Q-Expansion: Decoupled Context Enrichment via Hierarchical Query Synthesis

joshuah.rainstar@gmail.com

Abstract

Standard attention mechanisms function as ill-poised coordinate transport systems, relying on accidental high-rank restoration via feed-forward networks to maintain probability simplex targets. We assert that *long-context* capability is strictly a function of query support extension and value preservation, independent of the attention matrix's traditional density. We introduce **Q-Expansion**, a deterministic, parameter-free operator that recursively injects phase-transport centroids into the query stream. This constructs a pyramidal hierarchy of causal summaries. We further propose a **Decoupled Enrichment** mechanism where invariant kernel maps (K) retrieve optimal triggers from this expanded query support to synthesize a refined query map (Q'), which in turn enriches the value stream ($V^\star$). This effectively pre-loads geometric context into the transport payload before standard attention is applied.

1 Introduction

We posit the following functional distinctness in sequence modeling components:

- The Kernel Map (K): A fixed reference defining the "search" criteria. It must remain invariant to preserve lookup semantics.
- The Filter Map (Q): A plastic descriptor of local position and content. It acts as the "signal" stating presence.
- The Value Map (V): The content payload to be transported.

Conventional attention conflates these roles, demanding that a single query vector Q_t successfully interrogate the entire history of keys $K_{<t}$. We argue this is geometrically insufficient. Realistically, an optimal mixture must inform the future position about the past composition using a dense, multi-scale representation of that past.

We propose *Q-Expansion*, a methodology that computes centroids between adjacent temporal positions and injects them into the query stream. This process recurses to form a pyramidal structure, increasing the effective receptive field of the query set. We then introduce a synthetic retrieval stage where keys select their optimal context from this pyramid to generate a synthetic query map Q' and an enriched value map $V^\star$, upon which standard attention operates.

2 Preliminaries

Let $X = (x_1, \ldots, x_T)$ denote a sequence of input representations. Let $Q, K, V \in \mathbb{R}^{T \times d}$ be the query, key, and value projections respectively. We enforce that K is a kernel map (invariant) and Q is a filter map (transformable).

We utilize a robust geometric primitive for aggregating vectors, termed *Phase Transport*. This operator computes a centroid that respects the directional semantics of the embedding space, avoiding the magnitude collapse associated with simple arithmetic averaging in high-dimensional spheres.

3 The Phase Transport Centroid

We define the centroid construction method used for Q-Expansion. Let $x, y \in \mathbb{R}^d$ be two source vectors (e.g., adjacent queries).

3.1 Directional Decomposition

We define magnitudes and unit directions with numerical stabilization $\varepsilon = 10^{-12}$:

$$\|x\| = \max(\|x\|_2, \varepsilon), \qquad u = \frac{x}{\|x\|}, \tag{1}$$

$$\|y\| = \max(\|y\|_2, \varepsilon), \qquad v = \frac{y}{\|y\|}. \tag{2}$$

Let $w = x - y$ be the raw displacement and $c = \langle u, v \rangle$ be the cosine similarity.

3.2 Regime Classification and Auxiliaries

Let $\tau \ll 1$ be a semantic threshold (e.g., 10^{-6}).

Deterministic Antipodal Basis. If vectors are antipodal, we require a deterministic auxiliary vector p orthogonal to v. Let $k = \text{argmin}_j |v_j|$ be the index of the component of v with the smallest magnitude. Let e_k be the standard basis vector. We construct p via Gram-Schmidt:

$$\tilde{p} = e_k - v_k v, \quad p = \frac{\tilde{p}}{\max(\|\tilde{p}\|_2, \varepsilon)}. \tag{3}$$

Displacement Terms. For the general case, we define projections $a = \langle v, w \rangle$ and $b = \langle u, w \rangle$, and correction terms:

$$Kw = ua - vb, \tag{4}$$

$$K^2 w = u(ac - b) + v(bc - a). \tag{5}$$

The general phase-transported displacement is defined as:

$$y_{\text{gen}} = w - Kw + \frac{K^2 w}{\max(1 + c, \varepsilon)}. \tag{6}$$

The antipodal reflection is defined as:

$$y_{\text{neg}} = w - 2\langle w, v \rangle v - 2\langle w, p \rangle p. \tag{7}$$

3.3 Displacement Operator

We define the single piecewise displacement operator $\mathcal{T}(x, y)$ with strict precedence:

$$\mathcal{T}(x, y) = \begin{cases} y_{\text{neg}} & \text{if } c < -1 + \tau \quad \text{(Antipodal)}, \\ w & \text{if } (c > 1 - \tau) \vee (\|x\| < \tau) \vee (\|y\| < \tau) \quad \text{(Trivial)}, \\ y_{\text{gen}} & \text{otherwise} \quad \text{(General)}. \end{cases} \tag{8}$$

3.4 Centroid Definition

The centroid is defined as the midpoint derived from the transported displacement:

$$x \oplus y := \text{centroid}(x, y) = x - \frac{1}{2}\mathcal{T}(x, y). \tag{9}$$

4 Q-Expansion

We construct a pyramidal expansion of the query sequence Q.

Definition 4.1 (Base Layer)**.** Let $Q^{(0)} = (q_1^{(0)}, \dots, q_T^{(0)})$ where $q_t^{(0)}$ is the t-th row of Q.

Definition 4.2 (Recursive Expansion)**.** For recursion depth $r \geq 0$, and position t, the centroid at rank $r+1$ is computed from adjacent centroids at rank r:

$$q_t^{(r+1)} = q_t^{(r)} \oplus q_{t+1}^{(r)}. \tag{10}$$

The set of centroids at rank r is $Q^{(r)} = (q_1^{(r)}, \dots, q_{T-r}^{(r)})$. The span of a centroid $q_t^{(r)}$ covers the input interval $[t, t+r]$.

Definition 4.3 (Bounded Expanded Support)**.** We impose a maximum recursion depth R (budget). The expanded query set is:

$$Q^\star = \bigcup_{r=0}^{R} Q^{(r)}. \tag{11}$$

To preserve linear complexity $O(T)$, we require $R \ll T$ (e.g., R is constant or logarithmic).

5 Decoupled Enrichment and Attention

We define a two-stage mechanism: first, a kernel-driven retrieval of context from the expanded query set, followed by value enrichment and standard attention.

5.1 Strict Causal Admissibility

To ensure the retrieval relies solely on prior information, we enforce strict causality.

Definition 5.1 (Admissible Set)**.** For a key position i, the causally admissible query set is:

$$Q^\star_{<i} = \left\{ q_t^{(r)} \in Q^\star \mid t + r < i \right\}. \tag{12}$$

5.2 Kernel-Driven Query Synthesis

The kernel map K acts as a set of search criteria. For each time step i, we retrieve the most relevant information from the expanded history $Q^\star_{<i}$ to form a synthetic query q'_i.

Definition 5.2 (Synthetic Query Construction)**.** For each $i \in \{1, \dots, T\}$:

1. Identify the subset $\mathcal{M}_i \subset Q^\star_{<i}$ of size K_{top} that maximizes $\langle K_i, q \rangle$.
2. Compute the synthetic query vector q'_i as the centroid or sum of $\mathcal{M}_i$. For the top-1 case:

$$q'_i = \operatorname*{argmax}_{q \in Q^\star_{<i}} \langle K_i, q \rangle. \tag{13}$$

This yields a synthetic query map $Q' = (q'_1, \dots, q'_T)$.

5.3 Value Map Enrichment

The retrieved context Q' is used to enrich the original input X (or candidate values), ensuring that the values transported in the final attention step carry both the original signal and the expanded contextual hypothesis.

Definition 5.3 (Value Synthesis)**.** Let $V^\star \in \mathbb{R}^{T\times d}$ be the enriched value map:

$$V_i^\star = q_i' + x_i. \tag{14}$$

This composite structure allows residual passthrough of the input while effectively "tagging" the value with the search criteria that selected it.

5.4 Terminal Attention

Finally, we apply standard attention using the synthesized components. This step distributes the enriched values based on the alignment between the synthetic queries (what was found) and the kernels (what was sought).

$$O = \text{softmax}\left(\frac{Q'K^T}{\sqrt{d}}\right)V^\star. \tag{15}$$

Here, $Q', K, V^\star$ are all "normally shaped" ($T\times d$) matrices, allowing the use of standard optimized attention implementations.

6 Discussion

This architecture fundamentally separates the role of hypothesis generation from context distribution.

- **Decoupling:** The Q-Expansion and Kernel-Driven Synthesis stages ($Q \to Q^\star \to Q'$) handle the "search" for relevant historical patterns in a linear-complexity retrieval step.
- **Enrichment:** The Value Synthesis ($V^\star$) explicitly encodes the result of this search into the payload, ensuring that downstream heads receive context-aware representations.
- **Standard Interface:** By projecting the expanded, complex geometric search back into standard $(Q', K, V^\star)$ forms, the mechanism remains compatible with standard transformer backbones and attention optimizations.