LOTFORMER: DOUBLY STOCHASTIC LINEAR ATTENTION VIA LOW RANK OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard softmax attention is row stochastic and scales quadratically with sequence length. Doubly stochastic attention also controls aggregate key participation, but computing it is typically expensive. We introduce LOTFORMER, an efficient attention mechanism designed for bidirectional Transformers, where tokens act as both information sources and receivers. LOTFORMER routes interactions through a learned low rank pivot measure and composes two entropic transport couplings. The resulting ideal attention operator is doubly stochastic, has rank at most r, and applies to values in $O(nr)$ time without forming the full pairwise matrix. We evaluate the method both as a replacement for attention in pretrained embedding models and as an attention mechanism trained in vision Transformers. Across embedding models from 1.5B to 7B parameters, LOTFORMER preserves 98.3% to 98.9% of the original softmax retrieval quality. On four natural LongEmbed tasks, gte Qwen2 7B retains 98.9% of its original nDCG@10. In a separate throughput stress test at 128k tokens, it achieves a $7.5\times$ speedup over softmax. LOTFORMER also performs competitively across multiple vision backbones and under limited supervision. These results demonstrate a practical route to efficient bidirectional attention with doubly stochastic structure. Code is available at `https://github.com/anonymous13761234/LOTFormer`.

1 INTRODUCTION

Standard self attention maps query key interactions to a row stochastic attention matrix A, satisfying $A\mathbf{1} = \mathbf{1}$. Each query therefore distributes unit attention mass independently, while the aggregate mass assigned to individual keys remains unconstrained. Doubly stochastic attention additionally imposes $A^\top \mathbf{1} = \mathbf{1}$, providing joint normalization over both the query and key dimensions (Sander et al., 2022; Guo et al., 2023; Shahbazi et al., 2025).

This structure is particularly well suited to bidirectional self attention. Queries and keys are two learned representations of the same set of n tokens, and every query may interact with every key. The resulting square attention operator therefore admits normalization over both dimensions of the same token set, rather than normalizing each query independently. In contrast, causal attention imposes an asymmetric support constraint, while cross attention may operate between distinct query and key sets. We therefore focus on bidirectional self attention, where doubly stochastic normalization provides a natural structured alternative to standard softmax attention.

The remaining difficulty is computational. Existing mechanisms for doubly stochastic attention still depend on dense or otherwise expensive token interactions. Sinkhorn normalization (Sinkhorn, 1964) iteratively rescales a full pairwise interaction matrix and therefore retains quadratic complexity. ESPFORMER (Shahbazi et al., 2025) instead exploits the connection between optimal transport (Peyré & Cuturi, 2019) and doubly stochastic attention together with expected sliced transport constructions (Liu et al., 2025), but remains quadratic during training and superlinear at inference. Meanwhile, efficient attention mechanisms obtain linear or near linear complexity through kernel feature maps, projections, sparsity, or low rank structure (Katharopoulos et al., 2020; Wang et al., 2020; Choromanski et al., 2021; Shen et al., 2021; Chen et al., 2021; Xiong et al., 2021; Meng et al., 2025), but do not in general preserve two sided stochastic normalization.

This leaves a specific gap between structured normalization and efficient attention. To our knowledge, existing attention mechanisms have not simultaneously provided doubly stochastic structure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and linear complexity in sequence length. We therefore ask: *Can doubly stochastic attention be constructed for bidirectional Transformers without forming dense token to token interactions?*

We answer this question with LOTFORMER. Rather than constructing a direct coupling between n queries and n keys, LOTFORMER introduces a learnable pivot measure with support size $r \ll n$. It computes two entropic optimal transport couplings, one between queries and the pivot measure and another between keys and the same pivot measure. Because the two couplings share the pivot marginal, they can be composed into an attention operator with the desired query and key marginals.

This factorization changes the computational structure of doubly stochastic attention. The resulting $n \times n$ operator is doubly stochastic and has rank at most r, yet it never needs to be explicitly materialized. Instead, its two thin factors can be applied sequentially to the values, reducing the attention computation to $O(nr)$. The learnable pivot locations and masses are optimized jointly with the model, allowing the intermediate support to adapt to the geometry of the attention problem.

We evaluate whether this construction is useful beyond its structural guarantees in three complementary settings. First, we train LOTFORMER within multiple vision Transformer architectures to test whether doubly stochastic linear attention remains competitive when learned from scratch. Second, we study controlled replacement of softmax attention in pretrained bidirectional language models while modifying only the attention operator. Third, we evaluate operator and model level efficiency across sequence lengths. Together, these experiments test whether the proposed transport factorization provides a practical alternative to dense doubly stochastic attention.

Our main contributions are:

- We introduce LOTFORMER, a bidirectional attention mechanism that factorizes attention through a learnable pivot measure with support $r \ll n$. The resulting operator is doubly stochastic, has rank at most r, and can be applied to values in $O(nr)$ time without materializing an $n \times n$ attention matrix.
- We develop a controlled conversion protocol for pretrained bidirectional Transformers that replaces only the attention operator. Across models ranging from 1.5B to 7B parameters, LOTFORMER preserves substantially more of the original softmax behavior than competing efficient attention replacements.
- Across DeiT, PVT, and Swin architectures, we show that LOTFORMER matches or improves over strong efficient attention baselines, illustrating that doubly stochastic linear attention remains effective when trained from scratch across diverse vision backbones.
- We characterize practical efficiency at both the operator and model levels, showing linear scaling with sequence length and increasing end to end acceleration at longer contexts. LOTFORMER is also directly compatible with optimized Sinkhorn implementations such as FlashSinkhorn (Ye et al., 2026).

2 RELATED WORK

Efficient and linear attention. Many methods reduce the quadratic cost of softmax attention through structured sparsity (Aguilera-Martos et al., 2024; Beltagy et al., 2020; Kitaev et al., 2020; Zaheer et al., 2020), feature maps (Katharopoulos et al., 2020; Choromanski et al., 2021), low rank approximations (Wang et al., 2020; Xiong et al., 2021; Chen et al., 2021), or alternative parameterizations (Zhang et al., 2024; Meng et al., 2025). Recent vision methods improve linear attention through injectivity and local modeling (Han et al., 2024), differential visual tokens (Pu et al., 2025), or norm and direction decomposition (Meng et al., 2026). Some methods also replace pretrained attention, with Hedgehog (Zhang et al., 2024), for example, using attention distillation for conversion. LOTFORMER instead achieves linear complexity while retaining two sided stochastic normalization.

Doubly stochastic attention. Doubly stochastic attention normalizes both dimensions of the attention operator. SinkFormer (Sander et al., 2022) uses iterative Sinkhorn normalization, while related work studies two sided normalization for attention structure and robustness (Guo et al., 2023). ESPFORMER (Shahbazi et al., 2025) constructs doubly stochastic attention using expected sliced transport plans (Liu et al., 2025), and QDSFormer (Born et al., 2025) explores quantum and quantum inspired constructions. These methods remain difficult to scale to large models and long sequences because of dense interactions, iterative procedures, superlinear complexity, or evaluation primarily

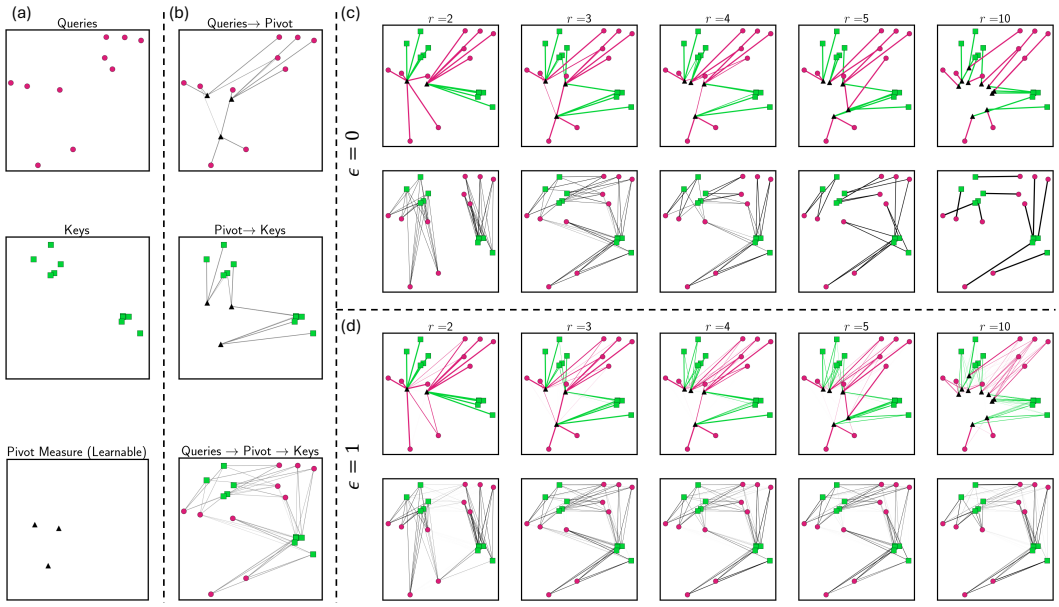

Figure 1: Illustration of LOTFormer. **(a)** Queries, keys, and a learnable pivot measure with $r \ll n$ support points. **(b)** Rather than constructing a full $n \times n$ query key coupling, LOTFormer computes query to pivot and pivot to key transport couplings and composes them through the shared pivot marginal. The resulting attention operator is low rank and doubly stochastic. **(c,d)** Effective query key couplings induced by different pivot sizes. Smaller r imposes a stronger low rank communication bottleneck, while increasing r allows progressively richer token interactions. Entropic regularization controls the smoothness of the resulting couplings.

at smaller scales. LOTFORMER instead enforces marginals through two thin transport couplings with linear sequence complexity.

Low rank and efficient optimal transport. Low rank optimal transport represents transport plans through compact factorizations. Low Rank Sinkhorn Factorization (Scetbon et al., 2021; Scetbon & Cuturi, 2022) composes two subcouplings through a shared intermediate marginal. LOTFORMER adapts this principle to attention by connecting queries and keys through a learnable pivot measure and two thin entropic transport couplings whose composition preserves the endpoint marginals. Unlike low rank OT methods that optimize over a constrained coupling family, LOTFORMER directly parameterizes the attention operator. Efficient implementations such as FlashSinkhorn (Ye et al., 2026) are complementary, while LOTFORMER reduces dense $n \times n$ interactions to two $n \times r$ interactions.

3 METHOD

We first review the optimal transport normalization used by LOTFormer and then derive its factorized attention operator.

Entropic optimal transport. For probability vectors $a \in \Delta^{n-1}$ and $b \in \Delta^{m-1}$, define

$$U(a,b) = \{\Gamma \in \mathbb{R}_+^{n\times m} : \Gamma \mathbf{1} = a,\ \Gamma^\top \mathbf{1} = b\}.$$

Given similarities $S \in \mathbb{R}^{n\times m}$, entropic optimal transport computes

$$\Gamma^\star = \arg\max_{\Gamma \in U(a,b)} \{\langle S, \Gamma\rangle + \varepsilon H(\Gamma)\}, \qquad H(\Gamma) = -\sum_{i,j} \Gamma_{ij}(\log \Gamma_{ij} - 1), \tag{1}$$

where $\varepsilon > 0$. For strictly positive marginals, the entropy term yields a unique coupling that can be computed by Sinkhorn scaling. Unlike row softmax, both marginals are prescribed.

Setup. Let $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times d_{\text{in}}}$. A single attention head forms

$$Q = XW_Q \in \mathbb{R}^{n\times d_k}, \qquad K = XW_K \in \mathbb{R}^{n\times d_k}, \qquad V = XW_V \in \mathbb{R}^{n\times d_v}.$$

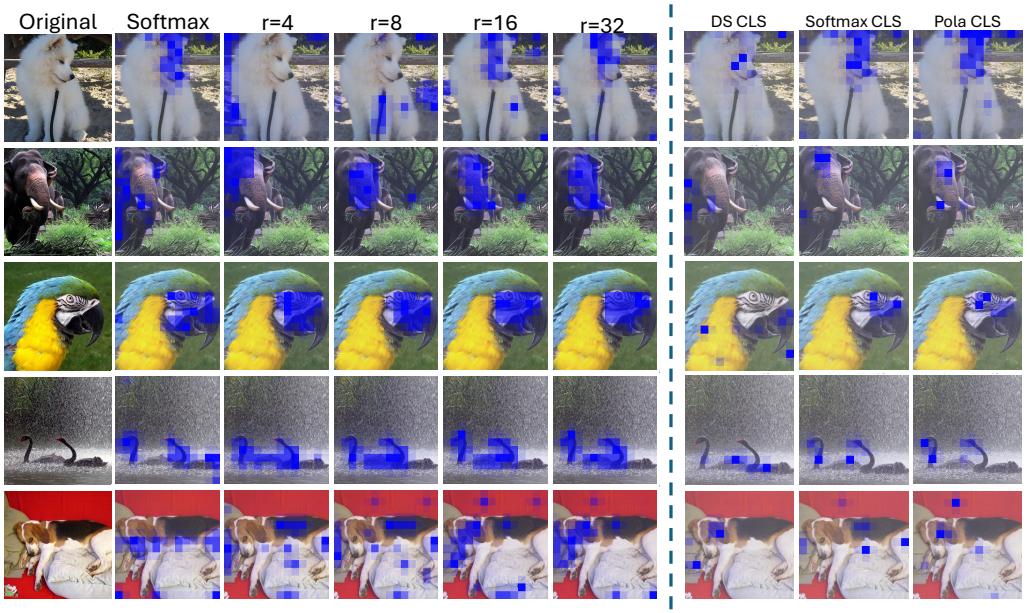

Figure 2: Patch level visualizations of [CLS] attention. **Left:** Softmax and LOTFormer with $r \in \{4, 8, 16, 32\}$ pivots. Increasing r provides a richer interaction structure and produces progressively sharper attention maps. **Right:** Different [CLS] treatments without DWC. Applying the doubly stochastic constraint to [CLS] limits its role as a global aggregator, while a separate softmax [CLS] row restores broad aggregation; polarization further sharpens its selectivity. The doubly stochastic guarantee applies to the patch to patch attention block when [CLS] is treated separately.

We associate queries and keys with the empirical measures

$$\mu = \sum_{i=1}^{n} p_i^1 \delta_{q_i}, \qquad \nu = \sum_{j=1}^{n} p_j^2 \delta_{k_j}, \qquad p^1, p^2 \in \Delta^{n-1},$$

and use $p^1 = p^2 = \frac{1}{n}\mathbf{1}$ unless stated otherwise.

LOTFormer introduces a learnable pivot measure

$$\sigma = \sum_{t=1}^{r} p_t^0 \delta_{z_t}, \qquad Z = [z_1, \ldots, z_r]^\top \in \mathbb{R}^{r \times d_k}, \qquad p^0 \in \Delta^{r-1}, \tag{2}$$

where $p_t^0 > 0$ and $r \ll n$. We reserve σ for the pivot measure and p^0 for its mass vector.

Transport through the pivot. Rather than constructing an $n \times n$ query key coupling, LOTFormer computes two thin transport plans using

$$S^Q = \frac{QZ^\top}{\sqrt{d_k}} \in \mathbb{R}^{n \times r}, \qquad S^K = \frac{ZK^\top}{\sqrt{d_k}} \in \mathbb{R}^{r \times n}.$$

The corresponding entropic couplings are

$$\Gamma^Q = \arg\max_{\Gamma \in U(p^1, p^0)} \left\{ \langle S^Q, \Gamma \rangle + \varepsilon H(\Gamma) \right\}, \tag{3}$$

$$\Gamma^K = \arg\max_{\Gamma \in U(p^0, p^2)} \left\{ \langle S^K, \Gamma \rangle + \varepsilon H(\Gamma) \right\}, \tag{4}$$

with $\Gamma^Q \in \mathbb{R}^{n \times r}$ and $\Gamma^K \in \mathbb{R}^{r \times n}$. By construction,

$$\Gamma^Q \mathbf{1} = p^1, \quad (\Gamma^Q)^\top \mathbf{1} = p^0, \quad \Gamma^K \mathbf{1} = p^0, \quad (\Gamma^K)^\top \mathbf{1} = p^2.$$

Glued coupling. Because the two plans share the pivot marginal p^0, they compose into

$$\Gamma = \Gamma^Q \operatorname{Diag}(p^0)^{-1} \Gamma^K \in \mathbb{R}^{n \times n}. \tag{5}$$

The endpoint marginals are preserved:

$$\Gamma \mathbf{1} = \Gamma^Q \operatorname{Diag}(p^0)^{-1} p^0 = p^1, \qquad \Gamma^\top \mathbf{1} = (\Gamma^K)^\top \operatorname{Diag}(p^0)^{-1} p^0 = p^2.$$

Under uniform token masses, $\Gamma \mathbf{1} = \Gamma^\top \mathbf{1} = \frac{1}{n}\mathbf{1}$. The corresponding attention matrix is therefore

$$A = n\Gamma = n\Gamma^Q \operatorname{Diag}(p^0)^{-1} \Gamma^K, \qquad A\mathbf{1} = A^\top \mathbf{1} = \mathbf{1}. \tag{6}$$

Thus, A is doubly stochastic. The scaling by n converts the probability coupling Γ into an attention operator with unit row and column sums.

LOTFormer attention. The attention output is evaluated directly through the two thin factors:

$$\mathrm{LOTAttn}(Q, K, V) = n\Gamma^Q \operatorname{Diag}(p^0)^{-1} (\Gamma^K V). \tag{7}$$

Associative evaluation avoids materializing the full $n \times n$ matrix A. Since $\Gamma^Q \in \mathbb{R}^{n \times r}$ and $\Gamma^K \in \mathbb{R}^{r \times n}$, $\operatorname{rank}(A) \le r$.

Complexity. Forming $QZ^\top$ and $ZK^\top$ costs $O(nrd_k)$, and T Sinkhorn iterations over the two thin couplings cost $O(Tnr)$. Applying the factorized operator to V costs $O(nrd_v)$. The total per head complexity is therefore

$$O(nr(d_k + d_v) + Tnr)\,, \tag{8}$$

which is linear in n for fixed r, T, d_k, and d_v. For $d_k = d_v = d$, this simplifies to $O(nrd + Tnr)$.

Learning. The pivot locations Z, masses p^0, and projections (W_Q, W_K, W_V) are learned end to end through (3) and (4). We parameterize p^0 with a softmax over learned logits, ensuring $p^0_t > 0$, and use separate pivots per attention head. Unless stated otherwise, both couplings share ε and the number of Sinkhorn iterations T.

Scope. LOTFormer defines a family of low rank, doubly stochastic attention operators rather than approximating the unrestricted quadratic entropic OT coupling. We evaluate it as a plug and play replacement in pretrained bidirectional models and as a native attention mechanism in vision Transformers. The formulation targets bidirectional attention, since doubly stochasticity under a standard triangular causal mask reduces to the identity; see Appendix F.

Relation to low rank optimal transport. LOTFormer differs from optimal low rank transport (Scetbon & Cuturi, 2022; Scetbon et al., 2021), which optimizes a transport objective over couplings of constrained nonnegative rank and generally depends on pairwise costs between the input measures. LOTFormer instead learns a shared pivot measure and computes only $QZ^\top$ and $ZK^\top$, avoiding the full $n \times n$ query key interaction matrix. The resulting $n \times r$ couplings compose into a doubly stochastic attention operator that can be applied in linear time for fixed r.

The pivot measure also admits a soft clustering interpretation. Each pivot acts as a latent representative inducing soft assignments over queries and keys, with attention mediated through the shared pivot. These assignments and pivot locations are learned jointly with the model; see Appendix C.

4 Experiments

We organize our evaluation around four questions. First, can a doubly stochastic linear attention operator replace softmax attention in pretrained bidirectional Transformers while preserving downstream quality (Section 4.1)? Second, does the same structure remain effective when trained natively in vision Transformers (Section 4.2)? Third, does its linear complexity translate into practical end to end gains as sequence length grows (Section 4.3)? Finally, we study the roles of pivot size, Sinkhorn normalization, pivot parameterization, and amortization (Section 4.4). Unless stated otherwise, we use rank $r = 32$, $T = 5$ Sinkhorn iterations, and entropy regularization $\varepsilon = 0.1$.

4.1 Replacing Pretrained Bidirectional Attention

A central goal of LOTFormer is to make doubly stochastic attention usable in existing pretrained bidirectional models without retraining the backbone from scratch. We therefore replace only the pretrained attention operator and apply the same conversion and adaptation budget to each efficient attention method. We first evaluate modern long context embedding models and then use BERT (Devlin et al., 2019) as an additional encoder scale validation. The approximation analysis in Appendix E further shows that LOTFormer remains relatively close to the pretrained softmax operator, helping explain why this conversion and distillation procedure can adapt the model effectively without retraining the backbone from scratch.

Table 1: Long context retrieval with gte Qwen2 7B. Scores are nDCG@10. Retention is relative to the original softmax model. Tasks: SummScreenFD (Chen et al., 2022; Shaham et al., 2022), QMSum (Zhong et al., 2021), 2WikiMQA (Ho et al., 2020), and NarrativeQA (Kočiský et al., 2018).

Method	SummScreenFD	QMSum	2WikiMQA	NarrativeQA	Avg.	Ret.
Softmax	94.2	68.5	71.0	52.3	71.5	100.0%
Hedgehog	90.1	64.2	66.8	48.1	67.3	94.1%
LOTFormer	**93.6**	**67.8**	**70.1**	**51.4**	$\mathbf{70.7}_{\pm 0.2}$	$\mathbf{98.9\%}_{\pm 0.1}$

Table 2: Retrieval quality through the attention conversion pipeline on gte Qwen2 7B.

Stage	Hedgehog	LOTFormer
Zero shot attention swap	41.0	**58.7**
Attention distillation	63.9	**68.2**
Contrastive LoRA fine tuning	67.3	**70.7**

Table 3: Softmax retrieval quality retained after conversion across pretrained embedding backbones.

Method	gte Qwen2 1.5B	gte Qwen2 7B	NV Embed v2	GritLM 7B
Hedgehog	93.2%	94.1%	93.8%	93.9%
LOTFormer	**98.3%**	**98.9%**	**98.6%**	**98.7%**

Table 4: Conversion of fine tuned BERT base on GLUE. Only bidirectional encoder attention is replaced. Recovery is relative to the original fine tuned model. T2R and T2R-HH baselines follow the protocol of Zhang et al. (2024).

Method	CoLA	SST2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	Rec.
BERT_{FT}	58.8	93.2	90.2	88.8	91.0	84.7	91.3	68.2	100.0
T2R	43.6	87.7	83.0	78.6	86.7	78.9	84.6	54.1	88.9
T2R-HH	56.9	90.9	89.1	77.7	90.0	77.4	84.5	56.3	93.5
Hedgehog	59.2	92.6	90.1	87.4	91.0	82.6	89.6	69.3	99.3
LOTFormer	59.1	92.5	90.3	89.7	90.5	83.0	90.2	68.5	$\mathbf{99.7}_{\pm 0.12}$

Long context retrieval. We evaluate pretrained embedding models spanning 1.5B to 7B parameters on LongEmbed (Zhu et al., 2024) retrieval tasks using nDCG@10. Table 1 reports detailed results for gte Qwen2 7B (Li et al., 2023). LOTFormer reaches an average nDCG@10 of 70.7, compared with 71.5 for the original softmax model and 67.3 for Hedgehog (Zhang et al., 2024), retaining 98.9% of the original retrieval quality while replacing quadratic attention with an $O(nr)$ operator.

Recovery through conversion. To understand this retention, we evaluate each model immediately after the attention replacement and after each adaptation stage. Table 2 shows that LOTFormer is substantially more compatible with the pretrained representation before adaptation, reaching 58.7 nDCG@10 compared with 41.0 for Hedgehog. Attention distillation raises LOTFormer to 68.2, while contrastive LoRA (Hu et al., 2022) fine tuning reaches 70.7.

Scaling across pretrained backbones. We repeat the same conversion protocol across four embedding models, gte Qwen2 1.5B and 7B (Li et al., 2023), NV Embed v2 (Lee et al., 2025), and GritLM 7B (Muennighoff et al., 2025), and report the fraction of each model's original softmax retrieval quality retained after conversion. As shown in Table 3, LOTFormer consistently retains 98.3–98.9%, compared with 93.2–94.1% for Hedgehog. The result therefore persists across both model scale and backbone.

BERT encoder conversion. We additionally evaluate task fine tuned BERT base. Following the conversion protocol used by Hedgehog (Zhang et al., 2024) and T2R (Kasai et al., 2021), we replace the encoder self attention, perform attention distillation, and fine tune on the original GLUE (Wang et al., 2019) task. LOTFormer recovers 99.7% of the fine tuned BERT baseline, compared with 99.3% for Hedgehog, indicating that the conversion behavior is not specific to long context embedding models.

Table 5: ImageNet 1K at 224^2 across four vision backbones. We report parameter count, FLOPs, and Top 1 accuracy.

Method	Input	Params	FLOPs	Top 1
DeiT Tiny				
Softmax(Touvron et al., 2021)	224^2	5.7M	1.1G	72.2
EfficientAttn(Shen et al., 2021)	224^2	5.7M	1.1G	70.2
FLattenAttn(Han et al., 2023)	224^2	6.1M	1.1G	74.1
MobiAttn(Yao et al., 2024)	224^2	5.7M	1.2G	73.3
PolaFormer(Meng et al., 2025)	224^2	6.1M	1.2G	74.6
Sinkformer(Sander et al., 2022)	224^2	5.7M	1.2G	70.2
ESPFormer(Shahbazi et al., 2025)	224^2	5.7M	1.2G	70.1
LOTFormer	224^2	5.8M	1.2G	$\mathbf{74.8}_{\pm 0.08}$
DeiT Small				
Softmax(Touvron et al., 2021)	224^2	22.0M	4.6G	79.8
PolaFormer(Meng et al., 2025)	224^2	23.0M	4.8G	79.8
LOTFormer	224^2	22.3M	4.8G	$\mathbf{80.1}_{\pm 0.10}$
PVT Tiny				
Softmax(Wang et al., 2021)	224^2	13.0M	1.9G	75.1
PolaFormer(Meng et al., 2025)	224^2	12.0M	2.0G	78.8
LOTFormer	224^2	12.2M	2.1G	$\mathbf{79.6}_{\pm 0.3}$
Swin Tiny				
Softmax(Liu et al., 2021)	224^2	28.0M	4.4G	81.2
FLattenAttn(Han et al., 2023)	224^2	29.0M	4.5G	82.1
PolaFormer(Meng et al., 2025)	224^2	29.0M	4.5G	**82.6**
LOTFormer	224^2	28.0M	4.5G	81.8

4.2 Vision Transformers

We next test whether the same doubly stochastic linear attention structure remains effective when trained natively. We evaluate ImageNet-1K at 224^2 resolution using DeiT Tiny and Small (Touvron et al., 2021), PVT Tiny (Wang et al., 2021), and Swin Tiny (Liu et al., 2021). Table 5 compares LOTFormer with softmax and representative efficient attention methods. LOTFormer reaches 74.8, 80.1, 79.6, and 81.8 Top 1 accuracy across the four backbones, respectively.

For architectures containing a `[CLS]` token, we apply the doubly stochastic operator to patch interactions while retaining a separate softmax aggregator for `[CLS]`, whose role is to collect information globally. We study this choice together with optional `[CLS]` polarization and depthwise convolution in Section 4.4.

Limited supervision. We further test whether the structured LOTFormer parameterization remains effective when supervision is limited. Using DeiT Tiny on ImageNet-1K, we train Softmax, PolaFormer, and LOTFormer with different fractions of the training set while keeping the optimization protocol fixed. Table 6 shows that LOTFormer performs best across all four regimes. The largest gain over Softmax occurs at 1% data, where Top 1 accuracy improves from 22.3 to 26.8, while at 25% data it improves from 63.5 to 66.8. Because LOTFormer changes both the normalization structure and pivot factorization, these results support the overall structured parameterization rather than isolating the causal effect of doubly stochasticity.

4.3 Efficiency

The $O(nr)$ complexity of LOTFormer applies to the attention operator, so we also evaluate end to end retrieval with the same gte Qwen2 7B model used above. Table 7 reports retrieval quality and throughput across context lengths, while complete peak memory results are provided in Appendix D.5. LOTFormer retains 98.9% of the original Softmax retrieval quality, compared with 94.1% for Hedgehog. Its throughput advantage over Softmax increases with context length, from $1.2\times$ at 8k tokens to $3.4\times$ at 64k and $7.5\times$ at 128k.

Table 6: Generalization under limited ImageNet-1K supervision using DeiT Tiny at 224^2. All methods use the same training protocol.

Data fraction	Softmax	PolaFormer	LOTFormer
1%	22.3	19.8	**26.8**
10%	52.6	48.7	**53.5**
25%	63.5	58.2	**66.8**
100%	72.2	74.6	**74.8**

Table 7: End to end retrieval quality and throughput on gte Qwen2 7B. Left: retrieval quality relative to the original Softmax model. Right: throughput across context lengths, with speedup measured for LOTFORMER relative to Softmax. Peak memory results are reported in Appendix D.5.

Retrieval quality			**Throughput scaling**				
Method	nDCG	Retained	Context	Softmax	Hedgehog	LOTFormer	Speedup
Softmax	71.5	100.0%	8k	1.05	**1.31**	1.26	1.2×
Hedgehog	67.3	94.1%	16k	0.46	**0.72**	0.64	1.4×
LOTFormer	**70.7**	**98.9%**	32k	0.17	**0.34**	0.31	1.8×
			64k	0.045	**0.18**	0.15	3.4×
			128k	0.010	0.045	**0.075**	7.5×

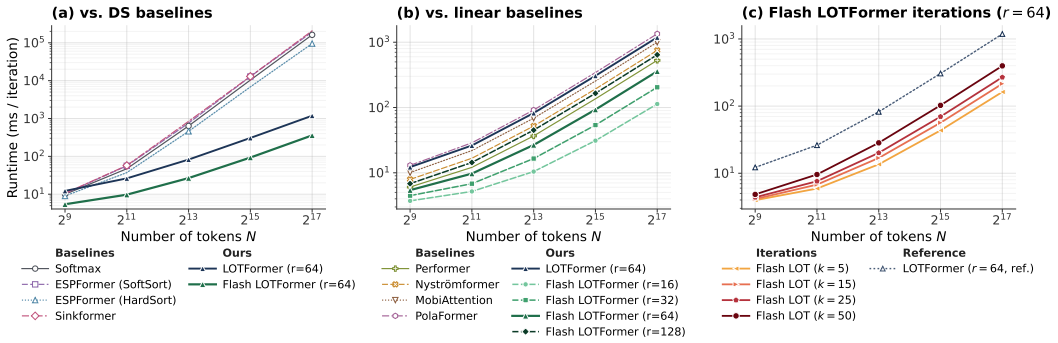

Figure 3: Operator runtime scaling with sequence length. Left: comparison with doubly stochastic attention methods. Middle: comparison with linear attention methods. Right: effect of Sinkhorn iterations at fixed $r = 64$. Plain LOTFORMER denotes the unfused implementation, while FlashSinkhorn accelerated LOTFORMER is reported for $r \in \{16, 32, 64, 128\}$.

Operator scaling and FlashSinkhorn. We next isolate the attention operator over sequence lengths $N \in \{2^9, \ldots, 2^{17}\}$. Figure 3 compares LOTFORMER with doubly stochastic and linear attention baselines and studies the effects of pivot size and Sinkhorn iterations. The pivot factorization avoids the quadratic scaling of dense doubly stochastic attention, while r and T provide direct controls over computation. FlashSinkhorn further reduces the overhead of the Sinkhorn iterations, making the linear scaling of the factorized construction practical in implementation. Complete peak memory measurements for both the unfused and FlashSinkhorn implementations are reported in Appendix D.5.

4.4 ABLATION STUDIES

Sinkhorn normalization and pivot size. We ablate entropy regularization ε, Sinkhorn iterations T, and pivot parameterization on ImageNet-100, and vary pivot size r on ImageNet-1K with DeiT Tiny. Table 8 shows that performance peaks near $\varepsilon = 1$, saturates around $T = 10$, and consistently favors learned pivots. Increasing r improves accuracy through $r = 32$, after which additional computation provides no further gain. We use $T = 5$ Sinkhorn iterations by default: increasing to $T = 10$ improves ImageNet-100 accuracy by at most 0.6 points (Table 8(ii)), while each additional iteration adds $O(nr)$ cost to every attention layer (see Figure 3(c)).

`[CLS]` aggregation in vision. Because `[CLS]` serves as a global aggregator, imposing the same doubly stochastic constraint used for patch interactions can limit its ability to collect information. Table 9 shows that decoupling `[CLS]` with a softmax row improves LOTFormer from 68.2 to 73.2.

Table 8: LOTFormer ablations. Panels (i, ii) report ImageNet-100 Top 1 accuracy; panel (iii) reports ImageNet-1K Top 1 accuracy and DeiT Tiny runtime. F and L denote fixed and learned pivots, respectively.

(i) EOT ε grid ($T = 10$)						(ii) Sinkhorn iterations						(iii) Pivot size					
ε	0.01	0.10	1.00	10.0	100	Pivot	ε	$T = 1$	$T = 5$	$T = 10$	$T = 20$	r	4	8	16	32	64
F	79.1	80.8	81.4	81.2	79.5	F	0.1	79.6	80.4	80.8	80.9	Top 1	64.5	66.4	68.8	**74.8**	73.9
L	79.2	81.1	**82.1**	81.3	79.6	F	1.0	80.2	81.2	81.4	81.7	Fwd.	29.6	36.2	45.1	63.3	91.7
						L	0.1	79.8	80.7	81.1	81.1	Bwd.	160.5	165.3	183.4	227.8	244.6
						L	1.0	80.5	81.5	**82.1**	82.0						

Table 9: Side by side ablation on DeiT Tiny / ImageNet 1K. "Pola type" indicates where polarization is applied: None, All tokens, or `[CLS]` only. Δ is computed within each block relative to the previous row.

PolaFormer						FlattenAttention						LOTFormer (r=32)				
`[CLS]`	Softmax	Pola type	DWC	Top 1	Δ	`[CLS]`	Softmax	Pola type	DWC	Top 1	Δ	`[CLS]`	Softmax	Pola type	DWC	Top 1	Δ
	✓	All	✗	61.9	0.0		✓	None	✗	71.8	0.0		✗	None	✗	68.2	0.0
	✓	All	✓	74.6	+12.7		✓	None	✓	74.1	+2.3		✓	None	✗	73.2	+5.0
													✓	`[CLS]`	✗	73.6	+0.4
													✓	`[CLS]`	✓	**74.8**	+1.2

Table 10: TV distance from softmax attention on gte Qwen2 7B. Ideal low rank DS optimizes pivots separately for each attention matrix; LOTFormer amortizes this construction. The corresponding stepwise approximation path is visualized in Figure 4.

Construction	TV from Softmax
DS projection of Softmax	11.7%
Ideal low rank DS	13.8%
LOTFormer	16.5%
Low Rank OT	24.5%

Adding `[CLS]` polarization raises this to 73.6, while depthwise convolution reaches 74.8. Under the same DeiT Tiny and ImageNet 1K setting with DWC, LOTFormer also slightly improves over PolaFormer (74.6) and FlattenAttention (74.1).

Amortizing the low rank DS construction. LOTFormer amortizes a low rank doubly stochastic construction through pivots learned across the training distribution rather than approximating unrestricted quadratic OT. We optimize the pivot separately for each attention matrix on gte Qwen2 7B to obtain an ideal low rank DS reference. Table 10 shows TV distances of 11.7% for the DS projection of softmax, 13.8% for the ideal low rank DS construction, and 16.5% for LOTFormer relative to the original attention. Crucially, LOTFormer is only 8.3% TV from the per attention ideal, isolating the cost of amortization. This proximity also explains why conversion and distillation from pretrained softmax attention are effective: adaptation only needs to close the remaining approximation gap. Low Rank OT is included for context and lies 10.1% TV from LOTFormer, but optimizes a different transport objective.

5 Conclusion

We introduced LOTFormer, an attention mechanism that routes interactions through a learned low rank pivot measure and applies the resulting operator in $O(nr)$ time. Composing two entropic transport couplings avoids the dense $n \times n$ attention matrix while giving the ideal operator prescribed marginals over both queries and keys. In practice, finite Sinkhorn iterations approximate those marginals. Across pretrained bidirectional embedding models, LOTFormer preserves substantially more retrieval quality than the evaluated efficient replacement, and it remains competitive when trained natively across multiple vision backbones. Its throughput advantage over softmax grows with sequence length, reaching $7.5\times$ in a separate 128k token stress test. The construction is also compatible with optimized Sinkhorn implementations, which reduce its implementation overhead. Together, these results show that structured transport can provide efficient bidirectional attention with control over both token marginals, without requiring a quadratic attention matrix.

ETHICS STATEMENT

This work introduces an efficient attention mechanism that reduces computational and memory costs while maintaining strong empirical performance across vision and language tasks. This may facilitate deployment in resource-constrained settings such as edge devices and real-time systems. As with other large-scale deep learning methods, potential risks include misuse in surveillance or automated decision-making systems, as well as environmental costs associated with training and inference. The method is architectural and does not introduce new data sources or supervision signals.

REPRODUCIBILITY STATEMENT

Implementation details, hyperparameters, and evaluation protocols are provided in Appendix D, together with additional theoretical material and ablations. Anonymous source code is available through the repository linked in the abstract. We report the training and conversion settings used for the main experiments so that the reported results can be reproduced from the released implementation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and LaTeX formatting. All AI-assisted text was reviewed and edited by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ignacio Aguilera-Martos, Andrés Herrera-Poyatos, Julián Luengo, and Francisco Herrera. Local attention: Enhancing the transformer architecture for efficient time series forecasting. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650762.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Jannis Born, Filip Skogh, Kahn Rhrissorrakrai, Filippo Utro, Nico Wagner, and Aleksandros Sobczyk. Quantum doubly stochastic transformers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 70224–70254. Curran Associates, Inc., 2025. doi: 10.52202/085713-2361. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/658e8b85ff800877233d544b7d4f7069-Paper-Conference.pdf.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8602–8615. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.589.

Yifan Chen, Qi Zeng, Heng Ji, and Yun Yang. Skyformer: Remodel self-attention with Gaussian kernel and Nyström method. *Advances in Neural Information Processing Systems*, 34:2122–2135, 2021.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with Performers. In *International Conference on Learning Representations*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17557–17568, 2023.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. FLatten Transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5961–5971, 2023.

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 79221–79245. Curran Associates, Inc., 2024. doi: 10.52202/079017-2515. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/90800f46f84381b7891e1378ee850013-Paper-Conference.pdf.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. Finetuning pretrained transformers into RNNs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10630–10643, 2021. URL https://aclanthology.org/2021.emnlp-main.830.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. URL https://aclanthology.org/Q18-1023.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Xinran Liu, Rocío Díaz Martín, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P7O1Vt1BdU.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. PolaFormer: Polarity-aware linear attention for vision transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kN6MFmKUSK.

Weikang Meng, Yadan Luo, Liangyu Huo, Yingjian Li, Yaowei Wang, Xin Li, and Zheng Zhang. Norm×Direction: Restoring the missing query norm in vision linear attention. In *Forty-third International Conference on Machine Learning*, 2026. arXiv:2506.21137.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Yifan Pu, Jixuan Ying, Qixiu Li, Tianzhu Ye, Dongchen Han, Xiaochen Wang, Ziyi Wang, Xinyu Shao, Gao Huang, and Xiu Li. Linear differential vision transformer: Learning visual contrasts via pairwise differentials. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 61255–61276. Curran Associates, Inc., 2025. doi: 10.52202/085713-2046. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5820ad65b1c27411417ae8b59433e580-Paper-Conference.pdf.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. URL https://aclanthology.org/2021.tacl-1.4.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6802–6814, 2022.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank Sinkhorn factorization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9344–9354. PMLR, July 2021.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 12007–12021. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.823.

Ashkan Shahbazi, Elaheh Akbari, Darian Salehi, Xinran Liu, Navid NaderiAlizadeh, and Soheil Kolouri. ESPFormer: Doubly-stochastic attention with expected sliced transport plans. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54186–54202. PMLR, 2025.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3531–3539, 2021.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, 2021.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A Nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14138–14148, 2021.

Zhiyu Yao, Jian Wang, Haixu Wu, Jingdong Wang, and Mingsheng Long. Mobile attention: Mobile-friendly linear-attention for vision transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56914–56926. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/yao24c.html`.

Felix X.-F. Ye, Xingjie Li, An Yu, Ming-Ching Chang, Linsong Chu, and Davis Wertheimer. FlashSinkhorn: IO-aware entropic optimal transport on GPU. In *Forty-third International Conference on Machine Learning*, 2026. arXiv:2602.03067.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The Hedgehog & the Porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4g02l2N2Nx.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.472.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 802–816. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.47.

A Notation and Conventions

Basic notation. Vectors are column vectors unless stated otherwise. Matrices are denoted by uppercase Roman letters, e.g., $A \in \mathbb{R}^{m\times n}$. We write $\mathbf{1}_m$ for the all ones vector in $\mathbb{R}^m$, $\mathrm{Diag}(a)$ for the diagonal matrix with diagonal a, and $|A| := \sum_{ij} A_{ij}$ for the sum of all entries of A. For a function f and a measure μ, we use the shorthand

$$\langle f, \mu\rangle := \int f \, d\mu.$$

The probability simplex and its strictly positive version are

$$\Delta_m := \{p \in \mathbb{R}^m : \ p_i \geq 0, \ \mathbf{1}_m^\top p = 1\}, \qquad \Delta_m^+ := \{p \in \Delta_m : \ p_i > 0\}.$$

Attention as transport. Let $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times d_{\text{in}}}$ be the input token matrix. We define

$$Q = XW_Q^\top, \qquad K = XW_K^\top, \qquad V = XW_V^\top,$$

where $Q, K \in \mathbb{R}^{n\times d_k}$ and $V \in \mathbb{R}^{n\times d_v}$; their rows are denoted q_i, k_i, v_i. LOTFormer interprets queries and keys as empirical measures and returns

$$\mathrm{LOTAttn}(Q, K, V) = AV,$$

where A is the low rank OT transport matrix.

Discrete measures and couplings. We use

$$\mu = \sum_{i=1}^n p_i^1 \delta_{q_i}, \qquad \nu = \sum_{j=1}^m p_j^2 \delta_{k_j}, \qquad \sigma = \sum_{\ell=1}^r p_\ell^0 \delta_{z_\ell},$$

with $p^1 \in \Delta_n$, $p^2 \in \Delta_m$, $p^0 \in \Delta_r^+$, and $Z = [z_1, \ldots, z_r]^\top \in \mathbb{R}^{r\times d_k}$. The transport polytope between $a \in \Delta_n$ and $b \in \Delta_m$ is

$$U(a, b) := \{\Gamma \in \mathbb{R}_+^{n\times m} : \Gamma \mathbf{1}_m = a, \ \Gamma^\top \mathbf{1}_n = b\}.$$

The entropy is

$$H(\Gamma) := -\sum_{ij} \Gamma_{ij}(\log \Gamma_{ij} - 1).$$

Throughout this section we use the attention cost $c(x, y) = -x^\top y$. The entropic OT problem is

$$\mathrm{OT}_\varepsilon(\mu, \nu) := \min_{\Gamma \in U(p^1, p^2)} \sum_{ij} c(q_i, k_j)\Gamma_{ij} - \varepsilon H(\Gamma).$$

Low rank OT, Linear OT, and clustering. For a nonnegative matrix $M \in \mathbb{R}_+^{n\times m}$, its nonnegative rank is

$$\mathrm{rk}_+(M) := \min\Big\{q : \ M = \sum_{\ell=1}^q a_\ell b_\ell^\top, \ a_\ell \in \mathbb{R}_+^n, \ b_\ell \in \mathbb{R}_+^m\Big\}.$$

The rank constrained transport set is

$$U(p^1, p^2; r) := U(p^1, p^2) \cap \{\Gamma \in \mathbb{R}_+^{n\times m} : \mathrm{rk}_+(\Gamma) \leq r\}.$$

A coupling in this set can be written through a pivot marginal $p^0 \in \Delta_r^+$ as

$$\Gamma = (\Gamma^{(1)})^\top \mathrm{Diag}(1/p^0)\Gamma^{(2)}, \qquad \Gamma^{(1)} \in U(p^0, p^1), \quad \Gamma^{(2)} \in U(p^0, p^2).$$

The factors $\Gamma^{(1)}$ and $\Gamma^{(2)}$ induce soft clusters. In particular, $C_i^\mu := \Gamma^{(1)}(s_i, \cdot)$ is a sub probability measure dominated by μ, and

$$\mu = \sum_{i=1}^r C_i^\mu.$$

Thus, LOT can be interpreted as transporting through a learned soft clustering induced by the pivot measure.

B Relation Between Low Rank OT and Linear OT

We now connect LOTFormer to the low rank OT formulation of Scetbon et al. (2021); Scetbon & Cuturi (2022). In the discrete setting,

$$\mu = \sum_{i=1}^{n} p_i^1 \delta_{x_i}, \qquad \nu = \sum_{j=1}^{m} p_j^2 \delta_{y_j},$$

the entropic low rank OT objective is

$$\mathrm{LrOT}_r(\mu,\nu) := \min_{\Gamma \in U(p^1,p^2;r)} \sum_{ij} c(x_i,y_j)\Gamma_{ij} - \varepsilon H(\Gamma).$$

Using the nonnegative rank factorization, its feasible set can equivalently be written as

$$U(p^1,p^2;r) = \Big\{\Gamma = (\Gamma^{(1)})^\top \mathrm{Diag}(1/p^0)\Gamma^{(2)} : \Gamma^{(1)} \in U(p^0,p^1),\ \Gamma^{(2)} \in U(p^0,p^2),\ p^0 \in \Delta_r^+\Big\}.$$

Hence,

$$\mathrm{LrOT}_r(\mu,\nu) = \min_{\substack{p^0 \in \Delta_r^+ \\ \Gamma^{(1)} \in U(p^0,p^1),\ \Gamma^{(2)} \in U(p^0,p^2)}} \sum_{ij} c(x_i,y_j)\Gamma_{ij} - \varepsilon H(\Gamma), \qquad \Gamma = (\Gamma^{(1)})^\top \mathrm{Diag}(1/p^0)\Gamma^{(2)}.$$

LOT as a restricted low rank construction. For a fixed pivot measure

$$\sigma = \sum_{\ell=1}^{r} p_\ell^0 \delta_{z_\ell},$$

LOT does not optimize freely over all feasible low rank factors. Instead, it obtains the two factors by solving two entropic OT problems to the same pivot:

$$\Gamma_\sigma^{(1)} \in \operatorname*{arg\,min}_{\Gamma^{(1)} \in U(p^0,p^1)} \sum_{\ell i} c(z_\ell,x_i)\Gamma_{\ell i}^{(1)} - \varepsilon H(\Gamma^{(1)}),$$

$$\Gamma_\sigma^{(2)} \in \operatorname*{arg\,min}_{\Gamma^{(2)} \in U(p^0,p^2)} \sum_{\ell j} c(z_\ell,y_j)\Gamma_{\ell j}^{(2)} - \varepsilon H(\Gamma^{(2)}).$$

These two couplings are glued into

$$\Gamma_\sigma = (\Gamma_\sigma^{(1)})^\top \mathrm{Diag}(1/p^0)\Gamma_\sigma^{(2)}.$$

The resulting LOT objective is

$$\mathrm{LOT}(\mu,\nu;\sigma) := \sum_{ij} c(x_i,y_j)(\Gamma_\sigma)_{ij} - \varepsilon H(\Gamma_\sigma), \qquad \mathrm{LOT}_r(\mu,\nu) := \inf_{\sigma \in \mathcal{P}_r(\mathbb{R}^d)} \mathrm{LOT}(\mu,\nu;\sigma).$$

Lemma B.1. *In the finite discrete setting, for any $\varepsilon \geq 0$,*

$$\mathrm{LrOT}_r(\mu,\nu) \leq \mathrm{LOT}_r(\mu,\nu).$$

Proof. Fix any pivot measure σ. The LOT plan

$$\Gamma_\sigma = (\Gamma_\sigma^{(1)})^\top \mathrm{Diag}(1/p^0)\Gamma_\sigma^{(2)}$$

belongs to $U(p^1,p^2;r)$ by the factorization above. Therefore, it is feasible for $\mathrm{LrOT}_r(\mu,\nu)$, which gives

$$\mathrm{LrOT}_r(\mu,\nu) \leq \sum_{ij} c(x_i,y_j)(\Gamma_\sigma)_{ij} - \varepsilon H(\Gamma_\sigma) = \mathrm{LOT}(\mu,\nu;\sigma).$$

Taking the infimum over all $\sigma \in \mathcal{P}_r(\mathbb{R}^d)$ yields $\mathrm{LrOT}_r(\mu,\nu) \leq \mathrm{LOT}_r(\mu,\nu)$. □

C LOTFORMER AND SOFT CLUSTERING.

LOTFormer effectively performs a soft clustering of queries and keys via the pivot measure, establishing correspondences between these clusters. Attention (message passing) is then mediated through the pivot.

Formally, given the optimal plan γ^1 for $OT(\sigma, \mu)$, for each pivot s_i, $i \in [1 : r]$, define the sub-probability measure

$$C_i^\mu \;=\; \gamma^1(s_i, \cdot) \;=\; \sum_{j=1}^{n} \gamma_{i,j}^1 \, \delta_{x_j}.$$

The collection $\{C_1^\mu, \ldots, C_r^\mu\}$ forms a soft clustering of queries. Similarly, from the optimal plan γ^2 for $OT(\sigma, \nu)$, we obtain

$$C_i^\nu \;=\; \gamma^2(s_i, \cdot) \;=\; \sum_{k=1}^{m} \gamma_{i,k}^2 \, \delta_{y_k},$$

yielding a soft clustering of keys. Crucially, C_i^μ and C_i^ν are coupled through their common pivot s_i.

In fact, the glued coupling can be written as a sum of outer products of the soft clusters. Specifically,

$$\Gamma \;=\; (\Gamma^{(1)})^\top \operatorname{Diag}(\sigma)^{-1} \, \Gamma^{(2)} \;=\; \sum_{i=1}^{r} \frac{1}{\sigma_i} \, C_i^\mu \otimes C_i^\nu,$$

where C_i^μ and C_i^ν are the soft clusters of queries and keys induced by the pivot s_i. This decomposition shows that attention is mediated through correspondences between query and key clusters.

The soft-clustering perspective connects our framework to a broad line of work that leverages clustering to define attention. For example, the *Routing Transformer* employs k-means clustering of queries for localized attention (Roy et al., 2021), while the *Reformer* uses LSH-based bucketing of queries and keys (Kitaev et al., 2020).

D IMPLEMENTATION DETAILS

General protocol. We evaluate LOTFORMER either by training from scratch or by converting a pretrained bidirectional model. In all settings, we modify only the bidirectional self attention operator and keep embeddings, FFNs, layer normalization, residual connections, and task specific heads unchanged.

LOTFormer defaults. Unless otherwise specified, we use $T = 5$ Sinkhorn iterations and a learnable pivot measure initialized uniformly. For architectures containing a `[CLS]` token, `[CLS]` aggregates through a standard softmax row, while the non `[CLS]` token block uses doubly stochastic attention as described in Sec. 4.2.

Depthwise convolution. When depthwise convolution is enabled, we use kernel size 3 and keep the remaining hyperparameters fixed.

D.1 FLASHSINKHORN INTEGRATION

FlashSinkhorn. For a kernel matrix $M \in \mathbb{R}^{n \times r}$, Sinkhorn scaling iterates

$$u \leftarrow p \oslash (Mv), \qquad v \leftarrow \sigma \oslash (M^\top u), \tag{9}$$

where p and σ denote the endpoint and pivot marginals, respectively, and $\oslash$ denotes elementwise division. A standard implementation repeatedly reads and writes the rectangular kernel during these iterations. FlashSinkhorn Ye et al. (2026) reduces this overhead using IO aware fused kernels that improve data reuse and reduce memory traffic across the Sinkhorn iterations.

Table 11: Training hyperparameters for ImageNet 1K classification.

Dataset	Optimizer	LR	Weight decay	Epochs	Warmup
ImageNet 1K	AdamW	3×10^{-4}	5×10^{-2}	400	10

Integration with LOTFormer. For each head, LOTFORMER constructs the two rectangular kernels

$$M^{(Q)} = \exp\big(QZ^\top/\varepsilon\big) \in \mathbb{R}^{n\times r}, \qquad M^{(K)} = \exp\big(KZ^\top/\varepsilon\big) \in \mathbb{R}^{n\times r}, \tag{10}$$

where $Z \in \mathbb{R}^{r\times d}$ denotes the learnable pivot support. FlashSinkhorn replaces the inner Sinkhorn computation for these two kernels and returns the scaling vectors after T iterations. The resulting couplings are

$$\Gamma^{(Q)} = \mathrm{Diag}(u^{(Q)})M^{(Q)}\mathrm{Diag}(v^{(Q)}), \qquad \Gamma^{(K)} = \mathrm{Diag}(u^{(K)})M^{(K)}\mathrm{Diag}(v^{(K)}). \tag{11}$$

The attention output is then computed as

$$O = \Gamma^{(Q)}\mathrm{Diag}(\sigma)^{-1}\big(\Gamma^{(K)}\big)^\top V, \tag{12}$$

as in the main formulation. FlashSinkhorn therefore changes only the implementation of the rectangular Sinkhorn computations and leaves the resulting LOTFORMER attention operator unchanged.

Runtime protocol. Operator scaling experiments use batch size 1 and head dimension $d_k = 64$. Reported runtimes are medians over 200 forward passes after 50 warmup iterations and are measured using CUDA events. We fix random seeds across all runs. Kernel construction uses `float32`, while the scaling vectors and value aggregation use `bfloat16`. Hardware for each efficiency experiment is reported with the corresponding experiment.

D.2 IMAGENET 1K

Backbones and resolution. We evaluate at 224^2 resolution using DeiT Tiny, DeiT Small, PVT Tiny, and Swin Tiny. For each backbone, we replace the original bidirectional self attention operator with LOTFORMER while keeping the remainder of the architecture unchanged.

Training recipe. We build on the official DeiT training framework and retain its standard augmentation and regularization recipe unless otherwise specified. ImageNet models are trained for 400 epochs using AdamW, with linear warmup for the first 10 epochs followed by cosine learning rate decay. The base learning rate is 3×10^{-4} for global batch size 1024, with weight decay 5×10^{-2}.

`[CLS]` handling. We use the dedicated `[CLS]` softmax aggregator described in Sec. 4.2. Table 8 reports controlled ablations of the `[CLS]` aggregator, polarization, and depthwise convolution under matched training settings.

Pivot size. For DeiT Tiny, we evaluate $r \in \{4, 8, 16, 32, 64\}$. The complete accuracy and efficiency ablation is reported in Table 9.

D.3 GLUE CONVERSION ON FINE TUNED BERT

Protocol and checkpoints. We follow the pretrained conversion protocol used by Hedgehog (Zhang et al., 2024). For each GLUE task, we start from a publicly available fine tuned BERT Base Uncased checkpoint and replace all bidirectional self attention blocks with the candidate attention mechanism.

Attention distillation. For methods requiring attention distillation, we optimize only the replacement attention modules for up to 5 epochs with early stopping, using AdamW with learning rate 10^{-2} and zero weight decay.

Table 12: Peak GPU memory across sequence lengths. Flash LOTFormer uses FlashSinkhorn. Overhead is measured relative to FlashAttention v3.

Seq. len	FA3 (GB)	PyTorch LOTFormer (GB)	Flash LOTFormer (GB)	PyTorch overhead	Flash overhead
4k	18.2	22.4	18.5	23.1%	1.6%
8k	20.8	26.1	21.1	25.5%	1.4%
16k	25.9	33.5	26.3	29.3%	1.5%
32k	36.2	48.3	36.8	33.4%	1.6%
64k	56.8	78.0	57.6	37.3%	1.4%
128k	98.0	137.3	99.5	40.1%	1.5%

Post conversion adaptation. After conversion, we adapt each model for up to 5 epochs with batch size 8, AdamW, learning rate 10^{-5}, zero weight decay, and cosine learning rate decay. We use the standard classification objective for classification tasks and the regression objective for STS–B.

D.4 Ablations

Sinkhorn and pivot measure. For the ViT ablations in Table 8, we use ImageNet 100 and a reduced ViT backbone with half the hidden dimension. We vary Sinkhorn iterations T, entropy regularization ε, and the pivot measure, comparing a fixed uniform measure with a learnable measure initialized uniformly.

Pivot size. For the DeiT Tiny pivot study, we use ImageNet 1K and evaluate $r \in \{4, 8, 16, 32, 64\}$. We report Top 1 accuracy together with forward and backward runtime under the same training configuration.

D.5 Peak GPU Memory

Measurement setup. We measure peak allocated GPU memory for FlashAttention v3, the unfused PyTorch implementation of LOTFormer, and LOTFormer accelerated with FlashSinkhorn under the same configuration. Overhead is measured relative to FlashAttention v3. These experiments are performed on H200. All methods use the same batch size, dtype, model dimensions, and sequence length for each comparison.

Takeaway. The unfused PyTorch implementation incurs 23.1% to 40.1% higher peak memory than FlashAttention v3 across the evaluated sequence lengths. With FlashSinkhorn, this overhead decreases to 1.4% to 1.6%. This indicates that much of the memory gap of the reference implementation comes from implementation overhead rather than the low rank factorization itself.

E Attention Error Decomposition

To identify the sources of approximation error in plug and play replacement, we decompose the path from pretrained softmax attention to LOTFormer. Starting from the original attention matrix, we first project onto the doubly stochastic set, then impose the low rank constraint, and finally replace the per matrix optimum with the amortized LOTFormer construction. We additionally compare with the corresponding low rank OT parameterization.

Figure 4 visualizes this sequence, while Table 10 reports total variation from the original softmax attention. A substantial portion of the discrepancy arises from the doubly stochastic and low rank constraints themselves, while the additional gap introduced by amortizing the low rank construction with LOTFormer is smaller.

F Bidirectional Scope and Causal Masking

Scope. LOTFormer enforces a doubly stochastic attention matrix

$$A\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top A = \mathbf{1}^\top, \qquad A \geq 0,$$

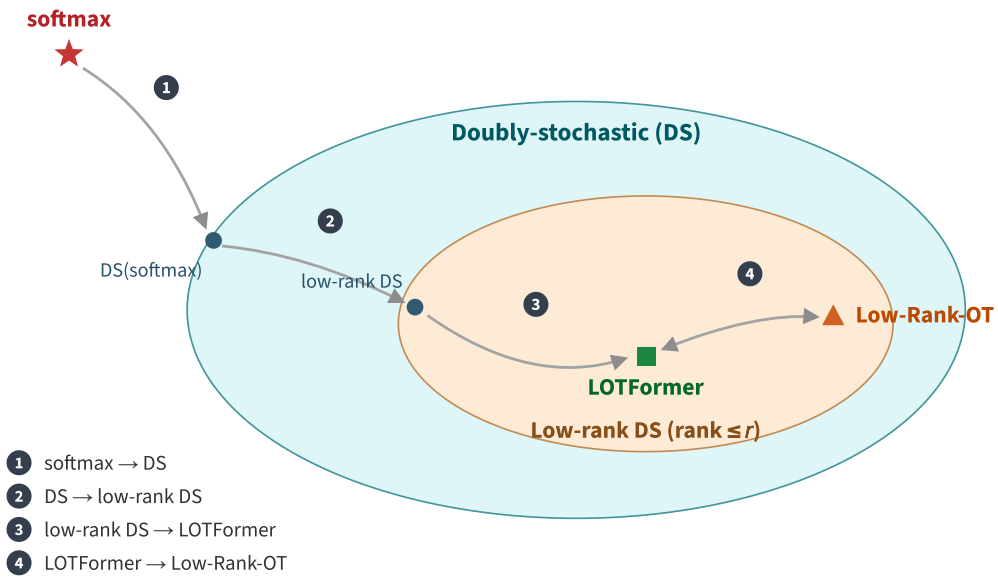

Figure 4: Attention error decomposition for plug and play replacement on gte Qwen2 7B. The path progressively introduces doubly stochastic structure, a low rank constraint, and the amortized LOTFORMER parameterization.

which is naturally suited to bidirectional self attention with full support.

Causal masking. Standard causal self attention restricts support to the lower triangular set,

$$A_{ij} = 0 \qquad \text{for } j > i.$$

Under this support constraint, doubly stochastic marginals force a unique degenerate solution. Column n can receive mass only from row n, so

$$\sum_{i=1}^{n} A_{in} = A_{nn}.$$

Since the column sum must equal 1, we obtain $A_{nn} = 1$. The row sum constraint then implies $A_{nj} = 0$ for every $j < n$. Repeating the argument for columns $n-1, n-2, \ldots, 1$ gives

$$A_{ii} = 1, \qquad A_{ij} = 0 \quad \text{for } i \neq j,$$

and therefore $A = I$.

Implication. LOTFORMER therefore targets encoder style bidirectional self attention. Extending doubly stochastic attention to autoregressive settings requires changing either the marginal constraints or the support structure, which we leave to future work.