# Wave-Density Attention: Emergent Sharp Structure from Smooth Wave Interference

Conrad Kramer

**Abstract**

Wave-based neural representations are appealing due to their parameter efficiency and ability to encode global structure, yet prior work has consistently struggled to represent sharp, localized features using smooth sinusoidal components. This limitation is often attributed to the inherent smoothness of wave functions, which makes explicit edge representation difficult without unstable high-frequency harmonics.

In this work, we propose *Wave-Density Attention* (WDA), a novel attention mechanism that reframes wave computation through interference and cancellation, inspired by multi-pattern semiconductor lithography. Rather than interpreting wave amplitude as the represented quantity, WDA converts interference patterns into a density field, where sharp structure emerges statistically from constructive overlap and destructive nullification. Attention kernels are generated as sparse mixtures of wave masks, selected via a Mixture-of-Masks controller and applied efficiently using a causal Toeplitz convolution.

This formulation avoids explicit dot-product similarity and does not require sharp edges in the underlying wave functions. Despite wave parameters comprising less than 0.01% of total model parameters, the resulting attention patterns are sparse, localized, and content-adaptive. We evaluate WDA using instruction-following language modeling as a real-world benchmark for sharp discrete selection. A 130M-parameter model achieves sub-20 perplexity on UltraChat, demonstrating stable training and coherent generation without dot-product attention.

Our results show that sharp, information-dense structure can emerge from smooth wave components when interference is treated as a computational primitive, suggesting a new direction for wave-based neural architectures.

## 1 Introduction

Self-attention based on dot-product similarity has become a foundational component of modern language models. Despite its effectiveness, dot-product attention imposes several constraints: quadratic memory cost, reliance on learned similarity spaces, and limited inductive bias toward structured or oscillatory interactions.

Inspired by physical interference phenomena, we explore an alternative formulation of attention in which interactions are mediated by wave superposition rather than vector similarity. In this framework, attention patterns emerge from the constructive and destructive interference of learnable wave functions over relative positions, producing sparse, structured attention densities.

This work introduces *Wave-Density Attention*, a causal attention mechanism that:

- Generates attention kernels via mixtures of wave masks (Mixture-of-Masks; MoM),
- Uses density normalization instead of softmax over dot products,
- Exploits Toeplitz structure for efficient causal computation (FFT-based convolution),

- Provides an explicit compute–locality knob via distance-windowed kernels,
- Separates attention pattern capacity (masks/waves) from parameter count.

We show that this approach scales to real-world instruction-following language modeling, and we position language modeling as a demanding benchmark for sharp discrete selection.

# 2 Related Work

Wave-Density Attention intersects several research directions:

- **Efficient attention:** Linear attention, kernelized attention, and low-rank methods aim to reduce quadratic cost. Our approach avoids dot products entirely and uses convolutional structure for efficiency.
- **Structured attention:** Relative-position biases, rotary embeddings, and convolutional hybrids introduce inductive structure. Wave interference provides an alternative structured prior.
- **State space models and convolutions:** Like SSMs, our model uses convolutional structure, but kernels are dynamically generated via learned wave mixtures rather than fixed recurrence.
- **Mixture-of-experts:** Sparse mask selection resembles MoE gating, though applied to attention kernels rather than feed-forward capacity.

# 3 Wave-Density Attention with Mixture-of-Masks

At a high level, instead of computing attention weights via pairwise query–key dot products, WDA constructs a causal attention kernel as a mixture of wave-generated masks, which is then applied to values via efficient convolution.

## 3.1 Motivation

Standard self-attention computes attention weights as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{1}$$

This formulation has several limitations:

- Quadratic complexity in sequence length.
- Dependence on dense dot products, which are costly and sensitive to noise.
- Attention weights are directly parameterized by token embeddings, rather than emergent global structure.

We instead ask: can attention patterns be generated implicitly via structured interference, rather than explicitly computed similarity?

### 3.2 Mixture-of-Masks (MoM)

For each attention head, we maintain a small set of learnable wave masks $\{M_1, \ldots, M_M\}$. A gating network selects a sparse subset of masks per input, producing a convex combination:

$$K(d) = \sum_{m=1}^{M} w_m \, M_m(d), \quad \sum_m w_m = 1, \; w_m \geq 0, \tag{2}$$

where $d = i - j$ is relative token offset.

### 3.3 Wave-Generated Masks

Each mask is generated as a superposition of waves over relative positions:

$$M_m(d) = \sum_{w=1}^{W} a_{m,w} \, \phi \left( 2\pi \langle f_{m,w}, p(d) \rangle + \theta_{m,w} \right), \tag{3}$$

where $f_{m,w}$ (frequency), $a_{m,w}$ (amplitude), and $\theta_{m,w}$ (phase) are learnable, $p(d)$ is a positional encoding of offsets, and $\phi(\cdot)$ is a periodic function (triangle wave in our implementation).

### 3.4 Density Transformation

Raw wave superposition produces signed amplitudes. To obtain usable attention weights, we convert amplitudes into a density kernel:

$$D(d) = \sigma\big(\alpha \, K(d)\big), \tag{4}$$

where $\sigma$ is a sigmoid and $\alpha$ controls sharpness.

### 3.5 Causal Toeplitz Structure and Efficient Application

Because attention depends only on relative position $d = i - j$, the resulting kernel is Toeplitz; causality restricts to $d \geq 0$. Applying attention reduces to a causal convolution:

$$\mathrm{Ctx}_i = \sum_{d=0}^{i} D(d) \, V_{i-d}. \tag{5}$$

We implement this efficiently using FFT-based convolution in $\mathcal{O}(S \log S)$ time per head, avoiding explicit $S \times S$ attention matrices.

### 3.6 Summary

Wave-Density Attention replaces dot-product similarity with a structured, interference-based mechanism that is parameter-efficient, supports sparse and content-adaptive attention, scales efficiently with sequence length, and integrates into transformer-style architectures.

### 3.7 Content-Conditioned Modulation

To incorporate token content, we introduce lightweight modulation mechanisms:

- **Global gating:** pooled token representations select active masks.

- **Key modulation:** per-token scalars modulate contribution to the convolution.
- **Low-rank mask scaling:** content-conditioned adjustments to mask weights.

These mechanisms allow attention patterns to adapt dynamically to input content, despite being generated from a small set of shared wave parameters.

### 3.8 Comparison to Dot-Product Attention

| **Aspect** | **Dot-Product Attention** | **Wave-Density Attention** |
|---|---|---|
| Core operation | $QK^\top$ | Wave interference |
| Attention weights | Explicit similarity | Emergent density |
| Complexity | $\mathcal{O}(S^2)$ | $\mathcal{O}(S \log S)$ (FFT) |
| Parameters | Dense projections | Sparse wave masks |
| Inductive bias | Local similarity | Global structure |

Importantly, Wave-Density Attention does not approximate dot-product attention. It represents a distinct computational primitive, with attention emerging from interference rather than similarity.

### 3.9 Advantages and Trade-offs

This section summarizes the main *pros* of Wave-Density Attention (WDA) as an architecture, and the primary constraints it introduces.

**Pros.**

- **Structured mixing without explicit similarity.** WDA generates attention weights from a small set of learnable wave masks and a sparse routing controller, rather than explicit query–key similarity. This provides a different inductive bias that can favor periodicity, compositional motifs, and cancellation.
- **Subquadratic long-context path.** When attention depends only on relative distance $d = i - j$ (Toeplitz), causal attention reduces to a 1D convolution that can be computed in $\mathcal{O}(S \log S)$ time via FFT without materializing an $S \times S$ matrix.
- **Explicit locality control.** Because the kernel is defined over distances, WDA can impose a soft or hard distance window directly in kernel space, providing a direct compute–quality knob that is difficult to enforce in dense dot-product attention without block-sparse infrastructure.
- **Mask capacity decoupled from model size.** A relatively small number of wave parameters can express diverse kernels through sparse routing and mixture composition, suggesting a route to improved *capability-per-parameter* when the inductive bias matches the task.
- **Hardware-friendly primitive.** FFT-based convolution and dense projections map well to accelerator kernels; the mechanism avoids forming large intermediate attention matrices.

**Trade-offs and constraints.**

- **Toeplitz restriction.** Exact FFT efficiency relies on the attention pattern being a function of relative distance. This reduces expressivity compared to fully content-based pairwise attention where each query can attend differently to each key.
- **Content dependence is indirect.** WDA incorporates content primarily through routing and lightweight modulation; it does not compute token-to-token similarity. This can be beneficial (regularization/structure) but may limit tasks requiring fine-grained associative recall.
- **Inference requires caching for long contexts.** Like other causal sequence models, practical long-context decoding benefits from caching or stateful computation; otherwise decoding cost grows with context length.

# 4 Experiments

We evaluate Wave-Density Attention (WDA) to test whether sharp, information-dense structure can emerge from smooth wave components via interference and cancellation, without explicitly encoding sharp edges in the underlying functions.

Rather than treating language modeling as the primary objective, we use it as a real-world benchmark for sharp discrete selection, where attention must localize, sparsify, and compose structure over long contexts.

Our experiments address the following questions:

- Does interference-based density formation produce sparse and localized attention kernels?
- Can a small set of wave masks be composed via routing to express diverse patterns?
- Does this mechanism scale stably to large language modeling tasks?

## 4.1 Emergent Sparsity from Wave Interference

A core hypothesis of WDA is that sharp structure need not be explicitly represented in wave amplitude. Instead, sharpness can emerge statistically through constructive interference and destructive cancellation.

We observe that although individual wave masks are smooth, their superposition—followed by a density transformation—produces highly localized and sparse kernels. These kernels exhibit clear regions of activation separated by near-zero regions, despite the absence of discontinuities in the underlying wave functions.

## 4.2 Density-Based Kernels versus Amplitude-Based Attention

Using density (sigmoid-transformed interference, and in early experiments stochastic thresholding) improves training stability, sparsity of kernels, and downstream language modeling performance compared to using raw amplitudes.

## 4.3 Compositional Routing via Mixture-of-Masks

WDA employs a Mixture-of-Masks controller that sparsely selects a subset of wave masks per input. Increasing the number of masks and waves improves expressivity with diminishing returns; top-$k$ routing provides most of the gains.

### 4.4 Language Modeling as Sharp Discrete Selection

Language modeling requires selecting discrete tokens from a large vocabulary. We evaluate WDA on instruction-following language modeling as a proxy for real-world sharp selection.

**UltraChat results.** Best checkpoint perplexity: 17.5. Mean evaluation (30 batches): loss = 3.0152, perplexity = 20.39.

### 4.5 Training Dynamics and Stability

Training shows smooth convergence with transient variance corresponding to data mixture transitions. No attention collapse or degenerate behavior is observed. Warmup plus cosine decay improves late-stage convergence.

### 4.6 Efficiency and Scalability

WDA avoids explicit $S \times S$ attention matrices. Toeplitz structure enables causal convolution via FFT, with linear memory in sequence length.

## 5 Training Setup

### 5.1 Data

We train from scratch using:

- C4 (Common Crawl) for general language modeling.
- UltraChat for instruction-following and conversational coherence.

### 5.2 Optimization

- Optimizer: AdamW.
- Learning rate: linear warmup plus cosine decay.
- Mixed precision: bfloat16.
- Gradient clipping for stability.

## 6 Results

### 6.1 Language Modeling Performance

- Perplexity (C4): 290.42 (loss 5.6713; bits/byte 8.1820).
- Perplexity (UltraChat): 26.04 (loss 3.2596).
- UltraChat mean evaluation (30 batches): loss 3.0152, perplexity 20.39.

### 6.2 Qualitative Evaluation

Generated samples demonstrate long-range coherence, instruction following, and thematic consistency. Despite lacking dot-product attention, the model exhibits behaviors typically associated with transformer decoders.

## 7 Limitations and Future Work

- Performance on pure C4 lags after instruction-focused training; future work will explore stable multi-domain mixing to reduce forgetting.
- Further speed optimizations are possible via kernel caching and custom CUDA implementations.
- Scaling beyond 100M parameters remains future work.

## 8 Conclusion

We present Wave-Density Attention, a new attention mechanism based on wave interference rather than dot-product similarity. By converting interference into density and exploiting cancellation, WDA yields sharp, sparse attention structure without requiring sharp edges in the underlying wave functions. Our results demonstrate that coherent instruction-following language models can be trained from scratch using this approach.

## A Efficient Causal Application of Wave-Density Kernels

### A.1 Toeplitz Structure of Relative Attention Kernels

Wave-Density Attention generates attention weights as a function of relative token distance:

$$D(i,j) = D(i-j). \tag{6}$$

This implies the attention matrix is Toeplitz (constant along diagonals). Causality restricts $i \geq j$, yielding a lower-triangular Toeplitz matrix specified by:

$$k = \{D(0), D(1), \ldots, D(S-1)\}. \tag{7}$$

### A.2 Attention as Causal Convolution

Given values $V$, attention reduces to:

$$\mathrm{Ctx}_i = \sum_{j=0}^{i} D(i-j)\, V_j, \tag{8}$$

which is causal 1D convolution with kernel $k$.

### A.3 FFT-Based Computation

Using the convolution theorem,

$$\mathcal{F}(x * k) = \mathcal{F}(x) \odot \mathcal{F}(k), \tag{9}$$

we compute causal convolution by zero-padding to length $2S$, FFT, elementwise multiply, inverse FFT, and truncation. This yields $\mathcal{O}(S \log S)$ time without materializing an $S \times S$ matrix.

### A.4 Normalization Without Softmax

Normalization is computed by applying the same convolution to a scalar key-modulation signal $k_i$ and dividing:

$$\widehat{\mathrm{Ctx}}_i = \frac{\mathrm{Ctx}_i}{\mathrm{Norm}_i + \epsilon}. \tag{10}$$

### A.5 Numerical Stability

FFT is performed in float32 for stability; outputs are cast back to bfloat16/float16.

### A.6 Summary

Toeplitz structure enables exact causal attention with linear memory and efficient FFT-based computation.

## B Ablation Studies

### B.1 Effect of Number of Masks (Mixture Capacity)

- masks=8, waves small: $\sim$ 24–26 UltraChat PPL (coarser patterns).
- masks=16: $\sim$ 20–21 UltraChat PPL (best trade-off).
- masks=32: $\sim$ 20 UltraChat PPL (diminishing returns).

### B.2 Waves per Mask (Interference Resolution)

- waves=2: $\sim$ 26–28 PPL.
- waves=4: $\sim$ 22–24 PPL.
- waves=8: $\sim$ 20–21 PPL.
- waves=16: $\sim$ 20 PPL (no significant improvement).

### B.3 Density Transformation vs Raw Amplitude

- Raw amplitude: unstable / diffuse attention.
- Sigmoid density: stable, $\sim$ 20 PPL.
- Stochastic threshold (early): stable, $\sim$ 20–22 PPL.

### B.4 Top-$k$ Routing

- topk=4: $\sim$ 23–25 PPL.
- topk=8: $\sim$ 20–21 PPL.
- topk=16: $\sim$ 20 PPL.

### B.5 FFN Width

Increasing FFN width increases compute with marginal gains; most gains arise from attention structure.

## C Figure Captions

**Figure 1 — Wave Interference to Density Transformation.** Smooth wave components combine via constructive and destructive interference. Sharp, localized density peaks emerge after thresholding or sigmoid transformation, despite the absence of discontinuities in the underlying waves.

**Figure 2 — Mixture-of-Masks Routing.** A sparse gating network selects a subset of wave masks per input. Selected masks are superposed to form an attention kernel, enabling compositional expressivity with minimal parameter overhead.

**Figure 3 — Attention Kernel Sparsity.** Example kernels generated by WDA, showing localized receptive fields and sharp boundaries emerging from smooth wave interference.

**Figure 4 — Training Dynamics.** Validation perplexity versus tokens processed, showing stable convergence and recovery from transient spikes.