed by Huffman coding with 2.5× further reduction. Each dataset retains 5 informative samples per spike.

compare two representative schemes: (1) only compress with Wired-OR a 1-wire 6-bit ramp configuration, achieving a compression ratio (CR) of 274.35, and (2) the full compression pipeline with a 10-bit ramp, eight interleaved wiring Wired-OR configuration, and the addition of digital post-compression, resulting in a total CR of 1098.23.

As shown in Fig. 9, both configurations achieve high spike recall across a broad SNR range, with the full compression pipeline consistently outperforming the baseline Wired-OR-only setup. Notably, even under aggressive compression, spike recall exceeds 95% for moderate SNR values (SNR $\geq$ 8) and approaches 100% for higher-quality signals. The shaded region represents variability across electrodes for different SNR ranges, reflecting electrode-specific SNRs.

These results demonstrate that combining Wired-OR readout with post-compression stages—including re-quantization, selective sampling, and Huffman-inspired encoding—can significantly increase compression efficiency without compromising spike detection fidelity.

We next evaluate how our proposed compression framework compares against other spike compression and sorting approaches [31] on the same 512-channel *ex vivo* primate retina datasets (referred to hereafter as ExVivo-1 and ExVivo-2) and Neuropixel dataset [48]. These datasets are chosen as representative cases of neural recordings with varying SNR distributions, which significantly affect compression performance—particularly in the re-quantization and entropy coding stages. Specifically, ExVivo-1 has a mean SNR of 16.16 with a standard deviation of 6.99, while ExVivo-2 exhibits a mean SNR of 12.64 with a narrower spread (standard deviation of 6.06).

Fig. 10 shows the spike clustering accuracy across three datasets using our method (with information-theoretic selective sampling), the spatial spike sorting method reported from [31], and a temporal spike sorting baseline.

The temporal method performs spike sorting independently on each electrode by using all Wired-OR spike samples and linearly interpolating missing ones before classification, following the methods that we previously proposed in [50]. The spatial method, as proposed in [31], combines spikes across electrodes and leverages spatial correlation for clustering using a self-organizing map approach. While the spatial method shows strong performance, it does not report the variability of sorting accuracy across units. Our method not only achieves the highest mean clustering accuracy in both datasets (90.0% for ExVivo-1, 92.6% for ExVivo-2 and 99.4% for Neuropixel data), but also shows lower variability compared to using only temporal features, indicating robust performance across all neural units.

To understand how each stage of the compression pipeline contributes to overall compression, we analyze the normalized data size after each component. As shown in Fig. 11, the Wired-OR stage alone achieves 35× and 19× compression on ExVivo-1 and ExVivo-2, respectively. With selective sampling and re-quantization, an additional 11× reduction is observed, followed by a further 3× from Huffman-inspired entropy coding. The overall compression rate exceeds 1000×, with acceptable loss in spike fidelity. In the Neuropixels dataset [48], the lower SNR (∼7) and long-shank electrode architecture result in significantly greater compression gains—achieving 358× from re-quantization and selective sampling, and an additional 4.8× from Huffman-inspired coding. Notably, the compressibility in the selective sampling stage depends on the number of retained samples; in this analysis, we select the top 5 most discriminative time points per spiking event on an electrode and its neighboring electrodes.

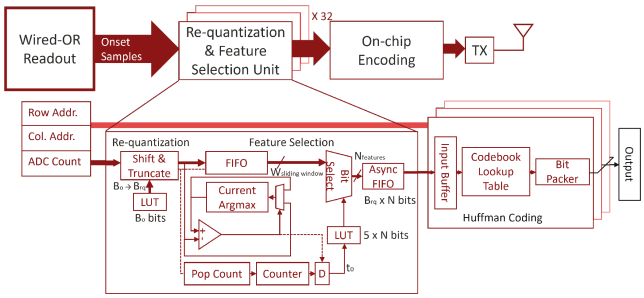

Fig. 12. Hardware architecture of the proposed digital compression core. The system receives sparse spike samples from the Wired-OR stage and performs re-quantization, sample selection, and Huffman-style encoding.

IV. Discussion

A. Implementation and Benchmarking of Digital Compression Core

Fig. 12 shows the proposed hardware architecture of the digital compression core that follows the Wired-OR analog-to-digital compressive readout. The system is divided into three main stages: (1) re-quantization, (2) discriminability-driven feature selection, and (3) approximate entropy coding via a lightweight Huffman encoder.

Incoming spike events from the Wired-OR stage are represented as sparse triplets: row address, column address, and ADC count. These are first fed into the "Re-quantization and Feature Selection Unit", which handles both amplitude compression and sample reduction.

In the re-quantization block, a programmable lookup table (LUT) determines the quantization resolution ($B_o \rightarrow B_{rq}$, where B_o is the length of the Wired-OR amplitude output, B_rq is the reduced bit width after re-quantization) based on the maximum SNR per channel. A simple shift-and-truncate operation compresses the amplitude values. This enables amplitude-specific precision tuning across electrodes with negligible computation overhead.

For sample selection, incoming spike waveforms are buffered in a FIFO and processed using a sliding window. A lightweight argmax engine computes the spike peak timing, and the most informative N samples are selected. The selection is driven by a second LUT that is trained offline during a calibration phase. Selected samples are forwarded via an asynchronous FIFO to the encoder.

Finally, the data passes through an on-chip Huffman encoder that uses a fixed codebook precomputed during the calibration phase. The encoder maps frequent sample patterns to shorter binary representations using a lookup table, and a bit packer serializes the output. This design enables approximate entropy-coded compression with low hardware complexity and avoids the need for runtime codebook updates.

To assess implementation feasibility, we provide an estimated breakdown of complexity and energy consumption for the digital compression core. Table I summarizes gate count[5] and energy cost per spike processed for each major module, based on standard logic primitives and published energy models for 28 nm CMOS [51], [52]. Gate counts are estimated by decomposing each module into standard digital building blocks such as flip-flops, adders, comparators, and control logic, using representative gate-equivalent costs derived from synthesis reports and textbook implementations [53], [54].

TABLE I
Estimated Memory, Gate Count, and Energy Per Re-Quantization and Feature Selection Unit (28 nm LP)

Module	Memory (bits)	Gates	Energy [51], [52] (pJ/spike)
Requantization Shift Reg.	10	∼80	0.2–0.3
Shift & Truncate	–	∼150	0.2
FIFO (30×7 samples)	132	∼6600	12.6
Argmax + Logic	23	∼310	2
Bit Selection via MUX + LUT (Shift Register)	40	∼3900	4.5
Async FIFO	25	∼850	0.3
Total per Unit	∼230	∼11.8k	**19.9**

The entire post-compression system is designed for shared processing across channels. A single Requantization & Feature Selection Unit is time-multiplexed across multiple channels. Based on spike event rates modeled empirically, only 32 such units are required to support a 1024-channel array. This architecture reduces the power and area overhead compared to per-channel buffering, while preserving the low-memory footprint that is a key advantage of the original Wired-OR design.

Including a conservative memory estimation of the Huffman lookup table of 4 KB,[6] the total on-chip memory required for the post-compression stage across 1024 channels is approximately 4.92 KB. The entire logic complexity is well within feasible bounds for modern implantable SoCs.

a) Power Savings: Assuming a baseline wireless transmission energy of $72.9pJ/bit$[55], the energy cost per bit after compression is broken down as follows:[7]

- $0.066pJ/b$ from re-quantization and selection,
- $\frac{1}{15} \times 3pJ/b = 0.2pJ/b$ from Huffman encoding [56],
- $\frac{1}{15\times2.4} \times 72.9pJ/b = 2.03pJ/b$ from transmission.

Total energy per transmitted bit: $2.296pJ/$ compared to transmitting all of the original Wired-OR outputs, this corresponds to an additional $31.75\times$ power reduction.

This analysis confirms that the post-compression stage delivers significant energy and bandwidth benefits with low hardware overhead, making it highly suitable for implantable or portable neural interface systems.

B. Comparison With Prior Work

Table II compares the proposed compression framework to previous neural signal compression and spike sorting

[5]For each potential spike, we consider the center electrode and its 6 adjacent electrodes and buffer 30 samples per electrode.

[6]A conservative 4 KB estimate accounts for two 1024-entry Huffman lookup tables of the channel addresses and recorded quantized ADC amplitudes.

[7]Here, we use the compression ratio results from the ExVivo-2 simulation as a representative example.

TABLE II
COMPARISON WITH PRIOR WORK

Compression Approach	Thresholding Spike Detection [21]	ML-based Autoencoder [27]	Neuromorphic Compression [23]	On-chip Spike Sorting with Temporal Feature [25]	On-chip Spike Sorting with Spatial Feature [31]	**Wired-OR + Post Compression (This Work)**
Memory Required	1kB	18kB	N/A	98.08kB	14kB	4.92kB
Low Computation Overhead	✓	79.25K MACs per spike	✓	x	✓	✓
No. of Channels	64–256	256	384	96–384	128–1024	512
Preserves Waveform	x	✓	✓	x	x	✓
Reported Accuracy	97.4%	N/A	92%	84.3%-98.7%	80.8%-99.7%	90%–99.4%
Compression Ratio	10–116×	20–500×	20-200×	240–39272×	1575–177768×	678-20621×

approaches. Compared to conventional thresholding-based spike detection [21], which incurs low overhead but discards waveform detail, we maintain full waveform shape and achieve 3–10× higher compression compared to this reported thresholding approach. While ML-based autoencoders [27], [29] can offer similar compression ratios, they require training, inference, and high memory resources—making them less suitable for ultra-low-power edge systems. On-chip spike sorting techniques [25], [31] extract spike times and unit labels in hardware, often discarding waveform information, and typically rely on fixed temporal or spatial features. In contrast, our system supports tunable feature selection via an information-theoretic approach and retains waveform fidelity.

Among all methods surveyed, Wired-OR is the only one to support high compression, waveform preservation, and scalable implementation without requiring large memory, floating-point operations, or machine learning. This makes it particularly well suited for large-scale neural recording systems targeting real-time closed-loop applications.

V. CONCLUSION

We have presented an end-to-end framework for adaptive compressive neural signal acquisition and digital compression. Building on our prior work on Wired-OR analog-to-digital compressive readout [33], [34], this paper introduces a post-compression architecture that combines re-quantization, selective sampling based on spike discriminability, and hardware-friendly entropy encoding. Together, these components form a scalable system that dramatically reduces data rates while preserving spike information critical for more fine-tuned spike sorting algorithms.

Evaluated on 512-channel *ex vivo* primate retina recordings, the system achieves over 1000× compression while preserving 90% spike sorting accuracy. The design balances spatial and temporal feature retention with hardware efficiency, and supports runtime adaptation to varying signal conditions.

REFERENCES

[1] Y. Wang et al., "Implantable intracortical microelectrodes: Reviewing the present with a focus on the future," *Microsyst. Nanoeng.*, vol. 9, Jan. 2023, Art. no. 7.

[2] T. Matsuo et al., "Simultaneous recording of single-neuron activities and broad-area intracranial electroencephalography: Electrode design and implantation procedure," *Neurosurgery*, vol. 73, no. 2, pp. ons146–ons154, Dec. 2013.

[3] A. L. Juavinett, G. Bekheet, and A. K. Churchland, "Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice," *eLife*, vol. 8, Aug. 2019, Art. no. e47188, doi: 10.7554/elife.47188.

[4] J. Putzeys et al., "Neuropixels data-acquisition system: A scalable platform for parallel recording of 10 000 electrophysiological signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1635–1644, Dec. 2019.

[5] T. Z. Luo et al.,, "An approach for long-term, multi-probe neuropixels recordings in unrestrained rats," *Elife*, vol. 9, Oct. 2020, Art. no. e59716.

[6] A. C. Paulk et al., "Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex," *Nat. Neurosci.*, vol. 25, no. 2, pp. 252–263, Feb. 2022.

[7] E. M. Trautmann et al., "Large-scale high-density brain-wide neural recording in nonhuman primates," *BiorXiv*, 2023, [Online]. Available: https://www.biorxiv.org/content/early/2023/05/04/2023.02.01.526664

[8] F. R. Willett et al., "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, Aug. 2023, doi: 10.1038/s41586-023-06377-x.

[9] P. D. Ganzer et al., "Restoring the sense of touch using a sensorimotor demultiplexing neural interface: 'disentangling' sensorimotor events during brain-computer interface control," in *Brain-Computer Interface Research: A State-of-the-Art Summary* (SpringerBriefs in Electrical and Computer Engineering Series). Berlin, Germany: Springer, 2021, pp. 75–85.

[10] N. A. Steinmetz et al., "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science*, vol. 372, no. 6539, Apr. 2021, Art. no. eabf4588.

[11] K. Sahasrabuddhe et al., "The argo: A high channel count recording system for neural recording in vivo," *J. Neural Eng.*, vol. 18, no. 1, Feb. 2021, Art. no. 015002.

[12] E. Musk and Neuralink, "An integrated brain-machine interface platform with thousands of channels," *J. Med. Internet Res.*, vol. 21, no. 10, Oct. 2019, Art. no. e16194.

[13] H. G. Rey, C. Pedreira, and R. Quian Quiroga, "Past, present and future of spike sorting techniques," *Brain Res. Bull.*, vol. 119, pp. 106–117, Apr. 2015.

[14] G. T. Einevoll et al., "Modelling and analysis of local field potentials for studying the function of cortical circuits," *Nat. Rev. Neurosci.*, vol. 14, no. 11, pp. 770–785, Nov. 2013.

[15] A. Nandi et al., "Single-neuron models linking electrophysiology, morphology, and transcriptomics across cortical cell types," *Cell Rep.*, vol. 40, no. 6, Aug. 2022, Art. no. 111176.

[16] N. W. Gouwens et al., "Classification of electrophysiological and morphological neuron types in the mouse visual cortex," *Nat. Neurosci.*, vol. 22, no. 7, pp. 1182–1195, Jun. 2019.

[17] O. Ophir, O. Shefi, and O. Lindenbaum, "Classifying neuronal cell types based on shared electrophysiological information from humans and mice," *Neuroinformatics*, vol. 22, no. 4, pp. 473–486, Oct. 2024, doi: 10.1007/s12021-024-09675-5.

[18] E. M. Trautmann et al., "Accurate estimation of neural population dynamics without spike sorting," *Neuron*, vol. 103, no. 2, pp. 292–308.e4, Jul. 2019.

[19] N. Even-Chen et al., "Power-saving design opportunities for wireless intracortical brain–computer interfaces," *Nature Biomed. Eng.*, vol. 4, no. 10, pp. 984–996, Aug. 2020.

[20] S.-Y. Park et al., "Dynamic power reduction in scalable neural recording interface using spatiotemporal correlation and temporal sparsity of neural signals," *IEEE J. Solid-State Circuits*, vol. 53, no. 4, pp. 1102–1114, Apr. 2018, doi: 10.1109/JSSC.2017.2787749.

[21] X. Guo, M. Shaeri, and M. Shoaran, "An accurate and hardware-efficient dual spike detector for implantable neural interfaces," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, Oct. 2022, pp. 70–74, doi: 10.1109/BioCAS54905.2022.9948602.

[22] M. A. Shaeri and A. M. Sodagar, "A method for compression of intracortically-recorded neural signals dedicated to implantable brain–machine interfaces," *IEEE Eng. Med. Biol. Soc.*, vol. 23, no. 3, pp. 485–497, May 2015, doi: 10.1109/TNSRE.2014.2355139.

[23] V. Mohan, W. P. Tay, and A. Basu, "Towards neuromorphic compression based neural sensing for next-generation wireless implantable brain machine interface," *Neuromorphic Comput. Eng.*, vol. 5, no. 1, Jan. 2025, Art. no. 014004, doi: 10.1088/2634-4386/adad10.

[24] D. Valencia and A. Alimohammad, "A real-time spike sorting system using parallel OSort clustering," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1700–1713, Dec. 2019.

[25] Y. Chen et al., "A 384-Channel online-spike-sorting IC using unsupervised Geo-OSort clustering and achieving 0.0013mm2/Ch and $1.78\mu\mathrm{W/ch}$," in *Proc. IEEE Int. Solid- State Circuits Conf.*, Feb. 2023, pp. 486–488, doi: 10.1109/ISSCC42615.2023.10067264.

[26] J. Li et al., "A 0.78-μ w 96-Ch deep sub-Vt neural spike processor integrated with a Nanowatt power management unit," in *Proc. IEEE 44th Eur. Solid State Circuits Conf.*, Sep. 2018, pp. 154–157, doi: 10.1109/ESSCIRC.2018.8494273.

[27] T. Wu et al.,, "Deep compressive autoencoder for action potential compression in large-scale neural recording," *J. Neural Eng.*, vol. 15, no. 6, Oct. 2018, Art. no. 066019, doi: 10.1088/2F1741-2552/2Faae18.

[28] M. Pagin and M. Ortmanns, "A neural data lossless compression scheme based on spatial and temporal prediction," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, Oct. 2017, pp. 1–4.

[29] J. Thies and A. Alimohammad, "Compact and low-power neural spike compression using undercomplete autoencoders," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1529–1538, Aug. 2019.

[30] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith, "Analog-to-digital compression: A new paradigm for converting signals to bits," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 16–39, May 2018.

[31] A. Akhoundi et al., "15.2 a 1024-channel 0.00029mm2/ch 74nw/ch online spatial spike-sorting chip with event-driven spike detection and self-organizing map clustering," in *Proc. IEEE Int. Solid-State Circuits Conf.*, vol. 68, 2025, pp. 268–270.

[32] Y. Chen et al., "An online-spike-sorting IC using unsupervised geometry-aware osort clustering for efficient embedded neural-signal processing," *IEEE J. Solid-State Circuits*, vol. 58, no. 11, pp. 2990–3002, Nov. 2023.

[33] D. G. Muratore et al., "A data-compressive wired-or readout for massively parallel neural recording," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1128–1140, Dec. 2019.

[34] P. Yan et al., "Data compression versus signal fidelity tradeoff in wired-or analog-to-digital compressive arrays for neural recording," *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 4, pp. 754–767, Aug. 2023.

[35] M. Jang et al., "A1024-channel 268-nw/pixel 36 μm2/channel data-compressive neural recording IC for high-bandwidth brain–computer interfaces," *IEEE J. Solid-State Circuits*, vol. 59, no. 4, pp. 1123–1136, Apr. 2024.

[36] J. J. Jun et al., "Fully integrated silicon probes for high-density recording of neural activity," *Nature*, vol. 551, no. 7679, pp. 232–236, Nov. 2017.

[37] J. L. Shobe et al., "Brain activity mapping at multiple scales with silicon microprobes containing 1,024 electrodes," *J. Neurophysiol.*, vol. 114, no. 3, pp. 2043–2052, Sep. 2015.

[38] B. J. Black et al., "Chronic recording and electrochemical performance of utah microelectrode arrays implanted in rat motor cortex," *J. Neurophysiol.*, vol. 120, no. 4, pp. 2083–2090, Oct. 2018.

[39] R. Bartolo et al., "Dimensionality, information and learning in prefrontal cortex," *PLoS Comput. Biol.*, vol. 16, no. 4, Apr. 2020, Art. no. e1007514.

[40] M. Pachitariu et al., "Spike sorting with kilosort4," *Nature Methods*, vol. 21, no. 5, pp. 914–921, May 2024, doi: 10.1038/s41592-024-02232-7.

[41] J. Ziv, "On universal quantization," *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 344–347, May 1985.

[42] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 518–521, Sep. 1980.

[43] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained source codes," in *Proc. Int. Symp. Inf. Theory*, 2004, Art. no. 464.

[44] S. Jalali and T. Weissman, "Denoising via MCMC-based lossy compression," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 3092–3100, Jun. 2012.

[45] I. Ochoa et al.,, "Effect of lossy compression of quality scores on variant calling," *Brief Bioinf.*, vol. 18, no. 2, pp. 183–194, Mar. 2017.

[46] B. Natarajan, K. Konstantinides, and C. Herley, "Occam filters for stochastic sources with application to digital images," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1434–1438, May 1998.

[47] Y. Ephraim and R. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inf. Theory*, vol. 34, no. 4, pp. 826–834, Jul. 1988.

[48] "Neuropixels datasets, 'sorting comparison results'," 2016. Accessed May 10, 2023. [Online]. Available: http://phy.cortexlab.net/data/sortingComparison/

[49] S. E. Paraskevopoulou et al., "Hierarchical adaptive means (ham) clustering for hardware-efficient, unsupervised and real-time spike sorting," *J. Neurosci. Methods*, vol. 235, pp. 145–156, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027014002489

[50] P. Yan et al., "Data compression versus signal fidelity trade-off in wired-or ADC arrays for neural recording," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2022, pp. 80–84.

[51] A. Pedram et al., "Dark memory and accelerator-rich system optimization in the dark silicon era," *IEEE Des. Test*, vol. 34, no. 2, pp. 39–50, Apr. 2017.

[52] K. Prabhu et al., "MINOTAUR: A posit-based 0.42–0.50-tops/w edge transformer inference and training accelerator," *IEEE J. Solid-State Circuits*, vol. 60, no. 4, pp. 1311–1323, Apr. 2025.

[53] D. Harris and S. Harris, *Digital Design and Computer Architecture*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.

[54] V. K. Kodavalla, "IP gate count estimation methodology during micro-architecture phase," *Des. Reuse*, Sep. 2008, Accessed: Apr. 23, 2025. [Online]. Available: https://www.design-reuse.com/articles/19171/ip-gate-count-estimation-micro-architecture-phase.html

[55] E. So and A. Arbabian, "6.1 12mb/s 4 ultrasound MIMO relay with wireless power and communication for neural interfaces," in *Proc. IEEE Int. Solid-State Circuits Conf.*, vol. 67, 2024, pp. 100–102.

[56] M. Aboelmaged, A. Shisha, and M. A. A. E. Ghany, "High-performance data compression-based design for dynamic IoT security systems," *Electronics*, vol. 10, no. 16, 2021, Art. no. 1989. [Online]. Available: https://www.mdpi.com/2079-9292/10/16/1989