A PROOF OF PROPERTY 3.1

Proof. Following Definition 3.1, because items are at the same position pos_a, $P^p_{pos_a,A} = P^q_{pos_a,A}$. Similarly for position pos_b, $P^p_{pos_b,A} = P^q_{pos_b,A}$. Then $P^q_{pos_b,A} = f(P^q_{pos_a,A}, pos_a, pos_b)$ holds for the function $f(\cdot,\cdot,\cdot)$. □

B PROOF OF PROPERTY 3.2

Proof. Following Definition 3.2, because item at pos_b for length p and item at pos_d for length q are at the reverse position $p - pos_b$, $P^p_{pos_b,B} = P^q_{pos_d,B}$. Similarly for position pos_a and pos_c, $P^p_{pos_a,B} = P^q_{pos_c,B}$. Then $P^q_{pos_c,B} = f(P^q_{pos_d,B}, pos_a, pos_b)$ holds for the function $f(\cdot,\cdot,\cdot)$. □

C PROOF OF COROLLARY 3.5

Proof. The RPE is explored by recent language models [7, 25, 36]. During the attention score calculation, the absolute positional encoding $\boldsymbol{P}$, e.g., SPE, is included as:

$$\boldsymbol{A} = (\boldsymbol{X}_i + \boldsymbol{P}_i)\boldsymbol{W}_{qry}\boldsymbol{W}_{key}^{\top}(\boldsymbol{X}_j + \boldsymbol{P}_j)^{\top}, \tag{13}$$

where $\boldsymbol{X}$ is the input feature and $\boldsymbol{W}$ is trainable weights.

For RPE in different work, they basically follow a format:

$$\boldsymbol{A} = \boldsymbol{X}_i\boldsymbol{W}_{qry}\boldsymbol{W}_{key}^{\top}\boldsymbol{X}_j^{\top} + g(\boldsymbol{P}_{ij}), \tag{14}$$

where $\boldsymbol{P}$ only represents the relative position between i and j. Because $\boldsymbol{P}_{ij}$ does not provide any information about the absolute position of a token, for different center tokens i_1 and i_2, $\boldsymbol{P}_{i_1 j} \neq \boldsymbol{P}_{i_2 j}$. Because j can be before or after i in position, RPE simultaneously is not *forward-aware* and *backward-aware.* □

D PROOF OF COROLLARY 3.6

Proof. Similar to the proof of Theorem 3.4, we firstly prove RSPE is *backward-aware* and then is not *forward-aware.* (1) According to Eq. (2), RSPE directly follows Definition 3.2 for *backward-aware.* (2) Take the first item of two sessions w.r.t. length 1 and 2 as example. For length 1 session, $P^1_{0,2i} = 0$ and $P^1_{0,2i+1} = 1$. If RSPE is *forward-aware*, for length 2 session, there should be a slice of $P^2_{0,A}$ is the same as $P^1_{0,2i}$ and $P^1_{0,2i+1}$. For $2i$ dimension of RSPE, $P^1_{0,2i} = P^2_{0,2i} = \sin(1/10000^{2i/d})$. It is clear that $1/10000^{2i/d} \in [0.00001, 1]$. Then $P^1_{0,2i} \neq P^2_{0,2i}$. Similarly, $P^1_{0,2i+1} \neq P^2_{0,2i+1}$. Therefore, RSPE is not *forward-aware.* □

E PROOF OF PROPERTY OF ASPE

Proof. From Shiv and Quirk [26], Vaswani et al. [30], SPE has a linear combination property as follows:

$$\begin{aligned} P^l_{x+y,2i} &= P^l_{x,2i}P^l_{y,2i+1} + P^l_{x,2i+1}P^l_{y,2i} \\ P^l_{x+y,2i+1} &= P^l_{x,2i+1}P^l_{y,2i+1} - P^l_{x,2i}P^l_{y,2i}. \end{aligned} \tag{15}$$

The modified RSPE is defined as:

$$\begin{aligned} P^l_{l-pos-1,2i} &= \cos((l - pos - 1)\,/10000^{2i/d}) \\ P^l_{l-pos-1,2i+1} &= \sin((l - pos - 1)\,/10000^{2i/d}). \end{aligned} \tag{16}$$

For simplicity, we redefine $POS = pos/10000^{2i/d}$ and $L = (l-1)\,/10000^{2i/d}$.

The addition of SPE and RSPE here is as follows:

$$\begin{aligned} P^l_{pos,2i} &= \sin(POS) + \cos(L - POS) \\ P^l_{pos,2i+1} &= \cos(POS) + \sin(L - POS). \end{aligned} \tag{17}$$

For position x and y:

$$\begin{aligned} P^l_{x,2i} &= \sin(X) + \cos(L - X) = (1 + \sin(L))\sin X + \cos L \cos X \\ P^l_{x,2i+1} &= \cos(X) + \sin(L - X) = (1 + \sin(L))\cos X + \cos L \sin X \\ P^l_{y,2i} &= \sin(Y) + \cos(L - Y) = (1 + \sin(L))\sin Y + \cos L \cos Y \\ P^l_{y,2i+1} &= \cos(Y) + \sin(L - Y) = (1 + \sin(L))\cos Y + \cos L \sin Y \\ P^l_{x+y,2i} &= \sin(X + Y) + \cos(L - (X + Y)) \\ &= (1 + \sin(L))\sin(X + Y) + \cos(L)\cos(X + Y) \\ P^l_{x+y,2i+1} &= \cos(X + Y) + \sin(L - (X + Y)) \\ &= (1 + \sin(L))\cos(X + Y) + \cos(L)\sin(X + Y). \end{aligned} \tag{18}$$

Similar to Eq. (15), we calculate the multiplication between encoding:

$$\begin{aligned} &P^l_{x,2i}P^l_{y,2i+1} \\ &= (1 + \sin(L))^2 \sin(X)\cos(Y) + (1 + \sin(L))\cos(L)\sin(X)\sin(Y) \\ &+ (1 + \sin(L))\cos(L)\cos(X)\cos(Y) + \cos^2(L)\cos(X)\sin(Y) \end{aligned} \tag{19}$$

$$\begin{aligned} &P^l_{x,2i+1}P^l_{y,2i} \\ &= (1 + \sin(L))^2 \sin(Y)\cos(X) + (1 + \sin(L))\cos(L)\cos(X)\cos(Y) \\ &+ (1 + \sin(L))\cos(L)\sin(X)\sin(Y) + \cos^2(L)\sin(X)\sin(Y) \end{aligned} \tag{20}$$

$$\begin{aligned} &P^l_{x,2i+1}P^l_{y,2i+1} \\ &= (1 + \sin(L))^2 \cos(X)\cos(Y) + (1 + \sin(L))\cos(L)\cos(X)\sin(Y) \\ &+ (1 + \sin(L))\cos(L)\sin(X)\cos(Y) + \cos^2(L)\sin(X)\sin(Y) \end{aligned} \tag{21}$$

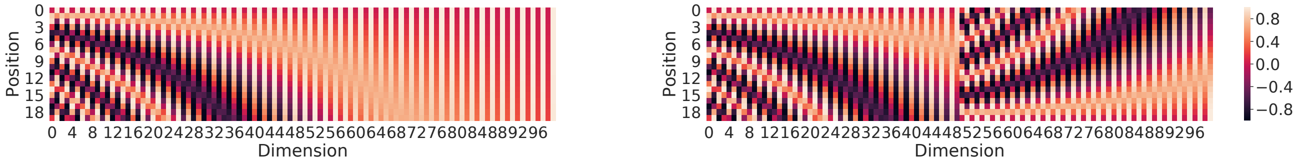
$$\begin{aligned} &P^l_{x,2i}P^l_{y,2i} \\ &= (1 + \sin(L))^2 \sin(X)\sin(Y) + (1 + \sin(L))\cos(L)\sin(X)\cos(Y) \\ &+ (1 + \sin(L))\cos(L)\cos(X)\sin(Y) + \cos^2(L)\cos(X)\cos(Y) \end{aligned} \tag{22}$$

For simplicity, we define A = Eq. (19) + Eq. (20), B = Eq. (19) − Eq. (20), C = Eq. (21) + Eq. (22) and D = Eq. (21) − Eq. (22)

Therefore, $P^l_{x+y,2i}$ and $P^l_{x+y,2i+1}$ can be computed based on A, B, C and D:

$$\begin{aligned} P^l_{x+y,2i} &= \frac{A - \cos(L)C}{2(1 + \cos^2(L))} + \frac{\cos(L)D}{2\sin(L)(1 + \sin(L))} \\ P^l_{x+y,2i+1} &= \frac{D}{2\sin(L)} + \frac{\cos(L)C - A}{s(1 + \sin(L))(\cos(L) - 1)}. \end{aligned} \tag{23}$$

Therefore, the relationship between the PE of two positions based on the addition of SPE and RSPE cannot be represented as linear combination because L is a variable among different sessions. □



(a) SPE of len. 20.

(b) DPE of len. 20.

Figure 6: Positional encoding visualization for the session length of 20.

F EXAMPLE OF 2DSPE

$$\begin{aligned} P^{l}_{pos,2i} &= \sin\left(pos/10000^{4i/d}\right) \\ P^{l}_{pos,2i+1} &= \cos\left(pos/10000^{4i/d}\right) \\ P^{l}_{pos,2i+d/2} &= \sin\left(l/10000^{4i/d}\right) \\ P^{l}_{pos,2i+1+d/2} &= \cos\left(l/10000^{4i/d}\right) \end{aligned} \tag{24}$$

G VISUALIZATION OF SPE AND DPE

From Figure 2(a) and 6(a), it is clear to see the *forward-awareness* of SPE. For example, no matter what length is, the encoding for the first position will always be the same (the first row). While for the last position of lengths 10 and 20, as the max length and the embedding size vary, there will not be a shared same part, which is not *backward-aware*. For RSPE, the case is just the reverse, i.e., the heatmap would be upside down of SPE. From Figure 2(b) and 6(b), both *forward-awareness* and *backward-awareness* are demonstrated. The same positions and reverse positions will always share the same half of the embedding respectively.