EFFICIENT ATTENTION VIA CONTROL VARIATES

Lin Zheng\textsuperscript{1}∗ Jianbo Yuan\textsuperscript{2} Chong Wang\textsuperscript{3}∗ Lingpeng Kong\textsuperscript{1}
\textsuperscript{1}The University of Hong Kong \textsuperscript{2}ByteDance Inc. \textsuperscript{3}Apple Inc.
{lzheng2,lpk}@cs.hku.hk jianbo.yuan@bytedance.com mr.chongwang@apple.com

ABSTRACT

Random-feature-based attention (RFA) is an efficient approximation of softmax attention with linear runtime and space complexity. However, the approximation gap between RFA and conventional softmax attention is not well studied. Built upon previous progress of RFA, we characterize this gap through the lens of control variates and show that RFA can be decomposed into a sum of multiple control variate estimators for each element in the sequence. This new framework reveals that exact softmax attention can be recovered from RFA by manipulating each control variate. Besides, it allows us to develop a more flexible form of control variates, resulting in a novel attention mechanism that significantly reduces the approximation gap while maintaining linear complexity. Extensive experiments demonstrate that our model outperforms state-of-the-art efficient attention mechanisms on both vision and language tasks

1 INTRODUCTION

Random-feature-based attention (RFA, also known as Performer; Choromanski et al., 2021; Peng et al., 2021b) is an established fast approximation to the conventional softmax attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017), which successfully scales Transformer models to processing much longer sequences (Choromanski et al., 2021). At its core is the usage of random features (RF; Rahimi & Recht, 2008) to linearize the exponential kernel in softmax attention, which reduces the computational cost from quadratic to linear runtime and space complexity. Despite its efficiency, recent studies have pointed out that such approximation suffers from substantial performance degeneration (Xiong et al., 2021a; Zheng et al., 2022b).

In this work, we generalize the formulation of RFA via control variates (Owen, 2013), which characterizes the approximation gap between RFA and softmax attention in theory. We first show that RFA can be decomposed from a \textit{global} approximation over the whole sequence into a sum of \textit{local} control variate estimators, each of which is applied to an individual element in the sequence. Under this formulation, RFA is equivalent to employing the same coefficient for all control variate estimators to scale their variance isotropically (§3.1). Besides, we prove that if we optimize the coefficient of each control variate to minimize the estimation variance individually, RFA estimation becomes exact, that is, softmax attention is recovered with zero bias and zero variance (§3.2).

Our key observation is that such formulation reveals a localized perspective of the RFA approximation. Instead of directly seeking a better estimate over the entire sequence, we can break down the problem into smaller problems that aim at improving the approximation for each subsequence (§4). The control variate estimator for each subsequence can be tuned separately and combined to yield better estimation, which provably reduces approximation error in the global sense (§4.1). Nevertheless, one caveat is that as the number of sub-problems increases, the approximation gap will be reduced but at the expense of higher computational complexity. For instance, if we optimize the control variate for every single element, softmax attention would be recovered as desired but with quadratic complexity. To attain a good trade-off between approximation quality and efficiency, we develop a new \textbf{E}fficient \textbf{A}ttention via \textbf{C}ontrol \textbf{V}ariates (EVA) that implements this divide-and-conquer strategy efficiently. In EVA, the sequence is partitioned into a fixed number of disjoint subsets. For the subset

\textsuperscript{*}The majority of this work was done while these authors were at Bytedance.

\textsuperscript{1}Our code and models are available at this link.
that might bear the highest correlations to the query, we explicitly optimize the control variate for each element, which recovers exact softmax attention probabilities; while for the others, the control variate coefficient is shared locally among all elements within the same subset. The resulting attention mechanism is not only highly effective but also runs with the same computational complexity as RFA (§4.2). Extensive experiments on both language and vision tasks demonstrate that EVA outperforms the state-of-the-art efficient attention methods (§5).

2 BACKGROUND

2.1 SOFTMAX ATTENTION MECHANISM

Assume there exist a set of \( \{q_n\}_{n=1}^N \) and \( M \) key-value pairs \( K = [k_1, \ldots, k_M] \) and \( V = [v_1, \ldots, v_M] \), where queries, keys and values are all \( d \)-dimensional vectors. The softmax attention mechanism (Bahdanau et al. [2014]; Vaswani et al. [2017]) is defined as an average over the value vectors weighted by the dot-product similarities of the queries and keys. For the \( n \)-th query, the attention mechanism outputs

\[
\text{SoftmaxAttn}(q_n, K, V) := \frac{\exp (q_n^\top k_m)}{\sum_{m'=1}^M \exp (q_n^\top k_{m'})} v_m. \tag{1}
\]

In the case of self-attention (Lin et al. [2017]; Vaswani et al. [2017]), we have \( M = N \), which results in quadratic computational complexity since we have to compute the similarity for each query-key pair explicitly.

2.2 RANDOM-FEATURE-BASED ATTENTION WITH SELF-NORMALIZED IMPORTANCE SAMPLING

Recently, Zheng et al. [2022b] identifies that softmax attention (Equation [1]) can be written as an expectation over an attention-like aggregating function,

\[
\text{SoftmaxAttn}(q_n, K, V) = \sum_{m=1}^M \frac{\exp (q_n^\top k_m)}{\sum_{m'=1}^M \exp (q_n^\top k_{m'})} v_m = E_{\omega \sim p_n(\omega)} [f_n(\omega)], \tag{2}
\]

where

\[
f_n(\omega) := \frac{\sum_{m=1}^M \xi(q_n, \omega)\xi(k_m, \omega)v_m}{\sum_{m'=1}^M \xi(q_n, \omega)\xi(k_{m'}, \omega)}, \quad p_n(\omega) := \frac{\mathcal{N}(\omega; 0, I) \sum_{m=1}^M \xi(q_n, \omega)^\top \xi(k_m, \omega)}{Z}. \tag{3}
\]

Here \( \xi(\cdot, \cdot) \) is the randomized mapping defined in such a way that \( \exp (q_n^\top k_m) = E_{\omega \sim \mathcal{N}(0, I)} [\xi(q_n, \omega)^\top \xi(k_m, \omega)] \), and \( Z = \sum_{m=1}^M \exp (q_n^\top k_m) \) denotes the normalizing constant of distribution \( p_n \). Throughout this paper, we consider the positive randomized mapping \( \xi(x, \omega) = \exp \left( \omega^\top x - \frac{1}{2} ||x||^2 \right) \) unless otherwise specified.

Random-Feature-based Attention (RFA) methods (Choromanski et al. [2021]; Peng et al. [2021]) can be interpreted as performing self-normalized importance sampling (SNIS; Hesterberg [1995]) to approximate Equation (Zheng et al. [2022b]). In SNIS, one draws Monte Carlo samples from some proposal distribution \( q(\omega) \) instead of the true distribution \( p_n(\omega) \) and estimates the target expectation as

\[
E_{\omega \sim p_n(\omega)} [f_n(\omega)] = E_{\omega \sim q(\omega)} \left[ \frac{p_n(\omega)}{q(\omega)} f_n(\omega) \right] \approx \sum_{s=1}^S \frac{p_n(\omega_s)}{q(\omega_s)} f_n(\omega_s), \quad \omega_1, \ldots, \omega_S \sim q(\omega).
\]

Vanilla RFA amounts to constructing the SNIS estimation with \( q(\omega) = \mathcal{N}(0; 0, I) \). The SNIS representation also turns out equivalent to the more established form of RFA,

\[
\text{RFA}(q_n, K, V) := \frac{\sum_{s=1}^S p_n(\omega_s) f(\omega_s)}{\sum_{s=1}^S p_n(\omega_s)} = \frac{\sum_{m=1}^M \phi(q_n, \omega)^\top \phi(k_m, \omega) v_m}{\sum_{m'=1}^M \phi(q_n, \omega)^\top \phi(k_{m'}, \omega)}, \tag{4}
\]

where the random feature, denoted by \( \phi(x, \omega) := 1/\sqrt{S} [\xi(x, \omega_1), \ldots, \xi(x, \omega_S)]^\top \), is proposed to approximate exponential kernels in its original motivation (see Appendix [A] for a detailed review).
2.3 Control Variates

Control variates aim to reduce the estimation variance of an expectation $E[g(\omega)]$. Assuming our original RFA estimation is $g(\omega) \in \mathbb{R}^d$ and there is some control variate $h(\omega) \in \mathbb{R}$ with a known expectation $E[h(\omega)]$, we can employ the control variate $h(\omega)$ with the coefficient $\beta \in \mathbb{R}^d$ as follows,

$$\tilde{g}(\omega) = g(\omega) - \beta h(\omega) + \beta E[h(\omega)].$$

Note that the resulting estimator remains unbiased since $E[\tilde{g}(\omega)] = E[g(\omega)] - \beta E[h(\omega)] + \beta E[h(\omega)] = E[g(\omega)]$. However, the estimation variance can be largely reduced if $g(\cdot)$ and the scaled control variate $\beta h(\omega)$ are positively correlated \cite{Owen2013}.

3 Dissecting RFA with Control Variates

In this section, we first go through the connections among RFA, importance sampling, and control variates, revealing a decomposed formulation of RFA \cite{Vlassis2021}, and then quantify the approximation gap between RFA and softmax attention \cite{Vlassis2021} from these connections.

3.1 RFA as a Sum of Local Control Variate Estimators

As shown in Equation 4, RFA estimation considers all key-value pairs and produces a global variate estimates following variates, revealing a decomposed formulation of RFA \cite{Vlassis2021}, and then quantify the approximation over the entire sequence. In contrast, our work develops a decomposed gap between RFA and softmax attention \cite{Vlassis2021} from these connections.

We now examine the formulation of $g(\cdot)$ and $h(\cdot)$ in the context of RFA. According to Equation 5,

$$g(\omega) = \frac{1}{S} \sum_{s=1}^{S} \left( \sum_{k=1}^{M} \xi(q_m, \omega_s) \xi(k_m, \omega_s) \right) v_m,$$

$$h(\omega) = \frac{1}{S} \sum_{s=1}^{S} \sum_{k=1}^{M} \xi(q_m, \omega_s) \xi(k_m, \omega_s),$$

where $\alpha(\omega_s) := \frac{1}{S} \sum_{s=1}^{S} \frac{p_n(\omega_s) f(\omega_s)}{q(\omega_s) \xi(q_m, \omega_s) \xi(k_m, \omega_s) v_m}$ collects terms that is constant w.r.t. queries, keys, and values. Our key observation is that by changing the order of summations, both $g(\cdot)$ and $h(\cdot)$ can be decomposed as $g(\omega) = \sum_{m=1}^{M} g_m(\omega)$ and $h(\omega) = \sum_{m=1}^{M} h_m(\omega)$ respectively, where

$$g_m(\omega) = \sum_{s=1}^{S} \alpha(\omega_s) \xi(q_m, \omega_s) \xi(k_m, \omega_s) v_m,$$

$$h_m(\omega) = \sum_{s=1}^{S} \alpha(\omega_s) \xi(q_m, \omega_s).$$

As a result, we can decompose the entire RFA estimate in Equation 6 into a summation of $M$ control variate estimates following

$$\tilde{g}(\omega) = g(\omega) - \beta h(\omega) + \beta E[h(\omega)].$$

$$= \left( \sum_{m=1}^{M} g_m(\omega) - \tilde{\beta} h_m(\omega) \right) - \beta E\left[ \sum_{m=1}^{M} h_m(\omega) \right]$$

$$= \sum_{m=1}^{M} g_m(\omega) - \beta h_m(\omega) - \beta E[h_m(\omega)] := \sum_{m=1}^{M} \tilde{g}_m(\omega).$$

(7)
Here $\bar{g}_m(\omega) = g_m(\omega) - \hat{\beta}(\omega) (h_m(\omega) - \mathbb{E}[h_m(\omega)])$ denotes the corresponding control variate estimator of the $m$-th key-value pair and $\hat{\beta}(\omega)$ is the coefficient shared across the entire sequence.

### 3.2 Optimizing Coefficients in RFA Locally Recovers Softmax Attention

Based on the decomposition of RFA in Equation [7], we have one local control variate attached to each key-value pair. To see the benefit of such decomposition, we demonstrate that softmax attention is equivalent to associating each control variate with a locally optimized coefficient $\hat{\beta}_m$ in RFA.

**Proposition 1.** Let $\bar{g}_m(\omega) = g_m(\omega) - \hat{\beta}_m (h_m(\omega) - \mathbb{E}[h_m(\omega)])$. We denote the variance of some estimator $g(\omega)$ as $\text{Var}[g(\omega)] := \text{Cov}[g(\omega), g(\omega)]$. Then the optimal $\hat{\beta}_m$ that minimizes $\text{Tr}(\text{Var}[\bar{g}_m(\omega)])$ (i.e., the sum variance over all dimensions) is of the form

$$
\beta^*_m := \arg \min_{\beta} \text{Tr}(\text{Var}[\bar{g}_m(\omega)]) = v_m = \frac{g_m(\omega)}{h_m(\omega)}.
$$

Furthermore, by letting $\hat{\beta}_m = \beta^*_m$ for all $m = 1, 2, \ldots, M$, we have $\text{Tr}(\text{Var}[\bar{g}_m(\omega)]) = 0$. As a result, $\text{Tr}(\text{Var}[\bar{g}(\omega)]) = 0$ and thus RFA($q_b, K, V$) = $\text{SoftmaxAttn}(q_b, K, V)$.

The proof is deferred to Appendix B.4. This proposition implies optimizing $\hat{\beta}_m$ for each key-value pair in the decomposed formulation of RFA recovers the exact softmax attention. It not only characterizes the theoretical gap introduced by RFA but also sheds light on how to improve RFA towards softmax attention from a localized perspective. Furthermore, it delineates the trade-off between estimation quality and computational costs. On the one hand, if we use a distinct $\hat{\beta}_m$ for each estimator, we could achieve a perfect estimation, albeit at the expense of computing $2M$ estimators, each estimator achieves a perfect estimation, albeit at the expense of computing $2M$ systems, each system achieves a perfect estimation, albeit at the expense of computing $2M$ systems. On the other hand, if a single shared coefficient is employed, it degrades to conventional RFA, where all the control variate estimators can be merged and computed together in linear complexity (Choromanski et al. [2021], Peng et al. [2021b], Zheng et al. [2022b]).

### 4 EVA: Efficient Attention via Control Variates

In this section, we demonstrate that the control variate formulation offers a natural way to improve RFA with a finer-grained treatment over control variates. We describe the improved efficient attention mechanism EVA in [8.1] and its practical implementation in [8.2].

#### 4.1 Control Variates with Locally Shared Coefficients

We denote $[M] := \{1, 2, \ldots, M\}$ as the set of all key-value indices. Instead of employing the same coefficient for all control variates as in RFA, we propose to partition $[M]$ into $C$ subsets $P_1, P_2, \ldots, P_C$ and allocate a locally shared $\beta_c$ for each subset $P_c$. For all $\beta_c$ and their optimum $\beta^*_c$ for each token, define the weighted mean squared error (weighted MSE) as $\sum_{c=1}^C \sum_{m \in P_c} \alpha_m \|\beta_c - \beta^*_c\|^2$, where $\alpha_m > 0$ and $\sum_{c=1}^C \sum_{m \in P_c} \alpha_m = 1$. To see the benefit of partitioning, we demonstrate that there always exists some $\{\beta_c\}_{c=1}^C$ that achieves lower weighted MSE than any globally shared coefficient (see Appendix B.5 for a formal argument). The next question is how to determine $\{\beta_c\}_{c=1}^C$. According to Proposition 1, a natural choice is to adapt the optimal coefficients (Equation [8]) to the case of partitioned subsets. We justify this choice here and it is also optimal in minimizing the MSE above weighted by the true attention probabilities.

**Proposition 2.** Suppose $U$ is a set of key-value indices, $\beta^*_m$ is the optimal coefficient for each $m \in U$ as defined in Proposition 7 and $P_1, P_2, \ldots, P_C$ are an arbitrary partition of $U$, where each subset $P_c$ is associated with a distinct $\beta_c$. We consider the following weighted mean squared error,

$$
J(\beta_1, \ldots, \beta_C) := \sum_{c=1}^C \sum_{m \in P_c} \frac{\exp(q_m^\top k_m)}{\sum_{m' \in U} \exp(q_m^\top k_{m'})} \|\beta_c - \beta^*_m\|^2.
$$

Note that the expectation of individual control variates $h_m(\cdot)$ is still in closed form as $\mathbb{E}[h_m(\omega)] = \exp(q_m^\top k_m)/Z$. The derivation can be found in Appendix B.3.
Then for each \( c = 1, \ldots, C \) we have

\[
\beta_c^* := \arg \min_{\beta_c} J(\beta_1, \ldots, \beta_C) = \frac{\mathbb{E} \left[ \sum_{m \in \mathcal{P}_c} g_m(\omega) \right]}{\mathbb{E} \left[ \sum_{m \in \mathcal{P}_c} h_m(\omega) \right]}.
\] (10)

As a consequence, with \( \beta_c = \beta_c^* \), the partition scheme must achieve lower weighted mean squared error than any globally shared \( \beta \), that is, \( J(\beta_1 = \beta_1^*, \ldots, \beta_C = \beta_C^*) \leq J(\beta_1 = \beta, \ldots, \beta_C = \beta) \).

The proof can be found in Appendix B.6. Apart from measuring the squared errors for all coefficients, Equation 9 also governs the significance of each error by its corresponding softmax weights, which attains closer alignment with true softmax attention. Therefore, this proposition implies that it is much easier for the partitioned control variate estimators to obtain coefficients closer to their optimum while faithfully respecting softmax attention.

The optimal coefficients \( \beta_c^* \) could be estimated via Monte Carlo samples as \( \beta_c^* \approx \hat{\beta}_c(\omega) = \frac{\left( \sum_{m \in \mathcal{P}_c} g_m(\omega) \right)}{\left( \sum_{m \in \mathcal{P}_c} h_m(\omega) \right)} \), which is a widely adopted strategy in the control variate literature (Wang et al., 2013; Owen, 2013). The resulting estimator for each subset \( \mathcal{P}_c \) takes the form

\[
\sum_{m \in \mathcal{P}_c} \left( g_m(\omega) - \hat{\beta}_c(\omega) h_m(\omega) + \hat{\beta}_c(\omega) \frac{\exp(q_m^\top k_m)}{Z} \right) = \frac{\sum_{m \in \mathcal{P}_c} \exp(q_m^\top k_m)}{Z} \hat{\beta}_c(\omega).
\] (11)

**Partially Optimized Coefficients.** Given the optimality of using a separate coefficient for each key-value pair, we could further improve the estimation by selecting some subset \( E \subseteq [M] \) and employ \( \hat{\beta}_m = \beta_m^* = \nu_m \) for each \( m \in E \). Without loss of generality, we assume \( E \cap \mathcal{P}_c = \emptyset \) for all \( c = 1, \ldots, C \) and \( [M] = \left( \bigcup_{c=1}^C \mathcal{P}_c \right) \cup E \). According to Proposition 1, for each \( m \in E \) we have

\[
g_m(\omega) = g_m(\omega) - \hat{\beta}_m h_m(\omega) + \hat{\beta}_m \frac{\exp(q_m^\top k_m)}{Z} = \frac{\exp(q_m^\top k_m) \nu_m}{Z}.
\] (12)

We choose \( E \) by running an additional sparse attention mechanism (e.g., local window attention (Child et al., 2019) or Reformer (Kitaev et al., 2020)), which tend to select tokens that are more relevant to the query in sub-quadratic complexity. Since estimates on these critical tokens are exact, this strategy not only reduces the overall squared error (Equation 9), but also produces a more informative context for queries, which often translates into better empirical performance. Combining Equations 12 and 11 together, we obtain an improved Efficient attention via control VAriates (EVA),

\[
\text{EVA}(q_m, K, V) := \tilde{g}(\omega) = \sum_{m \in E} \tilde{g}_m(\omega) + \sum_{m \notin E} \tilde{g}_m(\omega) = \sum_{m \in E} \frac{\exp(q_m^\top k_m)}{Z} \nu_m + \sum_{c=1}^C \sum_{m \in \mathcal{P}_c} \frac{\exp(q_m^\top k_m)}{Z} \beta_c(\omega).
\] (13)

**Comparison with Vanilla RFA.** EVA and vanilla RFA can be re-written in a similar way (see Appendix B.7 for a detailed derivation),

\[
\text{RFA}(q_m, K, V) = \frac{\sum_{m=1}^M g_m(\omega)}{\sum_{m=1}^M h_m(\omega)},
\] (14)

\[
\text{EVA}(q_m, K, V) = \sum_{m \in E} \frac{\exp(q_m^\top k_m)}{Z} g_m(\omega) + \sum_{c=1}^C \sum_{m \in \mathcal{P}_c} \frac{\exp(q_m^\top k_m)}{Z} g_m(\omega) \frac{\sum_{m \in \mathcal{P}_c} g_m(\omega)}{\sum_{m \in \mathcal{P}_c} h_m(\omega)}.
\] (15)

Intuitively, we can think of EVA as a calibrated version of RFA. Instead of directly computing and aggregating the random feature approximation for all tokens as in RFA (Equation 14), EVA (Equation 15) first constructs local estimation for either a single token \( m \in E \) or a subset (e.g., \( \mathcal{P}_c \)), and then corrects these approximations by their corresponding true attention scores (e.g., \( \sum_{m \in \mathcal{P}_c} \exp(q_m^\top k_m) \) for \( \mathcal{P}_c \)). These adjusted local estimates are finally aggregated and globally normalized. Thanks to the decomposed representation of RFA, we can realize this divide-and-conquer strategy in a principled manner, which imposes finer-grained control on the whole estimation accuracy and enjoys increased approximation fidelity.
Table 1: Classification accuracy on ImageNet1k in comparison to different RF-based approximations. vanilla PVT-v2-b3 (Wang et al., 2021b) uses a convolutional kernel to downsample key and value vectors, resulting in fewer FLOPs but with significant performance degradation.

<table>
<thead>
<tr>
<th>Model</th>
<th># Param.</th>
<th>FLOPs</th>
<th>Top-1 Acc.</th>
<th># Param.</th>
<th>FLOPs</th>
<th>Top-1 Acc.</th>
<th># Param.</th>
<th>FLOPs</th>
<th>Top-1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>5.7M</td>
<td>1.1G</td>
<td>67.10</td>
<td>22.0M</td>
<td>4.4G</td>
<td>74.06</td>
<td>36.0M</td>
<td>7.2G</td>
<td>83.34</td>
</tr>
<tr>
<td>Performer</td>
<td>5.7M</td>
<td>1.2G</td>
<td>65.92</td>
<td>22.0M</td>
<td>4.4G</td>
<td>74.09</td>
<td>39.9M</td>
<td>7.7G</td>
<td>83.47</td>
</tr>
<tr>
<td>LARA</td>
<td>5.8M</td>
<td>1.2G</td>
<td>73.00</td>
<td>22.2M</td>
<td>4.4G</td>
<td>80.65</td>
<td>36.1M</td>
<td>7.4G</td>
<td>83.71</td>
</tr>
<tr>
<td>EVA (Ours)</td>
<td>5.8M</td>
<td>1.2G</td>
<td>72.98</td>
<td>22.0M</td>
<td>4.3G</td>
<td>74.06</td>
<td>39.9M</td>
<td>7.7G</td>
<td>83.47</td>
</tr>
<tr>
<td>Softmax</td>
<td>5.7M</td>
<td>1.3G</td>
<td>72.98</td>
<td>22.0M</td>
<td>4.6G</td>
<td>80.36</td>
<td>45.2M</td>
<td>6.9G</td>
<td>83.14†</td>
</tr>
</tbody>
</table>

4.2 Practical Implementation

According to the formulation (Equation 13) of EVA, the terms within \( E \) could be computed efficiently due to its limited size; however, the partitioning requires computing \( \sum_{m \in P_c} \exp(q_n^\top k_m) \) explicitly for each subset, which again builds up to quadratic computational complexity. As discussed above, \( \sum_{m \in P_c} \exp(q_n^\top k_m) \) serves as a weight to correct the contribution from each subset \( P_c \). In this regard, we propose to approximate such control by \( \sum_{m \in P_c} \exp(q_n^\top k_m) \approx \exp(q_n^\top k_c) \), where \( k_c \) is an adaptive vector summarizing the information of all keys belonging to \( P_c \) (see Appendix C for more details). Such heuristic not only avoids computing the exponential dot product of each query-key pair explicitly, but also induces a fast approximation of the normalizing constant,

\[
Z = \sum_{m \in E} \exp(q_n^\top k_m) + \sum_{c=1}^{C} \sum_{m \in P_c} \exp(q_n^\top k_m) \approx \sum_{m \in E} \exp(q_n^\top k_m) + \sum_{c=1}^{C} \exp(q_n^\top k_c).
\]

Equipped with these results, our EVA estimator (Equation 13) can be reduced as follows,

\[
EVA(q_n, K, V) \approx \frac{\sum_{m \in E} \exp(q_n^\top k_m) v_m + \sum_{c=1}^{C} \exp(q_n^\top k_c) \beta_c(\omega)}{\sum_{m \in E} \exp(q_n^\top k_m) + \sum_{c=1}^{C} \exp(q_n^\top k_c)}. \tag{16}
\]

Parameterization Details. We define \( E \) in the same way as a simple block-wise local attention (Xiong et al., 2021a). The input sequence is first chunked into multiple blocks (or 2D windows for images), and each query \( q_n \) is associated with a specific \( E_n \) that only contains tokens within the same block as the query. For the remaining indices \([ M ] \setminus E_n \), we evenly split it into \( C \) contiguous chunks \( \{ P_1, \ldots, P_C \} \). Note that we add the superscript \( n \) here to denote the dependence on the query position; however, for notational brevity, we omit the notation when there is no ambiguity. The pseudo-code of EVA is provided in Algorithm 1 of Appendix. More implementation details, including the definition of \( k_c \) and \( \beta_c(\omega) \) in Equation 16 are deferred to Appendix C.

Extension to Autoregressive Modeling. The decoder (or causal) self-attention, where each query can only attend to previous tokens, is the key ingredient in Transformer-based generative modeling (Vaswani et al., 2017; Brown et al., 2020). We demonstrate that it is straightforward to extend EVA to support such auto-regressive modeling with few modifications. Thanks to the decomposed formulation of EVA, we only need to incorporate two triangular mask matrices into the computation, which eliminate the information from future singletons \( m \in E \) and entire future subsets \( P_c \) respectively. Unlike previous RFA methods, which are slow during training due to their recurrent computation (Choromanski et al., 2021; Peng et al., 2021b), the resulting causal variant remains highly efficient. More details can be found in Appendix D, including a pseudo-code Algorithm 2.

5 Experimental Results

In this section, we evaluate our proposed method on various tasks, including image classification (§5.1), language tasks (§5.2), and Long Range Arena benchmark (Appendix E). Details of experimental protocols and baselines can be found in Appendix E.
We explore the ability to learn visual representations for different attention mechanisms in vision transformers (ViTs; Dosovitskiy et al., 2021). In particular, we replace softmax attention used in ViTs with its efficient variants and evaluate their performance on the ImageNet1k dataset (Deng et al., 2009), which contains over 1,280K and 50K images of 1,000 classes for training and validation splits, respectively. For the transformer model, we consider both a plain ViT (DeiT; Dosovitskiy et al., 2020; Touvron et al., 2021) and a pyramidal ViT (PVT; Wang et al., 2021b) to test the performance. The former maintains the same sequence length (which is set to 196 by default) across all transformer layers, while the latter processes much longer sequences (up to 3136 tokens) at early layers and progressively reduces the sequence length to form a hierarchical structure. Detailed experimental settings could be found in Appendix E.2.

### Results

We first compare the performance of EVA against our main baselines on the standard ViT architectures. As shown in Table 1, EVA significantly improves the performance of previous RFA approaches (including Performer (Choromanski et al., 2021) and LARA (Zheng et al., 2022b)) and local attention by a large margin, and even outperforms the conventional softmax attention. We then consider a more challenging setting, where the plain architecture DeiT-Tiny is used but the sequence length is scaled up to 784 (denoted as DeiT-Tiny-784). We compare EVA against other attention variants in this setting and report the classification results in Table 2. EVA outperforms most previous baselines and remains highly competitive with softmax attention, illustrating its effectiveness.

### 5.2 Machine Translation and Language Modeling

We further evaluate EVA on the natural language domain. Specifically, we consider three tasks:

- **Masked language modeling (MLM)** on a pretrained-scale book corpus Books3 in the Pile dataset suite (Presser et al., 2020; Gao et al., 2020), consisting of over 196,640 published books.
- **Machine translation (MT)** on WMT14 En–De benchmark (Bojar et al., 2014).
- **Autoregressive language modeling (Autoregressive LM)** on a large-scale token-level LM benchmark Wikitext-103 (Merity et al., 2016).

### Results

We report MLM validation perplexity in Table 5, where the sequence length is 2048 by default. EVA substantially improves previous methods based on random features (including Performer and LARA) and outperforms the other efficient attention mechanisms. Thanks to the linear
In this section, we conduct an ablation study on image classification and MLM tasks to investigate the effects of main hyper-parameters in EVA (see Table 8 for more comprehensive analysis). In particular, we vary $|E|$ and the partition size $C$ and evaluate their performance on both image classification and masked language modeling. As presented in Table 6 and Table 7, increasing $|E|$ amounts to obtaining exact estimates for more key-value pairs, which greatly improves empirical performance; besides, increasing $C$ would process control variates at a finer scale, also translating into better modeling quality, consistent with our theoretical analysis (§4.1).
6 RELATED WORK

Control Variates. Control variates are a widely used variance reduction technique in reinforcement learning (Greensmith et al., 2004; Grathwohl et al., 2018; Vlassis et al., 2021), stochastic optimization (Wang et al., 2013), variational inference (Paisley et al., 2012; Ranganath et al., 2014; Geffner & Domke, 2013; Tucker et al., 2017; Grathwohl et al., 2018), Markov chain Monte Carlo (Baker et al., 2019) and many other topics. Our construction with control variates provides a new perspective on designing faster yet more accurate attention approximations.

Efficient Attention Mechanisms. A lot of research work has put the focus on reducing the quadratic complexity of conventional softmax attention. A widely used approach is to define a sparse attention pattern so that each query is limited to only attending to a subset of tokens. The sparse pattern could be either learnable (Kitaev et al., 2020; Vyas et al., 2020; Tay et al., 2020; Roy et al., 2021; Madaan et al., 2022) or simply fixed (Liu et al., 2018; Parmar et al., 2018; Child et al., 2019; Beltagy et al., 2020; Ansie et al., 2020; Zhaeber et al., 2020; Liu et al., 2021; Xiong et al., 2021a; Wang et al., 2022; Chen et al., 2022; Hutchins et al., 2022). Another paradigm is to adopt low-rank approximations, including via the Nyström method (Xiong et al., 2021b), down-sampling with learnable projections (Wang et al., 2020; Peng et al., 2021a), or explicitly compressing sequences (Rae et al., 2020; Dai et al., 2020; Ma et al., 2021; Jaegle et al., 2021). There are also studies improving both sparse and low-rank methods for better attention matrix approximation (Nguyen et al., 2021; Zhu et al., 2021; Chen et al., 2021a; Ren et al., 2021; Zhu & Soricut, 2021; Hua et al., 2022; Zheng et al., 2022). Instead of adopting approximate methods, a recent line of work (Rabe & Staats, 2021; Dao et al., 2022) proposes to compute the exact softmax attention in an online manner (Milakov & Gimelshein, 2018) without materializing the full attention matrix. In this way, softmax attention can be computed in linear memory complexity, and the runtime can also be greatly improved by further minimizing memory accesses (Dao et al., 2022).

Random-Feature-based Attention. Random-feature-based methods are a popular alternative that uses random features (Rahimi & Recht, 2008) to linearize exponential kernels in softmax attention (Katharopoulos et al., 2020; Choromanski et al., 2021; Peng et al., 2021b). Recent work attempts to improve RFA approximation from several aspects, such as designing more accurate random feature maps (Choromanski et al., 2022; Likhosherstov et al., 2022; Chowdhury et al., 2022), incorporating relative positional or other task-specific biases (Lutkus et al., 2021; Luo et al., 2021; Chen, 2021; Zheng et al., 2022a; Qin et al., 2022b; Wu et al., 2022; Qin et al., 2022a), or leveraging connections to fast weight programmers (Peng et al., 2021b; Schlag et al., 2021; Irie et al., 2021). Prior work closely related to ours includes Zheng et al. (2022b), which reinterprets RFA using self-normalized importance sampling (Hesterberg, 1995) and theoretically extends the random feature approximation from individual exponential kernels to the whole softmax attention. Our work further generalizes this result via control variates and characterizes the approximation gap caused by RFA. Scatterbrain (Chen et al., 2021a) is also similar to our work in that it also refines RF approximation on critical local regions. However, it is developed based on a different motivation that attempts to approximate the attention matrix with a combination of sparse and low-rank matrices. Interestingly, we find that Scatterbrain can be cast as a special case under our framework; see Appendix G for a detailed discussion about connections between EVA and previous attention mechanisms.

7 CONCLUSION AND LIMITATIONS

In this work, we develop an efficient attention mechanism EVA via control variates. Our framework reveals a localized perspective of RFA approximation, which not only bridges the gap between RFA and exact softmax attention but also attains a good trade-off between modeling quality and efficiency. We evaluate our method on both vision and language tasks and demonstrate substantial improvements over previous baselines. There are some limitations of our framework. For instance, the approximation in computing control variate estimation for each partitioned subset is crude and might limit the potential modeling capacity; in addition, we only explore the most straightforward partitioning strategy that evenly splits the sequence into multiple contiguous chunks; while in general, the partition could contain arbitrary subsequences or be adaptive to inputs via clustering methods, which can be guided by task-specific inductive biases. It is interesting to investigate these limitations to unleash the expressiveness of EVA further, which we leave for future work.
ACKNOWLEDGMENTS

We would like to thank the HKU NLP group, the Shark-NLP group, and the anonymous reviewers for their valuable suggestions that greatly helped improve this work. This work is partially supported by the joint research scheme of the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) under grant number N_HKU714/21.

REFERENCES


Appendix

A  A Brief Review of Vanilla Random Feature Attention  

B  Proofs & Derivations  
   B.1  An Extended Review of Control Variates  
   B.2  Derivation of SNIS as Control Variate Estimation  
   B.3  Derivation of the Expectation of Per-term Control Variates  
   B.4  Proof of Proposition 1  
   B.5  A Formal Analysis of the Advantage of Partitioning  
   B.6  Proof of Proposition 2  
   B.7  Derivation of Equations 14 and 15  

C  More Implementation Details for EVA  

D  A Causal Variant of EVA  

E  Experimental Details  
   E.1  Efficient Attention Baselines  
   E.2  Image Classification  
   E.3  Machine Translation and Language Modeling  
   E.4  Experimental Settings of Efficiency Comparison  

F  Experiments on Long Range Arena  

G  Connections to Other Attention Mechanisms  
   G.1  RFA, Softmax Attention, and EVA  
   G.2  Connections to LARA  
   G.3  Connections to Clustered Attention  
   G.4  Connections to Combiner  
   G.5  Connections to Scatterbrain
A Brief Review of Vanilla Random Feature Attention

Vanilla random feature attention methods, such as Performer (Choromanski et al., 2021; Peng et al., 2021b), seek to approximate the softmax attention mechanism through random features (Rahimi & Recht, 2008) \( \phi(x, \omega) := 1/\sqrt{S} [\xi(x, \omega_1), \ldots, \xi(x, \omega_S)]^\top. \) Here, \( \omega_1, \ldots, \omega_S \sim \mathcal{N}(0, I), \) and \( \xi(x, \omega) \) is the randomized mapping such that

\[
\exp(\mathbf{q}_n^\top \mathbf{k}_m) = \mathbb{E}_{\omega \sim \mathcal{N}(0, I)} [\xi(\mathbf{q}_n, \omega)^\top \xi(\mathbf{k}_m, \omega)] .
\]  

(17)

Therefore, we can draw multiple Monte Carlo samples to estimate the exponential kernel,

\[
\exp(\mathbf{q}_n^\top \mathbf{k}_m) \approx \frac{1}{S} \sum_{s=1}^{S} \xi(\mathbf{q}_n, \omega_s)^\top \xi(\mathbf{k}_m, \omega_s) := \phi(\mathbf{q}_n, \omega)^\top \phi(\mathbf{k}_m, \omega),
\]

and then approximate the attention mechanism as

\[
\sum_{m=1}^{M} \exp(\mathbf{q}_n^\top \mathbf{k}_m) \mathbf{v}_m \approx \frac{1}{S} \sum_{s=1}^{S} \xi(\mathbf{q}_n, \omega_s)^\top \xi(\mathbf{k}_m, \omega_s) \mathbf{v}_m.
\]

(18)

It is recently generalized as a self-normalized importance sampling estimator to approximate softmax attention (Zheng et al., 2022b), as described in §2.2. We refer the generalized random feature based approximations as RFA.

B Proofs & Derivations

B.1 An Extended Review of Control Variates

The control variate method takes the following form,

\[
\tilde{g}(\omega) = g(\omega) - \beta h(\omega) + \beta \mathbb{E} [h(\omega)],
\]

(19)

Given the particular forms of \( g(\cdot) \) and \( h(\cdot) \), \( \beta \) can be optimized to minimize the estimation variance. For notational convenience, we denote the covariance between a scalar and a random vector as \( \text{Cov} [h(\omega), g(\omega)] := \mathbb{E} [(h(\omega) - \mathbb{E} [h(\omega)])(g(\omega) - \mathbb{E} [g(\omega)])] \) and the variance of a random vector as \( \text{Var} [g(\omega)] := \text{Cov} [g(\omega), g(\omega)]. \) In particular, we have

\[
\text{Var} [\tilde{g}(\omega)] = \text{Var} [g(\omega) - \beta h(\omega)]
= \text{Var} [g(\omega)] - 2 \text{Cov} [\beta h(\omega), g(\omega)] + \text{Var} [\beta h(\omega)]
= \text{Var} [g(\omega)] - 2 \text{Cov} [h(\omega), g(\omega)] \beta^\top + \text{Var} [h(\omega)] \beta \beta^\top.
\]

We hope an optimal \( \beta \) would minimize \( \text{Tr} (\text{Var} [\tilde{g}(\omega)]) \), that is, the sum of estimating variance for each dimension. By differentiating, we obtain

\[
\beta^* = \arg \min_{\beta} \text{Tr} (\text{Var} [\tilde{g}(\omega)]) = \frac{\text{Cov} [h(\omega), g(\omega)]}{\text{Var} [h(\omega)]}.
\]

(20)

Since both the covariance and the variance may be intractable to compute, the optimal \( \beta^* \) is generally not available in closed form. Nevertheless, with the optimal coefficient, the variance of such control variate estimate would never be larger than the plain estimator \( g(\cdot) \).
B.2 Derivation of SNIS as Control Variate Estimation

For notational convenience, we denote the importance weight as $W(\omega_s) := p_n(\omega_s)/q(\omega_s)$. Then we have

$$
\bar{g}(\omega) = \frac{\sum_{s=1}^{S} p_n(\omega_s) f(\omega_s)}{\sum_{s=1}^{S} p_n(\omega_s) q(\omega_s)} = \frac{\sum_{s=1}^{S} W(\omega_s) f(\omega_s)}{\sum_{s=1}^{S} W(\omega_s)} = \frac{\sum_{s=1}^{S} W(\omega_s) f(\omega_s)}{S} = \frac{1}{S} \sum_{s=1}^{S} W(\omega_s) f(\omega_s) + \frac{1}{S} \sum_{s=1}^{S} W(\omega_s) f(\omega_s)
$$

For $N$ large enough, $\bar{g}(\omega)$ becomes asymptotically unbiased since $\bar{g}(\omega)$ is consistent and converges to a constant $\beta$ w.r.t. $\omega$.

Note that the expectation of importance weights equals 1, that is,

$$
E[h(\omega)] = E\left[\frac{1}{S} \sum_{s=1}^{S} W(\omega_s)\right] = E_{\omega_1, \ldots, \omega_S \sim q(\omega)} \left[\frac{1}{S} \sum_{s=1}^{S} p(\omega_s) \right] = \frac{1}{S} \sum_{s=1}^{S} E_{\omega_s \sim q(\omega)} \left[p(\omega_s) \right] = 1.
$$

Same as SNIS, this estimator is still biased due to the dependence of $\beta(\omega)$ on $\omega$. However, it would asymptotically become unbiased since $\beta(\omega)$ is consistent and converges to a constant $\beta$ w.r.t. $\omega$ given a large number of samples,

$$
\hat{\beta}(\omega) = \frac{g(\omega)}{h(\omega)} = \frac{E[g(\omega)]}{E[h(\omega)]} = \frac{E[p_n(\omega)] f(\omega)}{\text{constant}} := \beta.
$$

B.3 Derivation of the Expectation of Per-term Control Variates

According to the definition of randomized mappings, we have

$$
E[h_m(\omega)] = E_{\omega_1, \ldots, \omega_S \sim q(\omega)} \left[\frac{1}{S} \sum_{s=1}^{S} \mathcal{N}(\omega_s; 0, 1) \xi(q_m, \omega_s) \xi(k_m, \omega_s)\right] = \frac{1}{S} \sum_{s=1}^{S} \int \xi(q_m, \omega_s) \xi(k_m, \omega_s) \mathcal{N}(\omega_s; 0, 1) d\omega_s = \frac{\exp(q_m^T k_m)}{Z}.
$$

B.4 Proof of Proposition

Proof. We start with the formulation of $g(\cdot)$ and $h(\cdot)$,

$$
g_m(\omega) = \sum_{s=1}^{S} \mathcal{N}(\omega_s; 0, 1) \xi(q_m, \omega_s) \xi(k_m, \omega_s) v_m = h_m(\omega).
$$
As a result, we have $g_m(\omega) = h_m(\omega)v_m$ and $E[g_m(\omega)] = E[h_m(\omega)]v_m$. We now investigate the optimal $\beta^*_m$ according to Equation 20:

$$\beta^*_m = \arg\min_\beta \text{Tr}(\text{Var}[g_m(\omega)])$$

$$= \frac{\text{Cov}[h_m(\omega), g_m(\omega)]}{\text{Var}[h_m(\omega)]}$$

$$= \frac{E[(h(\omega) - E[h(\omega)])(h(\omega) - E[h(\omega)])]v_m}{E[(h(\omega) - E[h(\omega)])(h(\omega) - E[h(\omega)])]}$$

$$= v_m = \frac{g_m(\omega)}{h_m(\omega)}.$$ 

In terms of the variance, we again use $g_m(\omega) = h_m(\omega)v_m$ to obtain

$$\bar{g}_m(\omega) = g_m(\omega) - \bar{\beta}_m(h_m(\omega) - E[h_m(\omega)])$$

$$= g_m(\omega) - v_mh_m(\omega) + v_mE[h_m(\omega)]$$

$$= \exp(q_m^\top k_m)\frac{v_m}{Z}.$$

Since this holds true for every term $m = 1, \ldots, M$, our estimate becomes exactly softmax attention, $\bar{g}(\omega) = \sum_{m=1}^{M} \bar{g}_m(\omega) = \sum_{m=1}^{M} \frac{\exp(q_m^\top k_m)}{Z}v_m = \sum_{m=1}^{M} \frac{\exp(q_m^\top k_m)}{\sum_{m'=1}^{M} \exp(q_{m'}^\top k_{m'})}v_m$.

Since all randomness is eliminated, the estimate is exact with zero bias and variance. That is, \(\text{RFA}(q_m, K, V) = \bar{g}(\omega) = \text{SoftmaxAttn}(q_m, K, V)\).

\[\square\]

### B.5 A Formal Analysis of the Advantage of Partitioning

In this section, we demonstrate the advantage of partitioning by showing that there always exists some set $\{\beta_c\}_{c=1}^C$ that achieves lower weighted MSE than any globally shared coefficient, as discussed in §4.1.

**Lemma 3.** Suppose $\beta^*_m$ is the optimal coefficient for each $m \in [M]$ as defined in Proposition 7 and $P_1, P_2, \ldots, P_C$ are an arbitrary partition of $[M]$, where each subset $P_c$ is associated with a distinct $\beta_c$. We consider the following weighted mean squared error:

$$J(\beta_1, \ldots, \beta_C) := \sum_{c=1}^{C} \sum_{m \in P_c} \alpha_m \|\beta_c - \beta^*_m\|^2,$$  

(24)

where $\alpha_m > 0$ for each $m \in [M]$ and $\sum_{c=1}^{C} \sum_{m \in P_c} \alpha_m = 1$. Then for any choice of $\{\alpha_m\}_{m=1}^{M}$ and any globally shared coefficient $\beta$, there exists some $\{\beta^*_c\}_{c=1}^C$ so that $J(\beta_1 = \beta, \ldots, \beta_C = \beta) \geq J(\beta_1 = \beta^*_1, \ldots, \beta_C = \beta^*_C)$.

**Proof.** Let $\beta^*_c = \frac{\sum_{m \in P_c} \alpha_m \beta^*_m}{\sum_{m \in P_c} \alpha_m}$ for each $c = 1, \ldots, C$. Then we have

$$\sum_{m \in P_c} \alpha_m \left(\beta^*_c - \beta^*_m\right) = \beta^*_c \left(\sum_{m \in P_c} \alpha_m\right) - \sum_{m \in P_c} \alpha_m \beta^*_m$$

$$= \frac{\sum_{m \in P_c} \alpha_m \beta^*_m}{\sum_{m \in P_c} \alpha_m} \left(\sum_{m \in P_c} \alpha_m\right) - \sum_{m \in P_c} \alpha_m \beta^*_m$$

$$= \sum_{m \in P_c} \alpha_m \beta^*_m - \sum_{m \in P_c} \alpha_m \beta^*_m = 0.$$  

(25)
According to Equations \[25\] and \[24\] for any \( \beta \) we have the following inequality,

\[
J(\beta_1 = \beta, \ldots, \beta_C = \beta) = \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta - \beta_m^* \|^2 \\
= \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta - \beta_c^* + \beta_c^* - \beta_m^* \|^2 \\
= \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \left( \| \beta - \beta_c^* \|^2 + 2 (\beta - \beta_c^*)^T (\beta_c^* - \beta_m^*) + \| \beta_c^* - \beta_m^* \|^2 \right) \\
= \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta - \beta_c^* \|^2 + 2 \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m (\beta - \beta_c^*)^T (\beta_c^* - \beta_m^*) + \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta_c^* - \beta_m^* \|^2 \\
= \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta - \beta_c^* \|^2 + \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta_c^* - \beta_m^* \|^2 \\
\geq \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \alpha_m \| \beta_c^* - \beta_m^* \|^2 \\
= J(\beta_1^* = \beta_1^*, \ldots, \beta_C^* = \beta_C^*).
\]

As a result, for any choice of \( \{\alpha_m\}_{m=1}^{M} \) and any globally shared coefficient \( \beta \), there always exists some \( \{\beta_c^*\}_{c=1}^{C} \) that achieves lower (or equal) weighted MSE, and a solution can be simply \( \beta_c^* = \frac{\sum_{m \in \mathcal{P}_c} \alpha_m \beta_m^*}{\sum_{m \in \mathcal{P}_c} \alpha_m} \).

\[ \blacksquare \]

B.6 Proof of Proposition \[2\]

Proof. We first consider the case of partitioned indices, where each subset \( \mathcal{P}_c \) is associated with some specific \( \beta_c \). To see the global minimum of \( J \), we differentiate on both sides and obtain

\[
\frac{\partial J(\beta_1^*, \ldots, \beta_C^*)}{\partial \beta_c} = \frac{\partial}{\partial \beta_c} \sum_{c=1}^{C} \sum_{m \in \mathcal{P}_c} \sum_{m' \in U} \exp (\mathbf{q}_m^T \mathbf{k}_m) \| \beta_c - \beta_m^* \|^2 \\
= \sum_{m \in \mathcal{P}_c} \sum_{m' \in U} \exp (\mathbf{q}_m^T \mathbf{k}_m) \mathbf{q}_m^T \mathbf{k}_m \| \beta_c - \beta_m^* \|^2.
\]

By setting the partial derivative to zero, we obtain

\[
\beta_c^* = \frac{\sum_{m \in \mathcal{P}_c} \exp (\mathbf{q}_m^T \mathbf{k}_m) \beta_m^*}{\sum_{m \in \mathcal{P}_c} \exp (\mathbf{q}_m^T \mathbf{k}_m)} = \frac{\sum_{m \in \mathcal{P}_c} \exp (\mathbf{q}_m^T \mathbf{k}_m) \mathbf{v}_m}{\sum_{m \in \mathcal{P}_c} \exp (\mathbf{q}_m^T \mathbf{k}_m)} \\
= \frac{\sum_{m \in \mathcal{P}_c} \mathbb{E} [g_m(\omega)]}{\sum_{m \in \mathcal{P}_c} \mathbb{E} [h_m(\omega)]} = \mathbb{E} \left[ \frac{\sum_{m \in \mathcal{P}_c} g_m(\omega)}{\sum_{m \in \mathcal{P}_c} h_m(\omega)} \right].
\]

As a consequence, with \( \beta_c = \beta_c^* \), the partition scheme must achieve lower weighted mean squared error than any globally shared \( \bar{\beta} \), that is, \( J(\beta_1 = \beta_1^*, \ldots, \beta_C = \beta_C^*) \leq J(\beta_1 = \bar{\beta}, \ldots, \beta_C = \bar{\beta}) \).
In fact, with $\beta_c = \tilde{\beta}_c$, the partition scheme usually enjoys much lower error than adopting a globally shared coefficient. To see the error reduction of using the partitioned strategy, we first have

$$J(\beta_1 = \tilde{\beta}, \ldots, \beta_C = \tilde{\beta})$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \sum_{m' \in U} \exp (q_n^T k_m') \left\| \tilde{\beta} - \beta_c^* \right\|^2$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \sum_{m' \in U} \exp (q_n^T k_m') \left\| \tilde{\beta} - \beta_c + \beta_c - \beta_m^* \right\|^2$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \sum_{m' \in U} \exp (q_n^T k_m') \left\| \tilde{\beta} - \beta_c + \beta_c - \beta_m^* \right\|^2$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \sum_{m' \in U} \exp (q_n^T k_m') \left( \left\| \tilde{\beta} - \beta_c \right\|^2 + \left( \tilde{\beta} - \beta_c \right)^T (\beta_c - \beta_m^*) + \beta_c - \beta_m^* \right)$$

Since

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \sum_{m' \in U} \exp (q_n^T k_m) \left( \tilde{\beta} - \beta_c \right)^T (\beta_c - \beta_m^*)$$

$$= C \sum_{c=1}^{C} \left( \tilde{\beta} - \beta_c \right)^T \sum_{m \in P_c} \sum_{m' \in U} \exp (q_n^T k_m) (\beta_c - \beta_m^*)$$

$$= C \sum_{c=1}^{C} \left( \tilde{\beta} - \beta_c \right)^T \sum_{m \in P_c} \sum_{m' \in U} \exp (q_n^T k_m) \left( \sum_{m' \in U} \exp (q_n^T k_m) v_m - \beta_m^* \right)$$

$$= C \sum_{c=1}^{C} \left( \tilde{\beta} - \beta_c \right)^T \frac{\sum_{m \in P_c} \sum_{m' \in U} \exp (q_n^T k_m) \left( v_m - \beta_m^* \right)}{\sum_{m \in P_c} \sum_{m' \in U} \exp (q_n^T k_m)}$$

$$= 0,$$

plugging this result back we obtain

$$J(\beta_1 = \tilde{\beta}, \ldots, \beta_C = \tilde{\beta})$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \left( \left\| \tilde{\beta} - \beta_c \right\|^2 + \left( \tilde{\beta} - \beta_c \right)^T (\beta_c - \beta_m^*) + \beta_c - \beta_m^* \right)$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \left( \left\| \tilde{\beta} - \beta_c \right\|^2 + \beta_c - \beta_m^* \right)$$

$$= C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \left( \left\| \tilde{\beta} - \beta_c \right\|^2 + \sum_{c=1}^{C} \sum_{m' \in U} \exp (q_n^T k_m') \beta_c - \beta_m^* \right)$$

$$\geq C \sum_{c=1}^{C} \sum_{m \in P_c} \exp (q_n^T k_m) \left( \left\| \beta_c - \beta_m^* \right\|^2 \right).$$

The last inequality holds since the first term is always non-negative. Note that the first term computes the squared error between $\beta$ and each $\beta_c$, weighted by the sum of attention scores over the corresponding subset. As a result, it is usually positive and the error reduction is significant if each $\beta_c$ deviates from $\tilde{\beta}$ a lot. However, although the optimal coefficient in the partitioning always leads to lower error to the optimal individual coefficient, note that it does not necessarily yield lower estimation variance.
### Table 8: Classification results on ImageNet1k dataset under different hyper-parameter configurations of EVA. By default, we set $|E| = 49$ and $C = 49$ across all variants below.

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
<th>Top-1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition Scheme of {P_1, \ldots, P_C}</td>
<td>\begin{align*} \text{partition over} &amp; [M] \setminus E \ \text{partition over} &amp; [M] \end{align*}</td>
<td>76.53</td>
</tr>
<tr>
<td>Parameterization of $\sigma(\cdot)$</td>
<td>$\begin{align*} \sigma(\cdot) &amp;= \text{LN(Linear(\cdot))} \ \sigma(\cdot) &amp;= \text{Identity(\cdot)} \end{align*}$</td>
<td>76.67</td>
</tr>
<tr>
<td>Number of Groups ($C = 1$)</td>
<td>Number of Samples = 1</td>
<td>74.10</td>
</tr>
<tr>
<td></td>
<td>Number of Samples = 49</td>
<td>76.39</td>
</tr>
<tr>
<td>Number of Groups ($C = 49$)</td>
<td>Number of Samples = 1</td>
<td>74.10</td>
</tr>
<tr>
<td></td>
<td>Number of Samples = 49</td>
<td>76.39</td>
</tr>
<tr>
<td>Proposal Parameterization $q_c(\omega) := \mathcal{N}(\omega; \mu_c, I)$</td>
<td>$\begin{align*} \mu_c &amp;= \hat{q}_c + \hat{k}_c \ \mu_c &amp;= \hat{q}_c \ \mu_c &amp;= 0 \ \mu_c &amp;= \text{Trainable parameters} \end{align*}$</td>
<td>76.67</td>
</tr>
<tr>
<td></td>
<td>Softmax</td>
<td>77.16</td>
</tr>
</tbody>
</table>

### B.7 Derivation of Equations 14 and 15

According to the definition of $g_m(\cdot)$ and $h_m(\cdot)$ in §3.1 for the vanilla RFA (Equation 4), we have

$$\text{RFA}(q_n, K, V) = \sum_{s=1}^{S} \frac{p_n(\omega_s)}{q(\omega_s)} f(\omega_s) = \frac{\sum_{s=1}^{S} p_n(\omega_s)}{\sum_{s=1}^{S} q(\omega_s)} g(\omega) = \frac{\sum_{m=1}^{M} g_m(\omega)}{\sum_{m=1}^{M} h_m(\omega)}.$$  

Besides, since $v_m = g_m(\omega)/h_m(\omega)$ and $\beta_c(\omega) = (\sum_{m \in P_c} g_m(\omega)) / (\sum_{m \in P_c} h_m(\omega))$, we can re-write EVA as

$$\text{EVA}(q_n, K, V) := \tilde{g}(\omega) = \sum_{m \in E} \tilde{g}_m(\omega) + \sum_{m \notin E} \tilde{g}_m(\omega) = \sum_{m \in E} \frac{\exp(q_n^T k_m)}{Z} v_m + \sum_{c=1}^{C} \frac{\exp(q_n^T k_m)}{Z} \beta_c(\omega) = \sum_{m \in E} \frac{\exp(q_n^T k_m) g_m(\omega)}{Z h_m(\omega)} + \sum_{c=1}^{C} \frac{\sum_{m \in P_c} \exp(q_n^T k_m)}{Z} \sum_{m \in P_c} \frac{g_m(\omega)}{h_m(\omega)}.$$  

### C More Implementation Details for EVA

In this section, we provide more details of EVA. We also conduct a comprehensive ablation study to test the effect of different components in our implementation and report the results in Table 8. The pseudo-code for EVA is listed in Algorithm 1.

**Approximating** $\sum_{m \in P_c} \exp(q_n^T k_m)$ and **Parameterizing** $\hat{k}_c$. In our implementation, we approximate the sum of exponentials as $\sum_{m \in P_c} \exp(q_n^T k_m) \approx \exp(q_n^T \hat{k}_c)$. Here we provide an informal justification for this approximation.

Our main motivation for such approximation is based on the simple intuition that the sum of exponentials grows as fast as the maximum exponential value, as reflected by the following inequality,

$$\max_{m \in P_c} \exp(q_n^T k_m) \leq \sum_{m \in P_c} \exp(q_n^T k_m) \leq |P_c| \max_{m \in P_c} \exp(q_n^T k_m).$$

This means we can approximate the sum of exponentials by first computing the group representative $\hat{k}_c := \arg \max_{k_m \in [k_m, m \in P_c]} \exp(q_n^T k_m)$, evaluating the corresponding exponential $\exp(q_n^T \hat{k}_c)$
We can think of each coefficient $b$ where $e$ is calculated similarly to Equation 26. We refer readers to Zheng et al. (2022b) for more discussions about the parameterization choice of proposal distributions. We conduct further ablation studies to test the effect of proposal parameterizations in our proposed model, as shown in Table 8. In particular, we found our model is robust to different parameterization approaches.
The essence in making the algorithm memory-efficient is to use only one sample in calculating \( \tilde{\beta}_e(\omega) \). In this case, we have

\[
\tilde{\beta}_e(\omega) = \frac{\sum_{m \in P} g_m(\omega)}{\sum_{m \in P} h_m(\omega)} \equiv \frac{N(\omega;0,1)}{Z_q(\omega)} \sum_{m \in P} \xi(q_m, \omega) \xi(k_m, \omega) v_m \equiv \frac{N(\omega;0,1)}{Z_q(\omega)} \sum_{m \in P} \xi(q_m, \omega) \xi(k_m, \omega) v_m = \sum_{m \in P} \xi(k_m, \omega) v_m, \quad \omega_c \sim q_e(\omega).
\]

Since this degenerated formulation eliminates the dependence on individual queries \( q_n \), we could pre-compute \( \tilde{\beta}_e(\omega) \) for each \( P_c \), and then re-uses them for each query, which takes up \( \mathcal{O}(Cd) \) memory. If multiple samples are used instead, the influence of queries needs to be explicitly taken into account and thus we need to compute a distinct \( \tilde{\beta}_e(\omega) \) for each query, leading to \( \mathcal{O}(NCd) \) memory usage, which incurs a significant compute overhead.

On the other hand, if we set \( C = 1 \), that is, using a shared \( \tilde{\beta}_e(\omega) \) over all \( m \notin E \), our approach does not suffer from this issue, since the memory usage is at most \( \mathcal{O}(Nd) \). To investigate the effect of using larger \( C \) or increasing the number of samples, we conduct an ablative analysis as in Table 8 and find that 1) when \( C = 1 \), the performance usage degrades a lot when using one sample, which can be largely improved by adopting more samples; while when \( C > 1 \), our partitioning strategy dominates and increasing the number of samples only improves performance marginally. This also validates the effectiveness of adopting a finer-grained treatment over control variates.

**Partitioning Strategy.** EVA significantly improves random feature approximation by trying to locally estimate each subset of tokens, which is a much easier task than approximating the whole sequence as in previous RFA methods. To achieve this, EVA partitions the whole token sequence into multiple subsets according to the current query position \( n \), which is denoted by \( \{ E^n, P^n_1, P^n_2, \ldots, P^n_C \}_{n=1}^N \). For elements in subset \( E^n \), we optimize the control variate coefficient to give an exact estimate for each single token \( m \in E^n \). In addition, we impose T5-style relative positional encoding (Raffel et al., 2020a) over elements in \( E^n \). While for some other subset \( P_c \), we employ the shared coefficient to approximate all tokens belonging to \( P_c \). We assume all \( E^1, \ldots, E^N \) are of the same cardinality \( K \), and \( |P^n_c| \) is the same for any \( c = 1, \ldots, C \) and \( n = 1, \ldots, N \).

The partition strategy \( \{ E^n, P^n_1, P^n_2, \ldots, P^n_C \}_{n=1}^N \) is decided based on a simple criterion:

- for \( E^n \), it contains \( K \) local neighbors with respect to each query \( n \). To further simplify implementation and reduce memory usage, we chunk the whole sequence into contiguous blocks of size \( K \), and all adjacent queries belonging to the same block will share this block as the subset \( E^n \); 
- as for \( P^n_1, P^n_2, \ldots, P^n_C \), we follow a similar treatment by splitting the complement \( [M] \setminus E^n \) into \( C \) contiguous chunks of the same size. For ease of implementation, we simply partition the whole index set \( [M] \) into multiple groups instead of \( [M] \setminus E^n \), which circumvents the overload for explicitly performing set difference operations in practical implementation. Although this leads to extra approximation error, this amounts to putting more attention weights on tokens belonging to the subset \( E \) and we found this approximation does not lead to performance degradation (Table 8).

**D A CAUSAL VARIANT OF EVA**

In this section, we describe the causal variant of EVA, where each query can only attend to historical tokens. Thanks to the partitioning scheme, all future information with respect to the current query token can be masked conveniently. Following the formulation of EVA, we partition the whole sequence into \( C + 1 \) subsets \( \{ E^n, P^n_1, P^n_2, \ldots, P^n_C \} \) with respect to each query \( q_n \). To fulfill the
We compare our proposed attention mechanism EVA against various baselines:

- **Reformer** (Kitaev et al., 2020), a sparse approximation where hash functions are used to adaptively
  - **Local attention** (Child et al., 2019), a simple sparse approximation that splits the whole sequence
  - **Nyströmformer** (Xiong et al., 2021b), a low-rank approximation that adopts the Nyström method
  - **Linformer** (Wang et al., 2020), a low-rank approximation that uses a learnable matrix to project the
  - **LARA** (Zheng et al., 2022b), an advanced RF approximation that makes use of multiple adaptive

**E.1 Efficient Attention Baselines**

We compare our proposed attention mechanism EVA against various baselines:

- **Performer** (Choromanski et al., 2021), which uses the plain random features to approximate softmax attention;
- **LARA** (Zheng et al., 2022b), an advanced RF approximation that makes use of multiple adaptive
  - **Linformer** (Wang et al., 2020), a low-rank approximation that uses a learnable matrix to project the
  - **Nyströmformer** (Xiong et al., 2021b), a low-rank approximation that adopts the Nyström method
  - **Local attention** (Child et al., 2019), a simple sparse approximation that splits the whole sequence
  - **Reformer** (Kitaev et al., 2020), a sparse approximation where hash functions are used to adaptively

Algorithm 2 Pseudo-code for Causal EVA

Input: the randomized mapping $\xi(\cdot, \cdot)$, queries $Q := \{q_n\}_{n=1}^N$, keys $K := \{k_m\}_{m=1}^M$, values $V := \{v_m\}_{m=1}^M$, and partitions of the sequence $\{E^n, P_1^n, P_2^n, \ldots, P_C^n\}_{n=1}^N$.
Output: attention output $Y := \{y_n\}_{n=1}^N$.

for $c = 1, 2, \ldots, C$ do
    Compute $K_c$ according to Equation 26.
    Compute $q_c(\omega)$ according to Equation 28.
    Sample $\omega_c \sim q_c(\omega)$; \hfill $\triangleright$ During inference, simply set $\omega_c = \mathbb{E}_{q_c}(\omega)$.
    Compute $\beta_c(\omega) = \sum_{m \in P_c^n} \frac{\xi(k_m, \omega_c)}{Z} v_m$.
end for

Let $K \leftarrow \{E^n\}$; \hfill $\triangleright$ we assume all $E^n$ are the same in size.
Initialize $M^E \in \{0, 1\}^{K \times K}$ such that $M^E_{i,j} = 1_{i \leq j}$; \hfill $\triangleright$ Intra-$E$ masking matrix.
Initialize $M^P \in \{0, 1\}^{C \times C}$ such that $M^P_{c,t} = 1_{c \leq t}$; \hfill $\triangleright$ Inter-$P$ masking matrix.

for $n = 1, 2, \ldots, N$ do
    Find index $t$ such that $P^n_n$ is the most recent chunk on the left of $E$.
    Let $b^n \leftarrow \min\{i : i \in E^n\}$; \hfill $\triangleright$ The least position within $E^n$; used for shifting token indices.
    \hfill $\triangleright$ The same masking matrix $M^E$ can be reused across $n$ via shifting token positions by $b^n$.
    Compute $S = \sum_{m \in E^n} M^E_{m-b^n,n-b^n} \exp\left(q_n^\top k_m\right) v_m$.
    Compute $R = \sum_{c=1}^C M^P_{c,t} \exp\left(q_n^\top k_c\right) \beta_c(\omega)$.
    Compute $Z = \sum_{m \in E^n} M^E_{m-b^n,n-b^n} \exp\left(q_n^\top k_m\right) + \sum_{c=1}^C M^P_{c,t} \exp\left(q_n^\top k_c\right)$.
    Compute $y_n = (S + R) / Z$.
end for
Return $Y := \{y_1, \ldots, y_N\}$.

- Scatterbrain (Chen et al. 2021a), an approach that combines Performer and sparse attention. The details can be found in Appendix C. Here we implement the sparse module as a simple local attention to ensure a fair comparison.
- Combiner (Ren et al. 2021), a probabilistic approach that constructs a structured factorization over the softmax probability distribution via a sparse mechanism. Combiner allows both direct and indirect calculations of conditional probabilities, where the direct probability is implemented as the sparse mechanism while the indirect probability is implemented through a local abstraction over a group of tokens. Similarly, we implement the sparse mechanism as a simple local attention, which corresponds to the Combiner-Fixed variant (Ren et al. 2021);
- Transformer-LS, or Long-Short (Zhu et al. 2021), which is proposed to model long-term and short-term dependencies via low-rank structures and local attention respectively. The low-rank structure is defined as an input-dependent weight matrix that compresses the sequence into a shorter one; while the local attention is defined similarly as above.

Note that for all mechanisms that involve a local attention, we split the sequence into non-overlapping blocks (or 2D windows in terms of images) and each query can only attend to tokens within the same block. We also use the relative positional embedding (Raffel et al. 2020b; Liu et al. 2021) within the local attention computation. Unlike Transformer-LS (Zhu et al. 2021) that allows each query to attend to multiple blocks, we do not use this extension as we find greatly increases memory consumption, although it does improve the model performance.

E.2 IMAGE CLASSIFICATION

Through the experiments on image classification, we consider four different vision transformer (ViT) architectures:
Table 9: Our hyper-parameter configuration for different attention mechanisms on DeiT-Tiny-784.

<table>
<thead>
<tr>
<th>Attention</th>
<th>Hyper-parameter configuration on image classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local attention</td>
<td>Window size</td>
</tr>
<tr>
<td>Scatterbrain (Kitaev et al., 2020)</td>
<td>number of random feature samples</td>
</tr>
<tr>
<td></td>
<td>Local attention window size</td>
</tr>
<tr>
<td>Nystroemformer (Xiong et al., 2021b)</td>
<td>Number of landmarks</td>
</tr>
<tr>
<td>Performer (Choromanski et al., 2021)</td>
<td>Number of random feature samples</td>
</tr>
<tr>
<td></td>
<td>Type of random feature</td>
</tr>
<tr>
<td>Combiner (Ren et al., 2021)</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td>Span size</td>
</tr>
<tr>
<td></td>
<td>Conditional distribution parameterization</td>
</tr>
<tr>
<td>Transformer-LS (Zhu et al., 2021)</td>
<td>Dynamic projection dimension</td>
</tr>
<tr>
<td></td>
<td>Local window size</td>
</tr>
<tr>
<td>EVA (Ours)</td>
<td>Number of partitioned groups (C)</td>
</tr>
<tr>
<td></td>
<td>Size of E</td>
</tr>
</tbody>
</table>

- DeiT-Tiny (Touvron et al., 2021), which maintains the sequence length as 196 across all transformer layers. For the particular tiny variant, the number of transformer layers is set to 12, the embedding dimension is set to 196 and the number of heads is 3;
- DeiT-Small (Touvron et al., 2021), which scales the embedding dimension and number of attention heads in DeiT-Tiny up to 384 and 6, respectively;
- DeiT-Tiny-784, where the architecture is the same as DeiT-Tiny but the patch size in the tokenization step is decreased from 16 to 8. This effectively increases the sequence length from 196 to 784, which we found consistently improves predictive accuracy at the cost of significantly increased time and memory consumption. Under this setting, we also see clearer differences among these attention variants and it helps better evaluate the ability of different attention models to learn visual representations;
- PVT-v2-B3 (Wang et al., 2021b), a pyramidal transformer architecture that processes much longer token sequences at early layers and progressively reduces the sequence length to form a hierarchical structure. It patchifies input images into 3136 (56 \times 56) tokens, and then processes the sequence through 4 stages. Each stage contains several transformer layers and a down-sampling operation, which reduces the sequence length by a factor of 4 and increases the embedding dimension by 2×. Due to the prohibitively long sequences initially, PVT applies an additional down-sampling module on input sequences to obtain key and value vectors, which are then passed through a normal softmax attention mechanism. To evaluate different RF approximations, we remove the down-sampling operation and directly operate on the original sequence length, which results in much fewer model parameters than vanilla PVT-v2-B3. We refer readers to Wang et al. (2021b) for detailed architecture configurations.

For training, we do not use the [CLS] token for classification (Touvron et al., 2021); instead, we pool over the output of the last transformer layer to extract features and feed them into the classifier head. We followed the same protocol to train all model variants. Closely following DeiT (Touvron et al., 2021), we employ the AdamW (Loshchilov & Hutter, 2019) optimizer to train models for 300 epochs, where the number of warm-up epochs is 10, the learning rate is 0.001 with cosine learning rate decay (Loshchilov & Hutter, 2016), and batch size is set to 1024. The adopted augmentation and regularization are the same as DeiT, except that we remove repeated augmentation (Hoffer et al., 2020) in DeiT models as it often slows down convergence, as also observed in previous studies (Xiao et al., 2021). The specific configurations of each attention mechanism on DeiT-Tiny-784 are listed in Table 9. The hyper-parameter setup for each attention variant follows previous practices (Wang et al., 2021a; Zheng et al., 2022b) closely to ensure a similar computational cost.

Comparison to State-of-the-Art Model Architectures. We also compare our model against recent state-of-the-art (SOTA) model architectures with similar parameter sizes on ImageNet1k benchmark. As reported in Table 10, we observe that PVT-v2 (Wang et al., 2021b) with EVA greatly

---

4We retain the repeated augmentation technique in training PVT to be consistent with the original training protocol in Wang et al. (2021b).
Table 10: Results on ImageNet1k dataset compared with SOTA model architectures.

<table>
<thead>
<tr>
<th>Model</th>
<th># Param.</th>
<th>FLOPs</th>
<th>Top-1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVT-v1-M (Wang et al., 2021a)</td>
<td>44M</td>
<td>6.7G</td>
<td>81.2</td>
</tr>
<tr>
<td>RegNetY-8G (Radosavovic et al., 2020)</td>
<td>39M</td>
<td>8.0G</td>
<td>81.7</td>
</tr>
<tr>
<td>CvT-21 (Wu et al., 2021)</td>
<td>32M</td>
<td>7.1G</td>
<td>82.5</td>
</tr>
<tr>
<td>SOFT-M (Lu et al., 2021)</td>
<td>45M</td>
<td>7.2G</td>
<td>82.9</td>
</tr>
<tr>
<td>RegNetY-16G (Radosavovic et al., 2020)</td>
<td>84M</td>
<td>16.0G</td>
<td>82.9</td>
</tr>
<tr>
<td>UniFormer-S (Li et al., 2022)</td>
<td>22M</td>
<td>3.6G</td>
<td>82.9</td>
</tr>
<tr>
<td>Swin-S (Liu et al., 2021)</td>
<td>50M</td>
<td>8.7G</td>
<td>83.0</td>
</tr>
<tr>
<td>Swin-B (Liu et al., 2021)</td>
<td>88M</td>
<td>15.4G</td>
<td>83.3</td>
</tr>
<tr>
<td>RegionViT-M (Chen et al., 2021b)</td>
<td>42M</td>
<td>7.9G</td>
<td>83.4</td>
</tr>
<tr>
<td>ViL-M (Zhang et al., 2023)</td>
<td>40M</td>
<td>9.1G</td>
<td>83.5</td>
</tr>
<tr>
<td>Focal-S (Yang et al., 2021)</td>
<td>51M</td>
<td>9.1G</td>
<td>83.5</td>
</tr>
<tr>
<td>PVT-v2-b3 + LARA (Zheng et al., 2022b)</td>
<td>40M</td>
<td>7.7G</td>
<td>83.6</td>
</tr>
<tr>
<td>MaxViT-T (Li et al., 2022)</td>
<td>31M</td>
<td>5.6G</td>
<td>83.6</td>
</tr>
<tr>
<td>UniFormer-B (Li et al., 2022)</td>
<td>50M</td>
<td>8.3G</td>
<td>83.9</td>
</tr>
<tr>
<td>PVT-v2-b3 (Wang et al., 2021b)</td>
<td>45M</td>
<td>6.9G</td>
<td>83.1</td>
</tr>
<tr>
<td>PVT-v2-b3 + EVA</td>
<td>36M</td>
<td>7.4G</td>
<td>83.7</td>
</tr>
</tbody>
</table>

improves the predictive accuracy and performs competitively with recent SOTA architectures while using fewer parameters and FLOPs.

E.3 MACHINE TRANSLATION AND LANGUAGE MODELING

Our implementation for all language tasks is based on FairSeq toolkit (Ott et al., 2019). To compare different methods, we report BLEU scores on the test set as the main metric for MT and perplexity for both Autoregressive LM and MLM tasks. For the hyper-parameters $|E|$ and $C$ in EVA, we set $|E| = 2C$ by default, as we find that this choice attains a good trade-off between performance and computational costs across various tasks; while for $C$, it is determined based on previous practice for each task. Here we provide the detailed experimental protocol for each task.

**Masked Language Modeling.** Following the standard pretraining practice as in RoBERTa (Liu et al., 2019), in MLM, we aim to reconstruct a subset of tokens in the input sequence that are randomly masked out, which is the core element of BERT-style natural language pretraining (Devlin et al., 2019). This setting allows us to investigate the generalization ability of our model on larger model sizes and much more data. The task performance is measured with validation perplexity, which reflects how well the model fits the pretraining corpus and also exhibits good correlations with downstream task metrics. For the used corpus Books3, we randomly select 100 books without replacement for the validation split, similar to the setup in C4 dataset (Raffel et al., 2020b). For the model, we use the RoBERTa-base architecture (Liu et al., 2019), where all the layer normalization operations (Ba et al., 2016) are placed before attention and FFN blocks (i.e., we adopt the pre-norm architecture), which leads to more stable training for efficient attention mechanisms. We replace all softmax attention with EVA to test its effectiveness. The training setting and attention-specific parameters, which follow previous studies (Xiong et al., 2021a) to ensure a similar computational cost, can be found in Table 11 and Table 12 respectively.

**Machine Translation.** We follow Ott et al. (2018) to process WMT14 En–De dataset, resulting in around 4.5M/3K/3K English-German sentence pairs for training/validation/testing splits, respectively, and a shared vocabulary is obtained between the source and target language of around 32K BPE types. The architecture and training specifics closely follow Vaswani et al. (2017), as listed in Table 13. We follow the previous protocol (Zheng et al., 2022b) by replacing all encoder self-attention blocks in the encoder-decoder Transformer with EVA. For EVA, we find it beneficial to introduce an overlapping variant of $E$, where we allow $E$ to be overlapped with each other. Following previous practice in the context of local attention (Xiong et al., 2021a), $E$ not only contains all elements within the designated chunk but also additionally includes half the tokens in its neighboring chunks. As a result, EVA-32 corresponds to $|E| = 32$ with a contiguous chunk size of 16. During inference, we follow the same setup as Zheng et al. (2022b) and average the last 10 model checkpoints to obtain the final model parameters. We apply beam search with size 4, length penalty 0.6, and compound split.
Table 11: Our hyper-parameter configuration for Masked Language Modeling (MLM).

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>MLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transformer encoder layers</td>
<td>12</td>
</tr>
<tr>
<td>Hidden size</td>
<td>768</td>
</tr>
<tr>
<td>hidden size in FFN</td>
<td>3072</td>
</tr>
<tr>
<td>Number of attention heads</td>
<td>12</td>
</tr>
<tr>
<td>Batch size</td>
<td>256</td>
</tr>
<tr>
<td>Sequence length</td>
<td>[2048, 4096]</td>
</tr>
<tr>
<td>Number of training steps</td>
<td>200K</td>
</tr>
<tr>
<td>Number of warm-up steps</td>
<td>5K</td>
</tr>
<tr>
<td>Weight decay rate</td>
<td>0.01</td>
</tr>
<tr>
<td>Peak Learning Rate</td>
<td>1e-4</td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>Linear</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Adam $\epsilon$</td>
<td>1e-6</td>
</tr>
<tr>
<td>Adam $(\beta_1, \beta_2)$</td>
<td>(0.9, 0.98)</td>
</tr>
<tr>
<td>Gradient Clipping</td>
<td>0.0</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.1</td>
</tr>
<tr>
<td>Attention dropout (if applicable)</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 12: Our hyper-parameter configuration for different attention mechanisms on MLM task. * We used the exact positive random feature map [Choromanski et al., 2021] in our preliminary experiments. However, it failed to converge and exhibited substantial training instability. Therefore, we replace the positive random feature with a simple ReLU kernel function for MLM experiments, which yields better training performance.

<table>
<thead>
<tr>
<th>Attention</th>
<th>Hyper-parameter configuration on MLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local attention</td>
<td>Window size</td>
</tr>
<tr>
<td>Linformer [Wang et al., 2020]</td>
<td>Projected dimension</td>
</tr>
<tr>
<td>Reformer [Kitaev et al., 2020]</td>
<td>Number of hashes</td>
</tr>
<tr>
<td>Reformer [Kitaev et al., 2020]</td>
<td>Chunk size</td>
</tr>
<tr>
<td>Performer [Choromanski et al., 2021]</td>
<td>Number of random feature samples</td>
</tr>
<tr>
<td>Performer [Choromanski et al., 2021]</td>
<td>Type of random feature</td>
</tr>
<tr>
<td>LARA [Zheng et al., 2022b]</td>
<td>Number of landmarks</td>
</tr>
<tr>
<td>Combiner [Ren et al., 2021]</td>
<td>Mode</td>
</tr>
<tr>
<td>Combiner [Ren et al., 2021]</td>
<td>Span size</td>
</tr>
<tr>
<td>Combiner [Ren et al., 2021]</td>
<td>Conditional distribution parameterization</td>
</tr>
<tr>
<td>Transformer-LS [Zhu et al., 2021]</td>
<td>Dynamic projection dimension</td>
</tr>
<tr>
<td>Transformer-LS [Zhu et al., 2021]</td>
<td>Local window size</td>
</tr>
<tr>
<td>EVA (Ours)</td>
<td>Number of partitioned groups ($C$)</td>
</tr>
<tr>
<td>EVA (Ours)</td>
<td>Size of $E$</td>
</tr>
</tbody>
</table>

post-processing. Since the input sequences in WMT14 En–De benchmark are much shorter than the other tasks considered in this paper (with an average sequence length of around 25 tokens), we start with $C = 8$, $|E| = 16$ and gradually increase $|E|$ to test the translation performance, similar to the setup in [Ma et al., 2021; Zheng et al., 2022b]. Note that increasing $C$ also leads to better translation quality, although we found the performance gain is slightly less effective than that of increasing $|E|$ (c.f. Tables 6 and 7).

**Autoregressive Language Modeling.** We consider Wikitext-103 benchmark in this task, which consists of around 103M/218K/246K tokens for training/validation/testing splits, respectively. We adopt the vanilla transformer decoder architecture [Vaswani et al., 2017], replace all decoder self-attention modules in the Transformer with the causal EVA mechanism, and evaluate EVA under two different setups: 1) a standard 16-layer Transformer LM (with model sizes of around 247M) as in [Baevski & Auli, 2019], and 2) a larger 32-layer Transformer LM (with model sizes of around 450M) as in [Kasai et al., 2021]. We follow their hyper-parameter settings to train all models, where
Table 13: Our hyper-parameter configuration for machine translation.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Machine Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transformer encoder layers</td>
<td>6</td>
</tr>
<tr>
<td>Number of transformer decoder layers</td>
<td>6</td>
</tr>
<tr>
<td>Hidden size</td>
<td>512</td>
</tr>
<tr>
<td>hidden size in FFN</td>
<td>2048</td>
</tr>
<tr>
<td>Number of attention heads</td>
<td>8</td>
</tr>
<tr>
<td>Maximum number of tokens in a batch</td>
<td>32768</td>
</tr>
<tr>
<td>Number of training steps</td>
<td>300K</td>
</tr>
<tr>
<td>Number of warm-up steps</td>
<td>6K</td>
</tr>
<tr>
<td>Weight decay rate</td>
<td>0.0</td>
</tr>
<tr>
<td>Peak Learning Rate</td>
<td>0.0007</td>
</tr>
<tr>
<td>Label Smoothing</td>
<td>0.1</td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>Inverse square root</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Adam $\epsilon$</td>
<td>1e-6</td>
</tr>
<tr>
<td>Adam $(\beta_1, \beta_2)$</td>
<td>(0.9, 0.98)</td>
</tr>
<tr>
<td>Gradient Clipping</td>
<td>5.0</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.1</td>
</tr>
<tr>
<td>Attention dropout (if applicable)</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 14: Our hyper-parameter configuration for autoregressive language modeling.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>LM in Baevski &amp; Auli [2019]</th>
<th>LM in Kasai et al. [2021]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transformer decoder layers</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Hidden size</td>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>hidden size in FFN</td>
<td>4096</td>
<td>4096</td>
</tr>
<tr>
<td>Number of attention heads</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Number of tokens in a batch</td>
<td>65536</td>
<td>65536</td>
</tr>
<tr>
<td>Number of training steps</td>
<td>286K</td>
<td>286K</td>
</tr>
<tr>
<td>Number of warm-up steps</td>
<td>16K</td>
<td>16K</td>
</tr>
<tr>
<td>Weight decay rate</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Peak Learning Rate</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>cosine</td>
<td>cosine</td>
</tr>
<tr>
<td>Optimizer</td>
<td>nag</td>
<td>nag</td>
</tr>
<tr>
<td>Gradient Clipping</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>LayerDrop</td>
<td>–</td>
<td>0.2</td>
</tr>
<tr>
<td>Attention dropout</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The corresponding configurations are listed in Table 14. The vocabulary size is 267,744 with adaptive input embeddings (Baevski & Auli, 2019). During training, we set the sequence length to 512 and evaluate the validation/test PPL with various context window sizes in \{256, 480\}, aligning with previous work (Baevski & Auli, 2019; Kasai et al., 2021). For other random feature baselines, unfortunately, we failed to fully replicate their results as reported in Kasai et al. (2021), where RFA in our implementation achieved a test perplexity of 29.0 even under a 449M Transformer model. For EVA, we set $|E| = 128$ and $C = 64$ by default for both 16-layer and 32-layer settings, ensuring similar computational cost to previous work that also evaluates random feature methods (typically with 128 or 256 random-feature dimension size) on Wikitext-103 language modeling task (Schlag et al., 2021; Kasai et al., 2021).

E.4 EXPERIMENTAL SETTINGS OF EFFICIENCY COMPARISON

For the simulation experiment conducted in §5.3, we adopt the same transformer architecture across all attention variants. In particular, it uses 8 transformer layers, 192 embedding dimensions, and 2 attention heads so that longer sequences can fit into our devices. The batch size is set to 64 across

\footnote{The setup in Baevski & Auli (2019) can be found in the corresponding Fairseq training script: https://github.com/pytorch/fairseq/blob/master/examples/language_model/README.adaptive_inputs.md}
Figure 2: **Left** and **right**: Additional empirical memory consumption and running time comparison for different attention mechanisms under various sequence lengths.

8 V100 GPUs, and the statistics are computed by averaging the results of 30 runs. Besides, in our ablation study, the efficiency metrics reported in Table 6 and Table 7 are evaluated under the same setup used during training.

**Remark on Modeling Short Sequences.** Unfortunately, similar to most previous efficient attention baselines, EVA also runs slower than softmax attention under shorter sequences (e.g., length of 128 or 256), but it soon catches up in running speed, and the reduction of memory consumption is still significant. Besides, in short-sequence settings (such as the case of DeiT-Tiny/Small with sequences of 196 tokens), EVA often performs on par with or better than conventional softmax attention (see Table 1), whereas most previous attention variants usually perform much worse. This implies EVA can achieve a better trade-off between efficiency and quality: for short sequences, EVA is possible to achieve stronger performance competitive with softmax attention (despite in longer running time); while for long sequences, EVA can be run much faster with less memory.

**Comparison to Memory-efficient Attention Mechanisms.** In this section, we conduct an empirical efficiency comparison between efficient approximate attention methods and FlashAttention, one of the memory-efficient attention mechanisms [Rabe & Staats, 2021; Dao et al., 2022] with optimized memory accesses. FlashAttention computes the exact softmax attention in an online manner without materializing the full attention matrix, achieving linear memory complexity with respect to sequence lengths; besides, both runtime and memory usage are further improved by minimizing IO accesses. We benchmark different attention modules on one NVIDIA GeForce RTX 3090 GPU, where we measure the memory usage and runtime of running a single attention block, consisting of 8 attention heads with 512 embedding dimension size, for both a forward and backward pass. As shown in Figure 2, we observe that FlashAttention achieves significant memory usage reduction for softmax attention approximation and even consumes much less memory than all considered approximate baselines under all sequence lengths. In terms of runtime, we notice that FlashAttention runs faster than most attention baselines under sequence lengths less than 2048 despite scaling quadratically, but EVA, along with other more efficient approximate variants, begin to catch up at longer sequence lengths. This implies that the quadratic computational costs of softmax attention still bottleneck its runtime performance, aligning with one of the main findings in Dao et al. (2022). According to this empirical study, we observe that FlashAttention offers a general and effective technique to speed up softmax attention; since many approximate variants (including EVA) exhibit a similar formulation to softmax attention (e.g., Equation 16), we expect they can also benefit from the optimized online softmax calculation technique and memory accesses of FlashAttention (Dao et al., 2022).

**F Experiments on Long Range Arena**

Long Range Arena (LRA; Tay et al., 2021) is a lightweight benchmark that assesses the ability of efficient attention methods to model long sequences in diverse domains. We follow the same hyper-parameter setup as Xiong et al. (2021b) to re-evaluate all attention baselines and report the
Table 15: Classification accuracy (%) on LRA benchmark with different efficient attention mechanisms.

<table>
<thead>
<tr>
<th>Model</th>
<th>ListOps</th>
<th>Text</th>
<th>Retrieval</th>
<th>Image</th>
<th>Pathfinder</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax</td>
<td>38.66</td>
<td>64.91</td>
<td>80.70</td>
<td>40.61</td>
<td>68.29</td>
<td>58.63</td>
</tr>
<tr>
<td>Linformer</td>
<td>38.21</td>
<td>53.91</td>
<td>77.66</td>
<td>39.40</td>
<td>66.44</td>
<td>55.12</td>
</tr>
<tr>
<td>Performer</td>
<td>29.84</td>
<td>65.30</td>
<td>77.70</td>
<td>38.29</td>
<td>66.39</td>
<td>55.50</td>
</tr>
<tr>
<td>Reformer</td>
<td>27.12</td>
<td>63.90</td>
<td>78.08</td>
<td>42.40</td>
<td>51.90</td>
<td>52.69</td>
</tr>
<tr>
<td>Scatterbrain</td>
<td>38.21</td>
<td>64.04</td>
<td>77.83</td>
<td>42.51</td>
<td>68.46</td>
<td>58.10</td>
</tr>
<tr>
<td>Combiner</td>
<td>38.26</td>
<td>63.98</td>
<td>81.47</td>
<td>42.80</td>
<td>55.94</td>
<td>56.49</td>
</tr>
<tr>
<td>LARA</td>
<td>37.10</td>
<td>64.62</td>
<td>80.82</td>
<td>38.99</td>
<td>68.96</td>
<td>58.57</td>
</tr>
<tr>
<td>Nyströmformer</td>
<td>38.46</td>
<td>65.28</td>
<td>80.44</td>
<td>39.71</td>
<td>68.98</td>
<td>58.57</td>
</tr>
<tr>
<td>Local</td>
<td>38.46</td>
<td>63.70</td>
<td>80.71</td>
<td>42.25</td>
<td>68.46</td>
<td>58.72</td>
</tr>
<tr>
<td>Long-short</td>
<td>38.56</td>
<td>63.46</td>
<td>81.73</td>
<td>40.54</td>
<td>71.28</td>
<td>59.11</td>
</tr>
<tr>
<td>EVA</td>
<td>38.61</td>
<td>64.31</td>
<td>80.21</td>
<td>43.24</td>
<td>70.90</td>
<td>59.45</td>
</tr>
</tbody>
</table>

comparison in Table 15. We observe that EVA largely improves previous RFA methods such as Performer (Choromanski et al., 2021) and LARA (Zheng et al., 2022b), and performs competitively with full softmax attention. Notably, EVA even achieves better average results over all tasks, with higher accuracy on Image and Pathfinder benchmarks, suggesting its capability of capturing long-term dependencies. For LRA benchmark, we set all attention-specific hyper-parameters to 128 (e.g., the number of landmarks in Nyströmformer (Xiong et al., 2021b) and LARA (Zheng et al., 2022b), the window size in local attention and Combiner (Ren et al., 2021), etc.). We set $|E| = 128$ and $C = 64$ by default for EVA without any further tuning and find this setup works well.

G CONNECTIONS TO OTHER ATTENTION MECHANISMS

G.1 RFA, SOFTMAX ATTENTION, AND EVA

As mentioned in our main text, one of the main contributions of this work is to develop a more general framework that bridges RFA and conventional softmax attention. To see how EVA (Equation 13) achieves this goal formally, note that if either $|E| = M$ or $C = M$, EVA would be equivalent to standard softmax attention; while if we set $|E| = 0$ and $C = 1$, EVA would recover vanilla RFA.

G.2 CONNECTIONS TO LARA

Notably, EVA and LARA (Zheng et al., 2022b) are two efficient attention mechanisms that are both built upon the self-normalized importance sampling (SNIS) formulation of RFAs. LARA (Zheng et al., 2022b) puts the main focus on the proposal distribution used in SNIS and tries to design importance sampling proposals that are closer to the true underlying distribution. The proposed usage of multiple proposals further improves the estimation quality of SNIS and achieves strong empirical performance while still keeping linear complexity.

In contrast to LARA, in this work we do not focus on the design choice of proposals used in importance sampling but aim to generalize the SNIS formulation further via control variates. As demonstrated in §3.2, our theory clearly delineates how the gap between such SNIS estimation and softmax attention can be closed by manipulating control variates. Since LARA and RFA are both SNIS estimators (their main difference lies in the choice of proposal distributions), our generalization also applies to LARA. To summarize, compared with LARA, EVA is a more general framework and improves conventional RFA from an orthogonal perspective.

G.3 CONNECTIONS TO CLUSTERED ATTENTION

Clustered attention (Vyas et al., 2020) is an efficient attention mechanism that first clusters the set of queries into multiple groups, computes the mean centroid of each group, and then performs attention between query centroids and original key-value pairs. This framework is fast and effective and enjoys well-bounded approximation error.
Clustered attention and EVA share some similarities in two aspects. First, both of them adopt the partitioning technique to reduce the computational complexity while remaining effective; and secondly, both observe that the efficient attention mechanism can be improved by refining the approximation over specific elements. For instance, clustered attention can be improved \cite{Vyas et al., 2020} by selecting top-$k$ key-value pairs that are most relevant to each centroid and then refining the approximation by recomputing attention weights over these keys using original queries; while EVA notices that we can directly employ the optimal control variate coefficient for a subset of key-value pairs ($m \in E$) while still remaining efficient, which yields a more accurate approximation.

Nevertheless, our main technical contribution is to develop a control variate formulation in the context of RFA and demonstrate that how RFA can be further improved locally. On the other hand, while clustered attention \cite{Vyas et al., 2020} clusters queries, EVA partitions key-value pairs. This property makes EVA more amenable to the case of autoregressive language modeling since we do not impose clustering structures over the query set, and thus the causal relation among queries can be well maintained.

\section*{G.4 \ Connections to Combiner}

Combiner \cite{Ren et al., 2021} is a recently proposed attention mechanism that also partitions the sequence into chunks combined with local attention. The key difference between EVA and Combiner is the motivation, where Combiner introduces a structured factorization over the attention probability distribution, while our approach is built from the control variate perspective.

\section*{G.5 \ Connections to Scatterbrain}

In this section, we show that Scatterbrain \cite{Chen et al., 2021a} can be cast as a special case of our framework EVA, although they are proposed based on quite different motivations.

\textbf{A Brief Review of Scatterbrain.} Scatterbrain \cite{Chen et al., 2021a} notes that sparse attention and RFA can approximate sharp and flat regions of the softmax attention matrix well, respectively. Based on this insight, Scatterbrain is proposed to first compute a Performer approximation to softmax attention and then cancel out the approximation error on critical regions via a sparse mechanism.

Specifically, Scatterbrain \cite{Chen et al., 2021a} defines a sparse matrix $S \in \mathbb{R}^{N \times M}$ so that for each $(n, m) \in S$ that indexes a non-zero entry. For notational simplicity, we also denote $\text{Supp}(S) = \{(i, j) | S_{ij} \neq 0\}$ and $\text{Supp}_n(S) = \{m | S_{nm} \neq 0\}$. With random features $\phi(\cdot, \cdot)$ defined in Appendix A we let

$$S_{nm} = \exp (q^\top_n k_m) - \phi(q_n, \omega) \phi(k_m, \omega).$$

We then add it back to the approximate output:

$$y'_n = \sum_{m=1}^{M} \phi(q_n, \omega)^\top \phi(k_m, \omega)v_m + SV$$

$$= \sum_{m=1}^{M} \phi(q_n, \omega)^\top \phi(k_m, \omega)v_m + \sum_{m' \in \text{Supp}_n(S)} S_{nm'}v_{m'}$$

$$= \sum_{m \not\in \text{Supp}_n(S)} \phi(q_n, \omega)^\top \phi(k_m, \omega)v_m + \sum_{m' \in \text{Supp}_n(S)} \exp (q^\top_n k_{m'}) v_{m'}. \quad (29)$$

The sparse mechanism can be thought of as modeling the error due to RFA and eliminating it on the support of $S$. After the correction step, Scatterbrain further adds a post-hoc normalization step to obtain a normalized attention output:

$$y_n = \frac{\sum_{m \not\in \text{Supp}_n(S)} \phi(q_n, \omega)^\top \phi(k_m, \omega)v_m + \sum_{m' \in \text{Supp}_n(S)} \exp (q^\top_n k_{m'}) v_{m'}}{\sum_{m \not\in \text{Supp}_n(S)} \phi(q_n, \omega)^\top \phi(k_m, \omega) + \sum_{m' \in \text{Supp}_n(S)} \exp (q^\top_n k_{m'})}. \quad (30)$$

Intuitively, Scatterbrain \cite{Chen et al., 2021a} produces accurate approximation in the support of the sparse matrix and remains the random feature approximation outside the support.
Scatterbrain is a Special Case of EVA. For notational convenience, we denote $E := \text{Supp}_n(S)$. According to Proposition 1, suppose we employ optimal coefficients $\beta_m$ for all entries in $\text{Supp}_n(S)$, and use the same coefficient $\beta$ for all the remaining entries (in other words, we let $C = 1$ and the whole index set is only partitioned into two subsets $\{E, [M] \setminus E\}$). Then we have

$$g_m(\omega) = \begin{cases} 
g_m(\omega) - \beta h_m(\omega) + \frac{\exp(q_n^\top k_m)}{Z}, & \text{if } m \in E, \\
g_m(\omega) - \beta h_m(\omega) + \frac{\exp(q_n^\top k_m)}{Z}, & \text{if } m \notin E. 
\end{cases}$$

And the resulting estimator overall becomes

$$\bar{g}(\omega) = \sum_{m=1}^{M} \bar{g}_m(\omega)$$

$$= \sum_{m \in E} \bar{g}_m(\omega) + \sum_{m \notin E} \bar{g}_m(\omega)$$

$$= \sum_{m \in E} \frac{\exp(q_n^\top k_m) v_m}{Z} + \sum_{m \notin E} \left( g_m(\omega) - \beta h_m(\omega) + \frac{\exp(q_n^\top k_m)}{Z} \right)$$

$$= \sum_{m \in E} \frac{\exp(q_n^\top k_m) v_m}{Z} + \sum_{m \notin E} \left( g_m(\omega) - \beta h_m(\omega) + \beta \sum_{m \notin E} \frac{\exp(q_n^\top k_m)}{Z} \right)$$

Scatterbrain (Chen et al., 2021a) can be a special case of this estimation algorithm if we set the proposal distribution to $q(\omega) = N(\omega; 0, I)$, and estimate the normalizing constant as follows.

$$Z = \mathbb{E}_{\omega \sim q(\omega)} \left[ \frac{N(\omega; 0, I) \left( \sum_{m \in E} \xi(q_n, \omega)^\top \xi(k_m, \omega) + \sum_{m \notin E} \xi(q_n, \omega)^\top \xi(k_m, \omega) \right)}{q(\omega)} \right]$$

$$= \sum_{m \in E} \exp(q_n^\top k_m) + \mathbb{E}_{\omega \sim q(\omega)} \left[ \frac{N(\omega; 0, I) \sum_{m \notin E} \xi(q_n, \omega)^\top \xi(k_m, \omega)}{q(\omega)} \right]$$

$$\approx \sum_{m \in E} \exp(q_n^\top k_m) + \frac{1}{S} \sum_{s=1}^{S} N(\omega; 0, I) \sum_{m \notin E} \xi(q_n, \omega)^\top \xi(k_m, \omega)$$

$$= \sum_{m \in E} \exp(q_n^\top k_m) + \frac{1}{S} \sum_{s=1}^{S} \sum_{m \notin E} \xi(q_n, \omega)^\top \xi(k_m, \omega)$$

$$= \sum_{m \in E} \exp(q_n^\top k_m) + \sum_{m \notin E} \bar{h}_m(\omega),$$

where we define $\bar{h}_m(\omega) = Z h_m(\omega)$, as in this case

$$g(\omega) = \frac{1}{S} \sum_{s=1}^{S} \frac{h_n(\omega_s)}{q(\omega_s)} f(\omega_s) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{Z} \sum_{m=1}^{M} \xi(q_n, \omega_s) \xi(k_m, \omega_s) v_m,$$

$$h(\omega) = \frac{1}{S} \sum_{s=1}^{S} \frac{h_n(\omega_s)}{q(\omega_s)} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{Z} \sum_{m=1}^{M} \xi(q_n, \omega_s) \xi(k_m, \omega_s).$$
With these specifications, we obtain

\[
\bar{g}(\omega) = \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} \left( g_m(\omega) - \hat{\beta} h_m(\omega) \right) + \hat{\beta} \left( 1 - \sum_{m \in E} \frac{\exp(q_m^T k_m)}{Z} \right)
\]

\[
= \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} \left( g_m(\omega) - \hat{\beta} h_m(\omega) \right) + \hat{\beta} \sum_{m \notin E} \frac{\bar{h}_m(\omega)}{Z}
\]

\[
\approx \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} \left( g_m(\omega) - \hat{\beta} h_m(\omega) \right) + \hat{\beta} \sum_{m \notin E} h_m(\omega)
\]

\[
= \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} g_m(\omega)
\]

\[
= \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} \frac{1}{S} \sum_{s=1}^{S} \xi(q_m, \omega_s) \xi(k_m, \omega_s) v_m
\]

\[
= \sum_{m \in E} \frac{\exp(q_m^T k_m) v_m}{Z} + \sum_{m \notin E} \frac{\phi(q_m, \omega)^T \phi(k_m, \omega) v_m}{Z}
\]

\[
\approx \sum_{m \notin E} \frac{\phi(q_m, \omega)^T \phi(k_m, \omega) v_m + \sum_{m' \in E} \exp(q_m^T k_m') v_m'}{\sum_{m \in E} \phi(q_m, \omega)^T \phi(k_m, \omega) + \sum_{m' \in E} \exp(q_m^T k_m')} \tag{31}
\]

which is equivalent to Scatterbrain (Equation 30). Note that this equivalence would hold irrespective of the choice of shared coefficients \( \hat{\beta} \), which possibly indicates that the formulation of Scatterbrain limits the potential benefit of optimizing control variates under our framework.