

470	Supplement to “Token Dynamics of Self-Attention”	
471	Contents	
472	A Related Works	15
473	B Dynamical Properties of Tokens in Self-Attention	15
474	B.1 Distances Between Tokens Over Time	15
475	B.2 Convergence Scenario	16
476	B.3 Divergence Scenario	19
477	C Effects of Absolute and Rotary Positional Encodings	22
478	C.1 Absolute Positional Encoding	22
479	C.2 Rotary Positional Encoding	22
480	D Additional Experimental Details	23
481	D.1 Parameter settings used in the simulations	23
482	D.1.1 Figure 1: Distances between tokens over time	23
483	D.1.2 Figure 2: Token trajectories under convergence and divergence scenarios	24
484	D.1.3 Figure 5: Token trajectories shift to divergence regime in RoPE	24
485	D.1.4 Figure 9: Token trajectories in convergence and divergence regime	24
486	D.2 Implementation of Convergence, Divergence, and Intermediate scenarios.	25
487	D.3 Details on Language Modeling experiments	26
488	D.3.1 Dataset	26
489	D.3.2 Wikitext103 Model and Training Configurations	26
490	D.3.3 EnWik8 Model and Training Configurations	26
491	D.4 Details on ImageNet-1K object recognition task	26
492	D.4.1 Dataset	26
493	D.4.2 Model and Training Configurations	26
494	D.5 Additional visualizations	27
495	D.5.1 Tokens’ trajectory in pre-trained Transformers	27
496	D.5.2 Token norm and distance in RoPE	27
497	D.5.3 Token trajectories with RoPE	27
498	D.6 Additional Results	27
499	D.6.1 Full Evaluation Results	27
500	E Broader Impact	29

A Related Works

Transformers as Interacting Particle Systems. Particle systems offer a novel perspective to understand the dynamics of Transformer models and inspire architectural innovations. In [33], the Transformer architecture is mathematically framed as a numerical ODE solver for a convection-diffusion equation in a multi-particle dynamic system. Similarly, leveraging insights from numerical ODE solvers, [14] introduces TransEvolue, a temporal evolution scheme inspired by interacting particle dynamics. Furthermore, the ODE governing Transformer dynamics is closely connected to the extensive literature on nonlinear systems, including flocking phenomena [19], the Kuramoto model [25, 1], consensus formation [24, 35], opinion formation [21, 42], and systems of self-driven particles [52].

Clustering Effects of Transformers. Research on interacting particle systems has shown that tokens in Transformer models exhibit a long-time clustering phenomenon. Geshkovski et al proved in [17] that tokens cluster around limiting objects determined by their initial values, highlighting their context awareness. This was extended to the study of metastability of self-attention with layer normalization in [16, 15]. Additionally, [3] proved that tokens in pure-attention hardmax Transformer models asymptotically converge to clustered equilibria.

B Dynamical Properties of Tokens in Self-Attention

Recall that the dynamical system governing the continuous-time dynamic of the self-attention is given by:

$$\frac{dx_l(t)}{dt} = V^\top \cdot \sum_{i=1}^L \left(\frac{e^{x_l(t)^\top \cdot W \cdot x_i(t)}}{\sum_{j=1}^L e^{x_l(t)^\top \cdot W \cdot x_j(t)}} \right) \cdot x_i(t), \quad l = 1, \dots, L, \quad (9)$$

with the initial condition $(x_1(0), \dots, x_L(0)) = (x_{10}, \dots, x_{L0}) \in (\mathbb{R}^D)^L$. In this system, $W = Q \cdot K^\top$. The matrices Q, K, V are learnable matrices and they are called queries, keys and values matrices, respectively. In our setting, the parameters W and V are assumed to be time-independent.

Set $A = W \cdot (V^\top)^{-1}$. We will prove the following observations in the subsequent subsections.

Distances Between Tokens. If $A \prec 0$, then all tokens will tend to move closer and closer to each other as time t approaches infinity. In contrast, if $A \succ 0$, the tokens will either tend to maintain constant distances or move farther away from each other as time t approaches infinity (see Theorem B.2).

Convergence Scenario. If $A \prec 0$ and $W \succ 0$, then all tokens will tend to zero as the time t tends to infinity (see Theorem B.6).

Divergence Scenario. If V has at least one positive eigenvalue and W is arbitrary, then all tokens will diverge to infinity under certain assumption on the initial data (see Theorem 3.6).

B.1 Distances Between Tokens Over Time

In this subsection, we analyze the dynamical properties of the distances between tokens as the time t increases. We start with the following lemma, which can be seen as a refinement of [17, Lemma 7.1].

Lemma B.1. *Let W be an arbitrary matrix in $\mathbb{R}^{D \times D}$. For each x_1, \dots, x_L in \mathbb{R}^D , the function $f: \mathbb{R}^D \rightarrow \mathbb{R}$ defined by*

$$f(u) = \log \left(\sum_{j=1}^L e^{u^\top \cdot W \cdot x_j} \right), \quad u \in \mathbb{R}^D,$$

is a convex function.

538 *Proof.* For arbitrary $u, v \in \mathbb{R}^D$, we have

$$\begin{aligned}
e^{f(u)+f(v)} - e^{2f\left(\frac{u+v}{2}\right)} &= \left(\sum_{i=1}^L e^{u^\top \cdot W \cdot x_i} \right) \cdot \left(\sum_{j=1}^L e^{v^\top \cdot W \cdot x_j} \right) - \left(\sum_{k=1}^L e^{\left(\frac{u+v}{2}\right)^\top \cdot W \cdot x_k} \right)^2 \\
&= \sum_{i,j} \frac{1}{2} \left(e^{(u+v)^\top \cdot W \cdot x_i} + e^{(u+v)^\top \cdot W \cdot x_j} \right) - \sum_{i,j} e^{\frac{u+v}{2} \cdot W \cdot x_i + \frac{u+v}{2} \cdot W \cdot x_j} \\
&= \sum_{i,j} \frac{1}{2} \left(e^{\frac{u+v}{2} \cdot W \cdot x_i} - e^{\frac{u+v}{2} \cdot W \cdot x_j} \right)^2 \\
&\geq 0.
\end{aligned}$$

539 Therefore, $f(u) + f(v) \geq 2f\left(\frac{u+v}{2}\right)$. Hence, f is convex. \square

540 The function f defined in the above lemma is not strictly convex as the equality happens when
541 $u + v = 0$.

542 **Theorem B.2.** Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be a solution of the dynamical system (9).
543 Assume that V is invertible and $A = W \cdot (V^\top)^{-1}$ is symmetric. Then the map $t \mapsto \mathbf{q}_A(x_i(t) - x_j(t))$
544 is non-decreasing on $[0, +\infty)$.

545 As a consequence,

- 546 (a) if $A \succ 0$, then $\|x_i(t) - x_j(t)\|_A$ is non-decreasing on $[0, +\infty)$;
547 (b) if $A \prec 0$, then $\|x_i(t) - x_j(t)\|_A$ is non-increasing on $[0, +\infty)$.

548 *Proof.* Fix an arbitrary time $s \in (0, +\infty)$. Let f be the function defined in Lemma B.1 (with x_j
549 there is replaced by $x_j(s)$ in this proof). Then we see that

$$\frac{\partial f}{\partial u}(u) = W \cdot \sum_{i=1}^L \frac{e^{u^\top \cdot W \cdot x_i(s)}}{\sum_{j=1}^L e^{u^\top \cdot W \cdot x_j(s)}} \cdot x_i(s).$$

550 Therefore, from the dynamical system (9), we have

$$\frac{\partial f}{\partial u}(x_l(s)) = A \cdot \frac{d}{ds} x_l(s), \quad l = 1, \dots, L. \quad (10)$$

551 Since f is convex, we have

$$(x_i(s) - x_j(s))^\top \left(\frac{\partial f}{\partial u}(x_i(s)) - \frac{\partial f}{\partial u}(x_j(s)) \right) \geq 0.$$

552 From equation (10), we have

$$(x_i(s) - x_j(s))^\top \cdot A \cdot \left(\frac{d}{ds} x_i(s) - \frac{d}{ds} x_j(s) \right) \geq 0.$$

553 Since A is symmetric, it follows from the above inequality that:

$$\frac{d}{ds} \mathbf{q}_A(x_i(s) - x_j(s)) \geq 0.$$

554 This shows that the map $t \mapsto \mathbf{q}_A(x_i(t) - x_j(t))$ is non-decreasing on $[0, +\infty)$. The items (a) and
555 (b) are obtained due to the definition of $\|\cdot\|_A$. \square

556 B.2 Convergence Scenario

557 In this section, we will prove that when $A \prec 0$ and $W \succ 0$, the solution of the dynamical system (9)
558 tends to zero as t approaches infinity. The special case where $A = -I_D$ and $W = I_D$ was already
559 proved in [17, Section 8.2]. We borrow the proof technique from [17, Section 8.2] and carefully
560 refine it so that it applies to a broader generalization.

561 **Proposition B.3.** Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be a solution of the dynamical sys-
 562 tem (9). Assume that V is invertible and $A = W \cdot (V^\top)^{-1}$ is symmetric. Then, we have

$$\frac{d}{dt} \mathbf{q}_A(x_l(t)) \geq 2 - \frac{2L}{e^{\mathbf{q}_W(x_l(t))}},$$

563 for all $l = 1, \dots, L$ and $t \in [0, +\infty)$. As a consequence,

564 (a) If $A \prec 0$ and $W_{\text{sym}} \succ 0$, then $\|x_l(t)\|_A$ is bounded for all $l = 1, \dots, L$.

565 (b) If $A \succ 0$ and $W_{\text{sym}} \succ 0$, then $\lim_{t \rightarrow +\infty} \|x_l(t)\|_A = +\infty$ for all $l = 1, \dots, L$.

566 *Proof.* Since A is symmetric, we have

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \mathbf{q}_A(x_l(t)) &= x_l(t)^\top \cdot W \cdot (V^\top)^{-1} \cdot \frac{d}{dt} x_l(t) \\ &= \frac{\sum_{i=1}^L e^{x_l(t)^\top \cdot W \cdot x_i(t)} \cdot x_l(t)^\top \cdot W \cdot x_i(t)}{\sum_{j=1}^L e^{x_l(t)^\top \cdot W \cdot x_j(t)}} \\ &\geq \frac{\sum_{i=1}^L e^{x_l(t)^\top \cdot W \cdot x_i(t)} - L}{\sum_{j=1}^L e^{x_l(t)^\top \cdot W \cdot x_j(t)}} \\ &= 1 - \frac{L}{\sum_{j=1}^L e^{x_l(t)^\top \cdot W \cdot x_j(t)}}. \end{aligned}$$

567 In the above estimation, to obtain the inequality in the third line, we used the fact that $e^\lambda \lambda \geq e^\lambda - 1$
 568 for all $\lambda \in \mathbb{R}$. As a consequence, we have

$$\frac{d}{dt} \mathbf{q}_A(x_l(t)) \geq 2 - \frac{2L}{e^{\mathbf{q}_W(x_l(t))}}, \quad (11)$$

569 as claimed.

570 Next, we will prove (a) and (b) below.

571 (a) Assume that $A \prec 0$ and $W_{\text{sym}} \succ 0$. There is a constant $c > 0$ such that $\|\cdot\|_A^2 \geq c \|\cdot\|_W^2$.
 572 Then it follows from equation (11) that

$$-\frac{d}{dt} \|x_l(t)\|_A^2 \geq 2 - \frac{2L}{e^{c\|x_l(t)\|_A^2}},$$

573 or equivalently,

$$\frac{d}{dt} \|x_l(t)\|_A^2 \leq -2 + \frac{2L}{e^{c\|x_l(t)\|_A^2}},$$

574 Hence, $\|x_l(t)\|_A^2 \leq \frac{1}{c} \log \left(e^{-2ct} \left(e^{c\|x_l(0)\|_A^2} - L \right) + L \right)$, which is bounded.

575 (b) Assume that $A \succ 0$ and $W_{\text{sym}} \succ 0$. There is a constant $c' > 0$ such that $\|\cdot\|_A^2 \leq c' \|\cdot\|_W^2$.
 576 The equation (11) yields

$$\frac{d}{dt} \|x_l(t)\|_A^2 \geq 2 - \frac{2L}{e^{c'\|x_l(t)\|_A^2}},$$

577 which results in $\|x_l(t)\|_A^2 \geq \frac{1}{c'} \log \left(e^{2c't} \left(e^{c'\|x_l(0)\|_A^2} - L \right) + L \right)$. Thus
 578 $\lim_{t \rightarrow +\infty} \|x_l(t)\|_A = +\infty$.

579 □

580 The following lemma studies the limitation of the derivative of tokens in case $A \prec 0$ and $W \succ 0$.

Lemma B.4. Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be the unique solution of the dynamical system (9). Set $A = W \cdot (V^\top)^{-1}$. If $A \prec 0$ and $W \succ 0$, then

$$\int_0^{+\infty} \left\| \frac{d}{ds} x_l(s) \right\|_A^2 ds < +\infty,$$

581 for all l . In particular, we have $\lim_{t \rightarrow +\infty} \frac{d}{dt} x_l(t) = 0$ for all l .

Proof. Consider the function $h: [0, +\infty) \rightarrow \mathbb{R}$ defined by

$$h(t) = \sum_{i=1}^L \sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)}.$$

582 Then h is a positive function. Since W is symmetric, the derivative of h is

$$\begin{aligned} \frac{d}{dt}h(t) &= 2 \sum_{i=1}^L \sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)} \cdot \frac{d}{dt}x_i(t)^\top \cdot W \cdot x_j(t) \\ &= 2 \sum_{i=1}^L \frac{d}{dt}x_i(t)^\top \cdot W \cdot \left(\sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)} \cdot x_j(t) \right) \\ &= 2 \sum_{i=1}^L \frac{d}{dt}x_i(t)^\top \cdot A \cdot \frac{d}{dt}x_i(t) \cdot \left(\sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)} \right). \end{aligned}$$

583 Since $A \prec 0$, we have $\frac{d}{dt}x_i(t)^\top \cdot A \cdot \frac{d}{dt}x_i(t) = -\left\| \frac{d}{dt}x_i(t) \right\|_A^2$. Therefore, we can proceed the above
584 expression as

$$\frac{d}{dt}h(t) = -2 \sum_{i=1}^L \left\| \frac{d}{dt}x_i(t) \right\|_A^2 \cdot \left(\sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)} \right), \quad (12)$$

585 which is nonpositive. As a consequence, $h(t)$ is non-increasing. Thus, $\lim_{t \rightarrow +\infty} h(t)$ exists and
586 finite.

Next, since $A \prec 0$ and $W \succ 0$, it follows from Proposition B.3 that $x_l(t)$ are bounded for all $l = 1, \dots, L$. Therefore, there exists $\epsilon > 0$ such that

$$\sum_{j=1}^L e^{x_i(t)^\top \cdot W \cdot x_j(t)} \geq \epsilon,$$

for all $t \in [0, +\infty)$ and $i = 1, \dots, L$. It follows from equation (12) that

$$\frac{d}{dt}h(t) \leq -2\epsilon \left\| \frac{d}{dt}x_l(t) \right\|_A^2.$$

587 By taking the integral both sides, we see that

$$\int_0^{+\infty} \left\| \frac{d}{ds}x_l(s) \right\|_A^2 ds \leq \frac{1}{2\epsilon} (h(0) - \lim_{s \rightarrow +\infty} h(s)) < +\infty.$$

588 The lemma is then proved. □

589 We will also require the following lemma, which holds for any matrix W whose symmetric component
590 W_{sym} is either positive definite or negative definite. The special case where $W = I_D$ was established
591 in [17, Lemma 8.8]. Our proof, however, is simpler and extends to more general matrices W .

Lemma B.5. Assume that W_{sym} is (either positive or negative) definite. Let x_1^*, \dots, x_L^* be point in \mathbb{R}^D such that

$$\sum_{j=1}^L e^{(x_l^*)^\top \cdot W \cdot x_j^*} \cdot x_j^* = 0, \quad \forall l = 1, \dots, L.$$

592 Then $x_1^* = \dots = x_L^* = 0$.

Proof. Consider the function $g: \mathbb{R}^D \rightarrow \mathbb{R}$ defined by

$$g(u) = \sum_{l=1}^L e^{u^\top \cdot W \cdot x_l^*}, \quad \forall u \in \mathbb{R}^D.$$

593 Then for arbitrary $u, v \in \mathbb{R}^D$, we have

$$g(u) + g(v) - 2g\left(\frac{u+v}{2}\right) = \sum_{l=1}^L \left(e^{\frac{1}{2}u^\top \cdot W \cdot x_l^*} - e^{\frac{1}{2}v^\top \cdot W \cdot x_l^*} \right)^2 \geq 0.$$

Therefore g is convex. From the hypothesis, we have

$$\nabla g(x_1^*) = \dots = \nabla g(x_L^*) = 0.$$

This means that x_1^*, \dots, x_L^* are global minimum of g and the values of g at these points are all equal. Since g is convex, g achieves the global minimal value on the convex hull $\text{conv}(\{x_l^*\}_{l=1}^L)$. As a consequence, we have

$$g(x_i^*) = g(x_j^*) = g\left(\frac{x_i^* + x_j^*}{2}\right),$$

594 for all i, j . Therefore,

$$0 = g(x_i^*) + g(x_j^*) - 2g\left(\frac{x_i^* + x_j^*}{2}\right) = \sum_{l=1}^L \left(e^{\frac{1}{2}x_i^\top \cdot W \cdot x_l^*} - e^{\frac{1}{2}x_j^\top \cdot W \cdot x_l^*} \right)^2.$$

This happens only when

$$\frac{1}{2}x_i^\top \cdot W \cdot x_l^* = \frac{1}{2}x_j^\top \cdot W \cdot x_l^*,$$

or equivalently,

$$(x_i - x_j)^\top \cdot W \cdot x_l^* = 0,$$

for all $l = 1, \dots, L$. In particular, we have

$$\mathbf{q}_W(x_i^* - x_j^*) = (x_i^* - x_j^*)^\top \cdot W \cdot x_i^* - (x_i - x_j)^\top \cdot W \cdot x_j^* = 0.$$

595 Since W_{sym} is definite, \mathbf{q}_W is nondegenerate and $x_i^* = x_j^*$. Thus $x_1^* = \dots = x_L^*$. The only possibility
596 is $x_1^* = \dots = x_L^* = 0$. \square

597 We are ready to prove the main result of the convergence scenario.

598 **Theorem B.6.** *Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))$ be a solution of the dynamical system (9). If
599 $A \prec 0$ and $W \succ 0$, then $\lim_{t \rightarrow +\infty} x_l(t) = 0$ for all $l = 1, \dots, L$.*

600 *Proof.* Set $X(t) = (x_1(t), \dots, x_L(t))$. We need to prove that $\lim_{t \rightarrow +\infty} X(t) = 0$. Assume that
601 this is not the case. According to item (a) of Proposition B.3, $X(t)$ lies in a compact subspace
602 of $(\mathbb{R}^D)^L$. Therefore, there exists a sequence $\{t_k\}_k$ in $[0, +\infty)$ such that $\lim_{k \rightarrow +\infty} t_k = +\infty$
603 and $\lim_{k \rightarrow +\infty} X(t_k) = X^*$ for some $X^* = (x_1^*, \dots, x_L^*) \in (\mathbb{R}^D)^L$ and $X^* \neq 0$. As a conse-
604 quence, we have $\lim_{k \rightarrow +\infty} x_l(t_k) = x_l^*$ for each $l = 1, \dots, L$. While, according to Lemma B.4,
605 $\lim_{k \rightarrow +\infty} \frac{d}{dt} x_l(t_k) = 0$. Therefore, from the dynamical system (9), we obtain

$$\sum_{i=1}^L \frac{e^{(x_l^*)^\top \cdot W \cdot x_i^*}}{\sum_{j=1}^L e^{(x_l^*)^\top \cdot W \cdot x_j^*}} \cdot x_i^* = 0, \quad l = 1, \dots, L. \quad (13)$$

606 Then it follows from Lemma B.5 that $x_1^* = \dots = x_L^* = 0$. However, this contradict to the fact that
607 $X^* \neq 0$. Hence, $\lim_{t \rightarrow +\infty} X(t) = 0$ and the theorem is proved. \square

608 B.3 Divergence Scenario

609 To simplify the technical details, we will consider the case where $A = \lambda W$, i.e. $V = \frac{1}{\lambda} I_D$, for some
610 positive real number λ . The case where A and W have the same signs but $A \neq \lambda W$ may require
611 certain adaptations. In particular, we will prove that when $V = \lambda I_D$ and W is arbitrary, all tokens
612 tend to infinity at an exponential rate (under certain assumptions on the initial conditions). The case
613 where $V = W = I_D$ was already solved in [17, Section 8]. We borrow the technique from there and
614 modify it to ensure it works for all $V = \lambda I_D$ and arbitrary W .

In the following, for each subset $H \subseteq \mathbb{R}^D$, we denote by $\text{conv}(H)$ the convex hull of H , which is the smallest convex set containing H in \mathbb{R}^D . For a point u , the notation $\mathbf{d}(u, H)$ represents the Euclidean distance from u to H , which is defined as

$$\mathbf{d}(u, H) = \inf_{v \in H} \|u - v\|.$$

If H is a closed half-space of \mathbb{R}^D with an outer normal vector \mathbf{n} and $u \notin H$, then

$$\mathbf{d}(u, H) = \mathbf{n}^\top \cdot (u - \mathbf{proj}_H(u)),$$

615 where $\mathbf{proj}_H(u)$ is the projection of u onto H .

616 We begin with the following lemma, whose proof can be found in [17].

Lemma B.7. *Let H be a closed half-space of \mathbb{R}^D with an outer unit normal vector \mathbf{n} . Let $u_1, \dots, u_L: [0, +\infty) \rightarrow \mathbb{R}^D$ be an arbitrary sequence of differentiable functions. For each $t \in [0, +\infty)$, set*

$$\mathbf{Min}(t) = \{1 \leq l \leq L \mid \mathbf{d}(u_l(t), H) = \min_{1 \leq i \leq L} \mathbf{d}(u_i(t), H)\}.$$

Then we have

$$\frac{d}{dt} \min_{1 \leq i \leq L} \mathbf{d}(u_i(t), H) = \min_{i \in \mathbf{Min}(t)} \left(\mathbf{n}^\top \cdot \frac{d}{dt} u_i(t) \right).$$

617 *Proof.* This lemma is already proved in the proof of [17, Proposition 8.2]. □

618 The following proposition refines [17, Proposition 8.2], where the condition $V = W = I_D$ was
 619 assumed. Here, we extend the proof by relaxing this condition and demonstrating that the result
 620 remains valid for matrices of the form $V = \lambda I_D$ with $\lambda > 0$, without imposing any constraints on W .
 621 Another proof of this proposition can also be found in [21, Proposition 2.1].

Proposition B.8. *Assume that $V = \lambda I_D$ for some real number $\lambda > 0$ and W is an arbitrary matrix in $\mathbb{R}^{D \times D}$. Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be the unique solution of the dynamical system (9). Then*

$$e^{-|\lambda|t} x_l(t) \in \mathbf{conv}(\{x_{i0}\}_{i=1}^L),$$

622 for all $l = 1, \dots, L$ and $t \in [0, +\infty)$.

623 *Proof.* Let $z_l(t) = e^{-\lambda t} x_l(t)$. Then we have

$$\begin{aligned} \frac{d}{dt} z_l(t) &= e^{-\lambda t} \left(\frac{d}{dt} x_l(t) - \lambda x_l(t) \right) \\ &= e^{-\lambda t} \left(\lambda \sum_{i=1}^L \frac{e^{2\lambda t} z_l^\top \cdot W \cdot z_i}{\sum_{j=1}^L e^{2\lambda t} z_l^\top \cdot W \cdot z_j} e^{\lambda t} z_i(t) - \lambda e^{\lambda t} z_l(t) \right) \\ &= \lambda \sum_{i=1}^L \left(\frac{e^{2\lambda t} z_l^\top \cdot W \cdot z_i}{\sum_{j=1}^L e^{2\lambda t} z_l^\top \cdot W \cdot z_j} \right) (z_i(t) - z_l(t)). \end{aligned}$$

624 Therefore, the function $(z_1(t), \dots, z_L(t))$ satisfies the dynamical system:

$$\frac{dz_l(t)}{dt} = \sum_{i=1}^L P_{l,i}(t, z_1(t), \dots, z_L(t)) \cdot (z_i(t) - z_l(t)), \quad l = 1, \dots, L, \quad (14)$$

with the initial conditions $z_l(0) = x_{l0}$, where

$$P_{l,i}(t, z_1, \dots, z_L) = \frac{e^{2\lambda t} z_l^\top \cdot W \cdot z_i}{\sum_{j=1}^L e^{2\lambda t} z_l^\top \cdot W \cdot z_j} \cdot \lambda.$$

We claim that, for every closed half-space H of \mathbb{R}^D such that $\mathbf{conv}(\{x_{i0}\}_{i=1}^L) \cap H = \emptyset$, the map $\alpha: [0, +\infty) \rightarrow \mathbb{R}$ defined by

$$\alpha(t) = \min_{1 \leq i \leq L} \mathbf{d}(z_i(t), H)$$

625 is non-decreasing. Indeed, using item (a) of Lemma B.7, we have

$$\begin{aligned} \frac{d}{dt} \alpha(t) &= \min_{i \in \mathbf{Min}(t)} \left(\mathbf{n} \cdot \frac{d}{dt} z_i(t) \right) \\ &= \min_{i \in \mathbf{Min}(t)} \left(\sum_{j=1}^L P_{l,i}(t, z_1(t), \dots, z_L(t)) \cdot \mathbf{n}^\top \cdot (z_j(t) - z_i(t)) \right). \end{aligned}$$

626 On the right hand side, we have

$$\begin{aligned}\mathbf{n}^\top \cdot (z_j(t) - z_i(t)) &= \mathbf{n}^\top \cdot (z_j(t) - \mathbf{proj}_H(z_j(t))) - \mathbf{n}^\top \cdot (z_i(t) - \mathbf{proj}_H(z_i(t))) \\ &\quad + \mathbf{n}^\top \cdot (\mathbf{proj}_H(z_j(t)) - \mathbf{proj}_H(z_i(t))) \\ &= \mathbf{d}(z_j(t), H) - \mathbf{d}(z_i(t), H)\end{aligned}$$

which is nonnegative since $i \in \mathbf{Min}(t)$. Therefore, $\frac{d}{dt}\alpha(t) \geq 0$ and α is non-decreasing. As a consequence, we have

$$\mathbf{d}(z_l(t), H) \geq \min_{1 \leq i \leq L} \mathbf{d}(x_{i0}, H) > 0,$$

for all $l = 1, \dots, L$ and $t \in [0, +\infty)$. This means that, $z_l(t)$ is outside H as long as $H \cap \mathbf{conv}(\{x_{i0}\}_{i=1}^L) = \emptyset$. Hence,

$$z_l(t) \in \bigcap_{\substack{H \text{ closed half-space} \\ H \cap \mathbf{conv}(\{x_{i0}\}_{i=1}^L) = \emptyset}} H = \bigcap_{\substack{H' \text{ open half-space} \\ H' \supset \mathbf{conv}(\{x_{i0}\}_{i=1}^L)}} H' = \mathbf{conv}(\{x_{i0}\}_{i=1}^L).$$

627 The proposition is then proved. \square

628 In the following, for each nonzero vector $\mathbf{n} \in \mathbb{R}^D$, we denote by $H_{\mathbf{n}}$ the closed half-space of \mathbb{R}^D
629 with normal vector \mathbf{n} such that zero belongs to its boundary $\partial H_{\mathbf{n}}$. In this case, $\partial H_{\mathbf{n}}$ represents the
630 hyperplane containing zero with normal vector \mathbf{n} .

631 **Theorem B.9.** Assume that the value matrix V has at least one positive eigenvalue. Let \mathbf{n} be an
632 eigenvector of V corresponding to a positive eigenvalue. Then

$$\min_{1 \leq i \leq L} (\mathbf{n}^\top x_{i0}) \leq \mathbf{n}^\top e^{-tV^\top} x_l(t) \leq \max_{1 \leq i \leq L} (\mathbf{n}^\top x_{i0}),$$

633 for all $t \in [0, +\infty)$ and $l = 1, \dots, L$.

634 As a consequence, if the initial points x_{10}, \dots, x_{L0} are all on one side of the hyperplane $\partial H_{\mathbf{n}}$, and if
635 V has only positive eigenvalues, then

$$\lim_{t \rightarrow +\infty} \|x_l(t)\| = +\infty, \quad \text{for all } l.$$

636 *Proof.* Let $z_l(t) = e^{-tV^\top} x_l(t)$. Then the function $(z_1(t), \dots, z_L(t))$ satisfies the dynamical system:

$$\frac{dz_l(t)}{dt} = \sum_{i=1}^L P_{l,i}(t, z_1(t), \dots, z_L(t)) \cdot V^\top \cdot (z_i(t) - z_l(t)), \quad l = 1, \dots, L, \quad (15)$$

with the initial conditions $z_l(0) = x_{l0}$, where

$$P_{l,i}(t, z_1, \dots, z_L) = \frac{e^{z_l^\top \cdot e^{tV} \cdot W \cdot e^{tV^\top} \cdot z_i}}{\sum_{j=1}^L e^{z_l^\top \cdot e^{tV} \cdot W \cdot e^{tV^\top} \cdot z_j}}.$$

637 Let $\lambda > 0$ be the eigenvalue associated to \mathbf{n} . By multiplying \mathbf{n}^\top into both sides of equation (15), and
638 set $y_l(t) = \mathbf{n}^\top \cdot z_l(t)$, we obtain

$$\frac{dy_l(t)}{dt} = \sum_{i=1}^L P_{l,i}(t, z_1(t), \dots, z_L(t)) \cdot \lambda \cdot (y_i(t) - y_l(t)), \quad l = 1, \dots, L, \quad (16)$$

with the initial conditions $y_l(0) = \mathbf{n}^\top \cdot x_{l0} \in \mathbb{R}$. By using the same argument in Proposition B.8,
with z_l is replaced by y_l here, we see that

$$y_l(t) \in \mathbf{conv}(\{y_{i0}\}_{i=1}^L),$$

and hence,

$$\min_{1 \leq i \leq L} y_{i0} \leq y_l(t) \leq \max_{1 \leq i \leq L} y_{i0},$$

639 for all $t \in [0, +\infty)$ and $l = 1, \dots, L$. Therefore,

$$\min_{1 \leq i \leq L} (\mathbf{n}^\top \cdot x_{i0}) \leq \mathbf{n}^\top \cdot e^{-tV^\top} x_l(t) \leq \max_{1 \leq i \leq L} (\mathbf{n}^\top \cdot x_{i0}),$$

640 as claimed.

In case the initial points x_{10}, \dots, x_{L0} are all one side of the hyperplane ∂H_n , then either $\min_{1 \leq i \leq L} (\mathbf{n}^\top \cdot x_{i0}) > 0$ or $\max_{1 \leq i \leq L} (\mathbf{n}^\top \cdot x_{i0}) < 0$. In both case, there is $\epsilon > 0$ such that

$$\left| \mathbf{n} \cdot e^{-tV^\top} \cdot x_l(t) \right| > \epsilon$$

641 for all $t \in [0, +\infty)$ and l . Hence, when V has only positive eigenvalues, we must have

$$\lim_{t \rightarrow +\infty} \|x_l(t)\| = +\infty, \quad \text{for all } l.$$

642 The theorem is then proved. \square

643 C Effects of Absolute and Rotary Positional Encodings

644 In this section, we analyze the influence of absolute and rotary positional encodings on the dynamical
645 behavior of tokens within self-attention mechanisms.

646 C.1 Absolute Positional Encoding

647 Recall that the dynamical system governing the continuous-time limit of the self-attention with
648 absolute positional encoding is given by

$$\frac{dx_l(t)}{dt} = V^\top \cdot \sum_{i=1}^L \left(\frac{e^{(x_l(t)+p_l)^\top \cdot W \cdot (x_i(t)+p_i)}}{\sum_{j=1}^L e^{(x_l(t)+p_l)^\top \cdot W \cdot (x_j(t)+p_j)}} \right) \cdot (x_i(t) + p_i), \quad (17)$$

649 for $l = 1, \dots, L$, with the initial conditions

$$(x_1(0), \dots, x_L(0)) = (x_{10}, \dots, x_{L0}) \in (\mathbb{R}^D)^L.$$

650 Here $p_i = [p_{i,1}, \dots, p_{i,D}]^\top \in \mathbb{R}^D$. A common choice is to either learn the positional embeddings p_i
651 jointly with the model parameters, or to fix them using a sinusoidal scheme: with

$$p_{i,j} = \begin{cases} \sin(i \cdot 10000^{-\frac{j}{D}}), & \text{if } j \text{ is even,} \\ \cos(i \cdot 10000^{-\frac{j-1}{D}}), & \text{if } j \text{ is odd.} \end{cases}$$

652 This differential system can be easily transformed into equation (2) by the transition $x_l \mapsto x_l + p_l$.
653 Therefore, the dynamical properties of the self-attention with and without absolute are the similar as
654 we will see in the following corollaries.

655 **Corollary C.1.** *Let $(x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))$ be a solution of the dynamical system (17).
656 If $A \prec 0$ and $W \succ 0$, then $\lim_{t \rightarrow +\infty} x_l(t) = -p_l$ for all $l = 1, \dots, L$.*

657 **Corollary C.2.** *Let $X(t) = (x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be a solution of the differential
658 system (17). Assume that the value matrix V has at least one positive eigenvalue. Let \mathbf{n} be an
659 eigenvector of V corresponding to a positive eigenvalue. Then*

$$\min_{1 \leq i \leq L} (\mathbf{n}^\top (x_{i0} + p_i) - e^{-tV^\top} p_l) \leq \mathbf{n}^\top e^{-tV^\top} x_l(t) \leq \max_{1 \leq i \leq L} (\mathbf{n}^\top (x_{i0} + p_i)) - e^{-tV^\top} p_l,$$

660 for all $t \in [0, +\infty)$ and $l = 1, \dots, L$.

661 As a consequence, if the points $x_{10} + p_1, \dots, x_{L0} + p_L$ are all on one side of the hyperplane
662 $\partial H_n + e^{-tV^\top} p_l$, and if V has only positive eigenvalues, then

$$\lim_{t \rightarrow +\infty} \|x_l(t)\| = +\infty, \quad \text{for all } l.$$

663 C.2 Rotary Positional Encoding

664 In this section, we assume that the token dimension D is an even number. In contrast to absolute po-
665 sitional encoding, the continuous-time dynamics of the self-attention with rotary positional encoding
666 can be described via the differential system [46]

$$\frac{dx_l(t)}{dt} = V^\top \cdot \sum_{i=1}^L \left(\frac{e^{x_l(t)^\top \cdot W_{li} \cdot x_i(t)}}{\sum_{j=1}^L e^{x_l(t)^\top \cdot W_{lj} \cdot x_j(t)}} \right) \cdot x_i(t), \quad (18)$$

for $l = 1, \dots, L$, where

$$W_{li} = \frac{1}{\sqrt{D_k}} \left(Q \cdot K^\top + \bar{Q} \cdot R_{\theta, i-l}^D \cdot \bar{K}^\top \right), \quad (19)$$

with $Q, K, \bar{Q}, \bar{K} \in \mathbb{R}^{D \times D}$ are two additional learnable matrices and

$$R_{\Theta, m}^D = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{D/2} & -\sin m\theta_{D/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{D/2} & \cos m\theta_{D/2} \end{pmatrix}$$

and $\Theta = \{\theta_i = 10000^{-2(i-1)/D}, i \in [1, 2, \dots, D/2]\}$.

Rotary positional encoding is an essential component of latent attention, which lies at the core of DeepSeek [31]. Unlike absolute positional encoding, the differential system governing rotary positional encoding exhibits markedly different behavior. In cases where token trajectories diverge to infinity under absolute positional encoding, we observe a similar divergence scenario in the presence of rotary positional encoding, as demonstrated below:

Corollary C.3. *Let $X(t) = (x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty))^L$ be a solution of the differential system (18). Assume that the value matrix V has at least one positive eigenvalue. Let \mathbf{n} be an eigenvector of V corresponding to a positive eigenvalue. Then*

$$\min_{1 \leq i \leq L} (\mathbf{n}^\top x_{i0}) \leq \mathbf{n}^\top e^{-tV^\top} x_l(t) \leq \max_{1 \leq i \leq L} (\mathbf{n}^\top x_{i0}),$$

for all $t \in [0, +\infty)$ and $l = 1, \dots, L$.

As a consequence, if the points x_{10}, \dots, x_{L0} are all on one side of the hyperplane $\partial H_{\mathbf{n}}$, and if V has only positive eigenvalues, then

$$\lim_{t \rightarrow +\infty} \|x_l(t)\| = +\infty, \quad \text{for all } l.$$

Proof. Apply the same argument as in the proof of Theorem B.9. \square

Remark C.4. In contrast to absolute positional encoding, rotary positional encoding induces notably different dynamical behavior by promoting token divergence (even for the cases when tokens converge to a finite point in the absolute positional encoding). Specifically, the presence of the additional term $\bar{Q} \cdot R_{\theta, i-l}^D \cdot \bar{K}^\top$ in the query-key interaction matrix W_{li} , as defined in equation (19), can hinder the system from transitioning into a convergence regime. Consequently, self-attention equipped with rotary positional encoding tends to exhibit divergence behavior more frequently than those using absolute encoding or no positional encoding at all.

D Additional Experimental Details

D.1 Parameter settings used in the simulations

D.1.1 Figure 1: Distances between tokens over time

On the left side of Figure 1, we choose

- $A = \begin{pmatrix} 1.72628 & -3.79592 \\ -0.914069 & 3.49779 \end{pmatrix}$ whose symmetric component A_{sym} has positive eigenvalues 5.12809 and 0.0959758;
- $W = \begin{pmatrix} 0.534636 & -0.798866 \\ -1.17152 & -1.92153 \end{pmatrix}$; and
- the initial values $x_{10} = (-1.17525, 1.99834)$, $x_{20} = (-0.0231564, 0.591678)$, $x_{30} = (-0.94811, -1.37996)$, $x_4 = (1.00246, -1.69335)$.

On the right side of Figure 1, we choose

- $A = \begin{pmatrix} -1.43778 & -1.10989 \\ 0.563455 & -0.401696 \end{pmatrix}$ whose symmetric component A_{sym} has negative eigenvalues -1.50541 and -0.334061 ;
- $W = \begin{pmatrix} 0.433083 & -0.0371911 \\ -0.715343 & -1.53568 \end{pmatrix}$; and
- the initial values $x_{10} = (0.123688, 0.20691)$, $x_{20} = (0.53086, 1.47281)$, $x_{30} = (-0.78388, -1.24115)$, $x_{40} = (1.63476, 0.321809)$.

704 D.1.2 Figure 2: Token trajectories under convergence and divergence scenarios

705 In Figure 2a, we choose the following parameters:

- $A = \begin{pmatrix} -2.94058 & -2.12076 \\ -5.14498 & -4.58104 \end{pmatrix}$, thus the symmetric component A_{sym} has negative eigenvalues -7.48513 and -0.0364942 ;
- $W = \begin{pmatrix} 0.902496 & -2.37879 \\ 4.36478 & 3.84768 \end{pmatrix}$, thus the symmetric component W_{sym} has positive eigenvalues 4.15119 and 0.598979 .

Figure 2b illustrates the divergence scenario with $A = 2W$ (thus $V = 2I$) and

$$W = \begin{pmatrix} -0.404078 & 0.982735 \\ -0.567909 & 0.600242 \end{pmatrix}.$$

710 D.1.3 Figure 5: Token trajectories shift to divergence regime in RoPE

711 On the left of Figure 5, we choose

- $Q = \begin{pmatrix} 0.07331137 & 0.17647239 \\ -0.32738218 & -0.43457359 \end{pmatrix}$
- $K = \begin{pmatrix} -2.54009796 & 1.82991692 \\ -0.95688637 & 0.60349328 \end{pmatrix}$, thus W_{sym} have positive eigenvalues 0.03766541 and 0.15005164
- $V = -1.5I$

716 On the right of Figure 5, we choose the additional parameters

$$\begin{aligned} 717 \quad & - \bar{Q} = \begin{pmatrix} -3.01517413 & 2.4430872 \\ 2.11630464 & 1.40111342 \end{pmatrix} \\ 718 \quad & - \bar{K} = \begin{pmatrix} 5.03454859 & -3.12492845 \\ 4.58643881 & -2.00780098 \end{pmatrix}, \end{aligned}$$

719 D.1.4 Figure 9: Token trajectories in convergence and divergence regime

720 On Figure 9a (convergence case), we choose the following parameters:

- $Q = \begin{pmatrix} -1.18765511 & 0.8975229 \\ -0.7793589 & 0.79105257 \end{pmatrix}$
- $K = \begin{pmatrix} -1.97520362 & -1.98198651 \\ 2.24167927 & 2.93460903 \end{pmatrix}$
- $\bar{Q} = \begin{pmatrix} 1.04027991 & -0.12991073 \\ -1.32542484 & 1.08074871 \end{pmatrix}$
- $\bar{K} = \begin{pmatrix} -1.00477795 & -0.48804888 \\ -0.42151108 & 0.02556926 \end{pmatrix}$
- $V = -I$

726 On Figure 9b (divergence case), we choose the following parameters:

$$\begin{aligned}
727 \quad & \bullet Q = \begin{pmatrix} 2.068739 & -1.83750201 \\ -0.75622145 & 0.4784381 \end{pmatrix} \\
728 \quad & \bullet K = \begin{pmatrix} -1.12583337 & -1.40120114 \\ -2.79629618 & -3.24668939 \end{pmatrix} \\
729 \quad & \bullet \bar{Q} = \begin{pmatrix} -0.67880222 & 1.21234986 \\ -0.67132474 & 0.90406252 \end{pmatrix} \\
730 \quad & \bullet K = \begin{pmatrix} 0.57406142 & 2.88899216 \\ -1.10421806 & 0.75603913 \end{pmatrix} \\
731 \quad & \bullet V = 1.5I
\end{aligned}$$

732 D.2 Implementation of Convergence, Divergence, and Intermediate scenarios.

733 We propose a strategy to guarantee that the model falls into the three scenarios described in Section 5.1
734 and Section 3.4. In particular, we introduce a reparameterization of the matrices Q , K , and V that
735 enforces the positive or negative definiteness of the matrices

$$736 \quad W_{\text{sym}} = \frac{1}{2} (W + W^\top), \quad A_{\text{sym}} = \frac{1}{2} (A + A^\top), \quad (20)$$

736 where $W = QK^\top$ and $A = W(V^\top)^{-1}$. We first describe the approach to guarantee the positive
737 definiteness of A_{sym} and W_{sym} and subsequently extend the framework to accommodate negative
738 definiteness and intermediate cases.

739 **Ensuring the Positive Definiteness of A_{sym} .** To enforce the positive definiteness of A_{sym} , we leverage
740 the LDL^\top decomposition [18]. Specifically, we parametrize:

$$A_{\text{sym}} = L_a D_a L_a^\top, \quad (21)$$

741 where L_a is a lower triangular matrix with ones on the diagonal, and D_a is a diagonal matrix with
742 strictly positive elements. The positivity of D_a is ensured by applying the Softplus function:

$$D_a = \text{diag}(\text{Softplus}(d_a)). \quad (22)$$

743 **Parametrization of A and W .** Given A_{sym} , from Eq. 20 we obtain

$$A = A_{\text{sym}} + X_a, \quad (23)$$

744 where X_a is an antisymmetric matrix satisfying $X_a = -X_a^\top$. A natural parametrization for such a
745 matrix is:

$$X_a = T_a - T_a^\top, \quad (24)$$

746 Consequently, the final parametrization of A ensuring positive definiteness of A_{sym} is

$$A = L_a \text{diag}(\text{Softplus}(d_a)) L_a^\top + T_a - T_a^\top. \quad (25)$$

747 A similar parametrization applies to W to enforce the positive definiteness of W_{sym} :

$$W = L_w \text{diag}(\text{Softplus}(d_w)) L_w^\top + T_w - T_w^\top. \quad (26)$$

748 The above formulations ensures that both W_{sym} and A_{sym} maintain the desired definiteness properties
749 while allowing for a flexible parametrization of W and A .

750 **Parametrization of Q , K , and V .** In terms of Q , K , and V , the aforementioned formulations
751 equivalent to:

$$\begin{cases} Q \cdot K^\top & = L_w \text{diag}(\text{Softplus}(d_w)) L_w^\top + T_w - T_w^\top, \\ Q \cdot K^\top \cdot (V^\top)^{-1} & = L_a \text{diag}(\text{Softplus}(d_a)) L_a^\top + T_a - T_a^\top. \end{cases} \quad (27)$$

752 We designate Q , L_w , d_w , L_a , d_a , T_w , and T_a as learnable parameters. The key and value projection
753 matrices for self-attention are then computed as:

$$\begin{cases} K & = [Q^{-1} \cdot L_w \text{diag}(\text{Softplus}(d_w)) L_w^\top + T_w - T_w^\top]^\top, \\ V & = [(L_a \text{diag}(\text{Softplus}(d_a)) L_a^\top + T_a - T_a^\top)^{-1} \cdot (L_w \text{diag}(\text{Softplus}(d_w)) L_w^\top + T_w - T_w^\top)]^\top. \end{cases} \quad (28)$$

754 This parametrization guarantees that A_{sym} and W_{sym} are positive definite.

755 **Adjustments for Negative Definite and Intermediate Cases.** The definiteness of A_{sym} and W_{sym}
756 is determined by the sign of the elements in D_a and D_w , respectively. In the positive definite case,
757 these elements are constrained to be positive using the Softplus function. To adapt to the negative
758 definite and intermediate cases, we control the sign of the elements in D_a and D_w by multiplying
759 the output of Softplus with a vector containing only -1 for the negative definite case or a vector
760 containing an equal number of 1 and -1 for the intermediate case.

761 This approach ensures that our parametrization is flexible enough to accommodate different definite-
762 ness requirements while maintaining learnability and numerical stability.

763 **D.3 Details on Language Modeling experiments**

764 **D.3.1 Dataset**

765 **WikiText-103.** [34] The WikiText-103 dataset comprises approximately 268,000 unique words. Its
766 training set includes around 28,000 articles, totaling 103 million tokens. On average, this corresponds
767 to text blocks of about 3,600 words per article. The validation and test sets each consist of 60 articles,
768 containing 218,000 and 246,000 tokens, respectively.

769 **EnWik8.** [20] The Enwik8 dataset is a byte-level corpus comprising 100 million bytes extracted
770 from Wikipedia. In addition to standard English text, it includes markup, special characters, and
771 content in multiple languages. The standard split provides 90 million bytes for training and 5 million
772 for testing.

773 **D.3.2 Wikitext103 Model and Training Configurations**

774 **Model.** We utilize the Transformer-XL [9] (<https://github.com/kimiyoung/transformer-xl>) architec-
775 ture for word-level language modeling on the WikiText-103 dataset. The model comprises 16 layers,
776 each with a hidden size of 410. Multi-head attention is implemented with 10 heads, each having a
777 dimensionality of 41. The position-wise feedforward networks have an inner dimension of 2100.
778 Regularization is applied via a dropout rate of 0.05 on residual connections. For Rotary positional
779 embedding, the rotational dimension used is 16.

780 **Training Configurations.** Training is conducted using the Adam optimizer with a learning rate of
781 0.00025. A linear warmup is applied for the first 1,000 steps, followed by a cosine annealing schedule
782 over a total of 200,000 training steps. The model is trained with a target sequence length of 150 tokens
783 and no memory length, effectively disabling the segment-level recurrence mechanism. Evaluation
784 is performed with a slightly longer target length of 156 tokens. Training utilizes 2 NVIDIA A100
785 SXM4 80GB GPUs with a total batch size of 60.

786 **D.3.3 EnWik8 Model and Training Configurations**

787 **Model.** We trained an autoregressive Transformer model on the Enwik8 dataset using the x-
788 transformers (<https://github.com/lucidrains/x-transformers>) library. The model follows a GPT-style
789 architecture, comprises a 6-layer Transformer decoder. Each layer uses 8 attention heads and a model
790 dimension of 512. For tokenization, byte-level encoding was used, resulting in a vocabulary size of
791 256 unique tokens. Both training and generation sequences were fixed at 1024 tokens.

792 **Training Configurations.** Data was sampled into overlapping sequences of length 1025 (1024 input
793 tokens plus 1 target token). We used the Adam optimizer with a learning rate of $\times 10^{-4}$, and applied
794 gradient clipping with a maximum norm of 0.5. Gradients were accumulated over 4 steps to simulate
795 larger batch sizes. A batch size of 4 was used, with gradient accumulation yielding an effective batch
796 size of 16. The model was trained for 100,000 iterations. Validation was performed every 100 steps,
797 and text samples were generated every 500 steps with a generation length of 1024 tokens. Training
798 was performed on a NVIDIA A100 SXM4 80GB GPU.

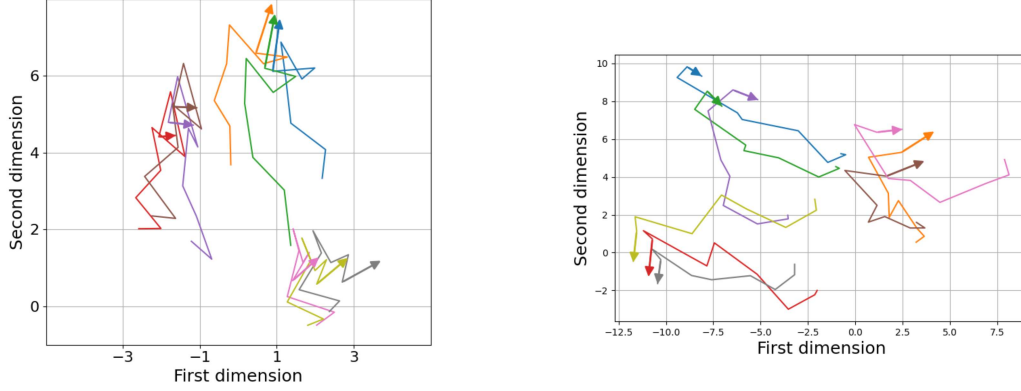
799 **D.4 Details on ImageNet-1K object recognition task**

800 **D.4.1 Dataset**

801 **ImageNet-1K.** [11] This dataset spans 1000 object classes and contains 1,281,167 training images,
802 50,000 validation images. The model learns to predict the class of the input image among 1000
803 categories. We report the top-1 and top-5 accuracy on all experiments.

804 **D.4.2 Model and Training Configurations**

805 **Models.** We adopted the DeiT [50] (<https://github.com/facebookresearch/deit>) architecture, a
806 lightweight Vision Transformer (ViT) variant designed for efficient image classification. The model



(a) With LayerNorm, without feedforward. (b) With both LayerNorm and feedforward.

Figure 7: Tokens’ trajectory in the later layers of a pre-trained GPT-2 model forms clusters with similar shapes.

used is Tiny version, which divides each 224×224 input image into non-overlapping 16×16 patches, resulting in a sequence of 196 tokens. Each patch is linearly projected into a 192-dimensional embedding space. The Transformer encoder consists of 12 layers, each employing multi-head self-attention with 3 heads, and an MLP block with a hidden dimension four times the embedding size (i.e., 768). Biases are included in the query, key, and value projections, and layer normalization is applied with an epsilon value of $1e-6$. The model includes a learnable [CLS] token and absolute positional embeddings. No distillation token or teacher model was used, and no architectural modifications (e.g., convolutional stems or hybrids) were introduced.

Training Configurations. The model was trained on the full ImageNet-1k training set for 300 epochs using the AdamW optimizer with a base learning rate of 5×10^{-4} , and weight decay of 0.05. The learning rate followed a cosine decay schedule, with a linear warmup phase over the first 5 epochs, and a minimum learning rate of 10^{-5} . A stochastic depth rate (drop path) of 0.1 were used for regularization. Data augmentation included RandAugment, Mixup with $\alpha = 0.8$, CutMix with $\alpha = 1.0$, label smoothing with $\epsilon = 0.1$, and random erasing with a probability of 0.25. The model was trained with a batch size of 64 across 4 NVIDIA A100 SXM4 80GB GPUs using mixed precision. Exponential Moving Average (EMA) of model weights was maintained with a decay factor of 0.99996.

D.5 Additional visualizations

D.5.1 Tokens’ trajectory in pre-trained Transformers

We provide additional visualization on pre-trained GPT-2 model to observe the trajectory of tokens. As illustrated in Figure 7, tokens trajectories in later layers still forms clusters with similar shapes.

D.5.2 Token norm and distance in RoPE

We visualize the evolution of token norms and pairwise L2 distances in a pre-trained Transformer model without LayerNorm and feed-forward layers, comparing Rotary Positional Encodings (RoPE) with sinusoidal positional encodings. Figure 8 shows that both token norms and token distances diverge faster with RoPE than with sinusoidal encodings, suggesting that RoPE mitigates the convergence aspect in self-attention.

D.5.3 Token trajectories with RoPE

We provide additional simulations of token trajectories in the system of Equation 8 for the convergence (Figure 9a) and divergence case (Figure 9b).

D.6 Additional Results

D.6.1 Full Evaluation Results

In this section, we demonstrate the means and standard deviations of the experiments we executed. Table 4 and 5 show the means and standard deviations of our experiments in Section 5.2 and Section 5.1.

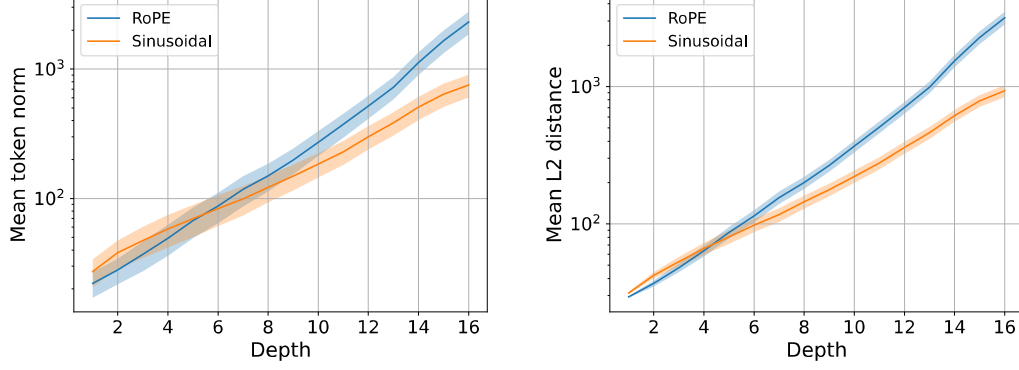
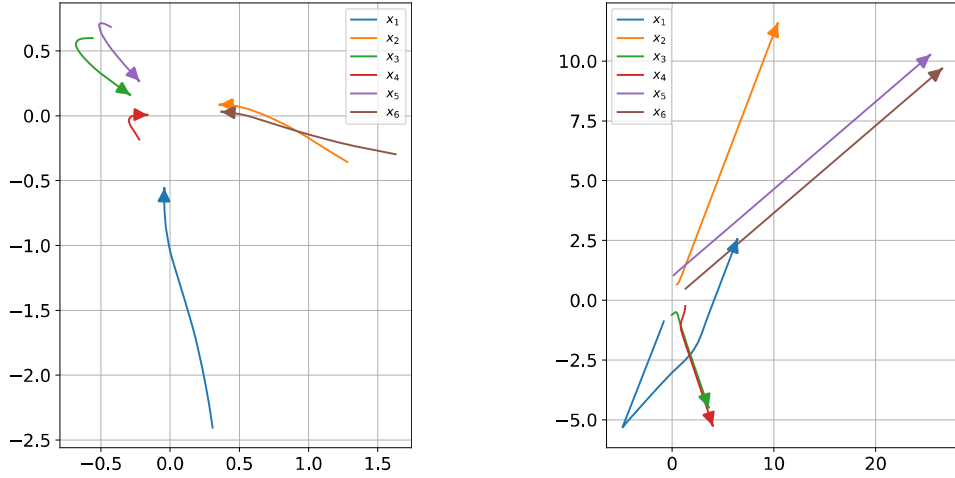


Figure 8: Token norm and distance throughout layers of a pre-trained model with RoPE and Sinusoidal positional encoding.



(a) Convergence case of Self-Attention + RoPE. Tokens tend to converge to zero when $t \rightarrow \infty$. (b) Divergence case of Self-Attention + RoPE. Tokens tend to diverge to ∞ when $t \rightarrow \infty$.

Figure 9: Self-Attention with RoPE under (a) convergence and (b) divergence settings.

Table 4: Bits Per Characters (BPC) and Perplexity (PPL) of Transformers with Rotary Positional Encoding across different scenarios on EnWik8 and WikiText-103 language modeling.

Scenario	EnWik8 Pretrain	WikiText-103 Pretrain	
	Test BPC (\downarrow)	Valid PPL (\downarrow)	Test PPL (\downarrow)
<i>Transformer + RoPE</i>	1.295 ± 0.003	31.37 ± 0.14	32.35 ± 0.18
Transformer + RoPE + λI_D	1.288 ± 0.002	31.10 ± 0.16	32.09 ± 0.17
Transformer + RoPE + λA	1.281 ± 0.002	31.06 ± 0.11	32.04 ± 0.14

Table 5: Bits Per Characters (BPC) and Perplexity (PPL) of Transformers with sinusoidal positional encoding across scenarios on EnWik8 and WikiText-103 language modeling.

Scenario	EnWik8 Pretrain	WikiText-103 Pretrain	
	Test BPC (\downarrow)	Valid PPL (\downarrow)	Test PPL (\downarrow)
<i>Baseline</i>	1.331 ± 0.002	31.63 ± 0.12	32.37 ± 0.10
Convergence	1.350 ± 0.003	32.24 ± 0.15	33.05 ± 0.13
Intermediate	1.345 ± 0.002	31.91 ± 0.11	32.78 ± 0.12
Divergence	1.324 ± 0.002	31.12 ± 0.11	32.07 ± 0.13

842 **E Broader Impact**

843 This work advances the theoretical understanding of token dynamics in Transformer models, providing
844 insights that could enhance model stability and performance across various applications. While
845 primarily theoretical, these findings may inform the development of more robust AI systems. However,
846 as with any advancement in AI, there is a potential for misuse in applications. We encourage the
847 community to consider these implications and promote responsible use of such technologies.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have mentioned the Limitations in Section 6

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We have provided the full set of assumptions and complete proof for the theoretical results in Appendix B

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided all experimental setup in Appendix D. Besides, we also provide the code to reproduce the results in the paper, which can be found in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

954 Question: Does the paper provide open access to the data and code, with sufficient instruc-
955 tions to faithfully reproduce the main experimental results, as described in supplemental
956 material?

957 Answer: [Yes]

958 Justification: We have provided the code in the supplemental material, with sufficient
959 instructions to reproduce the results.

960 Guidelines:

- 961 • The answer NA means that paper does not include experiments requiring code.
- 962 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/
963 public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 964 • While we encourage the release of code and data, we understand that this might not be
965 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
966 including code, unless this is central to the contribution (e.g., for a new open-source
967 benchmark).
- 968 • The instructions should contain the exact command and environment needed to run to
969 reproduce the results. See the NeurIPS code and data submission guidelines ([https:
970 //nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 971 • The authors should provide instructions on data access and preparation, including how
972 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 973 • The authors should provide scripts to reproduce all experimental results for the new
974 proposed method and baselines. If only a subset of experiments are reproducible, they
975 should state which ones are omitted from the script and why.
- 976 • At submission time, to preserve anonymity, the authors should release anonymized
977 versions (if applicable).
- 978 • Providing as much information as possible in supplemental material (appended to the
979 paper) is recommended, but including URLs to data and code is permitted.

980 6. Experimental setting/details

981 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
982 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
983 results?

984 Answer: [Yes]

985 Justification: We have specified all the training and test details, which can be found in
986 Appendix D.

987 Guidelines:

- 988 • The answer NA means that the paper does not include experiments.
- 989 • The experimental setting should be presented in the core of the paper to a level of detail
990 that is necessary to appreciate the results and make sense of them.
- 991 • The full details can be provided either with the code, in appendix, or as supplemental
992 material.

993 7. Experiment statistical significance

994 Question: Does the paper report error bars suitably and correctly defined or other appropriate
995 information about the statistical significance of the experiments?

996 Answer: [Yes]

997 Justification: We have provided error bars for an experiment in Section D.6.1. Besides, all
998 the experiments are averaged over 5 runs.

999 Guidelines:

- 1000 • The answer NA means that the paper does not include experiments.
- 1001 • The authors should answer "Yes" if the results are accompanied by error bars, confi-
1002 dence intervals, or statistical significance tests, at least for the experiments that support
1003 the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have provided information about computing resources needed in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the societal impact of the work performed in Appendix E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper has credited all code and data that has been used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The source code of the paper is based on existing source (which has been credited properly). We also provide README file to run the code.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

1159 • We recognize that the procedures for this may vary significantly between institutions
1160 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1161 guidelines for their institution.
1162 • For initial submissions, do not include any information that would break anonymity (if
1163 applicable), such as the institution conducting the review.

1164 **16. Declaration of LLM usage**

1165 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1166 non-standard component of the core methods in this research? Note that if the LLM is used
1167 only for writing, editing, or formatting purposes and does not impact the core methodology,
1168 scientific rigorousness, or originality of the research, declaration is not required.

1169 Answer: [NA]

1170 Justification: The core method development in this research does not involve LLMs as any
1171 important, original, or non-standard components.

1172 Guidelines:

1173 • The answer NA means that the core method development in this research does not
1174 involve LLMs as any important, original, or non-standard components.
1175 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1176 for what should or should not be described.