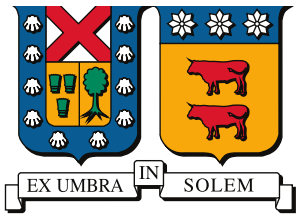

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Memoria de Título/Tesis de Magister

Attention Maps Reveal Stimulus-Dependent Retinal Population Codes

Tesis para optar al grado/título de
Magister en Ciencias de la Ingeniería Electrónica/Ingeniero Civil Electrónico

Alumno
Francisco Miqueles Varela

Guía de Tesis/Profesor Supervisor
Dra. María José Escobar Silva

Revisores/Comisión Co-Referente
Dr. Mauricio Araya López
Dr. Patricio Orio Alvarez

August 7, 2026, Valparaíso, Chile

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ☒ Tesis de Postgrado

Título del trabajo: Attention Maps Reveals Stimulus-Dependent Retinal Population Codes

Nombre del candidato(a): Francisco Miqueles Varela

Carrera / Grado: Magíster en Ciencia de la Ingeniería Electrónica

Campus: Casa Central **Departamento:** Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, ___María José Escobar Silva___, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 2/9/2026_____ **Firma:** _____

Estudiante o Candidato(a):

Fecha: 2/9/2026_____ **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

To discover a truth, it is necessary to suffer... Truth is an arduous acquisition, reached only after long and painful trials.
-Santiago Ramón y Cajal

Dedicatoria

AGRADECIMIENTOS

EL camino que he trazado y cruzado a lo largo de mi Magíster es, sin duda, el resultado del esfuerzo acumulado de las personas que me rodean, me quieren y me apoyan. Mi formación no se limita únicamente a lo académico, sino que se nutre de mis interacciones diarias, conversaciones y vivencias. En ese sentido, me siento sinceramente en deuda con todo mi círculo: amigos, colegas, familiares y conocidos.

Quiero agradecer profundamente a mi profesora guía, María José Escobar Silva, por ser parte fundamental de este largo trayecto desde el pregrado hasta el postgrado. Le agradezco por su tiempo y dedicación, por brindarme oportunidades únicas y por impulsarme a enfrentar desafíos que me hicieron crecer en lo profesional y en lo personal. Gracias por estar siempre presente y por ser una guía no solo en lo académico, sino también en lo humano.

Agradezco de corazón a mis amigos por ser un pilar incondicional. Gracias por escuchar, aconsejar y acompañarme en cada paso de este proceso. Un agradecimiento especial a mis colegas del AC3E, ya que las constantes conversaciones, discusiones científicas y conocimientos compartidos me permitieron ampliar mi visión. Todos ustedes fueron una parte esencial de mi formación, y gran parte de las proyecciones y objetivos que hoy persigo nacieron de nuestras interacciones.

Finalmente, quiero agradecer a mi familia. A mi madre Claudia, por su esfuerzo constante a lo largo de los años y por ser un ejemplo vivo de que siempre se puede seguir luchando; gracias por mantenerte fuerte en los momentos difíciles, por tu amor incondicional y por estar siempre para nosotros. A mi hermano Andrés, por ser mi partner de vida, mi mejor amigo y la persona con la que puedo compartirlo todo. A mi polola Annette, por ser mi compañera de vida, por construir un futuro a mi lado y acompañarme en cada aventura. A mi primo Joaquin, por demostrarme que la distancia es solo un matiz y que la hermandad no depende de la sangre. Y, por sobre todo, a mi padre Raúl, que en paz descanse: gracias por tu fortaleza y tu actitud frente a la vida; sé que desde donde estés, te sientes orgulloso de mí.

A todos ustedes, que de una u otra forma hicieron posible la culminación de este trabajo y de esta etapa de mi vida, mi más profundo y sincero agradecimiento.

Francisco Miqueles Varela

RESUMEN

COMPRENDER cómo los modelos de aprendizaje profundo mapean la actividad de la población neural con los estímulos requiere tanto de una alta precisión predictiva como de mecanismos internos interpretables. En este trabajo, empleamos el framework POYO, una arquitectura transformer escalable basada en la tokenización de spikes y el modelado latente, para decodificar registros a gran escala de células ganglionares de la retina. Nos preguntamos si los mecanismos de atención del modelo pueden proporcionar información biológicamente significativa mediante la evaluación de dos condiciones contrastantes: un estímulo de flash uniforme y un estímulo de una bola en movimiento estructurada espaciotemporalmente. Mostramos que el modelo decodifica ambos estímulos de manera confiable y se adapta rápidamente a nuevas preparaciones mediante fine-tuning, lo que sugiere la captura de códigos poblacionales transferibles. Luego, analizamos la organización interna del modelo, revelando que los patrones de atención del codificador se adaptan a la complejidad del estímulo: los cabezales de atención aparecen sincronizados y ampliamente distribuidos para el estímulo de flash, mientras que exhiben estrategias de asignación heterogéneas y especializadas para la bola en movimiento. Al agregar los pesos de atención para identificar las neuronas más relevantes para cada tarea, demostramos que estas unidades de alta atención poseen firmas fisiológicas distintas, concentrando respuestas sostenidas con altas tasas de disparo para el flash, frente a cinéticas diversas para la entrada estructurada. Confirmamos la validez causal de estos hallazgos mediante ablaciones guiadas por atención, donde la eliminación progresiva de estas unidades de mayor rango genera pérdidas sistemáticas en el rendimiento de la decodificación. Además, expandimos el análisis a la atención del decodificador, descubriendo estrategias de recuperación específicas del estímulo donde los cabezales individuales exhiben preferencias distintas de sintonización direccional. Concluimos que los mecanismos de atención genéricos pueden recuperar espontáneamente estrategias de codificación biológica, identificando subpoblaciones neurales funcionalmente distintas sin supervisión, validando así la utilidad de las arquitecturas basadas en transformers para el descubrimiento neurocientífico.

Palabras Clave

decodificación neural, retina, células ganglionares de la retina, transformers, interpretabilidad, atención, representaciones latentes.

ABSTRACT

UNDERSTANDING how deep learning models map neural population activity to stimuli requires both high predictive accuracy and interpretable internal mechanisms. In this work, we employ the POYO framework, a scalable transformer architecture based on spike tokenization and latent modeling, to decode large-scale retinal ganglion cell recordings. We ask whether the model's attention mechanisms can provide biologically meaningful insight by evaluating two contrasting conditions: a uniform flash stimulus and a spatiotemporally structured moving ball stimulus. We show that the model decodes both stimuli reliably and adapts rapidly to new preparations via fine-tuning, suggesting the capture of transferable population codes. We then analyze the model's internal organization, revealing that encoder attention patterns adapt to stimulus complexity: attention heads appear synchronized and broadly distributed for the flash stimulus, whereas they exhibit heterogeneous, specialized allocation strategies for the moving ball. By aggregating attention weights to identify the most relevant neurons for each task, we demonstrate that these high-attention units possess distinct physiological signatures—concentrating sustained, high-firing rates responses for the flash versus diverse kinetics for the structured input. We confirm the causal validity of these findings via attention-guided ablations, where the progressive removal of these top-ranked units yields systematic losses in decoding performance. Furthermore, we expand the analysis to the decoder's attention, uncovering stimulus-specific retrieval strategies where individual heads exhibit distinct directional tuning preferences. We conclude that generic attention mechanisms can spontaneously recover biological coding strategies, identifying functionally distinct neural subpopulations without supervision, thus validating the utility of transformer-based architectures for neuroscientific discovery.

Keywords

neural decoding, retina, retinal ganglion cells, transformers, interpretability, attention, latent representations

ÍNDICE

List of Figures

List of Tables

ABBREVIATIONS

Uppercase

RGCs	: retinal ganglion cells
RF	: receptive fields
ANN	: artificial neural networks
MEA	: multi-electrode array
ISI	: interspike interval
SNR	: signal-to-noise ratio
MLP	: multi-layer perceptron
MSE	: mean squared error
K-S	: Kolmogorov-Smirnov
OLE	: Optimal Linear Estimator
NDT2	: Neural Data Transformer 2
POSSM	: Population State-Space Model
GPFA	: Gaussian Process Factor Analysis

SYMBOLOGY

Matrices

$\mathbf{A}_{h,i}$	: encoder attention matrix, $\mathbf{A}_{h,i} \in \mathbb{R}^{L \times N_i}$
$\mathbf{D_h}$	: decoder attention matrix, $\mathbf{D}_h \in \mathbb{R}^{M \times L}$

Scalars

f_{BR}	: bias index
S_i	: sustain index
$\bar{w}_{h,i,k}$	: average attention weight
$\hat{w}_{h,i,k}$	: normalized average attention weight
H	: head dimension
L	: latent dimension
W	: context window dimension
N_i	: number of context windows
N_i^u	: number of spikes of the unit in a windows
M	: timestamps dimension
R^2	: coefficient of determination
E	: Shannon entropy
I	: importance score
K	: number of important units per seed

Subscripts

h	: head
i	: context window
j	: latent
k	: spike
u	: unit
m	: timestamp

Superscripts

h : head
i : context window

Chapter 1

INTRODUCTION

THe retina is a highly organized neural tissue located at the back of the eye, responsible for capturing incoming light and converting it into electrical signals that can be interpreted -through a dynamic learning processes mechanism- by the brain. This process begins when photoreceptors absorb photons and, through interactions with inter-neurons and retinal ganglion cells (RGCs), generate sequences or trains of action potentials that travel via the optic nerve to higher visual centers (6; 7). Understanding how these signals encode the diverse visual stimuli found in nature is fundamental to revealing the mechanisms by which sensory input is transformed into perception. A natural counterpart to this question is whether one can invert the process: given the neural responses, can the original stimulus be reconstructed? This challenge defines the problem of neural decoding, that seeks to reconstruct or infer a stimulus from the electrical responses of a population of neurons. In the visual system, this involves estimating features such as shape, motion, or contrast of the original stimulus from the recorded spiking activity of RGCs (1; 8; 9). Precise decoding models provide a dual benefit: they help determine how information about the external world is preserved in neural activity, and they aid in the design of visual prostheses and brain–computer interfaces.

Early decoding approaches relied on linear models, valued for their simplicity and direct interpretability. These methods allowed clear links between model parameters and physiological features, helping to identify the contribution of specific neurons or receptive field (RF) properties to stimulus reconstruction (1; 10). As datasets became richer and stimuli more complex, however, linear approaches proved to be insufficient. To capture nonlinear dependencies, researchers turned to artificial neural networks (ANN), Bayesian decoders, and later deep learning architectures, which achieved markedly higher accuracies in decoding natural images and dynamic visual scenes (11; 12; 13).

More recently, transformer-based architectures have reshaped the decoding landscape. Building on advances in language (14) and vision (15), they provide scalable representations and excel at capturing long-range dependencies. This attention-based mechanism offers an inherent path toward interpretability; by quantifying how a model dynamically prioritizes specific input features, researchers can move beyond black-box predictions to uncover the underlying structure of neural data. The POYO framework exemplifies this

progress by introducing a unified tokenization scheme and leveraging PerceiverIO-style attention to decode neural activity across sessions, animals, and tasks with unprecedented robustness (16). A key distinction of this approach is that, unlike traditional decoders that rely on binarizing spike trains into discrete time bins, POYO processes individual spikes as continuous-time events. This point-process framework allows the model to preserve the sub-millisecond precision of the retinal output, which is fundamental for capturing the rapid neural dynamics triggered by moving stimuli.

Its extension, POYO+, further scaled this idea, showing that training on highly diverse datasets spanning cell types, brain regions, and behaviors not only improves decoding performance but also enables cross-task and cross-region transfer (17). Importantly, the authors explored the interpretability of their model by analyzing the latent space trained on large-scale datasets, such as the Allen Brain Observatory. They observed that, even without explicit supervision, the learned unit embeddings spontaneously organized into clusters corresponding to distinct anatomical brain regions and cellular sub-types. While this provides promising evidence that large-scale models can capture biologically meaningful patterns, the analysis remains largely descriptive and does not yet yield strong physiological conclusions.

POYO capability raises a deeper question: Can the internal mechanisms of these high performing models, specifically their attention and latent representations, be used to reverse engineer the computational strategies of the neural population itself? In this work, we test the hypothesis that transformer-based decoders can reveal distinct population coding regimes. We propose that the attention mechanism, which dynamically weights the contribution of different inputs, provides a direct readout of how the neural population allocates its computational resources. We trained the POYO model on large-scale RGC recordings under two contrasting stimuli: a uniform flash and a complex moving ball. Beyond benchmarking accuracy, we analyzed whether the model's internal dynamics would expose a shift from a concentrated, specialized code for the simple, global stimulus to a distributed, diverse representation for the complex, spatiotemporal one. Our goal was to determine if interpretable artificial intelligence can move beyond performance and generate testable hypotheses about adaptive neural computation.

1.1 Hypothesis

We test the hypothesis that transformer-based decoders operating on continuous point-process events can spontaneously reveal distinct sensory population coding regimes through their internal attention mechanisms. Specifically, we propose that the model's internal attention allocation adapts dynamically to stimulus complexity, shifting from a synchronized, redundant pooling strategy under global, spatially uniform inputs to a desynchronized, functionally specialized multi-channel representation under complex spatiotemporal environments.

1.2 Objectives

1.2.1 General Objective

To investigate whether the internal attention mechanisms and latent representations of transformer-based neural decoders can resolve and explain adaptive neural population coding regimes in the visual system, using a data-driven, unsupervised spike-tokenization

framework trained under contrasting sensory conditions.

1.2.2 Specific Objectives

- **O1 - Benchmark Predictive Accuracy:** Implement and evaluate the POYO framework using continuous-time spike tokenization against classical linear estimators, state-space models, and binned neural transformers using large-scale multielectrode array retinal ganglion cell recordings under uniform flash and moving ball stimuli.

- **O2 - Assess Cross-Preparation Generalization:** Quantify the network's transferability and efficient adaptation capabilities across multiple independent mouse retinal preparations via a structured cross-session pretraining and fine-tuning regime.

- **O3 - Characterize Encoder Attention Dynamics:** Quantify and compare the temporal focus and synchronization of encoder attention heads across contrasting visual paradigms using Shannon entropy metrics and non-parametric statistical divergence tests.

- **O4 - Causally Validate Neural Attribution Rankings:** Establish a direct functional link between attention weights and predictive relevance by performing sequential, attention-guided causal unit ablations against a stochastic control baseline.

- **O5 - Map High-Attention Subpopulations to Physiological Profiles:** Classify the biological response signatures (bias index, sustain index, latency, and firing rate) of the neurons consistently prioritized by the encoder to reveal task-dependent selection criteria.

- **O6 - Decompose Decoder Information Retrieval Spaces:** Evaluate the decoder's cross-attention mechanisms using directional decomposition and Pearson correlation traces to identify stimulus-specific directional tuning preferences across distinct attention heads.

1.3 Scope and Limitations

1.3.1 Scope

- **Point-Process Resolution:** The analysis explicitly processes individual action potentials as continuous-time event coordinates, preserving sub-millisecond precision and bypassing traditional rate-binning limitations.

- **Causal Attribution Validation:** The validation extends beyond descriptive interpretations of attention weights by incorporating functional, closed-loop progressive unit exclusions to measure true decoding degradation.

- **Experimental Dataset Boundaries:** The multi-retina evaluation is consistently executed using large-scale microelectrode array recordings from five independent wild-type mouse retinal preparations exposed to identical full-field luminance and fractional Brownian motion motion protocols.

1.3.2 Limitations

- **Abstract Latent Dimensions:** While the encoder and decoder attention layers provide interpretable profiles via entropy and correlation metrics, the underlying latent projection space vectors lack explicit physiological or anatomical labels.
- **Attribution Redundancy Risks:** High attention weights can capture inherent network correlation structures, shared sensory drives, or input firing densities; therefore, direct causal interpretations must be managed with caution.
- **Stimulus Paradigm Bounds:** The study is strictly constrained to the characterization of full-field luminance flashes and fractional Brownian motion trajectories of a single object, excluding natural visual scenes or highly correlated semantic textures.
- **Unconstrained Network Architecture:** Receptive-field alignment or bio-inspired structural wiring constraints are not explicitly enforced during training, representing a data-driven computational abstraction rather than a rigid anatomical replica of the retinal circuitry.

1.4 Chapter Outline

The manuscript is structured into three main parts: the literature review, the proposed methodology, and our empirical contributions.

To establish the state of the art, Chapter 2 provides the theoretical foundation by reviewing traditional and contemporary neural decoding architectures, ranging from classical linear estimators and state-space models to modern binned neural transformers, contrastive learning, and the continuous-time POYO framework.

To detail our core methodological contributions, Chapter 3 introduces our novel interpretability frameworks, formalizing the mathematical tools used for encoder-decoder attention entropy, causal neuron ranking, and latent space exploration. Chapter 4 describes the experimental protocols, including large-scale multielectrode array retinal recordings, visual stimulation paradigms, and our algorithmic design for progressive temporal and spatial spike-shuffling perturbations.

The second half of the manuscript presents our empirical findings and neuroscientific discoveries. Chapter 5 evaluates quantitative decoding performance benchmarks (R^2) across model architectures and analyzes network robustness under structural code degradations. Chapter 6 explores the internal dynamics of the trained network, demonstrating unsupervised latent cell-type stratification, adaptive encoder attention allocation, causal validation via guided unit ablations, and decoder directional tuning. Finally, Chapter 7 discusses the neuroscientific implications and biological analogies of these findings, and Chapter 8 summarizes our main conclusions while outlining directions for future research.

Chapter 2

NEURAL DECODING ARCHITECTURES

The extraction of meaningful lower-dimensional structures from high-dimensional neural population recordings is bounded by a fundamental trade-off: balancing computational expressivity, real-time decoding efficiency, and biological interpretability. Historically, architectures developed for neural state estimation have evolved non-linearly to overcome the structural assumptions of their predecessors. This chapter maps the conceptual trajectory of neural decoding, tracking how the field transitioned from simple additive linear filters to modern, unstructured foundation transformers. We analyze this progression through three evolutionary axes: the shift from fixed temporal bins to continuous point-process events, the transition from explicit generative assumptions to data-driven contrastive objectives, and the move from rigid single-session alignments toward open-vocabulary architectures capable of generalizing across diverse brain regions and cellular sub-types.

2.1 Optimal Linear Estimator (OLE)

The Optimal Linear Estimator (OLE), classically introduced in the context of multineuronal populations, represents a fundamental benchmark in the field of neural decoding (1). Rooted in visual reconstruction frameworks originally pioneered for single-unit recordings in invertebrate systems, the OLE operates as an inverse filter that maps the collective, time-dependent spiking activity of a neural population back onto a continuous external stimulus variable, establishing a direct, open-loop pipeline from multiunit events to continuous intensity estimation (see Figure 2.1).

The underlying premise of linear reconstruction is that while the forward encoding process—how the retina converts photons into discrete action potentials—is highly non-linear and rectified, the reverse decoding operation can be efficiently approximated via a linear summation of additive postsynaptic contributions. Under this paradigm, each individual action potential fired by a given neuron is treated as an independent packet of information that injects a fixed, time-varying waveform into the continuous estimate of the stimulus. The objective of the OLE is to analytically derive a set of population filters

that minimize the mean squared error (MSE) between this reconstructed estimate and the objective ground-truth stimulus.

2.1.1 Mathematical Formulation

To mathematically formalize the multineuronal linear decoder, the continuous external stimulus $I(t)$ and the raw point-process spike trains $r^{\nu}(t)$ of a population of neurons (where $\nu \in \{1, 2, \ldots, U\}$ denotes the unit index) must be discretized into uniform temporal bins of width Δt. Following standard normalization conventions, the stimulus intensity is scaled to zero mean and unit variance:

$$s_i = \frac{I_i - M}{C} \tag{2.1}$$

where I_i is the empirical light intensity within the i-th temporal interval $[i\Delta t, (i+1)\Delta t]$, M is the global mean intensity of the stimulus ensemble, and C represents its standard deviation. Concurrently, the neural response vector for each temporal bin is computed by counting the integer number of action potentials generated by each individual unit:

$$r_i^{\nu} = \int_{i\Delta t}^{(i+1)\Delta t} \sum_k \delta(t - t_k^{\nu})\, dt \tag{2.2}$$

where $\delta(t)$ represents the Dirac delta function for spike times t_k^{ν} corresponding to unit ν. The continuous-time reconstruction estimate u_i is obtained by performing a discrete convolution between each binned spike train and its corresponding localized linear filter f_j^{ν}, integrated along a finite temporal memory window $N\Delta t$, plus a global scalar offset term a:

$$u_i = a + \sum_{\nu=1}^{U} \sum_{j=0}^{N-1} f_j^{\nu} r_{i-j}^{\nu} \tag{2.3}$$

This relationship can be translated into a highly compact matrix formalism suitable for numerical optimization. Let $s \in \mathbb{R}^M$ represent the ground-truth stimulus vector and $u \in \mathbb{R}^M$ represent the reconstructed estimate across M evaluation bins. We define a global population response matrix $R \in \mathbb{R}^{M \times (UN+1)}$, where each row contains the concatenated binned spike histories of all U units spanning back N steps, alongside a constant column of ones to handle the offset a. This matrix mapping directly reflects the transformation block executed by the decoder to convert discrete binned counts into the continuous profile u illustrated in Figure 2.1. The forward reconstruction is thus expressed as a linear algebraic mapping:

$$u = R \cdot f \tag{2.4}$$

where $f \in \mathbb{R}^{UN+1}$ is the unified population filter vector:

$$f^T = \begin{bmatrix} a & f_0^1 & f_1^1 \cdots f_{N-1}^1 & \cdots & f_0^U \cdots f_{N-1}^U \end{bmatrix} \tag{2.5}$$

To compute the optimal filter coefficients, the objective function minimizes the residual sum of squares of the reconstruction error, formulated as $\mathcal{L}_{MSE} = (s-u)^T(s-u)$. Minimizing this objective with respect to the filter array f yields the classic closed-form solution via the normal equations:

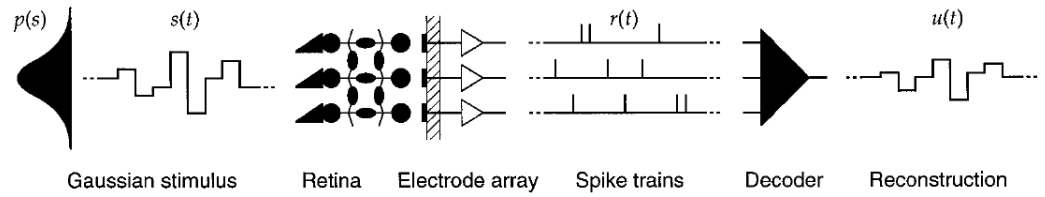

Figura 2.1: **Schematic overview of the multineuronal linear decoding process.** A continuous time-varying stimulus $s(t)$ is encoded by a population of retinal ganglion cells into point-process spike trains $r(t)$, which are monitored via a multielectrode array. The linear decoder filters and sums these co-active responses to produce an analytical reconstruction $u(t)$ (1).

$$f = (R^T R)^{-1} \cdot (R^T s) \tag{2.6}$$

The analytical utility of Equation 2.6 lies in its deterministic mapping of physical cross-correlations. The term $R^T s$ represents the raw cross-correlation vectors between the stimulus and the population's spike trains, which directly corresponds to the spike-triggered average (STA) of each neuron. Conversely, the matrix $R^T R$ captures the auto-correlation functions of individual spike trains alongside the cross-correlation profiles across different pairs of units within the population.

2.1.2 Population Context and Conditional Spike Meaning

The inversion of the correlation matrix $(R^T R)^{-1}$ plays a critical role in population decoding. In the limiting case where neural firing is exceptionally sparse and spikes are spaced further apart than the filter duration $N\Delta t$, the cross-correlation matrix $R^T R$ collapses into a strictly diagonal structure. Under this sparse firing regime, each action potential acts as an isolated, independent message completely decoupled from the rest of the network, meaning that the optimal decoding filter f^ν becomes perfectly proportional to that single cell's spike-triggered average stimulus.

However, inside dense, highly correlated biological populations, multiple neurons fire simultaneously within the same integration window. Under these conditions, the off-diagonal elements of $R^T R$ become highly pronounced, and its inversion drastically reshapes the morphology of the decoding filters compared to their single-unit isolated counterparts. When an individual neuron is decoded within a co-active population, its optimal linear filter exhibits a substantial drop in absolute amplitude and transforms into a highly oscillatory, biphasic or triphasic waveform.

This systematic alteration proves that the optimal interpretation of an individual action potential cannot be determined in isolation; instead, its functional meaning is heavily conditioned by the simultaneous presence or absence of spikes in neighboring channels. The OLE implicitly leverages these multiunit dependencies through the $(R^T R)^{-1}$ operator to decorrelate redundant population activity, effectively suppressing shared global noise and preventing the overcounting of overlapping sensory signals.

2.1.3 Performance Boundaries and Information Saturation

When evaluated under temporally broadband, uniform visual flicker, the linear reconstruction framework exhibits distinct frequency-specific coding constraints. Reconstruction quality drops off sharply at low frequencies below 1 Hz due to strong temporal differentiation and high-pass filtering introduced by amacrine circuits in the inner retina, meaning the population retains virtually no information about constant, static light intensity ($f = 0$ Hz). Concurrently, the decoder hits a strict performance ceiling at higher temporal frequencies, with reconstruction power rolling off sharply above 10 Hz. This high-frequency limitation is primarily imposed by the biophysical low-pass kinetics of phototransduction within the upstream cone photoreceptors, compounded by flat, broadband network noise introduced in the downstream retinal layers.

A key challenge documented in multiunit visual estimation is the rapid saturation of decoded information as more units are provided to the linear filter. Because neighboring retinal ganglion cells of the same functional response type (such as fast-OFF units) share overlapping receptive fields and are driven by identical stimulus trajectories, they transmit highly redundant information. Consequently, combining multiple cells of the same class yields only marginal tracking improvements over a single well-isolated neuron.

In contrast, units of opposite functional polarities—specifically ON and OFF channels—convey highly independent, non-overlapping features of the visual environment, resulting in an additive summation of their information spectral densities. Due to this parallel complementary organization, the total information rate rapidly asymptotes, with a local heterogeneous cluster of just four distinct functional cell types recovering up to 79% of the absolute maximal information available to the entire multielectrode array.

2.2 Neural Data Transformer

While the Optimal Linear Estimator establishes an analytical, closed-form baseline for neural reconstruction, its formulation introduces severe biological constraints by treating each action potential as an additive, isolated message, rendering it fundamentally incapable of resolving higher-order non-linear dependencies or dynamic functional interactions within dense co-active populations. To capture these complex non-linear dynamics without sacrificing single-trial temporal evolution, researchers traditionally turned to Recurrent Neural Networks (RNNs); however, explicit recurrence necessitates sequential data processing, creating computational bottlenecks that heavily restrict real-time inference efficiency.

The Neural Data Transformer (NDT1) bypassed these sequential constraints by introducing a parallel, non-recurrent alternative based on self-attention mechanisms that operate directly on multivariate sequences of binned spiking activity (2). Instead of iteratively updating a hidden state across timesteps, this architecture aggregates global contextual information across the entire sequence window simultaneously, enabling high-fidelity rate inference with substantially reduced latency suitable for real-time state estimation. The structural transition from traditional sequential processing graphs to this parallel transformer workflow, along with single-trial baseline alignments, is explicitly illustrated in Fig. 2.2.

Building upon this non-recurrent foundation, the Neural Data Transformer 2 (NDT2) was developed to enable large-scale, multi-context pretraining over highly heterogeneous datasets that span different experimental sessions, subjects, and behavioral tasks (3). To

scale effectively across shifted neural distributions and overcome the data instability inherent to microelectrode recordings, NDT2 fundamentally restructures the original framework through three structural innovations: spatiotemporal attention via neural patching, learned context metadata embeddings, and an asymmetric encoder-decoder design.

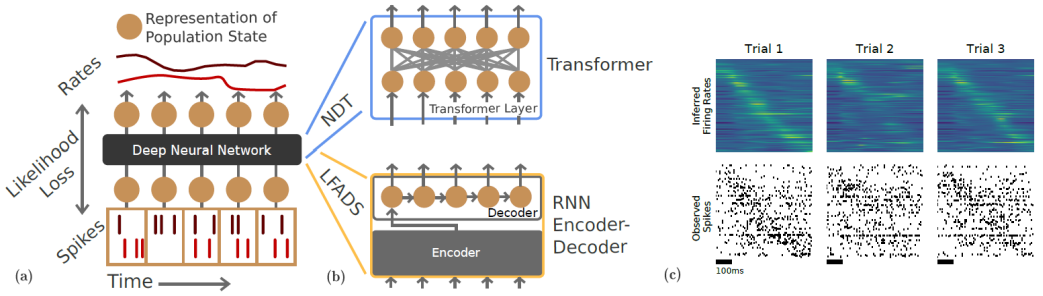

Figura 2.2: **Schematic paradigm of sequential versus parallel neural models.** (a) Unsupervised network framework mapping binned point-processes onto log-firing rates via a likelihood training objective. (b) Parallel transformer layer context execution compared directly against step-by-step recurrent encoders. (c) Empirical input spike coordinates sorted chronologically against internal model-inferred continuous factor profiles (2).

2.2.1 Architectural Framework and Spatiotemporal Attention

The mathematical implementation of NDT1 transforms a sequence of population spike counts $X \in \mathbb{R}^{T \times C}$ across T timesteps and C recording channels into inferred, continuous-valued log-firing rates by allocating attention purely across the temporal axis, embedding the full population as a single token per timestep.

In contrast, to balance expressivity and computational cost across heterogeneous environments where neuron counts and channel identities vary, NDT2 introduces a spatiotemporal tokenization strategy. Spiking activity is decomposed into localized units by grouping a fixed number of K neurons into discrete spatial patches, akin to pixel patches in Vision Transformers. These neural patches are embedded by concatenating linear projections of the spike counts within the patch. For an input representation matrix X, the attention operation projects the patched data onto three distinct spaces via learned query, key, and value weight matrices ($W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$):

$$Q = W^Q X, \quad K = W^K X, \quad V = W^V X \tag{2.7}$$

Using these representations, a single-token output y_i is assembled by evaluating the scaled dot-product similarity between the query vector q_i at token index i and the key vectors k_j across all tokens $j \in \{1, 2, \ldots, N\}$. These similarity scores are normalized using the softmax function to yield the attention weights w_i^j:

$$w_i^j = \frac{\exp\left(\frac{q_i \cdot k_j}{\sqrt{d}}\right)}{\sum_{l=1}^{N} \exp\left(\frac{q_i \cdot k_l}{\sqrt{d}}\right)} \tag{2.8}$$

where $\sqrt{d}$ serves as the attention scaling factor to stabilize gradient flow, and N represents the total number of spatiotemporal tokens within the context window. To account for

temporal sequencing without recurrence, a learned position embedding vector is additively combined with each token representation before entering the transformer layers. The complete parallel transformation for the entire sequence window is summarized algebraically as:

$$Y = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2.9}$$

Following the self-attention blocks, the representations pass through layer normalization and a feedforward network block containing a multi-layer perceptron (MLP) with rectified linear unit (ReLU) activations. While NDT1 projects encoder outputs directly through a linear readout layer followed by an exponentiation operation to estimate log-firing rates, NDT2 adopts an asymmetric architecture. The deep multi-layer encoder processes only the unmasked spatiotemporal tokens, drastically reducing memory overhead. The resulting refined latent representations are then mapped back to the full sequence space by a shallower, lightweight neural decoder dedicated to rate reconstruction. This comprehensive multicontext architecture, demonstrating the asymmetric masking framework alongside the grouping of populations into discrete patches, is schematized in Figure 2.3.

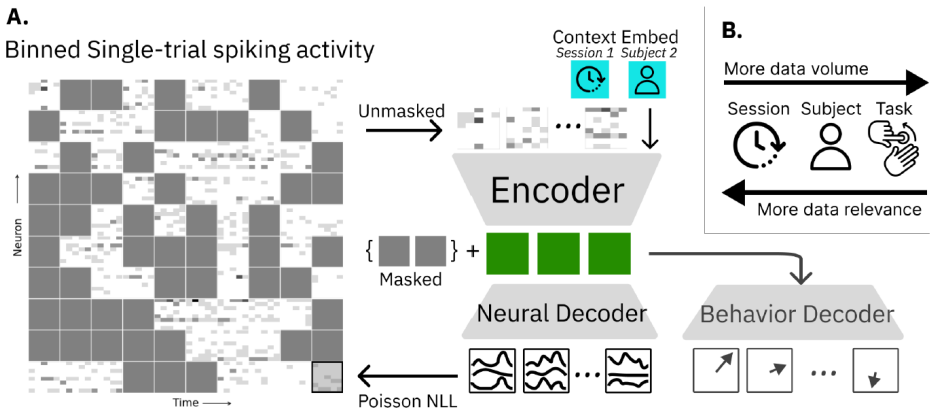

Figura 2.3: **Architectural mechanics of NDT2 multi-context pretraining.** (A) Spatiotemporal autoencoding workflow on a single-trial layout, where unmasked patches are passed through a deep encoder, and a shallow decoder subsequently reconstructs masked coordinates via Poisson likelihood optimization. (B) Information affinity mapping showing the optimization boundary between absolute data volume and downstream data target relevance (3).

Furthermore, NDT2 supplements the token sequence with learnable context tokens. These context embeddings ingest metadata parameters—specifically unique session, subject, and task identifiers—acting as prompt tuning vectors that absorb hidden experimental shifts and enable cheap functional specialization within a centralized, shared model backbone.

2.2.2 Masked Training Paradigm and Regularization

Training the transformer architecture in an unsupervised configuration necessitates an adaptation of self-supervised masked autoencoding. During the training phase, a random

subset of the spatiotemporal sequence tokens—typically a fixed ratio such as 20% in NDT1 and up to 50% in NDT2—is designated for masking. To minimize the distribution shift between training and real-time inference settings, the network avoids artificial placeholder tokens in favor of a "zero mask" strategy, where the spike counts of the selected masked intervals are completely zeroed out.

The optimization objective forces the network to accurately reconstruct the unobserved spiking density of the masked intervals given the intact context provided by the surrounding unmasked tokens. Rather than optimizing a cross-entropy loss, the objective function minimizes a negative log-likelihood (NLL) based on a Poisson emission model:

$$\mathcal{L}_{\text{NLL}} = \sum_{i=1}^{T} \sum_{c=1}^{C} (\hat{r}_{i,c} - r_{i,c} \ln(\hat{r}_{i,c}) + \ln(r_{i,c}!)) \tag{2.10}$$

where $r_{i,c}$ represents the empirical spike count and $\hat{r}_{i,c}$ denotes the network-inferred firing rate for a given channel c and bin i. Because neuroscientific datasets are significantly smaller than typical machine learning corpora, the models require heavy regularization to prevent overfitting and ensure numerical stability in log space. This is achieved by embedding deep dropout layers at the entry and exit points of the transformer body, sweeping dropout ratios across a heavy $[0.2, 0.6]$ boundary for single-session networks, while NDT2 utilizes a regularized dropout floor of 0.1 to facilitate generalized representation learning during large-scale multi-context pretraining.

2.2.3 Inference Efficiency and Real-Time Scaling

The primary advantage of the non-recurrent transformer framework over recurrent autoencoders is its execution speed during continuous online estimation. For real-time applications such as brain-machine interfaces (iBCIs) or closed-loop neuromodulation, rolling windows of neural data must be processed with sub-millisecond latencies, as delays severely degrade closed-loop behavioral control. Recurrent architectures scale their computation sequentially with window length, as they require a full pass over the temporal trajectory to compute initialization inputs for the dynamical decoder.

By contrast, because the transformer body processes all historical input blocks in parallel, its forward pass execution speed exhibits a near-constant relationship with respect to the sequence length. This architectural parallelism enables a stable 3.9 ms inference latency when estimating population activity within a standard 70-bin motor cortical window, yielding a 6.7-fold speedup over state-of-the-art recurrent autoencoders operating under identical sequence parameters.

Furthermore, by capitalizing on this non-recurrent parallelism, the model training constraints can be dynamically scaled. While a standard 6-layer encoder provides maximum expressivity for complex dynamics, the structural footprint can be compressed into a shallow 2-layer or 1-layer configuration. This architectural scaling significantly reduces the total parameter volume, accelerating training down to under 30 minutes while completely bypassing the costly backpropagation-through-time operations mandatory for recurrent networks.

Downstream, this pretrained foundation enables highly efficient transfer learning; by freezing the core spatiotemporal transformer weights and optimizing only the lightweight context embeddings or final behavioral readout heads, the model achieves rapid adaptation to entirely novel subjects, days, and tasks using a minimal volume of target-session calibration trials.

2.3 CEBRA

The Neural Data Transformer optimizes single-trial state estimation by processing binned spike sequences in parallel, dramatically improving inference speeds compared to sequential recurrent networks. Nevertheless, both NDT and traditional autoencoders remain bound by an explicit generative constraint: they require the direct optimization of a strict Poisson likelihood emission model against empirical binned spike counts. Defining rigid generative statistics limits model flexibility in highly volatile biological regimes where firing distributions inevitably deviate from standard point-process assumptions. To bypass the need for explicit data-generating distribution models, contemporary representation frameworks introduce self-supervised contrastive architectures, leading directly to the development of CEBRA (4). As a self-supervised dimensionality reduction framework, CEBRA maps high-dimensional neural population activity into consistent, lower-dimensional latent embedding spaces by flexibly conditioning data relationships on time, behavioral variables, or discrete experimental context through a contrastive representation learning scheme. This non-generative workflow, which implements feature extraction and contrastive optimization to produce lower-dimensional topologies, is schematized in Figure 2.4.

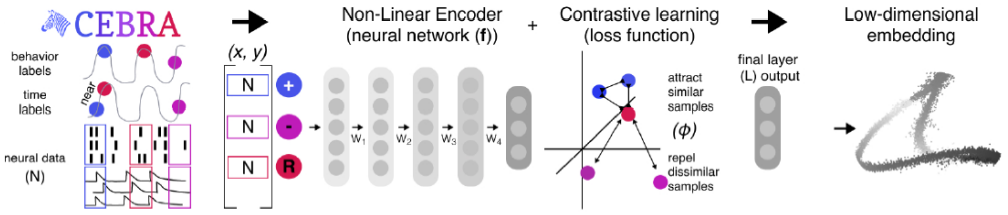

Figura 2.4: **Methodological workflow of the CEBRA representation learning pipeline.** Paired neural recordings and behavioral or temporal auxiliary labels are passed through a non-linear network encoder f. The resulting features are projected onto a unit hypersphere and optimized using a contrastive objective that attracts similar positive samples while repelling dissimilar distractors, generating a consistent low-dimensional latent embedding (4).

2.3.1 Mathematical Formulation and Contrastive Optimization

The mathematical architecture maps a multivariate time series of neural population signals $s_t \in \mathbb{R}^D$ into a lower-dimensional embedding space $Z \subset \mathbb{R}^E$. Depending on the experimental context, reference samples x and corresponding positive or negative samples y are extracted from the signal stream and passed through a non-linear feature encoder $f(\cdot)$—typically parameterized by a MLP or a deep convolutional neural network (CNN)—yielding normalized embedding states on a unit hypersphere.

The network optimization is formulated as a classification problem using a generalized InfoNCE objective function. In the limit of unlimited samples, the criterion maps reference and target vectors via an internal similarity measure $\phi(\cdot, \cdot)$, which is defined as the dot product scaled by a temperature hyperparameter τ:

$$\psi(x, y) = \frac{\phi(f(x), f(y))}{\tau} = \frac{f(x)^T f(y)}{\tau} \tag{2.11}$$

Given a reference sample x, the positive target conditional distribution $p(y|x)$ and multiple negative target distributions $q(y|x)$ are contrasted. The network minimizes the empirical loss function using batch or stochastic gradient descent by drawing a single positive target y_+ and n negative distractor samples y_i:

$$\mathcal{L}_{\text{InfoNCE}} = \mathbb{E}_{x\sim p_{\mathcal{D}}(x), y_+\sim p(y|x)} \left[-\psi(x, y_+) + \ln \sum_{i=1}^{n} \exp\left(\psi(x, y_i)\right) \right] \tag{2.12}$$

Upon full convergence, the value of the minimized objective maps directly to the negative Kullback-Leibler (KL) divergence between the pre-defined positive and negative sampling configurations, yielding a bounded goodness-of-fit metric independent of the batch size:

$$-\mathcal{D}_{\text{KL}}\left(p(y|x) \,\middle\|\, q(y|x)\right) \leq \mathcal{L}_{\text{InfoNCE}} - \ln n \leq 0 \tag{2.13}$$

2.3.2 Auxiliary-Variable Sampling Schemes

The structural versatility of the algorithm relies on the design of its data-sampling distributions, which are split into three operating regimes depending on how neighborhood associations are established:

- **Discovery-Driven (Time-Contrastive Mode):** The positive sampling distribution isolates local temporal proximity, setting $y_+ = s_{t+\Delta}$ within a tight time window Δ. The negative distribution draws stochastic baseline distractors uniformly from the full dataset empirical distribution. This self-supervised configuration assumes that neural states vary continuously over time, capturing fundamental manifold topologies without behavior labels.

- **Hypothesis-Driven (Behaviorally-Guided Mode):** The conditional distribution maps data relationships directly through observed auxiliary variables. For a discrete behavioral context k_t, positive samples are restricted to identical functional task parameters. For continuous context labels c_t (such as spatial tracking coordinates), the algorithm constructs empirical neighborhood sets where positive samples are chosen to minimize distance in behavioral feature space.

- **Hybrid Mode:** This configuration dynamically combines continuous behavior labels with temporal offsets. The integration forces the non-linear encoder to disentangle the latent geometry, segmenting the lower-dimensional manifold into behavioral coordinate tracking subspaces and separate orthogonal components dedicated to task-independent internal temporal variance.

The operational mechanics of these configuration regimes, including the selective routing of reference and negative point pairs, are explicitly mapped out in the timeline diagram of Fig. 2.5.

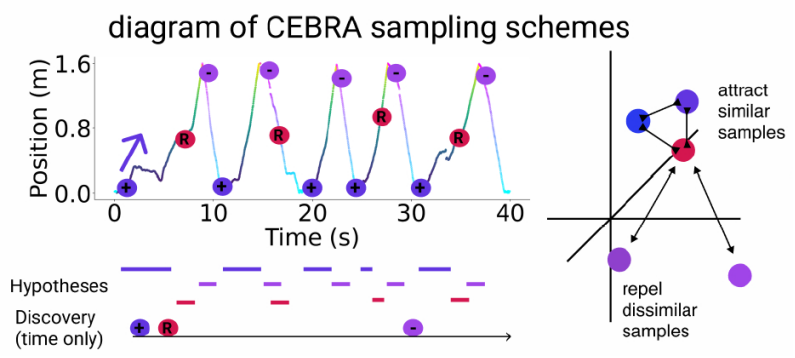

Figura 2.5: **Algorithmic diagram of user-defined CEBRA sampling distributions.** The horizontal timeline axis exemplifies how positive and negative pairs are selected to configure the topological neighborhood parameters of the latent space, contrasting behavior-guided hypotheses against discovery-driven choices (4).

2.3.3 Domain Generalization and Multi-Session Alignment

By decoupling latent estimation from individual data-generating statistics, the contrastive architecture can bridge distribution shifts across multiple experimental blocks or recording sessions. For multi-session scaling across separate subjects, independent encoders $f^{(i)}(\cdot)$ are initialized to match session-dependent channel dimensions, but they map the disjoint signals into a centralized, shared co-domain Z.

The optimization forces uniform pair sampling across sessions, irrespective of variable recording lengths. This constraint systematically suppresses idiosyncratic individual variance, eliminating session-specific batch effects or biological hardware alignment offsets without removing core anatomical feature representations. The resulting animal-invariant embedding space allows for pseudo-subject decoding and enables rapid transfer learning, where a pre-trained core encoder can be deployed and fine-tuned on completely unseen neural subjects within a few adaptation steps.

2.4 Population State-Space Model (POSSM)

By replacing generative reconstruction losses with a hypothesis-driven contrastive optimization, CEBRA successfully extracts latent embedding spaces that are highly consistent across independent subjects and experimental blocks. However, because contrastive learning relies heavily on the design of user-defined auxiliary variable sampling distributions, it requires explicit behavior or time labels to shape the underlying manifold geometry. When explicit behavioral constrain ts are unavailable, or when the goal is to evaluate the intrinsic system dynamics by extracting smooth, low-dimensional neural trajectories from simultaneous single-trial population recordings without inductive label biases, a fully probabilistic, continuous-time state-space alternative is required (5). The Population State-Space Model (POSSM), formalized mathematically through Gaussian-Process Factor Analysis (GPFA), addresses these limitations by unifying the temporal smoothing and spatial dimensionality reduction steps within a single generative probabilistic framework.

2.4.1 Mathematical Formulation and Generative Process

To formalize the generative framework, let $y_t \in \mathbb{R}^q$ represent the neural population observation vector at a discrete timestep $t \in \{1, 2, \ldots, T\}$, where q denotes the total number of recorded units, and let $x_t \in \mathbb{R}^p$ represent the corresponding low-dimensional latent state vector, under the structural constraint that $p \ll q$. The observation space is modeled as a linear transformation of the latent space corrupted by an independent Gaussian noise structure:

$$y_t \mid x_t \sim \mathcal{N}(Cx_t + d, R) \tag{2.14}$$

where $C \in \mathbb{R}^{q \times p}$ represents the factor loading matrix that maps latent states to neural coordinates, $d \in \mathbb{R}^q$ is a baseline scalar offset vector capturing the mean firing rate of each neuron, and $R \in \mathbb{R}^{q \times q}$ is a diagonal matrix containing the private variance σ_i^2 unique to each individual unit.

The temporal dynamics of the latent state are defined globally across the full sequence window rather than via step-by-step Markovian autoregressive loops. Let $\mathbf{x} \in \mathbb{R}^{pT}$ denote the concatenated latent trajectory vector, which stacks the p-dimensional states across all

T timesteps. The prior distribution of $\mathbf{x}$ is defined as a zero-mean multivariate Gaussian distribution:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, K) \tag{2.15}$$

where the global covariance matrix $K \in \mathbb{R}^{pT \times pT}$ possesses a block-diagonal structure. Each individual block $K_k \in \mathbb{R}^{T \times T}$ (for $k \in \{1, 2, \ldots, p\}$) governs the temporal evolution of the k-th latent dimension independently. The elements of each temporal matrix block are generated analytically using a squared exponential covariance kernel:

$$K_k(t_i, t_j) = \tau_k^2 \exp\left(-\frac{(t_i - t_j)^2}{2\gamma_k^2}\right) + \delta_{ij}\sigma_{\text{floor}}^2 \tag{2.16}$$

where τ_k^2 scales the overall signal variance, γ_k^2 acts as the characteristic temporal length-scale parameter that determines the smoothness of the k-th trajectory dimension, and σ_{floor}^2 is a fixed small noise floor to guarantee positive-definiteness during numerical matrix inversion operations.

2.4.2 Parameter Estimation via the Expectation-Maximization Algorithm

The complete set of learnable parameters $\theta = \{C, d, R, \gamma_1, \ldots, \gamma_p\}$ is optimized directly from empirical single-trial matrices using the Expectation-Maximization (EM) algorithm. Because the generative equations utilize strictly linear mappings and Gaussian noise fields, the joint distribution of the multiunit observations $\mathbf{y} \in \mathbb{R}^{qT}$ and stacked latents $\mathbf{x}$ remains exactly Gaussian:

$$\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{d}\end{bmatrix}, \begin{bmatrix}K & K\mathbf{C}^T \\ \mathbf{C}K & \mathbf{C}K\mathbf{C}^T + \mathbf{R}\end{bmatrix}\right) \tag{2.17}$$

where $\mathbf{C} \in \mathbb{R}^{qT \times pT}$ and $\mathbf{R} \in \mathbb{R}^{qT \times qT}$ are large block-diagonal matrices constructed by replicating C and R across all T steps. During the E-step, the algorithm computes the posterior conditional distribution of the latent trajectory given the empirical observations, which is expressed analytically as:

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}}) \tag{2.18}$$

$$\mu_{\mathbf{x}|\mathbf{y}} = K\mathbf{C}^T(\mathbf{C}K\mathbf{C}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{d}) \tag{2.19}$$

$$\Sigma_{\mathbf{x}|\mathbf{y}} = K - K\mathbf{C}^T(\mathbf{C}K\mathbf{C}^T + \mathbf{R})^{-1}\mathbf{C}K \tag{2.20}$$

During the M-step, the parameters are updated iteratively by maximizing the expected log-likelihood function conditioned on the posterior statistics calculated in the E-step. The update rule for the factor loading matrix, for example, is computed via:

$$C^{\text{new}} = \left(\sum_{t=1}^{T} y_t \mathbb{E}[x_t \mid y_t]^T\right)\left(\sum_{t=1}^{T} \mathbb{E}[x_t x_t^T \mid y_t]\right)^{-1} \tag{2.21}$$

The length-scale parameters γ_k are updated concurrently by applying gradient-based optimization routines to the log-likelihood function. The predictive goodness-of-fit

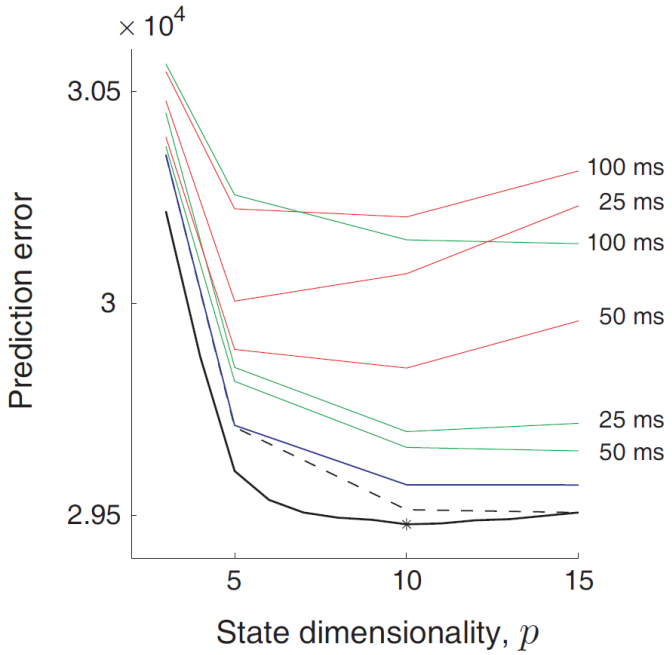

Figura 2.6: **Cross-validated cross-prediction errors across different state dimensionalities (p).** The profiles compare the leave-neuron-out execution performance of conventional two-stage frameworks (PPCA in red, FA in green) under varying localized Gaussian smoothing widths against a parametric first-order autoregressive baseline (blue) and the integrated GPFA formulations (dashed and solid black traces (5).

resulting from this simultaneous optimization framework, evaluated through cross-validated leave-neuron-out prediction errors against classic two-stage baselines, is quantitatively summarized in Fig. 2.6.

2.4.3 Orthogonalization and Trajectory Decoupling

The factor loading matrix C obtained upon final EM convergence is not structurally unique, as the generative equations are invariant to arbitrary linear coordinate rotations ($Cx_t = (CM)(M^{-1}x_t)$). To eliminate this geometric ambiguity and decouple the extracted latent paths, the POSSM framework implements a post-hoc orthogonalization pipeline.

First, singular value decomposition (SVD) is performed on the optimized factor loading matrix to yield $C = U \cdot D \cdot V^T$, where the columns of $U \in \mathbb{R}^{q \times p}$ define a set of mutually orthogonal spatial projection axes within the neural population space. The latent trajectories are then rotated into this new coordinate frame, which aligns the neural dimensions in descending order based on the percentage of shared population variance they capture.

Second, an additional rotation is applied to the covariance structures of individual trials to decouple the continuous trajectories over time. This orthogonalization ensures that the final latent coordinates are completely uncorrelated both across distinct neurons and across different execution steps. The resulting independent channels effectively isolate parallel

sub-components of the population dynamics, allowing a subsequent linear regressor (such as Ridge regression) to map the smooth trajectories onto continuous kinematic or stimulus outputs without suffering from high multi-channel collinearity. The systematic output of this post-processing decomposition pipeline, illustrating the isolation of structured single-trial paths ordered by variance magnitude, is displayed in Fig. 2.7.

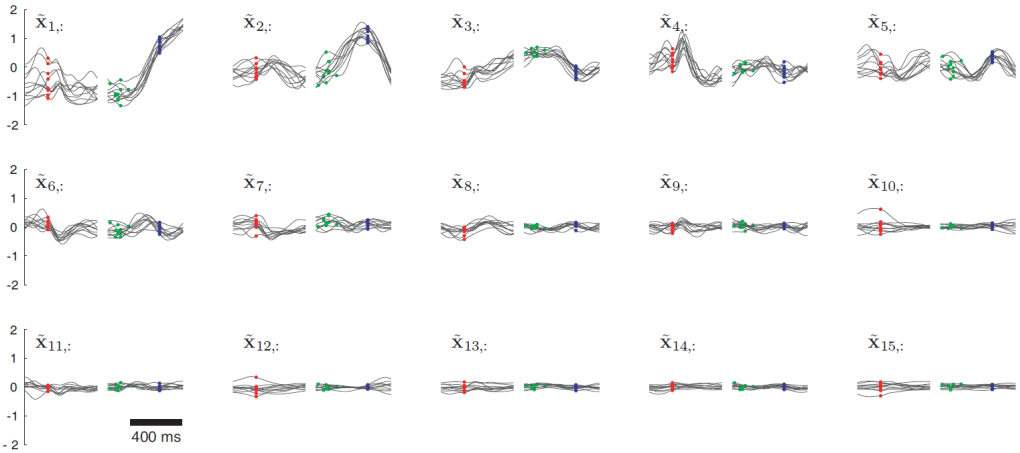

Figura 2.7: **Orthonormalized single-trial neural trajectories extracted via the integrated GPFA model framework ($p = 15$).** The stacked channels are ordered top-to-bottom by data covariance explained. Individual black traces map single experimental trials aligned dynamically to target onset (left block) and movement execution (right block), allowing the separation of structured single-trial dynamics from unshared noise dimensions (5).

2.5 The POYO Framework: Spike-Tokenization and Large-Scale Pretraining

The state-space frameworks reviewed thus far share a common architectural constraint: they are structurally restricted to a closed-vocabulary paradigm bound by a fixed, predefined number of recording channels. Consequently, they cannot ingest diverse sessions with differing neural arrays without undergoing extensive structural restructuring while simultaneously smoothing out fine sub-millisecond spike precision through arbitrary temporal binning dependencies (16). To bridge the gap between these isolated single-session approaches and a truly universal translator for neural dynamics, the decoding landscape required a paradigm shift toward an open-vocabulary framework that treats individual action potentials as sparse, continuous-time event coordinate tokens independent of static channel layouts. The POYO framework addresses these challenges by introducing a sparse, spike-based tokenization scheme combined with a flexible PerceiverIO cross-attention backbone, enabling a single unified deep learning model to be trained end-to-end across hundreds of highly heterogeneous multi-lab and multi-subject datasets simultaneously.

2.5.1 POYO Model Architecture

To enable open-vocabulary and continuous-time neural decoding, the POYO framework (16) introduces a transformer-based architecture designed to learn robust representations of neural population dynamics. Unlike standard approaches that bin neural activity—which smooths out fine temporal details and introduces arbitrary bin-size dependencies—POYO utilizes a spike-based tokenization strategy. Under this paradigm, the model input comprises tokens representing individual, continuous-time spike events within fixed context windows (e.g., 1-second windows), effectively preserving the sub-millisecond precision of the neural code while adopting a sparse and computationally efficient representation. Each individual spike is mapped into a D-dimensional coordinate-based token characterized by a tuple (x_i, t_i). Here, t_i represents the continuous event timestamp, and x_i represents a learnable unit embedding extracted from a centralized lookup table UnitEmbed$(\cdot)$, which encapsulates the functional identity and cell-type role of the generating neuron. For an arbitrary population of cells, all firing events are aggregated chronologically into a single multivariate sequence of variable token length M. The framework organizes the processing pipeline into two primary functional components: an encoder and a decoder.

In the encoder, high-dimensional input spike tokens are mapped onto a fixed set of latent tokens using a cross-attention mechanism. This step effectively reduces dimensionality from the variable-length spike sequence into a compressed bottleneck while preserving relevant contextual and temporal features across the population. Let $X = [x_1, \ldots, x_M] \in \mathbb{R}^{M \times D}$ denote the input sequence of spike tokens, and let $Z_0 = [z_{0,1}, \ldots, z_{0,N}] \in \mathbb{R}^{N \times D}$ represent a fixed sequence of learned latent tokens, where $N \ll M$. The cross-attention module projects these representations into queries, keys, and values through learnable linear weights $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{D \times d_k}$. The compression mapping is mathematically formulated as:

$$Z_1 = \text{Cross-Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.22}$$

where the queries are derived from the compressed space ($Q = \mathbf{W}_q Z_0$), and the keys and values are projected directly from the raw spike token sequence ($K = \mathbf{W}_k X$, $V = \mathbf{W}_v X$). Subsequently, a self-attention module is applied to the latent tokens, enabling the model to capture long-range temporal dependencies and complex population interactions across different neuronal units. Operating on the compressed sequence Z_l at layer l, the self-attention mechanism updates the latent representations with an $O(N^2)$ computational cost instead of the expensive $O(M^2)$ associated with raw spike sequences:

$$Z_{l+1} = \text{Self-Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.23}$$

where Q, K, V are all distinct projections of the same latent sequence Z_l (16), yielding a final contextualized latent tensor $Z_L = [z_{L,1}, \ldots, z_{L,N}]$ after L deep layers. To systematically encode relative timing relationships across both cross-attention and self-attention operations, Rotary Position Embeddings (RoPE) are integrated into the query and key spaces, replacing the standard affinity calculation with a localized rotation matrix $R(t)$ that enforces relative continuous-time dependency:

$$a_{ij} = \text{softmax}\left((R(t_i)q_i)^T(R(t_j)k_j)\right) \tag{2.24}$$

In the decoder, the refined latent representations are projected onto a sequence of output tokens through an additional cross-attention operation. A sequence of target output

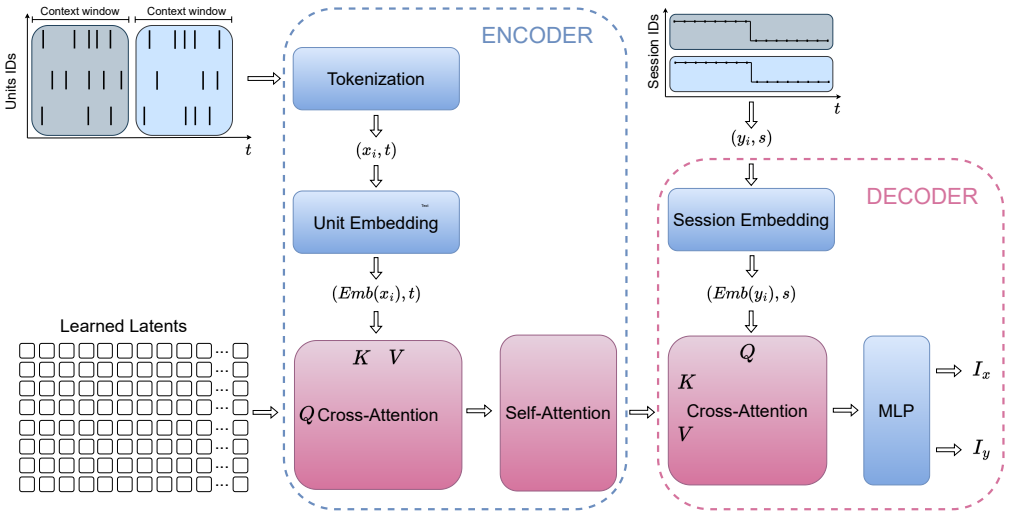

Figura 2.8: **Schematic of the POYO architecture adapted for retinal decoding.** The diagram illustrates the information flow. Input spikes (left) represent the neural activity, which the cross-attention mechanism compresses into a fixed set of learned latents (center). Self-attention blocks process these latents, and the final stage queries them to predict the temporal stimulus features (right).

tokens $Y_0 = [y_{0,1}, \ldots, y_{0,P}]$ is defined, where P corresponds to the number of temporal reconstruction query points, allowing the network to adapt flexibly to varying sampling frequencies across different recording platforms. These output tokens are integrated with session-level embeddings: learnable vector representations that capture the specific characteristics and hidden experimental variables of each independent recording block. This embedding injection accounts for inter-preparation variability, surgical day offsets, and distinct multi-electrode array properties, enabling a unified, multi-session decoding strategy across highly heterogeneous datasets. The readout queries the latent space by formulating the cross-attention interaction as:

$$Y = \text{Cross-Attn}(Q_y, K_y, V_y) = \text{softmax}\left(\frac{Q_y K_y^T}{\sqrt{d_k}}\right) V_y \tag{2.25}$$

where the behavior-driven queries Q_y are projected from the session-conditioned output tokens Y_0, and the keys and values are pulled directly from the refined encoder latent space ($K_y = \mathbf{W}_{k2} Z_L$, $V_y = \mathbf{W}_{v2} Z_L$). Finally, the decoded output tokens Y pass through a multi-layer perceptron (MLP) (18), which transforms these high-level, continuous-time representations into continuous-valued predictions corresponding to underlying stimulus features. The exact configuration of this final linear-projection head can be tailored to match the specific dimensional constraints of various visual paradigms: for full-field visual stimuli consisting of global luminance variations, the MLP condenses each frame into a single time-varying scalar value, whereas for complex spatiotemporal motion trajectories, the MLP maps the representations into a multi-dimensional output space to simultaneously predict positional coordinates (x, y) as a function of time. A schematic illustration of the POYO architecture is shown in Figure 2.8.

2.5.2 The POYO+ Extension: Multi-Task and Multi-Modal Architecture

Building upon the core spike-tokenization foundation of POYO (16), the POYO+ architecture (17) generalizes the framework to handle dense continuous time-series (such as two-photon calcium imaging $\Delta F/F$ traces) and joint multi-task, multi-modal decoding across diverse experimental contexts. While the original POYO tokenizes sparse, discrete point-process events (x_i, t_i), POYO+ adapts tokenization to regular, continuous neural activity traces across P recorded neurons and T timesteps. A sequence of $P \times T$ tokens is constructed where each token tuple (x_{ij}, t_{ij}) at timestep i for neuron j is defined as:

$$x_{ij} = [\mathbf{u}_j, \mathbf{W}_f f_{ij}] \tag{2.26}$$

where $f_{ij} \in \mathbb{R}$ represents the continuous scalar amplitude (e.g., fluorescence change), $\mathbf{W}_f \in \mathbb{R}^N$ is a learnable projection matrix, $\mathbf{u}_j \in \mathbb{R}^N$ is a static, learned unit-level embedding extracted from a lookup table, and t_{ij} represents a continuous 2D sinusoidal time embedding.

The primary architectural innovation of POYO+ lies in its decoder, which replaces single-target queries with a flexible cross-attention mechanism capable of decoding arbitrary continuous regression, sequence classification, and visual frame segmentation tasks simultaneously (17). To query the contextualized latent space Z_L, an output query vector o_{lk} for decoding task k in recording session l is constructed by additively combining two learned representations:

$$o_{lk} = \boldsymbol{\phi}_k + s_l \tag{2.27}$$

where $\boldsymbol{\phi}_k \in \mathbb{R}^D$ is a learned task embedding unique to the target variable (e.g., locomotion speed, pupil position, or visual orientation), and $s_l \in \mathbb{R}^D$ is the session embedding. At continuous output timestep t_i, the output token y_{lki} is retrieved via cross-attention with RoPE $R(t)$:

$$y_{lki} = o_{lk} + \sum_{n=1}^{N} \sum_{m=1}^{M} \text{softmax}\left(\frac{(R(t_i)\mathbf{W}_Q o_{lk})^T (R(\tau_{mn})\mathbf{W}_K z_{mn}^L)}{\sqrt{d_k}} \right) \mathbf{W}_V z_{mn}^L \tag{2.28}$$

Following the cross-attention block, a custom router feeds output tokens corresponding to task k into a task-specific linear projection head $\mathbf{W}_k : \mathbb{R}^D \to \mathbb{R}^{D_k}$, where D_k matches the target dimension of the task. Unlike the single-target MSE objective in standard POYO, POYO+ optimizes the network end-to-end using a weighted combination of task-specific loss functions:

$$\mathcal{L}_{\text{total}} = \sum_{k \in \text{Tasks}} w_k \mathcal{L}_k \tag{2.29}$$

where $\mathcal{L}_k$ corresponds to MSE for continuous behavioral regressions (e.g., running speed, pupil location) and Binary Cross-Entropy (BCE) for discrete classifications or segmentation frame predictions (e.g., drifting gratings, natural movies).

2.5.3 Justification and Core Advantages

While recent extensions such as POYO+ demonstrate remarkable capabilities for multi-task decoding across broad cortical areas using continuous signals, the fundamental objective

of this thesis requires decoding high-resolution electrophysiological spike trains recorded directly from RGCs. Because retinal information processing relies heavily on sub-millisecond spike timing and fast population transients, POYO+'s continuous time-series tokenization is less suited for raw point-process electrophysiology. Consequently, we select the original POYO framework as the primary decoding and interpretability engine for this research, driven by three core architectural advantages:

- **Sub-millisecond Point-Process Precision:** Unlike standard baseline decoders or binned transformer variants that discretize neural activity into arbitrary time bins, POYO's continuous-time event tokenization preserves the exact, sub-millisecond timestamps of action potentials. This temporal fidelity is mathematically essential for reverse-engineering adaptive retinal filtering kinetics under dynamic visual drive.

- **Sample Efficiency and Rapid Cross-Preparation Generalization:** Adapting the decoder to novel, unseen retinal preparations avoids the need for full architectural retraining or expensive hyperparameter sweeps. By leveraging learned session embeddings and freezing the foundational encoder, the model rapidly aligns its representations to new recording blocks via a few-shot fine-tuning regime, providing high adaptation efficiency across heterogeneous experimental datasets.

- **Emergent Biological Structure and Interpretability Bottleneck:** As established in the original POYO model and further validated through the scaling properties of POYO+, PerceiverIO-style cross-attention layers spontaneously capture underlying biological organization without explicit label supervision. The unconstrained, single-task attention bottleneck of POYO provides a direct mathematical window to audit internal attention allocation, compute entropy metrics, and validate the functional relevance of prioritized neural subpopulations via causal ablations.

In summary, the original POYO architecture provides the optimal balance of continuous point-process temporal precision, cross-session transferability, and mechanistic transparency, establishing it as the backbone model employed throughout the empirical investigations of this thesis.

2.5.4 Model Training and Implementation Details

To ensure reproducibility, we partitioned the dataset into training (70%), validation (10%), and test (20%) sets using non-overlapping temporal blocks. We implemented the architecture using the POYO framework with a latent dimension of 128, a latent sequence length of 64, and a depth of 6 layers. Both cross-attention and self-attention mechanisms utilized 8 heads each, with a head dimension of 64. To prevent overfitting, we applied a dropout rate of 0.3 across the feed-forward, linear, and attention layers. We trained the model to minimize the Mean Squared Error (MSE) between the reconstructed and ground-truth stimuli using the AdamW optimizer (initial learning rate 1×10^{-4}, weight decay 1×10^{-2}). Training lasted 100 epochs for single experiments and 400 for multiple experiments, with a batch size of 16, employing a scheduler that decayed the learning rate by a factor of 0.1 during the final 25% of the epochs.

2.6 Comparative Benchmarks

To systematically establish the performance boundaries of the point-process transformer paradigm evaluated in this thesis, the aforementioned baseline frameworks were rigorously implemented and contextualized within a unified evaluation pipeline. Rather than just serving as a historical literature review, these models provide the technical points of comparison against which the sparse tokenization of the POYO framework is benchmarked:

- **Optimal Linear Estimator (OLE):** Implemented via a continuous Ridge regression model targeting standardized stimulus coordinates. Multiunit spiking activity was discretized into non-overlapping temporal bins, and the analytical solution was computed using the closed-form normal equations.
- **Neural Data Transformer (NDT2):** Configured as a representation benchmark for modern binned transformers. The network featured a 4-layer stacked encoder optimized through a self-supervised masked modeling grid search using a 20% temporal "zero mask" ratio. Firing output projections were explicitly constrained to logspace representations via an exponential readout layer to stabilize training under Poisson negative log-likelihood constraints.
- **CEBRA:** Deployed as a metric-driven, non-generative manifold framework. The architecture was optimized using a contrastive InfoNCE criterion scaled by a learnable temperature parameter. Neighborhood structures were systematically evaluated across a parametric dimension sweep, contrasting discovery-driven temporal sampling with hypothesis-guided behavioral constraints to map the consistency of lower-dimensional neural projections.
- **Population State-Space Model (POSSM):** Structured as a generative baseline via standard Gaussian-Process Factor Analysis (GPFA). The joint multiunit observation fields were stabilized through variance-stabilizing square-root transforms, and the continuous trajectory manifolds were smooth-linked using a squared exponential covariance kernel. Model hyperparameters, specifically the independent channel noise variances and distinct dimension-specific timescales, were optimized iteratively using the Expectation-Maximization (EM) algorithm.

By evaluating this comprehensive landscape of neural data science tools this thesis directly contrasts the limits of closed-vocabulary, rate-binned decoders against the fluid, open-vocabulary point-process tracking capability offered by the POYO framework. Crucially, while these baseline architectures were originally established and validated using primate motor cortical recordings or hippocampal spatial maps, our work translates and benchmarks their conceptual formulations within the early visual pathways, evaluating their capacity to decode large-scale population dynamics recorded directly from the vertebrate retina.

Chapter 3

METHODS: INTERPRETABILITY ANALYSIS

Having established the landscape of neural decoding architectures and formalized the continuous-time, point-process POYO framework in the preceding chapter, the focus now shifts from model specification to internal interpretability. While achieving high predictive accuracy is fundamental, understanding how the transformer architecture allocates its computational resources to reconstruct sensory inputs requires inspecting its internal decision-making processes. Rather than evaluating the trained network as an opaque black-box regressor, this chapter presents the mathematical and analytical tools designed to audit its learned latent representations and dynamic attention mechanisms. Specifically, we detail the methodology for examining encoder latent space geometry, quantifying the temporal focus and synchronization of attention heads through Shannon entropy and non-parametric statistical divergence tests, establishing functional neural relevance via attention-guided causal ablations, and mapping latent-to-output interactions through decoder cross-attention coupling. Together, these interpretability frameworks set the methodological stage to discover whether generic attention mechanisms can spontaneously reverse-engineer the adaptive population coding strategies of the vertebrate retina.

3.1 Encoder Latent Representation

To gain insight into how the transformer architecture decodes the raw neural stimulus, we analyze the internal properties of the encoder's latent space. Specifically, we extract the representations immediately following the initial cross-attention layer, where continuous-time spike tokens interact dynamically with a fixed set of learned latent vectors. For any given 1-s context window containing N_i spike tokens, this operation maps the variable-length input sequence into a fixed-size latent matrix. When evaluated across the population, this process yields a high-dimensional tensor characterizing the model's localized feature extraction strategy.

To evaluate whether a unified data-driven optimization explicitly preserves biological cell-type identity within its shared latent space, we isolate the context windows from the

test sets of pure retinal ganglion cell subtypes: ON, OFF, and ON-OFF populations. The resulting specialized tensors are concatenated along the population dimension into a centralized high-dimensional matrix. Because raw latent representations reside in a highly abstract, high-dimensional space where direct visualization and clustering metrics are intractable, we implement a two-stage dimensionality reduction pipeline consisting of global linear projection followed by localized non-linear manifold mapping.

3.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an orthogonal linear transformation implemented as the initial stage of the reduction pipeline to handle the curse of dimensionality and filter out high-frequency noise while preserving global data variance. Let the concatenated encoder latent tensor be represented as a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n denotes the total number of extracted context windows and d represents the original dimensionality of the latent representation space ($d = 512 \times 128$). Assuming the data matrix $\mathbf{X}$ is centered such that its empirical mean vector satisfies $\mu = \mathbf{0}$, the empirical covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ is computed as:

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{3.1}$$

PCA aims to find a linear mapping matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$ (where $k \ll d$) that projects each row vector $\mathbf{x}_i$ into a lower-dimensional representation $\mathbf{z}_i \in \mathbb{R}^k$ by maximizing the variance of the projected coordinates. This optimization problem corresponds mathematically to finding the orthonormal eigenvectors associated with the largest eigenvalues of $\mathbf{\Sigma}$, satisfying the characteristic equation:

$$\mathbf{\Sigma}\mathbf{w}_j = \lambda_j \mathbf{w}_j, \quad \forall j \in \{1, 2, \ldots, k\} \tag{3.2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ denote the top k sorted eigenvalues, and $\mathbf{w}_j$ represents the j-th principal component. In this work, the latent matrix $\mathbf{X}$ is projected onto a 10-dimensional space ($k = 10$). This linear compression drastically reduces computational complexity for subsequent stages and stabilizes subsequent non-linear distance computations by maximizing the preservation of global covariance structures across RGC subpopulations.

3.1.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Following the linear variance maximization, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) to compress the 10-dimensional PCA projection into a lower-dimensional manifold suitable for visual stratification and neighborhood analysis. Unlike PCA, which prioritizes large distances to maintain global structures, t-SNE acts non-linearly to preserve localized structural topologies and affinities within the dataset.

Given the low-dimensional PCA vectors $\mathbf{z}_i \in \mathbb{R}^k$, t-SNE converts the Euclidean distances between data points into conditional probabilities that represent directional similarities. The asymmetric similarity of point $\mathbf{z}_j$ to point $\mathbf{z}_i$ is defined as the conditional probability $p_{j|i}$:

$$p_{j|i} = \frac{\exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{z}_i - \mathbf{z}_k\|^2 / 2\sigma_i^2)} \tag{3.3}$$

where σ_i represents the variance of the Gaussian distribution centered over $\mathbf{z}_i$, which is dynamically determined to satisfy a user-defined perplexity parameter. To handle

outliers and scale discrepancies, asymmetric affinities are converted into a symmetric joint probability distribution over the input space by setting:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{3.4}$$

In the low-dimensional target visualization embedding space $\mathbb{R}^m$ ($m \in \{2, 3\}$), the coordinates $\mathbf{y}_i, \mathbf{y}_j$ are mapped such that their low-dimensional affinity q_{ij} mimics the original distribution p_{ij}. To mitigate the crowding problem—where the volume of a sphere drops exponentially in lower dimensions—t-SNE employs a student-t distribution with one degree of freedom (which is equivalent to a Cauchy distribution) to model low-dimensional similarities:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} \tag{3.5}$$

The target coordinates $\mathbf{y}_i$ are systematically optimized by minimizing the discrepancy between the two probability distributions across all pairs, quantified by the Kullback-Leibler (KL) divergence objective function:

$$\mathcal{L}_{\text{KL}} = \sum_i \sum_{j \neq i} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) \tag{3.6}$$

This loss function is minimized via gradient descent techniques. Each point in the final optimized mapping corresponds to a single 1-s context window and is color-coded post-hoc according to its ground-truth physiological RGC classification (ON, OFF, or ON-OFF). This visualization paradigm enables an explicit, quantitative evaluation of whether cell-type-specific computational features are autonomously segregated by the transformer encoder without supervising label feedback during training.

3.2 Analysis of Encoder Attention

To quantify the contribution of individual retinal units, we analyze the cross-attention mechanism in the encoder. For a given context window i containing N_i spike tokens, the model computes an attention matrix $\mathbf{A}_{h,i}$ for each head $h \in \{1, \ldots, H\}$. These matrices are defined as:

$$\mathbf{A}_{h,i} \in \mathbb{R}^{L \times N_i} \tag{3.7}$$

where L is the fixed number of latent tokens. Each element $a_{j,k}^{h,i}$ represents the attention weight assigned by the j-th latent token to the k-th spike event for the h-th head. To obtain a spike-specific relevance score that is independent of the latent projection, we collapse the latent dimension by averaging the weights across all L tokens:

$$\bar{w}_{h,i,k} = \frac{1}{L} \sum_{j=1}^{L} a_{j,k}^{h,i} \tag{3.8}$$

To ensure that the importance scores are comparable across windows with varying spike densities and to prevent bias from high-firing intervals, we normalize the weights by the total spike count within the window:

$$\hat{w}_{h,i,k} = \bar{w}_{h,i,k} \cdot N_i \tag{3.9}$$

The resulting normalized weights $\hat{w}_{h,i,k}$ are used to color-code raster plots, where each spike k from unit u is plotted at its occurrence time t_k with intensity $\hat{w}$. This procedure allows for an objective comparison of how different heads prioritize specific neural subpopulations across different visual stimuli.

3.2.1 Entropy and Statistical Divergence

To quantify the temporal organization and focus of attention, we compute the Shannon entropy E for each attention head h within every 1-second context window i. Using the spike weights $\bar{w}_{h,i,k}$, the entropy (in nats) is defined as:

$$E(h,i) = -\sum_{k=1}^{N_i} \bar{w}_{h,i,k} \ln(\bar{w}_{h,i,k}) \tag{3.10}$$

This transformation yields a time-varying metric with a resolution of 1 Hz, where each value $E(h,i)$ is assigned to the timestamp corresponding to the end of its respective window. To determine whether individual attention heads developed specialized functional roles, we performed pairwise two-sample Kolmogorov-Smirnov (K-S) tests across all head combinations for both stimuli. This non-parametric test allowed us to assess significant differences in the cumulative distribution of entropy values, providing a robust measure of head divergence.

3.2.2 Causal Relevance and Unit Ranking

To assess the functional contribution of individual retinal units u to decoding performance, we define an importance score $I(u)$ based on the average attention weight assigned to its spikes. This score is calculated by averaging the spike-specific relevance scores across all attention heads and all spikes belonging to unit u:

$$I(u) = \frac{1}{HWN_i^u} \sum_{h\in H} \sum_{i\in W} \sum_{k\in N_i^u} \bar{w}_{h,i,k} \tag{3.11}$$

where W is the number of context windows with in the stimulus duration and N_i^u is the number of spikes of the unit u in the context window i. This metric identifies the neurons that the model consistently considers most relevant for the decoding task, independent of their total firing rate.

To validate the predictive relevance of this ranking, we performed a causal ablation analysis by comparing the decoding degradation under two exclusion strategies:

- **Attention-guided ablation:** Units are removed sequentially in descending order of $I(u)$.
- **Random control:** Units are removed in a stochastic order to establish a performance baseline.

For each removal step, we re-evaluated the model on the test set and computed the coefficient of determination R^2 to track the degradation in decoding integrity. For each stimulus type, we repeated this procedure for eight independent training runs (seeds).

3.2.3 Physiological Characterization of High-Attention Units

We further examined whether the units prioritized by the encoder share distinctive physiological properties. For each stimulus and each seed, we determined the number of units (K) whose cumulative removal reduced R^2 to 0.5 in the ablation curve, producing a top-K list per seed. To obtain a stimulus-specific set of consistently attended units, we aggregated the eight lists and retained only those units that appeared in at least half of the seeds. We then compared the distributions of light-response features between these consensus subsets and the global population of recorded units. To evaluate these differences, we employed a non-parametric statistical K-S test to detect differences in the overall shape and cumulative distribution of the features.

3.3 Analysis of Decoder Attention

To reconstruct the visual stimulus, the decoder employs a cross-attention mechanism that selectively retrieves information from the learned latent representations. For a given reconstruction sequence, the model computes a decoder attention matrix $\mathbf{D}_h$ for each head h. Unlike the encoder, which maps spikes to latents, the decoder operates on the output temporal grid:

$$\mathbf{D}_h \in \mathbb{R}^{M \times L} \tag{3.12}$$

where M is the number of output tokens (timestamps). Each element $d^h_{m,j}$ in the matrix represents the attention weight that the m-th output token assigns to the j-th latent vector. We visualize these matrices as heatmaps to observe the temporal dynamics of information integration.

3.3.1 Stimulus-Attention Coupling Analysis

To quantify the functional link between internal attention dynamics and stimulus reconstruction, we calculate the Pearson correlation coefficient r between the decoder entropy traces and the relevant external dynamics. The decoder entropy $E(h, m)$, of the hth head, measures the uncertainty in latent selection for each output token m. Using the weights $d^h_{m,j}$ from the decoder attention matrix $\mathbf{D}_h$, the entropy is defined as:

$$E(h, m) = -\sum_{j=1}^{L} d^h_{m,j} \ln(d^h_{m,j}) \tag{3.13}$$

This metric is computed at the sampling frequency of the reconstructed stimulus to capture high-resolution temporal dynamics. The correlation strategy then adapts to the specific structure of each task:

- **Flash:** We compute a single correlation value for each head by comparing its entropy trace directly with the light intensity profile.
- **Moving ball:** To account for spatial complexity, we decompose the trajectory into rectified Cartesian components $(+X, -X, +Y, -Y)$. For each attention head, we construct two directional correlation vectors: one representing positive displacements (X, Y) and another for negative displacements $(-X, -Y)$.

Chapter 4

METHODS: EXPERIMENTAL PROTOCOLS AND PERTURBATION

Having formalized the mathematical backbone of the POYO framework and defined the analytical tools required to inspect its inner attention layers in the previous chapter, the narrative now shifts from algorithmic abstractions to the empirical constraints of living tissue. This chapter serves as the experimental bridge of our research, detailly characterizing the large-scale multi-electrode array (MEA) recordings and functional light response structures obtained from isolated mouse visual pathways. Rather than evaluating our data under a fixed, static assumption, we unfold the methodological design behind our controlled code perturbations: progressive temporal jittering and spatial row permutations. By systematically injecting entropy into either precise spike-timing or physical retinotopic mappings, we establish a clean, predictive testing ground. This characterization provides the essential context needed to transition from theory into the upcoming chapter, where we will rigorously measure how distinct cellular channels handle structural code degradations under complex visual environments.

4.1 Electrophysiological recordings and Spike Sorting

We followed the complete electrophysiological recording protocol described in (19; 20). In brief, we used a MEA (USB256, Multichannel Systems GmbH, Reutlingen, Germany) with 252 electrodes and a sampling rate of 20 kHz to record RGCs in isolated retinal tissue. We stored all recordings on a computer for offline analysis. We obtained retinal tissue from 5 adult wild-type mice (3 males and 2 females, 3 months old). Before each experiment, we dark-adapted the animals for 30 minutes. Subsequently, animals were deeply anesthetized via inhalation of 2.5% isoflurane (Baxter, Deerfield, IL, USA) dissolved in O_2 (flow rate 300 mL/min in a 3L chamber) using a small animal anesthesia system (RWD Life Science Co., China), and immediately euthanized by decapitation. We quickly enucleated the eyes under

dim red light and immersed the eyecups in Ames medium with bicarbonate buffer (Sigma-Aldrich, St. Louis, MO, USA) at 32°C and pH 7.4, continuously oxygenated with 95% (O_2) and 5% (CO_2). We gently separated small pieces of retina from the retinal epithelium and placed them on a rod device supporting a dialysis membrane ring (MWCO-25000, Spectrumlabs, Rancho Dominguez, CA, USA), which we treated with polylysine (Product P4707, Sigma-Aldrich, St. Louis, MO, USA) to facilitate contact between the RGC side of the retina and the surface of the MEA electrodes.

We performed spike detection and classification using SpyKING Circus and SpikeInterface software (21). We retained units for further analysis only if their interspike interval (ISI) violation rate was less than 1.5%, their signal-to-noise ratio (SNR) exceeded 3 standard deviations, and they exhibited at least five trials with 10 or more spikes within the first 10 trials. This method ensures the reliability of unit classifications by stringent criteria. As responsiveness varied between the two visual protocols (detailed in Section 2.2), we obtained distinct unit counts for each. For the flash stimulus, we identified 512, 446, and 339 units for the males, and 373 and 339 for the females. For the moving ball stimulus, we recovered 490, 400, and 314 units for the males, and 377 and 366 for the females.

4.2 Light Stimuli

We generated and delivered light stimuli using MATLAB software and Psychtoolbox via a standard LED projector (PB 60G-JE, LG). We projected the images through a custom optical bench aligned with the photoreceptor layer. We calibrated spectral emissions at 460 and 520 nm using a USB4000 spectrophotometer (Ocean Optics Inc.) and measured the optical power at the sample plane using a Newport 1918-R optical power meter, yielding an average irradiance of 70 nW/mm^2. Finally, we displayed 400 x 400 pixel images spanning a 2 x 2 mm area on the MEA array under an inverted microscope (Eclipse T200, Nikon), where each pixel represented approximately 4 microns.

The stimulation protocol included two sets of visual stimuli. The first stimulus consisted of full-field flashes used to measure global luminance responses. The protocol comprised a 3-second flash ON followed by a 3-second flash OFF. We repeated this sequence for 10 trials.

The second stimulus consisted of a moving ball displayed against a black background. The ball was a cyan ellipsoid with a mean pixel intensity of 154.2 (scale 0–255). The ball subtended approximately 28.0×36.6 pixels throughout the presentation. We presented the stimulus for 20 seconds and repeated it 10 times. To ensure an irregular, persistent trajectory that activated different retinal regions, we generated the motion using fractional Brownian motion with a Hurst exponent of $H = 0.9$. The ball moved with a mean speed of 232 pixels/second.

4.3 Light Features

To characterize the light-evoked response properties of each RGC, we computed several indices: polarity, sustain index, and latency. We estimated a flash bias response to classify each RGC as ON, OFF, or ON-OFF based on their relative amplitudes, calculated as:

$$f_{BR} = \frac{f_{ON} - f_{OFF}}{f_{ON} + f_{OFF}} \tag{4.1}$$

where f_{ON} and f_{OFF} are the maximum spike counts during the ON or OFF parts of the flash stimuli. A value of 1 corresponds to a pure ON unit, a value of -1 corresponds to a

pure OFF unit, and 0 corresponds to a pure ON-OFF unit.
We calculated a sustain index (S_i) to evaluate the temporal profile of each response (sustained vs. transient):

$$S_i = \frac{f_{ON/OFF} - \bar{f}}{f_{ON/OFF} + \bar{f}} \tag{4.2}$$

where $\bar{f}$ represents the mean firing rate during the entire flash sequence and $f_{ON/OFF}$ is the total spike count during the ON or OFF parts of the stimuli. A sustain index of 1 corresponds to a pure transient response, while a value of 0 corresponds to a pure sustained response. Finally, we defined the flash latency as the time interval between stimulus onset (either ON or OFF phase) and the peak firing response. We also computed the mean firing rate for each unit as the total number of spikes divided by the total duration of the flash stimulus.

4.4 Spike Shuffling Analysis

To systematically evaluate the sensitivity of the pre-trained decoding model to both the fine-grained temporal precision and the macroscopic spatial organization of retinal population activity, we design and implement two distinct, parameterized shuffling procedures on the input spike rasters: *temporal shuffling* and *spatial shuffling*. Rather than treating noise as a binary condition, both perturbation strategies are implemented progressively. We define a perturbation parameter $p \in [0, 1]$, varied in discrete increments of 0.10 (10%), where $p = 0.0$ represents the pristine, unperturbed empirical recording (control baseline) and $p = 1.0$ represents complete randomization (maximum entropy).

By independently manipulating the temporal and spatial coordinates of the point-process spike tokens, this sensitivity analysis serves to isolate and identify the specific coding principles—such as synchronous firing, precise spike-timing, or localized retinotopical mappings—that the transformer architecture relies upon to reconstruct different visual features.

4.4.1 Temporal Perturbation and Jittering

Temporal shuffling is designed to evaluate the functional relevance of sub-millisecond spike timing and localized temporal correlations without altering the macroscopic firing rate or the global distribution of activity across the RGC population. For a given context window i of duration $T = 1$ s, let the spike train of a specific recorded unit u be defined as a set of timestamps $\mathcal{T}_u = \{t_1, t_2, \ldots, t_{N_u}\}$, where $t_k \in [0, T]$ and $t_k < t_{k+1}$.

When applying a temporal shuffle at a percentage p, a randomly selected subset containing $\lfloor p \cdot N_u \rfloor$ spikes is removed from $\mathcal{T}_u$. For each selected spike, its original timestamp t_k is replaced by a new timestamp t'_k drawn stochastically from a uniform continuous distribution across the entire window bounds:

$$t'_k \sim \mathcal{U}(0, T) \tag{4.3}$$

This operation preserves the exact row indexing within the raster and ensures that the total spike count N_u for each individual neuron remains completely invariant. At $p = 0.0$, the exact phase and spike-timing relationships are maintained intact. Conversely, at $p = 1.0$, the spike train is effectively transformed into a homogeneous Poisson process, removing higher-order temporal features such as inter-spike intervals (ISIs) and stimulus-locked transients.

This allows us to test whether the model utilizes the exact millisecond precision preserved by its tokenization framework or if it operates as a simplified rate-based integrator.

4.4.2 Spatial Reassignment and Retinotopic Disruption

Spatial shuffling is designed to isolate the importance of the spatial topology and arrangement of the recording units while keeping their individual temporal structures completely intact. This procedure explicitly disrupts the spatial relationships and structural dependencies across the multielectrode array layout.

Let the input raster within a context window be represented as an ordered set of parallel rows, where each row corresponds to a specific unit ID $u \in \{1, 2, \ldots, U\}$ associated with its respective temporal point process $\mathcal{T}_u$. To implement a spatial perturbation of magnitude p, we define a permutation mask. A subset of rows of size $\lfloor p \cdot U \rfloor$ is stochastically selected. We then apply a random bijective mapping $\pi : \mathcal{U}_{\text{sel}} \to \mathcal{U}_{\text{sel}}$ to permute the row positions within the raster matrix, effectively assigning the spike train $\mathcal{T}_u$ to a different unit embedding ID:

$$\mathcal{T}_u^{\text{shuffled}} = \mathcal{T}_{\pi(u)} \tag{4.4}$$

This spatial rearrangement leaves the fine-grained temporal structures, ISIs, and autocorrelation functions of individual neurons completely untouched. However, it destroys the topological mapping between the network's learned unit embeddings and the physical coordinates of the RGC receptive fields on the retinal tissue. At $p = 1.0$, the spatial coordinate map is completely randomized, meaning the network can no longer rely on localized population vectors or retinotopic coherence. This process reveals whether the decoder maps localized spatiotemporal features or if it operates on a spatially invariant, distributed population code.

Chapter 5

DECODING PERFORMANCE AND ROBUSTNESS ANALYSIS UNDER PERTURBATIONS

With the electrophysiological recording protocols and the algorithmic design of progressive temporal and spatial spike-shuffling perturbations fully established in the preceding chapter, we now translate our theoretical definitions into empirical evidence. This chapter marks a critical turning point in our narrative by evaluating how the tokenized transformer framework actually performs when confronted with the biological volatility of multiple distinct retinal preparations. By presenting raw decoding accuracy benchmarks against the classical and modern baselines reviewed at the beginning of our manuscript, we map out a consistent performance gap that validates our spike-tokenization approach. More importantly, we dive into the robustness of the pretrained models under structural code degradation, tracking how performance linear-decays or abruptly collapses when fine-grained temporal precision or macroscopic retinotopic mappings are systematically randomized, thereby isolating the true formatting principles that the network relies upon to track visual features. In doing so, this chapter directly addresses and satisfies Specific Objectives **O1 (Benchmark Predictive Accuracy)** and **O2 (Assess Cross-Preparation Generalization)**.

5.1 Decoding Performance and Adaptation across Retinal Preparations

To systematically evaluate the decoding capabilities of the point-process transformer architecture, we first establish a quantitative benchmark of reconstruction accuracy against classical linear estimators, state-space models, and modern binned transformers. We evaluated neural population recordings from five independent mouse retinas under two distinct regimes: single-session training to assess raw predictive fidelity, and a multi-retina pre-training setup followed by fine-tuning to evaluate cross-preparation generalization.

Table 5.1: Benchmark of decoding performance (R^2) across different machine learning architectures for both visual stimuli (Flash and Moving Ball). The comparison includes traditional linear baselines and state-of-the-art neural decoders.

Architecture	**Decoding Strategy**	**Flash (R^2)**	**Moving Ball (R^2)**
OLE	Optimal Linear Estimator	0.327	0.749
POSSM	State-Space Model	0.426	0.690
CEBRA	Contrastive Learning	0.821	0.744
NDT2	Binned Transformer	0.844	0.975
POYO	**Spike-Token Attention**	**0.994**	**0.978**

To address Specific Objective **O1**, we compare in Table 5.1 the stimulus reconstruction performance (R^2) of the POYO framework against established machine learning baselines across both sensory conditions. For the spatiotemporally complex moving ball trajectory, POYO achieves a high accuracy ($R^2 = 0.978$), outperforming the linear OLE baseline ($R^2 = 0.749$) and the generative POSSM state-space model ($R^2 = 0.690$). Under the full-field flash stimulus, POYO maintains near-perfect reconstruction ($R^2 = 0.994$), surpassing both the binned NDT2 transformer ($R^2 = 0.844$) and the contrastive CEBRA manifold model ($R^2 = 0.821$). These empirical benchmarks establish a consistent performance advantage for continuous-time spike-token attention over traditional rate-binned or linear decoding strategies.

A qualitative inspection of the decoded time series is presented in Figure 5.1, revealing the underlying dynamic behaviors behind these R^2 differences. For the full-field flash stimulus, POYO achieves instantaneous, lag-free step transitions at luminance boundaries while maintaining flat, noise-free trajectories during static plateaus. In contrast, linear and state-space baselines (OLE, POSSM) exhibit temporal hysteresis and baseline variance, whereas binned models (CEBRA, NDT2) suffer from transient edge instabilities and high-frequency decoding jitter. Similarly, for the continuous moving ball trajectory, POYO seamlessly tracks smooth 2D spatial excursions across both orthogonal axes without phase delays or peak amplitude clipping. Conversely, classical decoders severely attenuate peak displacements, and modern manifold approaches introduce high-frequency fluctuations during sharp direction reversals.

Table 5.2 details the decoding metrics across all individual experimental sessions and combined preparations. Across single-retina experiments, the model consistently reached R^2 scores above 0.98 for the flash stimulus and between 0.97 and 0.98 for the moving ball. To verify whether the captured population codes generalize across biological preparations, we evaluated the fine-tuning efficiency on a held-out retina. To verify whether the captured population codes generalize across biological preparations —fulfilling Specific Objective **O2** - we evaluated the fine-tuning efficiency on a held-out retina. As illustrated by the learning trajectories in Figure 5.2, the pre-trained multi-retina encoder required only a few fine-tuning epochs to match or exceed the performance of models trained from scratch on unseen preparations. This rapid adaptation confirms the realization of **O2**, demonstrating that the network extracts transferable, preparation-invariant representations of retinal population dynamics, allowing efficient adaptation to novel biological substrates without full model retraining.

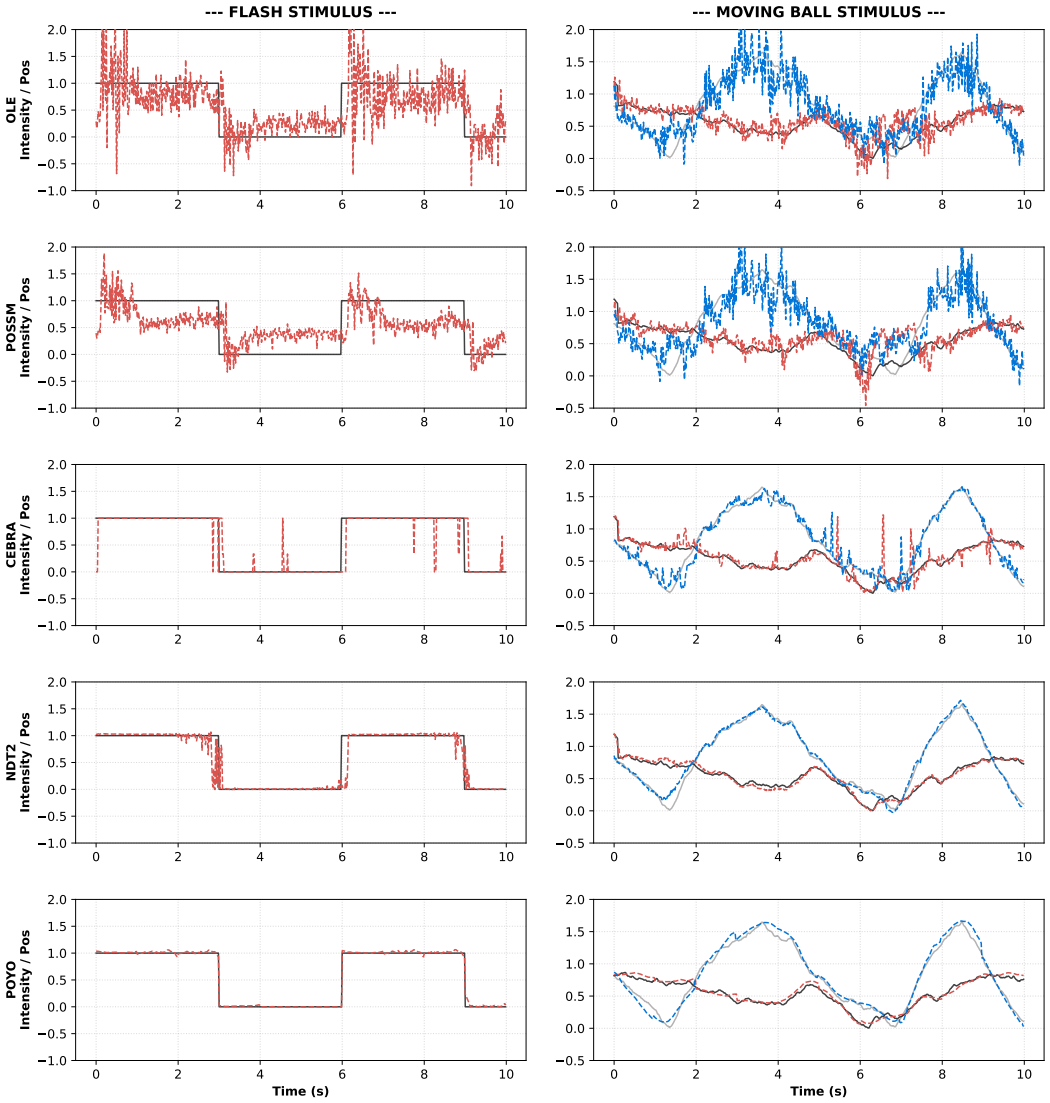

Figura 5.1: **Qualitative comparison of neural decoding architectures under visual stimulation regimes.** Reconstruction trajectories across five decoding paradigms arranged hierarchically from classical linear estimators to continuous-time transformers: OLE, POSSM, CEBRA, NDT2, and POYO. In the left column, the reconstruction of the full-field flash stimulus is shown. Continuous-time tokenization in POYO captures sharp luminance transitions without transient phase shifts or ringing artifacts, whereas binned approaches (NDT2, CEBRA) and linear baselines display high-frequency noise and delayed step edges. In the righth column the reconstruction of the moving ball two-dimensional trajectory (x-axis in red/black, y-axis in blue/gray) is shown. POYO tracks smooth spatial dynamics and peak amplitudes with high fidelity, while baseline decoders suffer from attenuation and high-frequency decoding jitter.

Table 5.2: Performance comparison across experiments for two visual stimuli: flash and moving ball. Metrics include coefficient of determination (R^2), number of units, and total spikes.

Experiment	**Flash**			**Moving Ball**		
	R^2	**Units**	**Spikes**	R^2	**Units**	**Spikes**
1st	0.994	512	219,908	0.978	490	644,050
2nd	0.997	446	313,084	0.971	400	557,731
3rd	0.993	399	270,732	0.974	377	656,751
4th	0.989	373	234,452	0.983	366	487,480
All	0.987	1,730	1,038,176	0.972	1,633	2,346,012
5th (fine-tuned)	0.995	339	197,143	0.889	314	348,988

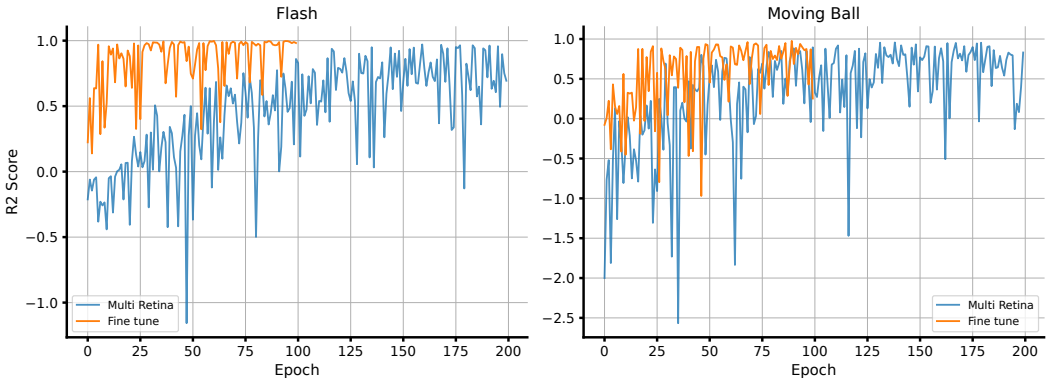

Figura 5.2: **Comparison of learning curves between fine-tuning and training from scratch.** The panels display the training trajectories for both stimuli (*left*: flash; *right*: moving ball), comparing the multi-retina pre-trained model against the fine-tuning process on a held-out retina. The pre-trained model requires only a few epochs of fine-tuning to match the performance of training from scratch, illustrating the presence of transferable representations that allow efficient adaptation to new preparations.

5.2 Impact of Temporal and Spatial Shuffling on Decoding Performance

To evaluate the structural integrity and code dependencies of the learned population representations, we tested the pre-trained models on input spike trains subjected to controlled, progressive shuffling perturbations applied directly to the rasters without network re-training. This sensitivity analysis systematically degrades two fundamental dimensions of the neural point process: (i) temporal shuffling, which disrupts sub-millisecond spike-timing precision by continuously randomizing spike timestamps within each unit independently, and (ii) spatial shuffling, which breaks retinotopic mapping and spatial organization by permuting unit identities across recording channels. For both visual paradigms (the temporally driving flash and the spatiotemporally structured moving ball), four model configurations were evaluated: a global model trained on the full heterogeneous population (All) and three models specialized on functionally isolated RGC subpopulations (ON, OFF, and ON-OFF).

Figure 5.3 provides a schematic overview of the shuffling pipeline alongside representative input raster windows under control conditions (a), maximum temporal randomization (b), and maximum spatial permutation (c). This illustrates how each procedure selectively disrupts one structural coordinate while rigorously preserving the global spike count statistics of the other.

The resulting decoding performance boundaries (R^2) under progressive levels of noise are presented in Figure 5.4. By plotting reconstruction accuracy against the shuffle percentage across stimuli and cell types, this analysis maps the functional resilience of the architecture and quantifies its reliance on fine-grained spatiotemporal neural structure.

5.2.1 Quantitative Effects of Temporal Randomization

As quantified in Figure 5.4a, the model architectures exhibit fundamentally distinct sensitivity profiles to temporal jittering depending on the complexity and spatial structure of the visual scene. For the full-field flash stimulus, precise spike timing represents the primary coding constraint for tracking global luminance changes. The global population model (All, red curve) maintains high decoding integrity ($R^2 > 0.90$) up to 20% noise, followed by an abrupt, non-linear threshold collapse between 20% and 40% shuffle, where performance plummets to its minimum baseline ($R^2 \approx -1.0$). This steep cliff-like degradation is closely mirrored by the ON (purple) and ON-OFF (yellow) subpopulations. Interestingly, the isolated OFF subpopulation model (blue curve) exhibits a shifted vulnerability profile under the flash condition, maintaining higher relative resilience up to 40% shuffle before undergoing a synchronized performance drop.

Conversely, the reconstruction of the moving ball trajectory demonstrates significantly higher robustness against pure temporal perturbations. Rather than suffering a sudden threshold collapse, decoding performance across the All, ON, and ON-OFF models decays in a highly predictable, gradual linear fashion as temporal entropy increases, retaining viable positive reconstruction capacities ($R^2 > 0.50$) even at a severe 60% shuffling level. The OFF cell pathway again represents a notable exception, showing an immediate and steeper susceptibility to temporal noise compared to its peers, emphasizing its specialized reliance on precise spike-timing vectors to resolve localized object motion.

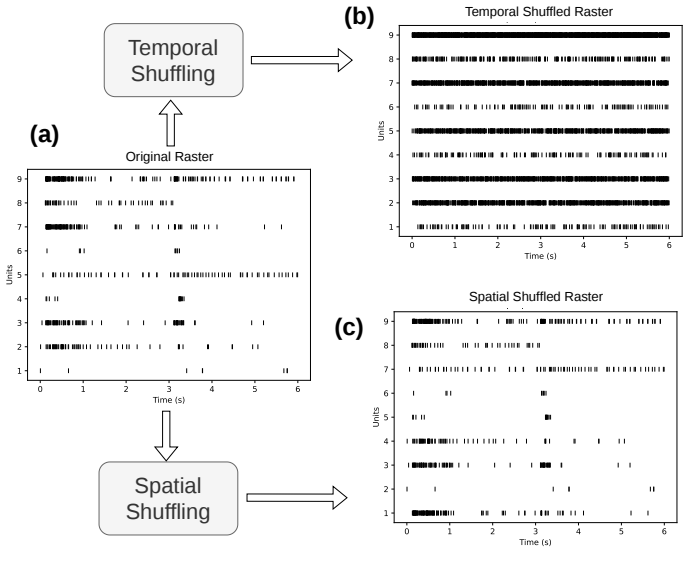

Figura 5.3: **Schematic pipeline of temporal and spatial shuffling perturbations.** (a) Original spiking activity across retinal ganglion cell units. (b) Temporal shuffling protocol, where spike times are stochastically permuted across the recording duration. (c) Spatial shuffling protocol, where unit identities are randomly swapped across the population.

5.2.2 Quantitative Effects of Spatial Disruption

The cross-over analysis of spatial row permutation, summarized in Figure 5.4b, reveals an inverted susceptibility landscape between the two visual paradigms. For the full-field flash stimulus, the spatial arrangement and physical coordinates of the recording units are largely redundant. The global population model and the isolated OFF subpopulation show virtually no performance degradation up to 50% spatial shuffle, maintaining an $R^2 \approx 0.90$ before gradually trailing off. Even under maximum spatial randomization (100% shuffle), the models retain partial decoding capacity, confirming that when spatial structure is absent from the stimulus, global temporal synchronization across the population is sufficient to bypass retinotopic mapping constraints.

In stark contrast, spatial shuffling proves catastrophic for the moving ball paradigm, where decoding performance depends heavily on the spatial layout and geometric arrangement of receptive fields. The specialized OFF subpopulation plummets immediately, dropping from an $R^2 \approx 0.97$ down to a non-functional baseline ($R^2 \approx 0.0$) at just 20% spatial shuffling. The All, ON, and ON-OFF models display a parallel, linear degradation profile, with decoding accuracy systematically breaking down as unit identities are randomized. At 100% spatial shuffling, all models approach a complete decoding failure ($R^2 \leq -0.50$), empirically validating that the transformer framework relies on stable, localized spatial configurations within its learned unit embeddings to successfully compute continuous direction and velocity vectors.

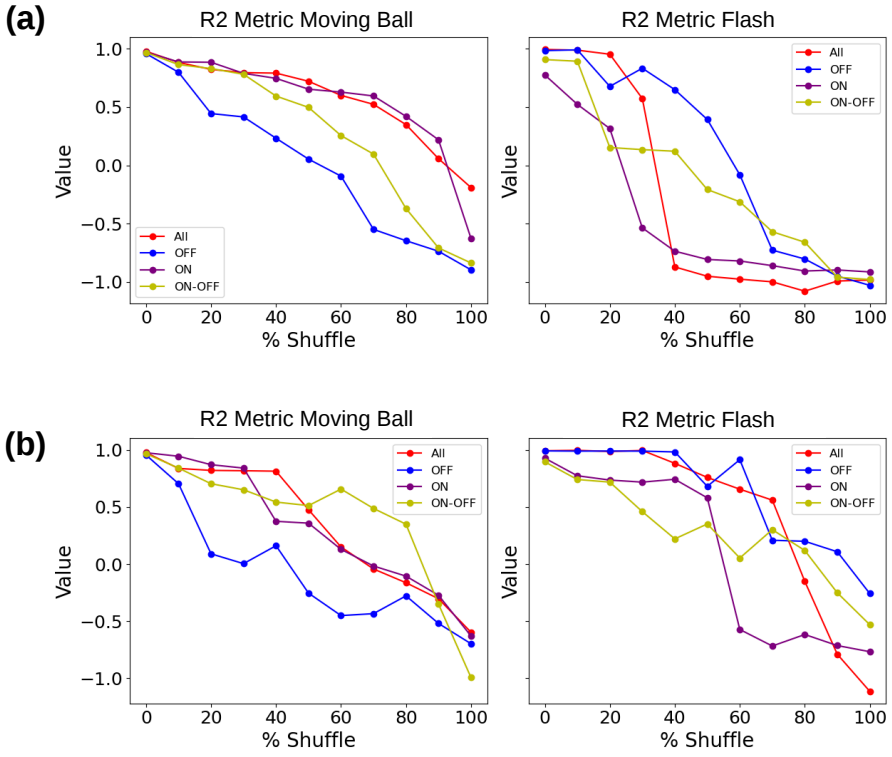

Figura 5.4: **Decoding performance degradation under progressive spike-train shuffling.** Impact of temporal (top row) and spatial (bottom row) shuffling on reconstruction accuracy (R^2) for the moving ball (left) and flash (right) stimuli as a function of perturbation percentage. Curves illustrate decoding resilience across different functional subpopulations: total recorded population (red), OFF units (blue), ON units (purple), and ON-OFF units (olive). A systematic decay in R^2 is observed across all conditions, demonstrating the model's reliance on precise spatiotemporal neural codes, with distinct sensitivity profiles emerging across cell types and stimulus complexities.

Chapter 6

ATTENTION DYNAMICS AND LATENT SPACE INTERPRETABILITY

The macroscopic performance benchmarks and sensitivity curves detailed in the previous chapter confirm that the transformer architecture excels at stimulus reconstruction, yet they leave its internal decision-making process unexplained. This chapter addresses that missing mechanical link by opening the trained network to look directly at the emergent biological structures captured within its layers. Moving beyond aggregate accuracy metrics, we guide the reader through an unsupervised exploration of the encoder's latent geometry, uncovering how functional retinal ganglion cell pathways spontaneously segregate into parallel computational channels as visual complexity scales from flat flashes to structured motion. By analyzing the time-varying focus of both encoder and decoder attention heads through entropy trace coupling and statistical divergence tests, we connect model attribution weights to distinct physiological signatures. Ultimately, through attention-guided causal ablations, we close the loop of our research, demonstrating that artificial attention serves as a principled proxy to understand how the living retina autonomously balances information capacity and parallel filtering demands. This chapter fulfills Specific Objectives **O3 through O6** by dissecting encoder attention dynamics, causally validating attribution rankings, mapping high-attention subpopulations to physiological profiles, and decomposing decoder cross-attention spaces.

6.1 Unsupervised Cell-Type Stratification within the Encoder Latent Space

To understand how the transformer backbone organizes neural population dynamics before generating continuous predictions, we examine the emergent geometry of the encoder latent space following the first cross-attention layer. We analyze context windows containing spikes from specific RGC subtypes (ON, OFF, and ON-OFF). This allows us to test whether a

unified objective function leads the model to naturally separate cell types based on their functional traits without explicit supervision.

For each visual paradigm, the latent matrix was extracted across test sequences, projected onto a 10-dimensional space using PCA to isolate dominant variance axes, and subsequently mapped to two dimensions via t-SNE to resolve localized manifold structures. Figure 6.1 illustrates these topological embeddings, where each data point represents a 1-second context window color-coded by its underlying functional cell class. Panels (a–c) show representative input rasters for pure ON, OFF, and ON-OFF subpopulations, while panels (d) and (e) compare the resulting latent space geometry across stimulus conditions.

6.1.1 Latent Structure Under Flash Stimulus

Under the full-field flash stimulus, which features global luminance shifts, the encoder latent space exhibits a continuous, highly integrated topology (Figure 6.1d). While pure ON (purple) and OFF (cyan) context windows occupy opposite sides of the mapping axis, they do not form completely isolated clusters. Instead, they remain connected by an overlapping intermediate region dominated by the ON-OFF subpopulation (yellow).

This configuration indicates that under uniform, non-spatial visual inputs, the network adopts a low-dimensional pooling strategy. Because a full-field flash activates the entire retinal array simultaneously, individual RGC types exhibit synchronized temporal transitions. Consequently, the cross-attention layers compress these firing patterns into a continuous, shared representation, reflecting functional redundancy across parallel pathways during uniform temporal stimulation.

6.1.2 Latent Structure Under Moving Ball Stimulus

A clear topological reorganization occurs when the model is evaluated on the moving ball paradigm (Figure 6.1e). As spatial boundaries and motion vectors are introduced, the latent distribution splits into a structured, multi-cluster manifold where functional cell types undergo clear, unsupervised segregation.

Context windows corresponding to pure OFF units (cyan) contract into a dense, isolated cluster that is detached from the rest of the population. This separation suggests that the encoder learns dedicated feature representations for tracking localized light decrements and trailing edges. Concurrently, pure ON units (purple) form an independent cluster on the opposite side of the embedding space.

In line with their biological role in detecting bidirectional transitions, ON-OFF units (yellow) form a bridge-like structure spanning the representational gap between the ON and OFF domains. This topological arrangement demonstrates that the transformer framework spontaneously mirrors the biological division of labor in the retina, mapping neural point processes into specialized computational channels when required by spatiotemporal stimulus complexity.

6.2 Encoder Attention Allocation Strategies under Different Stimuli

While the topological layout of the latent space confirms that cell-type identity is preserved during pretraining, it does not reveal the dynamic mechanisms used to extract those features. To address Specific Objective **O3**, the analysis shifts from static coordinate embeddings

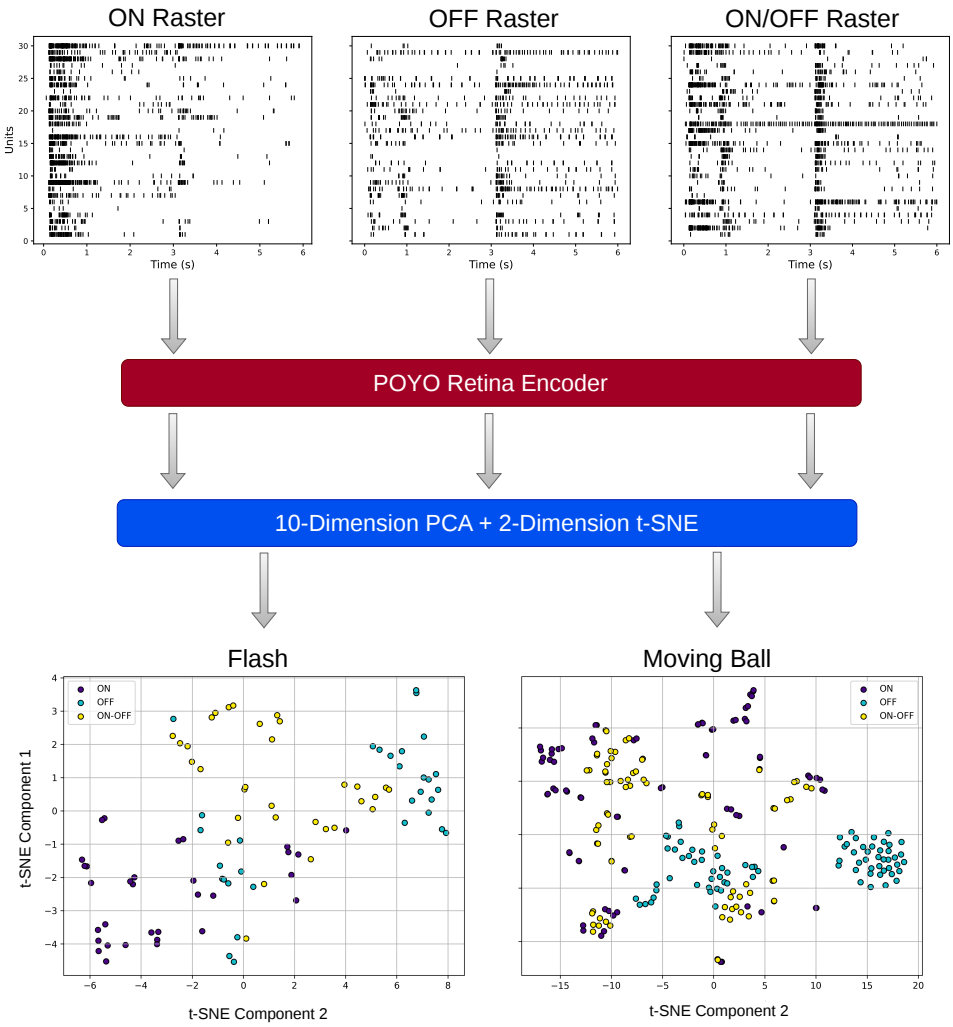

Figura 6.1: **Latent space visualization and dimensionality reduction pipeline of encoder representations.** Latent features from the POYO retina encoder are first aggregated across context windows and projected to a 10-dimensional space via Principal Component Analysis (PCA), after which a two-dimensional embedding is computed using t-SNE on the PCA scores. The resulting maps are shown for both stimuli (flash and moving ball), with points colored by RGC subtype (ON, OFF, ON-OFF).

to the time-varying focus of the encoder attention heads, measuring how the architecture dynamically weights individual retinal spikes when mapping inputs to the latent space. For each context window, we computed attention scores between latent tokens and individual spike tokens and normalized these values to allow fair comparison across varying spike counts.

Figure 6.2 displays raster plots where spike activity is color-coded by the normalized attention weights across all latents for a representative subset of four attention heads per stimulus. Under the uniform flash stimulus, we observed a broadly distributed attention pattern that lacks clear functional boundaries, reflecting the global temporal drive of the light intensity changes. By contrast, the moving ball stimulus triggers a more selective allocation strategy, where individual heads emphasize specific neuronal subsets and discrete time intervals that align with rapid changes in the trajectory.

To quantitatively characterize these dynamic patterns, we evaluated the temporal organization and information distribution of encoder weights through the Shannon entropy of attention distributions. As shown in Figure 6.3A, we analyzed the temporal evolution of attention entropy sampled at 1-second intervals across individual heads. For the flash stimulus, the entropy traces across all heads exhibit high temporal synchronization, with sharp, simultaneous drops in entropy corresponding to major luminance transitions. This demonstrates that the encoder adopts a unified, synchronized pooling strategy when processing global temporal shifts.

Conversely, the spatiotemporally structured moving ball stimulus elicits highly heterogeneous entropy profiles, where different attention heads independently adjust their temporal focus to capture distinct directional or kinematic features. This shift in functional specialization is further supported by the statistical distribution of entropy values (Figure 6.3B), where the broader variance and multimodal spread in the moving ball condition confirm that visual complexity forces the attention backbone to diversify its internal feature allocation strategies across heads.

Table 6.1: Pairwise comparison of attention head entropy distributions. This matrix combines p-values from two-sample Kolmogorov-Smirnov tests. The upper diagonal (blue) displays results for the moving ball stimulus, while the lower diagonal (orange) corresponds to the flash stimulus. Values in bold indicate $P < 0.05$.

		Heads							
		H0	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**
Heads	**H0**	-	0.426	**1.06e-07**	0.093	**0.035**	1.00	0.716	**1.20e-04**
	H1	**0.008**	-	**3.33e-09**	**0.001**	**0.003**	0.215	0.426	**1.65e-05**
	H2	0.536	**0.008**	-	**1.65e-05**	**0.003**	**3.33e-09**	**1.20e-04**	0.716
	H3	0.099	**0.008**	0.869	-	**0.035**	0.093	0.093	**1.20e-04**
	H4	**0.001**	**2.04e-04**	**0.008**	**0.008**	-	**0.035**	0.215	**0.003**
	H5	0.536	**0.008**	0.998	0.536	**0.008**	-	0.716	**1.20e-04**
	H6	0.869	0.099	0.099	**0.008**	**0.001**	0.099	-	**0.003**
	H7	0.536	**0.008**	0.536	0.869	**0.031**	0.536	0.256	-

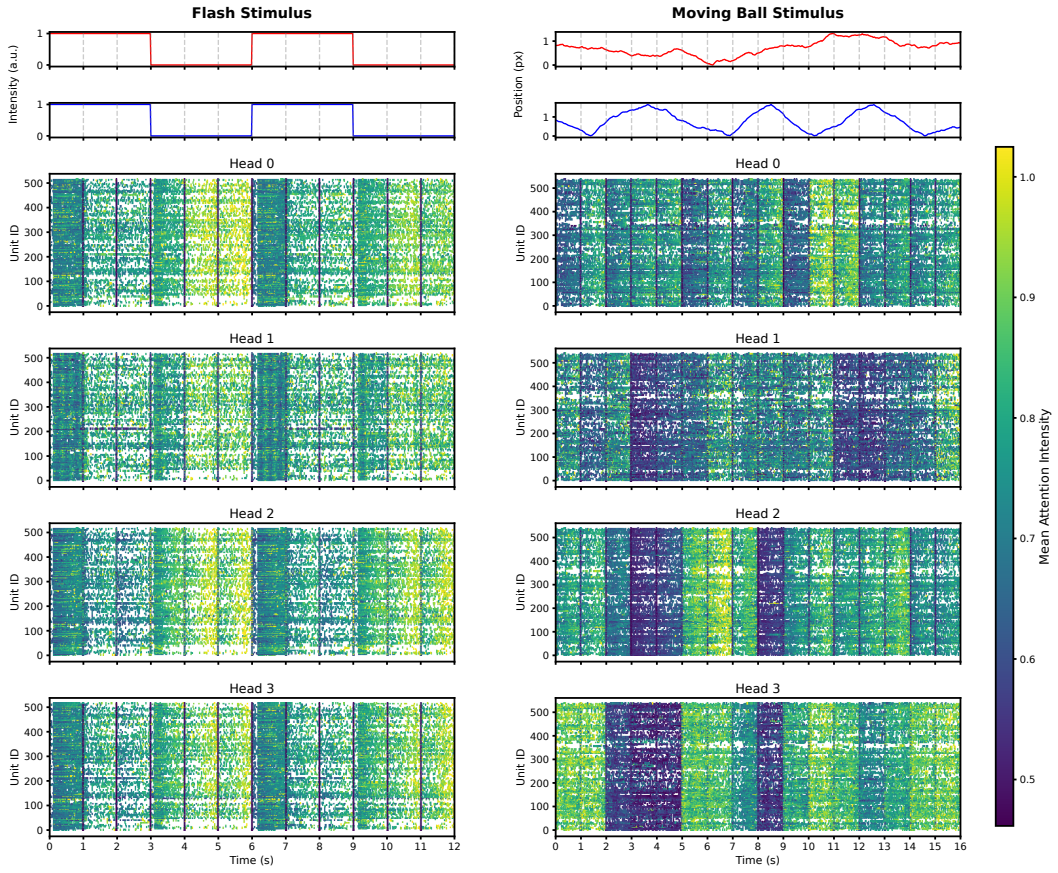

Figura 6.2: **Visualization of encoder attention allocation during flash and moving ball stimuli.** The raster plots display retinal spike activity color-coded by the normalized attention weights, averaged across the latent dimension for a representative subset of four attention heads per stimulus. (Left) Under the uniform flash stimulus, attention patterns appear temporally synchronized across heads, consistently prioritizing stimulus intervals aligned with global luminance changes. (Right) In contrast, during the moving ball stimulus, attention patterns exhibit higher heterogeneity across heads and time, with high-attention bands aligning with rapid changes in the spatiotemporal dynamics of the trajectory. The top panels for each column illustrate the ground-truth stimulus features: light intensity for the flash (left) and the x and y positional coordinates for the moving ball (right).

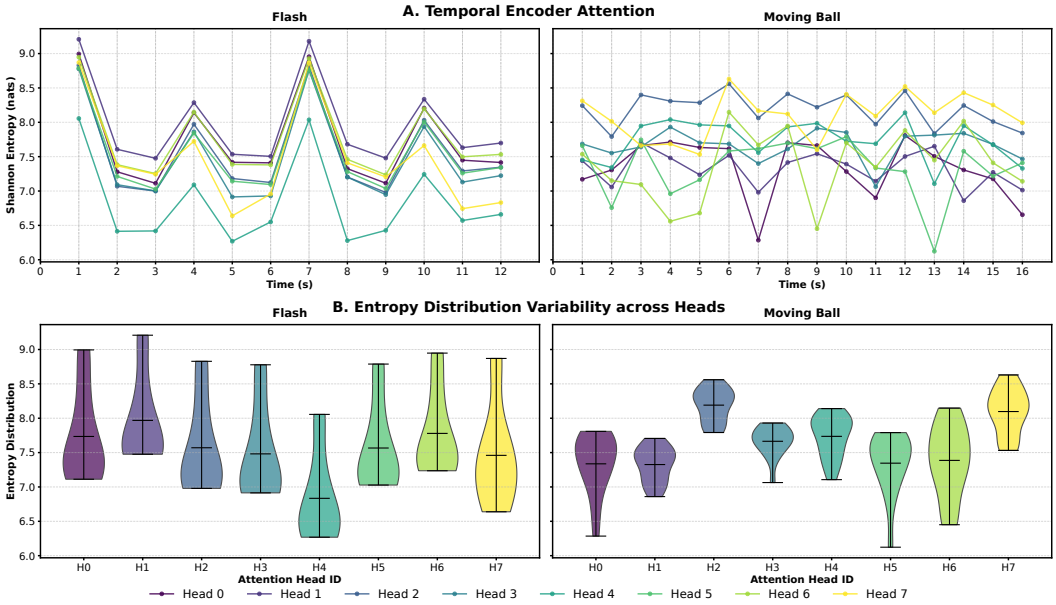

Figura 6.3: **Statistical analysis of encoder attention entropy.** (A) Temporal evolution of Shannon Entropy for all attention heads during the flash (left) and moving ball (right) stimuli. In the flash condition, entropy traces show high synchronization and sharp drops aligned with global luminance changes. In contrast, the moving ball stimulus induces heterogeneous entropy profiles, reflecting specialized attention allocation across heads. (B) Distribution of entropy values across individual heads. The violin plots illustrate the variability and range of attention focus, where diverse distributions in the moving ball condition highlight functional specialization among heads compared to the more uniform behavior observed during the flash stimulus.

We validated these observations through a pairwise statistical comparison summarized in Table 6.1. By consolidating the results into a dual-diagonal format, we contrasted the head specialization between both experimental conditions: the lower diagonal represents the flash stimulus, while the upper diagonal shows the moving ball results. The K-S test reveals that while a substantial portion of head pairs maintain redundant distributions under the flash stimulus ($P > 0.05$), the moving ball stimulus induces a profound functional divergence. Nearly all comparisons for the complex motion reached extreme significance levels (e.g., $P < 10^{-9}$ for Head 2). This marked difference in both the scale and frequency of statistical significance confirms that while global stimuli maintain a largely redundant processing regime, structured motion triggers a highly specialized and heterogeneous recruitment of attention heads.

6.3 Causal Relevance and Physiological Properties of High-Attention Units

6.3.1 Causal Validation of Predictive Neuron Rankings via Guided Ablations

In accordance with Specific Objective **O4**, we evaluated the causal relevance of the learned attention weights by examining the overlap among the top K attended units across eight independent training seeds per stimulus. We observed that several units were consistently selected across runs, indicating stable attention patterns and confirming that the model repeatedly identifies a core subset of neurons critical for decoding. To validate this hypothesis causally, we assessed how decoding performance (R^2) changed as we systematically removed these high-attention units compared to a stochastic, random removal baseline.

Figure 6.4 shows the R^2 values as a function of the number of removed units for both the attention-guided strategy and the random control. For both stimuli, we found that performance decreased sharply when we excluded the most-attended units, whereas the random removal of units resulted in a significantly slower decay in decoding quality. This gap confirms that attention weights accurately capture neurons with high predictive relevance rather than simply reflecting a general dependence on input density. We noted that this effect was more pronounced for the flash stimulus, where fewer units contributed disproportionately to performance. In contrast, in the moving ball condition, performance remained more robust to initial removals, indicating that information is distributed across a broader neuronal population.

6.3.2 Physiological Signatures of Consistently Prioritized Subpopulations

To fulfill Specific Objective **O5**, we characterized the functional profiles of the neurons selected by the encoder. We compared the joint distributions between the bias index and three light-response features (sustain index, latency, and firing rate) for the overall recorded population and for the consensus sets of most-attended units in each stimulus. For every seed, we identified the top-K units as those whose removal reduced R^2 to 0.5 in the ablation analysis; we then aggregated these lists across seeds and retained only units appearing in at least half of the seeds to yield stimulus-specific subsets of consistently attended neurons.

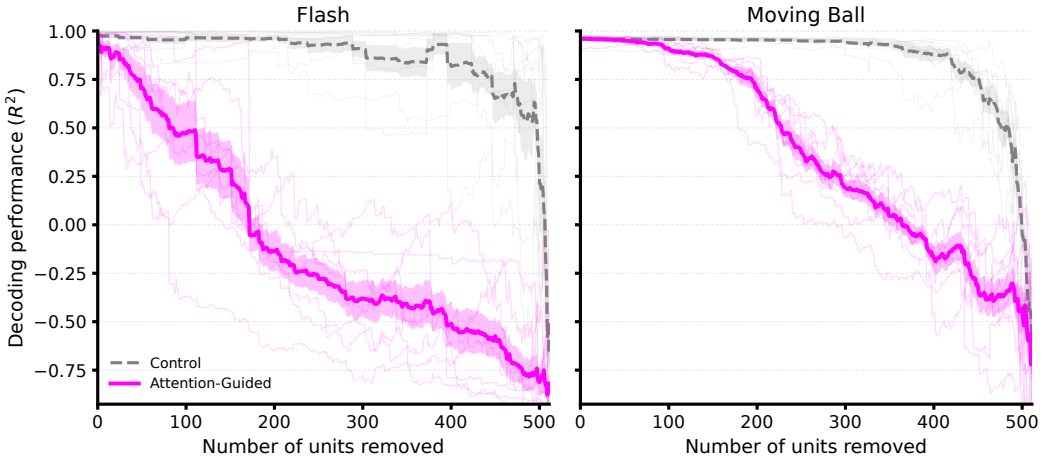

Figura 6.4: **Impact of progressive ablation of high-attention units on decoding performance.** Each panel displays the R^2 scores as a function of the number of units progressively removed from the input population for the flash (left) and moving ball (right) stimuli. The solid magenta curves illustrate the attention-guided ablation, where units are removed in descending order of their cumulative encoder attention weights averaged across independent seeds (n=8). The dashed gray curves represent the random control, where units are removed in a stochastic order, with the surrounding shaded gray area indicating the variability across multiple iterations. For both stimuli, the performance decline is significantly more rapid under attention-guided removal compared to the control, confirming that the attention mechanism effectively prioritizes neurons with high predictive relevance for stimulus reconstruction.

Table 6.2: Statistical analysis of physiological features in high-attention units. The table presents a comparison of light-response features between stimulus-specific high-attention subsets and the global retinal population using the two-sample Kolmogorov-Smirnov test. Values in bold indicate $P < 0.05$.

Comparison	**Bias Index**		**Sustain Index**		**Latency**		**Firing Rate**	
	D	P-value	D	P-value	D	P-value	D	P-value
Flash vs. All	0.142	0.476	0.238	**0.041**	0.127	0.617	0.248	**0.029**
Ball vs. All	0.048	0.830	0.047	0.847	0.072	0.361	0.039	0.955
Flash vs. Ball	0.130	0.629	0.201	0.150	0.103	0.872	0.255	**0.031**

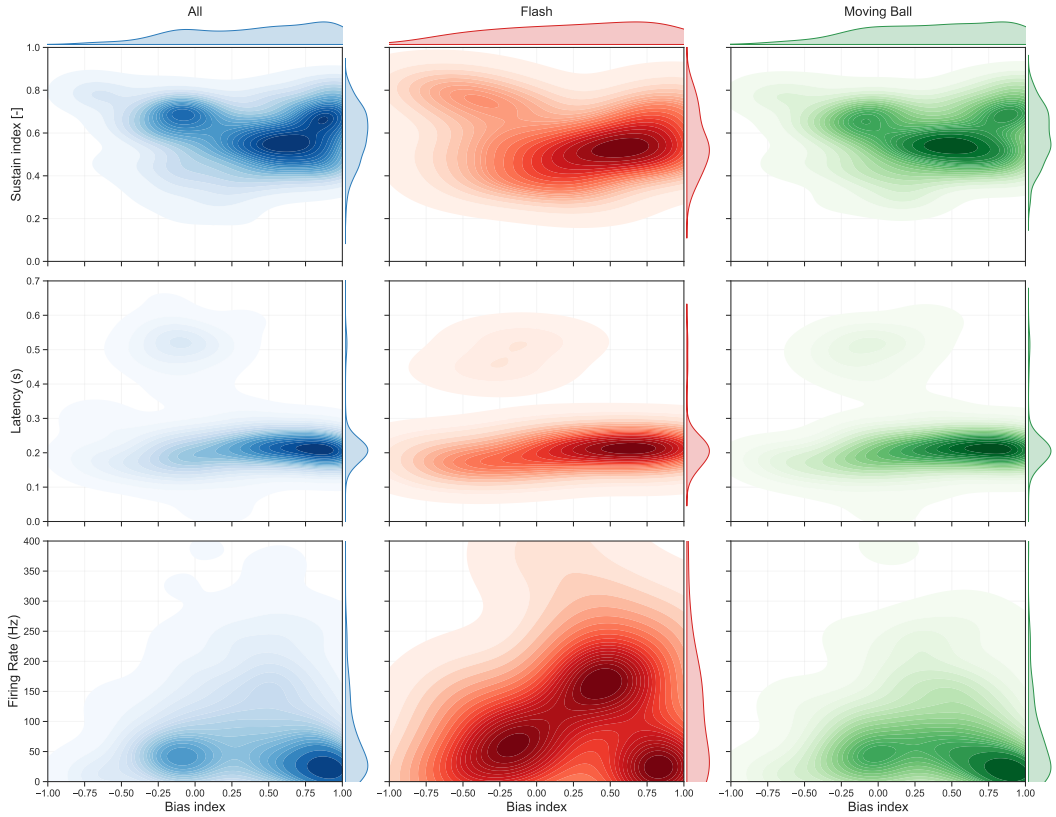

Figura 6.5: **Joint distributions of physiological properties for the global population versus high-attention units.** The panels display joint kernel density estimates (KDE) of three key response features—sustain index, latency, and firing rate—plotted as a function of the bias index. The plots compare the distributions for the global population of recorded units (blue, $n = 512$) with the subsets of units most strongly attended by the model under the flash (red, $n = 98$) and moving ball (green, $n = 230$) stimuli. The attended subsets comprise unique neurons that appeared in the top-attended sets of at least four of the eight independently trained models.

Figure 6.5 shows the resulting two-dimensional KDE distributions, with the corresponding statistical analysis detailed in Table 6.2. Under the flash stimulus, we found that the most-attended neurons display a distinct physiological profile characterized by significantly higher sustain indices and elevated firing rates compared to the global population. While the KDE visualizations suggest concentration modes in the bias–feature planes, the statistical comparisons for bias index and response latency revealed that these features remain indistinguishable from the overall recorded population. For the moving ball stimulus the attended units exhibit physiological distributions that closely follow the global population across all measured features. As shown in Table 6.2, no significant differences were found for the bias index, sustain index, latency, or firing rate in the moving ball condition. Furthermore, a direct comparison between the two stimuli confirms that the encoder selects units with significantly different firing rate distributions depending on the visual task. Together, these trends indicate that while the model leverages specialized, high-activity neurons to decode uniform global changes like the flash, it relies on a more diverse and representative sample of the retinal population to reconstruct complex motion-rich inputs.

6.4 Decoder Attention Patterns and Latent-Output Interactions

Finally, to resolve how the architecture transforms internal latent representations into continuous output reconstructions, we analyzed the decoder's cross-attention mechanism. For each stimulus type, we extracted attention matrices across all eight heads and visualized them as heatmaps with latent projection space indices on the vertical axis and output timepoints on the horizontal axis. Figure 6.6 presents these results alongside the stimulus trajectories for a representative subset of four attention heads per condition. In the flash condition, we observed that decoder attention is organized in repetitive, band-like patterns aligned with luminance changes. This structure suggests a uniform retrieval strategy where the model queries the latent space in synchronized intervals to track global intensity. In contrast, the moving ball condition yields more heterogeneous and distributed patterns. We noted that while some heads focus on narrow latent subsets during rapid position changes, others distribute attention more diffusely, reflecting a more complex integration of spatial and temporal dynamics.

Addressing Specific Objective **O6**, we analyzed the Shannon entropy of the decoder attention distribution. As shown in Figure 6.7, for the flash stimulus, all eight attention heads show entropy traces that are tightly coupled to the light intensity, resulting in consistently high negative Pearson correlation coefficients ranging from $r = -0.77$ to $r = -0.93$. This strong correlation confirms that the decoder systematically increases its focus (reducing entropy) whenever a significant luminance transition occurs.

In the moving ball condition, however, the relationship between attention entropy and stimulus position reveals a specialized functional organization across heads. We observed that attention heads exhibit distinct directional tuning preferences: certain heads show strong negative correlations specifically for motion in the lower visual field (e.g., Head 3), while others are tuned to horizontal or upper-quadrant trajectories. These quantitative results indicate that the decoder does not merely track the stimulus globally; instead, it adopts a spatially selective, head-specific strategy where individual attention channels specialize in monitoring specific regions of the visual field to reconstruct complex trajectories.

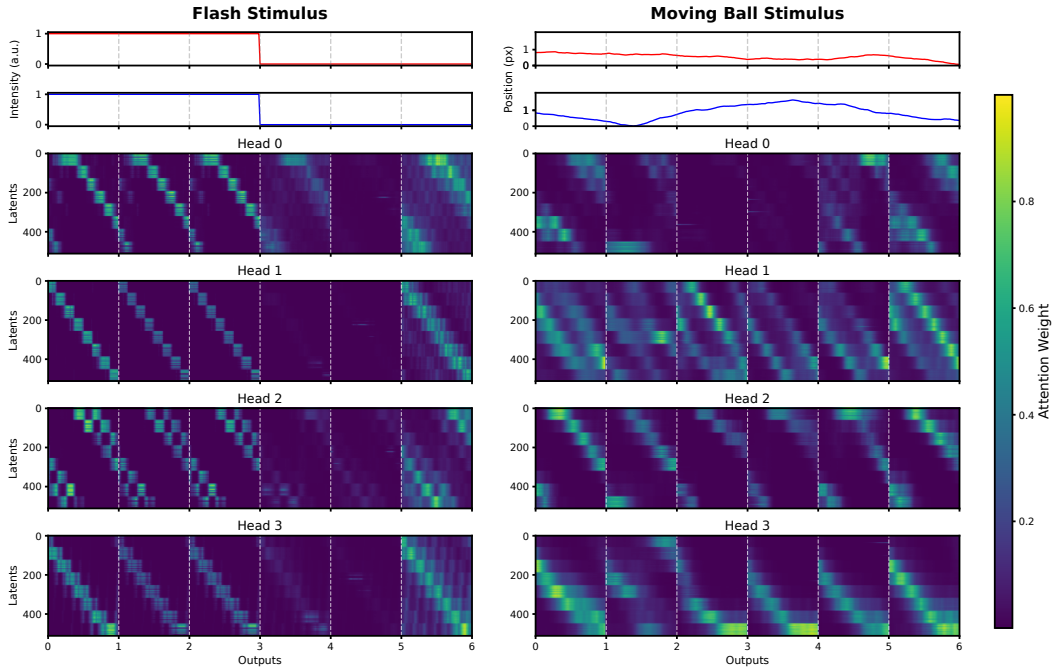

Figura 6.6: **Decoder cross-attention maps linking latent representations to output predictions.** The heatmaps display cross-attention patterns between latent projection space and output tokens for the two evaluated stimuli, using a representative subset of four attention heads per condition. Attention weights indicate how strongly each output token queries specific latent vectors. (Left) For the flash stimulus, characterized by purely temporal variations, attention is organized in repetitive, band-like patterns aligned with luminance changes. (Right) In contrast, for the moving ball stimulus, the attention maps exhibit a more heterogeneous and distributed structure across heads and time, reflecting the processing of combined spatial and temporal dynamics. The top panels for each column illustrate the ground-truth stimulus features: light intensity for the flash (left) and positional coordinates for the moving ball (right).

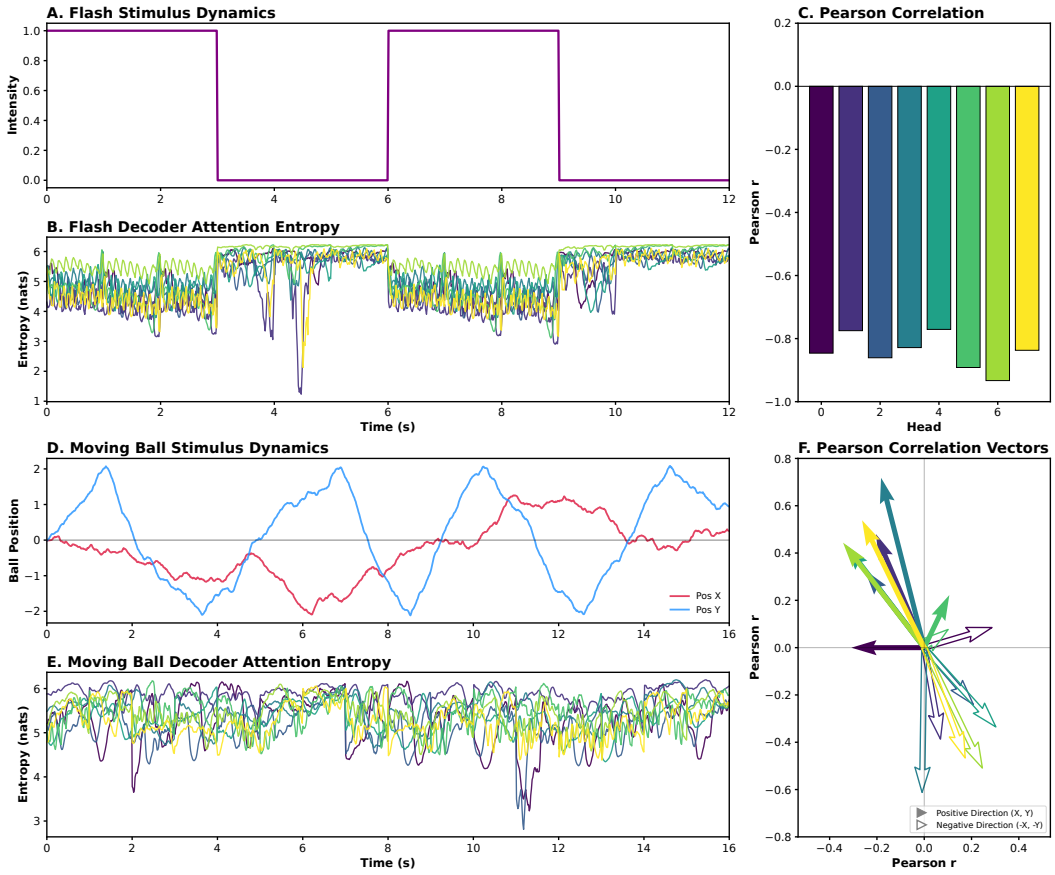

Figura 6.7: **Quantitative analysis of decoder attention entropy and stimulus coupling.** (A) Temporal dynamics of the flash stimulus light intensity. (B) Evolution of Shannon Entropy for the eight decoder attention heads during the flash sequence. (C) Pearson correlation coefficients for all heads under the flash condition. (D) Spatiotemporal dynamics of the moving ball stimulus, showing horizontal (x) and vertical (y) positions alongside the calculated eccentricity. (E) Evolution of Shannon Entropy for the decoder heads during the moving ball sequence. (F) Pearson correlation vectors for each head. Solid arrows represent tuning toward positive direction, while hollow arrows indicate tuning toward negative direction.

Chapter 7

DISCUSSION

In the present research, we evaluated the POYO architecture not only as a high-performing neural decoder but also as a potential tool for gaining insights into retinal coding. While we observed that the model consistently achieved high decoding accuracy across retinas and stimuli, our primary focus was to assess whether its internal mechanisms could provide interpretable structure that aligns with known neurophysiology. Ultimately, our findings support the central hypothesis of this work, demonstrating that transformer-based attention dynamics and continuous spike tokenization spontaneously recover fundamental biological principles of retinal population coding without explicit physiological supervision.

First, the fine-tuning procedure makes the model rapidly adapt to new retinas with only a few epochs, suggesting that POYO extracts transferable latent structures that capture general principles of retinal encoding rather than merely memorizing training data. Such flexibility is critical given the inherent volatility of biological sensory systems, where fluctuations in synaptic strength and neural noise can severely limit the information transmission of static, optimal linear decoders (22; 23).

Second, by utilizing a dynamic attention mechanism, our findings suggest that POYO's ability to rapidly align its internal representations allows it to bypass the SNR saturation typically found in volatile biological regimes, effectively mitigating the information loss imposed by such instability (22). Furthermore, by utilizing spike-tokenization instead of the traditional temporal binning found in architectures like NDT2, POYO preserves the sub-millisecond precision necessary to decode complex spatiotemporal dynamics. This enables the identification of an invariant-like manifold of neural activity that remains robust across different retinal preparations. Consequently, our approach shows that attention-like dynamics enable flexible population coding, bridging visual attention mechanisms and overcoming the limits of fixed optimal decoding under synaptic coarse-tuning (22; 24).

To further investigate the nature of these learned features, we examined how the model's internal priority, represented by its attention weights, aligns with known physiological response patterns. Specifically, the encoder attention analysis reveals that the model spontaneously recovers biological coding strategies by directly mapping individual spikes to latent representations. For the flash stimulus, we observed that attention appeared broadly distributed, consistent with the global and uniform nature of luminance-driven responses

(25). The highly synchronized entropy traces observed across all attention heads support this view, indicating that the model recruits the neural population into a redundant, unified representation for uniform inputs. Although the high sensitivity of the K-S test detects minor variations in these distributions, the resulting p-values remain relatively close to the significance threshold, confirming the absence of strong functional divergence.

In contrast, the moving ball stimulus induced highly heterogeneous entropy profiles, marked by a profound statistical divergence between attention heads ($P < 10^{-9}$). We interpret this massive statistical separation in entropy levels as evidence of a functional 'division of labor', where the heads desynchronize to capture the diverse spatiotemporal features—such as position and motion direction—inherent to complex trajectories (26; 27). Such transition from synchronized to desynchronized states demonstrates how the transformer's internal strategy dynamically adapts its computational complexity to the structure of the visual input (14; 2). This physiological alignment and ablation analysis support a key premise of our hypothesis: that the priority assigned by the encoder attention heads is not merely an optimization artifact, but reflects the functional relevance and predictive contribution of specific neural subpopulations within the decoding architecture.

Specifically, we propose that this mechanism functionally emulates dynamic gain control (28) and lateral inhibition mechanisms that characterize retinal circuits. By desynchronizing its attention, the model effectively suppresses global redundancy to prioritize specific, non-redundant visual features, a hallmark of efficient sensory coding. This interpretation is consistent with the findings of (29), who demonstrated that retinal ganglion cells act largely as independent encoders, with more than 90% of stimulus information being recoverable even when correlations are ignored. Our results suggest that by emulating these biological filtering strategies, the POYO framework spontaneously adopts a multi-channel weighting strategy that exploits this neuronal independence to resolve complex spatiotemporal dynamics. Nevertheless, we acknowledge that attention does not guarantee causal relevance, and high weights may also reflect correlation, redundancy, or shared stimulus drive (30). Thus, although encoder attention provides interpretable signals, we maintain that its biological significance requires further validation (31; 32).

To address this gap, we performed an attention-guided ablation analysis and quantified the seed-level stability of the top-ranked units across eight independent training runs. The recurrence of specific high-attention units across seeds confirms that the model consistently converges on a robust subset of informative neurons. Our results demonstrate that progressively removing units in attention rank order led to systematic and stimulus-specific performance degradation (Figure 6.4). In the flash stimulus, decoding performance declines rapidly and nearly monotonically, revealing a concentrated coding where a few units carry the majority of the signal (33). For the moving ball, performance remains robust despite initial removals, consistent with a distributed and functionally diverse representation where information is shared across a broader neuronal population (8). Importantly, we emphasize that this relationship between attention and functional relevance is not trivial. Many high-performance neural decoders rely on complex internal representations where attribution scores do not necessarily correspond to causal contribution (24; 31). Here, we provided the missing causal link via the ablation analysis: units that received consistently high attention were precisely those whose removal most impaired decoding. This causal validation strengthens the case for using attention as a principled guide to neuronal relevance, allowing us to investigate whether these prioritized units share distinctive physiological properties.

The statistical comparison of light-response features between the global population and high-attention subsets reveals distinct selection profiles that depend on the visual

task(Figure 6.5). Under the flash stimulus, the encoder's prioritization of a subset with significantly higher firing rates and a shift toward sustained temporal responses suggests a strategy aimed at maximizing total information capacity. This focus on temporal persistence allows the model to track strong global luminance modulations, exploiting the fact that higher discharge frequencies are associated with greater absolute information rates (34; 35). A different strategy emerges for the moving ball stimulus, where high-attention units do not deviate from the overall population in terms of bias, sustainability, firing rate, or latency diversity. We interpret this reliance on the population's intrinsic diversity as an adaptation to the greater spatial structure and directional requirements of complex motion, where a more heterogeneous population is necessary (36; 37). Collectively, these results indicate that the encoder allocates attention based on the spatiotemporal demands of the task by balancing the trade-off between absolute information rate and coding efficiency (35). This suggests that POYO's attention mechanism can autonomously reflect efficient computational principles of the vertebrate retina, shifting its selection criteria to match the information-theoretic demands of the visual input. (26; 38; 35)

In the decoder, while the functional meaning of the learned latent vectors remains abstract, our entropy-based analysis reveals that its internal logic is surprisingly interpretable. Under the flash stimulus, we observed that decoder attention organizes into repetitive, band-like patterns where all heads redundantly track the global luminance modulation. This is confirmed by the entropy traces, which show a strong, consistent correlation with the temporal profile of the light intensity (r between -0.77 and -0.93) . In the absence of spatial complexity, the model adopts a unified retrieval strategy, focusing its computational resources on the only available stimulus dimension.

A different regime emerges during the moving ball stimulus, where attention patterns become highly heterogeneous. Our directional decomposition reveals that these heads exhibit distinct directional tuning preferences, clearly visualized by the divergent Pearson correlation vectors in Figure 6.7F. The spatial orientation of these vectors demonstrates that individual heads specialize in monitoring specific regions and directions within the visual field, effectively partitioning the stimulus space. This specialized spatial organization provides a compelling functional analogy to the parallel processing streams of the vertebrate retina. Specifically, the observed "division of labor" among decoder heads mirrors the physiological organization of direction-selective ganglion cells (DSGCs). Just as DSGCs are organized into cardinal directions to encode motion vectors (39; 36), the POYO framework spontaneously adopts a multi-channel weighting strategy to resolve spatiotemporal complexities. This indicates that the model does not merely perform high-dimensional regression; it implicitly recovers biological coding principles to filter redundant information in favor of local motion features. While the lack of explicit physiological labels for the latents remains a challenge, these entropy-based correlation vectors offer a robust quantitative window into the logic of neural population codes. Taken together, these results provide compelling evidence in support of our overarching hypothesis. By spontaneously organizing into spatially tuned channels, segregating cell identities in latent space, and dynamically altering attention entropy based on visual complexity, the POYO framework demonstrates that attention-driven deep learning models can serve as promising, interpretable surrogate tools for probing population neural coding under controlled visual regimes.

Chapter 8

CONCLUSION

8.1 Conclusion

We conclude that transformer-based architectures, specifically the POYO framework, do not merely achieve state-of-the-art decoding performance but spontaneously recover fundamental biological coding strategies. Our analysis of encoder–decoder dynamics reveals that the model adaptively redistributes its computational resources based on stimulus complexity, shifting from a redundant, synchronized pooling for global flashes to a highly specialized, heterogeneous recruitment of attention heads for structured motion.

By establishing a direct causal link through attention-guided ablations, we demonstrate that the model's internal prioritization is physiologically grounded: it identifies and relies on specific neural subpopulations whose removal systematically degrades decoding integrity. Furthermore, the model's ability to selectively weight units with distinct signatures, such as sustained temporal responses for luminance tracking, confirms that artificial attention can serve as a principled proxy for understanding how the retina allocates information across parallel functional channels (34; 36).

While we acknowledge that the abstract nature of learned latents still poses challenges for full mechanistic transparency, our work demonstrates that when validated through entropy metrics and statistical hypothesis testing, these architectures offer a transparent window into the logic of neural population codes. The ability of POYO to autonomously prioritize relevant physiological features across different retinal preparations positions these models not just as decoders, but as powerful tools for automated biological discovery in large-scale neuroscience.

8.2 Future work

As further research, we propose that future progress will require integrating deep architectures with interpretability-focused approaches, including causal perturbation analyses, comparisons to receptive-field models, and frameworks that encourage biologically grounded latent structure (40). Only by bridging predictive power with mechanistic transparency will models such as POYO advance beyond black-box decoding and contribute

more directly to our understanding of retinal computation.

From the retinal coding perspective, it would be of interest to further investigate the different coding strategies implemented within each attentional encoder head, helping us elucidate the conditions that trigger either population, pairwise, or independent neural encoding (33; 41; 29).

Appendix A

GENERATED PUBLICATIONS

The following publications have been partially or completely derived from the research involved in the development of this thesis project.

ISI Journal Papers

1. F. Miqueles, A. G. Palacios, J. Atkinson and M.-J. Escobar, "Attention maps reveal stimulus-dependent retinal population codes," *Frontiers in Computational Neuroscience*, vol. 20, art. 1769478, May. 2026. doi: 10.3389/fncom.2026.1769478

Conference Abstracts

1. Maria Jose Escobar; Francisco Miqueles; John Atkinson; Adrian G. Palacios, "Decoding Retinal Responses: A Transformer-Based Model for Visual Stimulus Prediction," in *Vision Sciences Society Annual Meeting Abstract* , doi: 10.1167/jov.25.9.2844

Related Projects

1. FONDECYT Project 1230170, 1200880 and EXPLORACION 13220082 – Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

2. ANID-Basal Project FB0008 – Centro Interdisciplinario de Neurociencia de Valparaíso (CINV), Universidad de Valparaíso.

Bibliography

[1] D. K. Warland, P. Reinagel, and M. Meister, "Decoding visual information from a population of retinal ganglion cells," *Journal of Neurophysiology*, vol. 78, no. 5, pp. 2336–2350, 1997.

[2] J. Ye and C. Pandarinath, "Representation learning for neural population activity with neural data transformers," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 10 365–10 378, 2021.

[3] J. Ye, J. Collinger, L. Wehbe, and R. Gaunt, "Neural data transformer 2: multi-context pretraining for neural spiking activity," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 352–80 374, 2023.

[4] S. Schneider, J. H. Lee, and M. W. Mathis, "Learnable latent embeddings for joint behavioural and neural analysis," *Nature*, vol. 617, no. 7962, pp. 754–762, 2023.

[5] B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-process factor analysis for low-dimensional learned latent dynamics of neural population activity," *Journal of Neurophysiology*, 2009.

[6] J. Dowling, *The Retina: An Approachable Part of the Brain*. Belknap Press, 2012.

[7] H. Kolb, E. Fernandez, and R. Nelson, "The organization of the retina and visual system," 2023, accessed: 2025-08-11.

[8] O. Marre, V. Botella-Soler, K. D. Simmons, T. Mora, G. Tkacik, and M. J. Berry, "High accuracy decoding of dynamical motion from a large retinal population," *PLoS Computational Biology*, vol. 11, no. 7, p. e1004304, 2015.

[9] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. MIT Press, 1999.

[10] Z. Nichols, S. Nirenberg, and J. Victor, "Interacting linear and nonlinear characteristics produce population coding asymmetries between on and off cells in the retina," *Journal of Neuroscience*, vol. 33, no. 37, pp. 14 958–14 973, 2013.

[11] N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. Chichilnisky, and L. Paninski, "Neural networks for efficient bayesian decoding of natural images from retinal neurons," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017.

[12] V. Botella-Soler, S. Deny, G. Martius, O. Marre, and G. Tkačik, "Nonlinear decoding of a complex movie from the mammalian retina," *PLoS Computational Biology*, vol. 14, no. 5, p. e1006057, 2018.

[13] Z. Yu, T. Bu, Y. Zhang, S. Jia, T. Huang, and J. K. Liu, "Robust decoding of rich dynamical visual scenes with retinal spikes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3396–3409, 2025.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[16] M. Azabou, V. Arora, V. Ganesh, X. Mao, S. Nachimuthu, M. J. Mendelson, B. Richards, M. G. Perich, G. Lajoie, and E. L. Dyer, "A unified, scalable framework for neural population decoding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://poyo-brain.github.io

[17] M. Azabou, K. X. Pan, V. Arora, I. Knight, E. L. Dyer, and B. Richards, "Multi-session, multi-task neural decoding from distinct cell-types and brain regions," in *International Conference on Learning Representations (ICLR)*, 2025.

[18] G. E. H. David E. Rumelhart, "Learning representations by back-propagating errors," *Nature*, 1986.

[19] C. R. Ravello, L. U. Perrinet, M.-J. Escobar, and A. G. Palacios, "Speed-Selectivity in Retinal Ganglion Cells is Sharpened by Broad Spatial Frequency, Naturalistic Stimuli," *Scientific Reports*, vol. 9, p. 456, 12 2019. [Online]. Available: http://www.nature.com/articles/s41598-018-36861-8

[20] M.-J. Escobar, C. Reyes, R. Herzog, J. Araya, M. Otero, C. Ibaceta, and A. G. Palacios, "Characterization of Retinal Functionality at Different Eccentricities in a Diurnal Rodent," *Frontiers in Cellular Neuroscience*, vol. 12, p. 444, 2018.

[21] P. Yger, G. Spampinato, E. Esposito, B. Lefebvre, S. Deny, C. Gardella, M. Stimberg, F. Jetter, G. Zeck, S. Picaud, J. Duebel, and O. Marre, "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo," *eLife*, vol. 7, p. e34518, 2018.

[22] T. C. Sprague, S. Saproo, and J. T. Serences, "Visual attention mitigates information loss in small- and large-scale neural codes," *Trends in Cognitive Sciences*, vol. 19, no. 4, pp. 215–226, 2015.

[23] O. Hendler *et al.*, "Limits of optimal decoding under synaptic coarse-tuning," *bioRxiv*, 2026.

[24] J. I. Glaser, A. S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording, "Machine learning for neural decoding," *eNeuro*, vol. 7, no. 4, pp. ENEURO.0506–19.2020, 2020.

[25] E. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Network: Computation in Neural Systems*, vol. 12, no. 2, pp. 199–213, 2001.

[26] T. Gollisch and M. Meister, "What the retina tells the brain," *Annual Review of Neuroscience*, vol. 33, pp. 89–114, 2010.

[27] T. Baden, P. Berens, K. Franke, M. Román Rosón, M. Bethge, and T. Euler, "The functional specialization of retinal ganglion cells in the mouse," *Nature*, vol. 529, no. 7586, pp. 345–350, 2016.

[28] J. B. Demb, "Functional circuitry of visual adaptation in the retina," *Journal of Physiology*, vol. 586, no. 18, pp. 4377–4384, 2008.

[29] S. Nirenberg, S. M. Carcieri, A. L. Jacobs, and P. E. Latham, "Retinal ganglion cells act largely as independent encoders," *Nature*, vol. 411, pp. 698–701, 2001.

[30] L. Paninski and J. P. Cunningham, "Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience," *Neuron*, vol. 99, no. 4, pp. 732–761, 2018.

[31] L. Duncker, L. Driscoll, K. Shenoy, and M. Sahani, "Interpretable and flexible models for neural population activity," *Nature Neuroscience*, vol. 24, no. 9, pp. 1330–1341, 2021.

[32] N. Kriegeskorte and P. K. Douglas, "Interpreting deep neural networks in computational neuroscience," *bioRxiv*, p. 003038, 2015.

[33] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[34] K. Bollinger, H. V. Oviedo, and T. Gollisch, "The chirp stimulus revisited: Systematic evaluation of nonlinearities in the retina with global light modulations," *Journal of Neurophysiology*, vol. 124, no. 2, pp. 634–648, 2020.

[35] K. Koch, J. McLean, M. Berry, P. Sterling, V. Balasubramanian, and M. A. Freed, "Efficiency of information transmission by retinal ganglion cells," *Current Biology*, vol. 14, no. 17, pp. 1523–1530, 2004.

[36] M. Fiscella, F. Franke, K. Farrow, J. Müller, B. Roska, A. Hierlemann, and F. Franke, "Visual coding with a population of direction-selective neurons in the retina," *Nature Neuroscience*, vol. 18, no. 10, pp. 1331–1339, 2015.

[37] S. de Vries and D. Baylor, "Mosaic arrangement of ganglion cell receptive fields in rabbit retina," *Journal of Neurophysiology*, vol. 87, no. 4, pp. 1424–1435, 2002.

[38] M. Meister, L. Lagnado, and D. A. Baylor, "Concerted signaling by retinal ganglion cells," *Science*, vol. 270, no. 5239, pp. 1207–1210, 1995.

[39] S. B. . T. W. Vaney, D., "Direction selectivity in the retina: symmetry and asymmetry in structure and function," *Nat Rev Neurosci*, vol. 19, pp. 194–208, 2012.

[40] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," *arXiv preprint arXiv:1901.10912*, 2019.

[41] L. P. . P. Averbeck, B., "Neural correlations, population coding and computation," *Nat Rev Neurosci*, vol. 7, pp. 358–366, 2006.