EXPRESSIVITY AND LIMITATIONS OF PROJECTION TYING IN SELF-ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-attention typically uses separate query, key, and value projections, yet the effects of tying these projections on expressivity remain poorly understood. We provide a systematic theoretical comparison of query-key (Q=K) and key-value (K=V) tying. We first prove that, for fixed sequence length and hidden dimension, Transformers with positional encodings remain universal approximators even under full tying (Q=K=V), given sufficient depth. At the level of individual attention heads, however, the two tying schemes impose distinct limitations. Under Q=K, the Gram structure of unmasked attention logits precludes arbitrarily accurate approximation of cyclic-shift attention patterns, regardless of head dimension. With causal masking, Q=K can approximate any causally admissible attention pattern, but previous-token routing requires asymptotically larger projected-feature norms than in the untied case. In contrast, K=V preserves attention-pattern realizability while coupling addressing and content transmission through a shared representation. We prove that the minimum head dimension required for K=V to exactly reproduce a given untied head equals the dimension of the joint span of that heads effective key and value subspaces, and can possibly be twice as large as in the untied case. Controlled synthetic experiments corroborate these theoretical separations. Together, these results distinguish model-level universality from head-level expressivity and limitations.

1 INTRODUCTION

Self-attention (Vaswani et al., 2017), a core component of modern large language models, uses queries to specify what information to seek, keys to determine how tokens are addressed, and values to carry their content. Yet the role of independent projections remains unclear: when can they be tied without sacrificing expressive power, and what functional or resource costs arise?

Recent efforts toward architectural simplification include shared-weight self-attention, which uses a common projection with separate diagonal modulations (Kowsher et al., 2025), and Matrix Atom Sharing in Attention, which studies shared matrix dictionaries across layers and projection types (Zhussip et al., 2026). Meanwhile, multi-query and grouped-query attention share keys and values across query heads, reducing KV-cache storage and memory traffic (Shazeer, 2019; Ainslie et al., 2023). These studies primarily emphasize empirical performance and efficiency. Together with the distinct behavior observed across projection-tying schemes (Kayyam et al., 2026), they motivate a systematic theoretical understanding of the expressive consequences of sharing.

We study exact projection tying within individual attention heads, focusing on Q=K and K=V. Q=K imposes a Gram structure on the pre-mask logits, restricting directional routing in unmasked attention, whereas K=V preserves attention-pattern realizability but couples addressing and content. K=V also enables cache sharing within a head, complementary to GQA's sharing across heads. We leave Q=V outside the scope of this work, as it offers no analogous KV-cache savings: past queries are not cached during autoregressive decoding.

To the best of our knowledge, we provide the first systematic theoretical comparison of these tying schemes. For fixed sequence length and hidden dimension, we show that unmasked Transformers with positional embeddings remain universal approximators even under full tying (Q=K=V), given sufficient depth. Universality alone therefore does not distinguish these architectures. At the head

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.

level, however, distinct resource costs emerge: Q=K can require larger projected-feature norms under causal masking, whereas K=V can require greater head dimension when addressing and content occupy different feature subspaces. **Our contributions** are as follows.

- **Universality under projection tying.** We prove that Transformers remain universal approximators under Q=K=V, and hence also under Q=K and K=V. Thus, universal approximation does not characterize the functional role of independent projections.
- **Routing limitations of Q=K.** An untied head can realize arbitrary sparse attention patterns with head dimension controlled by local sparsity. In contrast, an unmasked Q=K head cannot realize certain simple directional patterns, regardless of head dimension. A causal mask removes this qualitative obstruction by breaking the symmetry, but a quantitative gap remains: Q=K can require asymptotically larger projected-feature norms than an untied head.
- **The address–content bottleneck of K=V.** Tying K=V does not fundamentally restrict the attainable attention patterns. Its cost instead arises in reproducing the full attention output: the minimum head dimension is exactly determined by the joint span of the effective addressing and content subspaces. Subspace overlap reduces the additional dimension required by tying and eliminates it precisely when one effective subspace contains the other. When the subspaces have equal dimension and trivial intersection, tying doubles the minimum required head dimension.

Together, these results clarify the distinct roles of the query, key, and value projections. Independent query and key projections enable flexible directional routing, while independent key and value projections allow addressing and content transmission to occupy distinct feature spaces. Universal approximation alone does not capture these benefits, which emerge through the geometry and resource requirements of individual attention heads. Our goal is to explain the value of independent projections and identify principles that may inform the design of more efficient architectures, rather than to advocate projection tying as a practical replacement for standard attention.

1.1 Related Work

Several lines of work have investigated the expressive power and limitations of Transformers, including (i) **approximation theory** (Yun et al., 2020; 2019; Kajitsuka & Sato, 2023; Takakura & Suzuki, 2023; Jiang & Li, 2024; Havrilla & Liao, 2024; Shen et al., 2025; Liu et al., 2025; Hu et al., 2025; Cheng et al., 2025; Jiao et al., 2025; 2026; He et al., 2026; Lai & Sun, 2026); (ii) **formal languages** (Hahn, 2020; Bhattamishra et al., 2020; Hao et al., 2022; Merrill et al., 2022; Chiang et al., 2023; Strobl et al., 2024); and (iii) **expressivity limitations** (Bhojanapalli et al., 2020; Likhosherstov et al., 2021; Sanford et al., 2023; Amsel et al., 2025; Yu et al., 2026).

While the QKV formulation has become standard, the necessity of using three separate projections remains an open question. Several works have studied this issue from both empirical and theoretical perspectives (Borji, 2023; Kowsher et al., 2025; Karbevski & Mijoski, 2025; Kayyam et al., 2026; Yao et al., 2026). However, a systematic theoretical characterization of the expressivity and limitations induced by different projection-tying schemes in self-attention is still lacking.

A more detailed discussion of related work is provided in Appendix B.

1.2 Preliminaries

Let $[n] := \{1, \dots, n\}$. We represent an input sequence of length L by $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$, where the i-th row is the hidden representation of the i-th token. To isolate the effect of tying the attention projection matrices, we omit layer normalization following the convention in existing theoretical work (Yun et al., 2019; Kajitsuka & Sato, 2023).

Multi-head self-attention. For any $\boldsymbol{S} \in (\mathbb{R} \cup \{-\infty\})^{L \times L}$ with at least one finite entry in each row, $\mathrm{Softmax}(\boldsymbol{S})$ is applied row-wise: $\big[\mathrm{Softmax}(\boldsymbol{S})\big]_{ij} := \frac{\exp(\boldsymbol{S}_{ij})}{\sum_{r=1}^{L} \exp(\boldsymbol{S}_{ir})}$, with $\exp(-\infty) := 0$. We consider either unmasked attention, for which $\boldsymbol{M}_{\mathrm{MASK}} = \boldsymbol{0}$, or causal attention, for which $(\boldsymbol{M}_{\mathrm{MASK}})_{ij} := \begin{cases} 0, & j \le i, \\ -\infty, & j > i. \end{cases}$ Consider H attention heads of dimension d. For each $h \in [H]$, let

$\boldsymbol{W}_Q^{(h)}, \boldsymbol{W}_K^{(h)}, \boldsymbol{W}_V^{(h)} \in \mathbb{R}^{d_{\mathrm{hid}} \times d}$ be the query, key, and value projection matrices, and define

$$\boldsymbol{A}^{(h)}(\boldsymbol{X}) := \mathrm{Softmax}\Big((\boldsymbol{X}\boldsymbol{W}_Q^{(h)})(\boldsymbol{X}\boldsymbol{W}_K^{(h)})^\top + \boldsymbol{M}_{\mathrm{mask}}\Big), \tag{1.1}$$

$$\mathrm{Head}^{(h)}(\boldsymbol{X}) := \boldsymbol{A}^{(h)}(\boldsymbol{X})\boldsymbol{X}\boldsymbol{W}_V^{(h)}. \tag{1.2}$$

We ignore the scaling factor $1/\sqrt{d}$ here as it can be absorbed into model parameters. Let $\boldsymbol{W}_O \in \mathbb{R}^{Hd \times d_{\mathrm{hid}}}$ be the output projection. The multi-head self-attention sublayer with a residual connection is

$$\mathrm{Attn}(\boldsymbol{X}) := \boldsymbol{X} + \mathrm{Concat}_{h=1}^{H}\big(\mathrm{Head}^{(h)}(\boldsymbol{X})\big)\boldsymbol{W}_O. \tag{1.3}$$

Position-wise feed-forward networks. Let W be the hidden width, and let $\boldsymbol{W}_1 \in \mathbb{R}^{W \times d_{\mathrm{hid}}}, \boldsymbol{W}_2 \in \mathbb{R}^{d_{\mathrm{hid}} \times W}, \boldsymbol{b}_1 \in \mathbb{R}^W, \boldsymbol{b}_2 \in \mathbb{R}^{d_{\mathrm{hid}}}$. The feed-forward sublayer is applied independently to each token and is given by

$$\mathrm{FFN}(\boldsymbol{X}) := \boldsymbol{X} + \sigma\big(\boldsymbol{X}\boldsymbol{W}_1^\top + \mathbf{1}\boldsymbol{b}_1^\top\big)\,\boldsymbol{W}_2^\top + \mathbf{1}\boldsymbol{b}_2^\top, \tag{1.4}$$

where $\mathbf{1} \in \mathbb{R}^L$ and $\sigma(t) := \max\{t, 0\}$ is applied entrywise. Equivalently, its action on a single token $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}}$ is $F(\boldsymbol{x}) := \boldsymbol{x} + \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2$. We use F for the vector map and $\boldsymbol{F}$ for its row-wise extension to matrices.

Transformer function classes. Let N be the number of Transformer blocks, and let $\boldsymbol{E} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be a fixed or learnable positional embedding matrix. It is added once, before the first Transformer block: $\mathrm{PE}_{\boldsymbol{E}}(\boldsymbol{X}) := \boldsymbol{X} + \boldsymbol{E}$. When $\boldsymbol{E}$ is learnable, it is included among the model parameters. For each $\ell \in [N]$, define $\mathrm{Block}^{(\ell)} := \mathrm{FFN}^{(\ell)} \circ \mathrm{Attn}^{(\ell)}$. Fix $L, d_{\mathrm{hid}}, N, H, d, W$, the choice of mask, and whether $\boldsymbol{E}$ is fixed or learnable. Let Θ denote the resulting parameter space. For $\theta \in \Theta$, define

$$T_\theta : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}, \qquad T_\theta := \mathrm{Block}_\theta^{(N)} \circ \cdots \circ \mathrm{Block}_\theta^{(1)} \circ \mathrm{PE}_{\boldsymbol{E}}, \qquad \mathcal{T} := \{T_\theta : \theta \in \Theta\}.$$

Finally, the tied Transformer classes are

$$\mathcal{T}_{Q=K} := \Big\{T_\theta : \theta \in \Theta,\ \boldsymbol{W}_Q^{(\ell,h)} = \boldsymbol{W}_K^{(\ell,h)} \text{ for all } (\ell,h) \in [N] \times [H]\Big\}, \tag{1.5}$$

$$\mathcal{T}_{K=V} := \Big\{T_\theta : \theta \in \Theta,\ \boldsymbol{W}_K^{(\ell,h)} = \boldsymbol{W}_V^{(\ell,h)} \text{ for all } (\ell,h) \in [N] \times [H]\Big\}, \tag{1.6}$$

$$\mathcal{T}_{Q=K=V} := \Big\{T_\theta : \theta \in \Theta,\ \boldsymbol{W}_Q^{(\ell,h)} = \boldsymbol{W}_K^{(\ell,h)} = \boldsymbol{W}_V^{(\ell,h)} \text{ for all } (\ell,h) \in [N] \times [H]\Big\}. \tag{1.7}$$

In particular, $\mathcal{T}_{Q=K=V} \subseteq \mathcal{T}_{Q=K} \cap \mathcal{T}_{K=V} \subseteq \mathcal{T}$.

2 Universal Approximation under Projection Tying

We show that tying the query, key, and value projections does not destroy the universal approximation property of Transformers. In fact, universality holds even under the strongest constraint Q=K=V, and hence also under Q=K and K=V.

Throughout this section, we consider unmasked self-attention with designed positional encodings. Positional information is necessary for approximating general sequence-to-sequence maps, since the unmasked architecture is permutation equivariant without it (Yun et al., 2019). Let $\mathcal{C}_{L,d_{\mathrm{hid}}} := \mathcal{C}\big([0,1]^{L \times d_{\mathrm{hid}}}; \mathbb{R}^{L \times d_{\mathrm{hid}}}\big)$, and, for a measurable map $G : [0,1]^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}$, define $\|G\|_{L^2} := \Big(\int_{[0,1]^{L \times d_{\mathrm{hid}}}} \|G(\boldsymbol{X})\|_F^2 \,\mathrm{d}\boldsymbol{X}\Big)^{1/2}$. We work in the fixed-width, variable-depth regime. For fixed numbers of heads H, head dimension d, and feed-forward width W, define

$$\mathfrak{T}_\star(H, d, W) := \bigcup_{N \geq 1} \mathcal{T}_\star(N, H, d, W), \qquad \star \in \{Q = K, K = V, Q = K = V\}.$$

Thus, only the number of blocks N may grow with the target function or approximation accuracy, consistent with the fixed-width design of practical Transformers. The unconstrained class $\mathfrak{T}(H, d, W)$ is defined analogously. Since $\mathfrak{T}_{Q=K=V} \subseteq \mathfrak{T}_{Q=K} \cap \mathfrak{T}_{K=V} \subseteq \mathfrak{T}$, it suffices to establish universality for the fully tied class.

Theorem 2.1 (Universality with tied projections)**.** *There exists a universal constant $C > 0$ such that, for every $L, d_{\mathrm{hid}} \in \mathbb{N}$, there are positive integers H, d, W satisfying $\max\{H, d, W\} \leq CLd_{\mathrm{hid}}$ for which the following holds: for every $F \in \mathcal{C}_{L,d_{\mathrm{hid}}}$ and every $\varepsilon > 0$, there exists $T \in \mathfrak{T}_{Q=K=V}(H, d, W)$ such that*

$$\|F - T\|_{L^2} < \varepsilon.$$

Consequently, the same conclusion holds for $\mathfrak{T}_{Q=K}(H, d, W)$ and $\mathfrak{T}_{K=V}(H, d, W)$.

The proof, deferred to Appendix E, adapts the three-stage construction: quantization, contextual mapping, and value mapping, used in existing Transformer approximation results (Yun et al., 2019; Kajitsuka & Sato, 2023; Jiao et al., 2025; Cheng et al., 2025). A position-wise feed-forward network first quantizes the input; self-attention then assigns distinct context-dependent representations to different token–context pairs; and a final feed-forward network maps these representations to the prescribed outputs. Our novel observation, formalized in Lemma E.2, is that the contextual-mapping step can already be implemented with $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$. The construction exploits the main techniques in (Kajitsuka & Sato, 2023). Thus, although tying all three projections couples information routing with content transmission within each head, it does not preclude universal approximation. Appendix D summarizes the proof-level novelty. This result is consistent with the empirical observation that tied projections can perform well on some static-length tasks (Kayyam et al., 2026), such as image classification, set processing, and provides a representational explanation for why tying need not be fatal in principle .

Parameter efficiency. Under the Hölder regularity assumption stated in Appendix E.5, our fully tied construction achieves a parameter-count upper bound of $\mathcal{O}(\varepsilon^{-d_0/\alpha})$ for fixed sequence length L and hidden dimension d_{hid}, where $d_0 = Ld_{\mathrm{hid}}$, $\alpha \in (0, 1]$ represents the Hölder smooth. This matches the upper bound established by He et al. (2026). A sharp comparison between the tied and untied Transformer classes requires matching upper and lower bounds on their parameter complexity, accounting for both the number of parameters and the precision with which they are represented. Establishing such bounds is left for future work.

3 Limitations of Tied Q and K

The universality result in the previous section does not distinguish among QKV, Q=K, and K=V, and thus does not isolate the effect of tying projection matrices. Rather than serving as a genuine separation result, it primarily provides a baseline. We therefore study the attention patterns realizable by a single head, isolating information routing from the subsequent value and feed-forward transformations.

We focus on Q=K. Queries specify what information tokens seek, whereas keys determine how tokens can be addressed. Tying them couples these distinct roles and constrains the pre-mask logits to the Gram form $\boldsymbol{Q}\boldsymbol{Q}^\top$, potentially restricting directional routing (Kayyam et al., 2026). Moreover, learned attention patterns often exhibit concentrated structures associated with syntax, coreference, and induction-like pattern matching (Clark et al., 2019; Voita et al., 2019; Olsson et al., 2022). Although not complete explanations of model predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), such patterns reflect how information is selectively retrieved and routed, especially in long-context tasks (Child et al., 2019; Zaheer et al., 2020; Jiang et al., 2024; Lai et al., 2025; Vasylenko et al., 2026; Ram et al., 2026) . Sparse stochastic matrices, which encode sharp directional dependencies on a small set of information sources, thus provide a natural benchmark for comparing tied and untied attention heads. We begin by specifying the target class of attention patterns.

Definition 3.1 ((k, γ)-sparse attention patterns)**.** Let $1 \leq k \leq L$ and $\gamma \geq 1$. A matrix $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ is called a (k, γ)-sparse attention pattern if (i) it is row-stochastic, that is, $A_{ij} \geq 0, \sum_{j=1}^{L} A_{ij} = 1, i \in [L]$, (ii) has at most k nonzero entries in each row and column, (iii) and satisfies $\gamma^{-1} \leq \frac{A_{i,j_1}}{A_{i,j_2}} \leq \gamma$ for every $i, j_1, j_2 \in [L]$ such that $A_{i,j_1}, A_{i,j_2} > 0$.

The sparsity parameter k controls both the number of interactions available to each token and the number of tokens that can route information to a common location. The parameter γ controls the dynamic range of the nonzero attention weights within each row, thereby providing explicit control over the projected-feature norms. This definition captures a broad class of practically relevant attention patterns while allowing our bounds to depend on the local sparsity k, rather than directly

on the sequence length L. In a setting closely related to ours, Likhosherstov et al. (2021); Sanford et al. (2023) study the expressivity of untied self-attention by examining its ability to approximate arbitrary sparse attention patterns.

3.1 The Ability to approximate Sparse Attention Patterns

We first consider an unmasked attention head with independent query and key projections. The value projection does not affect the attention matrix, so the relevant distinction at this stage is between an unconstrained pair $(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ and the tied choice $\boldsymbol{W}_Q = \boldsymbol{W}_K$.

The following theorem shows that untied projections provide a uniform realization mechanism for sparse attention patterns. Importantly, the projection matrices are fixed across all target patterns; only the input representation $\boldsymbol{X}$ is allowed to depend on $\boldsymbol{A}$.

Theorem 3.1. *Fix $L \geq 2$ and $1 \leq k \leq L$, and set $d := \min\{2k, L-1\}, d_{\mathrm{hid}} := 2d$. There exist fixed matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}} \times d}$, depending only on k and L, such that the following holds. For every $\gamma \geq 1$, every $\varepsilon > 0$, and every (k, γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L \times L}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ that satisfies*

$$\| \operatorname{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\big) - \boldsymbol{A} \|_2 < \varepsilon.$$

Theorem 3.1 gives a head dimension of at most $2k$, independent of L whenever $2k < L$. Thus, an untied attention head can realize sparse routing rules using a dimension governed by the local complexity of the target pattern. The result is stronger than a target-dependent factorization: the same pair $(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ works for the entire class of (k, γ)-sparse patterns.

At the level of attention-pattern realizability, K=V intuitively does not restrict the attention scores, because the value projection does not enter their computation. Appendix F.2 establishes a stronger fixed-projection statement: with sufficient hidden dimension, the key projection may be prescribed arbitrarily subject to full column rank, while a suitable fixed query projection and target-dependent input representations realize the same sparse routing class. This statement concerns the realizability of attention matrices, rather than the preservation of the full attention function across all inputs. We discuss the latter requirement in detail in Section 4.

We now turn to tied query and key projections. Writing $\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{W}_Q = \boldsymbol{X}\boldsymbol{W}_K$, the score matrix of an unmasked Q=K head is the Gram matrix $\boldsymbol{Z}\boldsymbol{Z}^\top$. Its symmetry and positive semidefiniteness impose global compatibility constraints on the directed routing patterns produced after the row-wise Softmax. The cyclic shift gives a simple and natural witness to this obstruction.

Proposition 3.1 (Cyclic-shift obstruction)**.** *Let $L \geq 2$, and define the cyclic-shift pattern $\boldsymbol{P}^{\mathrm{cyc}} \in \mathbb{R}^{L \times L}$ by $P^{\mathrm{cyc}}_{ij} := \mathbf{1}\{j \equiv i - 1 \pmod{L}\}$. Then, for every $d \geq 1$ and every $\boldsymbol{Z} \in \mathbb{R}^{L \times d}$,*

$$\big\| \operatorname{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top) - \boldsymbol{P}^{\mathrm{cyc}} \big\|_2 \geq 1.$$

The target $\boldsymbol{P}^{\mathrm{cyc}}$ is a $(1, 1)$-sparse attention pattern (see Figure 4). By Theorem 3.1, an untied head can approximate it to arbitrary accuracy using $d = \min\{2, L-1\}$. Proposition 3.1, in contrast, gives an error floor of 1 for every unmasked Q=K head, regardless of its dimension. This yields a strict separation: increasing the head dimension cannot compensate for the geometric constraint created by tying the query and key projections. The experimental verification is provided in Section 5.

3.2 The role of causal masking

The preceding impossibility is specific to unmasked attention. Let $\boldsymbol{M}_{\mathrm{MASK}}$ denote the causal mask, which sets all logits above the main diagonal to $-\infty$ and hence prevents tokens from attending to future positions. Although the pre-mask score matrix remains a symmetric Gram matrix under $\boldsymbol{W}_Q = \boldsymbol{W}_K$, the mask supplies directionality externally and thereby breaks the symmetry of the resulting routing rule. We call an attention pattern $\boldsymbol{A}$ *causal-compatible* if $A_{ij} = 0$ whenever $j > i$. The following theorem shows that the causal mask removes the qualitative obstruction identified in the unmasked setting.

Theorem 3.2 (Realizability under a causal mask)**.** *Let $L \geq 2$ and set $d = d_{\mathrm{hid}} = L - 1$ and $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{I}_d$. For every causal-compatible row-stochastic matrix $\boldsymbol{A} \in \mathbb{R}^{L \times L}$ and every*

$\varepsilon > 0$, there exists $\boldsymbol{X} \in \mathbb{R}^{L\times d}$ satisfying

$$\left\|\mathrm{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top + \boldsymbol{M}_{\mathrm{MASK}}\big) - \boldsymbol{A}\right\|_2 < \varepsilon.$$

Thus, once directionality is supplied by the architecture, tied query and key projections can approximate every causal-compatible row-stochastic routing pattern. The construction uses head dimension $L-1$. For a (k,γ)-sparse target, untied projections instead require only the sparsity-dependent dimension $\min\{2k, L-1\}$ in Theorem 3.1. Since Theorem 3.2 provides only an upper bound, it remains open whether a dimension depending only on k is sufficient under the constraint $\boldsymbol{W}_Q = \boldsymbol{W}_K$.

A norm separation. Realizability alone does not capture the full cost of tying the projections. We therefore quantify the size of the projected query and key representations by $\|\boldsymbol{X}\boldsymbol{W}_Q\|_F$ and $\|\boldsymbol{X}\boldsymbol{W}_K\|_F$. These norms control the magnitude of the dot-product scores and are invariant under compensating rescaling of the input and projection matrices. They also connect representation size to parameter size: whenever $\|\boldsymbol{X}\|_2 \le B$ with $B > 0$ (a condition met by layer normalization), the inequality $\|\boldsymbol{X}\boldsymbol{W}\|_F \le B\|\boldsymbol{W}\|_F$ shows that a lower bound on the projected-feature norm implies the corresponding lower bound, divided by B, on the Frobenius norm of the projection matrix.

Theorem 3.3 (Untied realization with controlled norm)**.** *Let $L > 1$, $1 \le k \le L$, $\gamma \ge 1$, $\varepsilon > 0$, let d be even, and suppose that $d_{\mathrm{hid}} \ge d$. Define*

$$\bar{\varepsilon} := \min\{\varepsilon, 1\}, \qquad \Lambda := \max\left\{\log\gamma + \log\frac{64L\sqrt{k}}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{\sqrt{k}}, 1\right\}.$$

Suppose that $d \ge \max\left\{\frac{32k^2}{\bar{\varepsilon}^2}\Lambda^2\log\big(8kL(L-1)\big),\, 64\log(16L)\right\}$. Then there exist fixed matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}}\times d}$, depending only on d_{hid} and d, such that, for every causal-compatible (k,γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L\times L}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L\times d_{\mathrm{hid}}}$ satisfying

$$\left\|\mathrm{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top + \boldsymbol{M}_{\mathrm{MASK}}\big) - \boldsymbol{A}\right\|_2 < \varepsilon$$

and $\max\{\|\boldsymbol{X}\boldsymbol{W}_Q\|_F, \|\boldsymbol{X}\boldsymbol{W}_K\|_F\} \le \sqrt{\frac{3}{2}}\,\sqrt{L}\,k^{1/4}\sqrt{\Lambda}$.

Proof idea. The proof, which is deferred to Appendix F.5, randomly projects a sparse logit representation of the target pattern into a low-dimensional query–key space while preserving the relevant score differences and controlling the projected norms. These score guarantees yield small off-support mass and accurate relative weights on the support. Using the row and column sparsity of the target then gives $\|M - A\|_2 \le \sqrt{\|M-A\|_1\|M-A\|_\infty} < \varepsilon$.

We allow $\boldsymbol{M}_{\mathrm{MASK}}$ in Theorem 3.3 to be either zero mask or causal mask. For fixed k, γ, and ε, we have $\Lambda = \Theta(\log L)$ as $L \to \infty$. The theorem therefore yields head dimension $\mathcal{O}((\log L)^3)$ and projected-feature norms $\mathcal{O}(\sqrt{L\log L})$.

We now show that tying the projections incurs a substantially larger norm, even for the elementary previous-token pattern $\boldsymbol{P}^{\mathrm{prev}} \in \mathbb{R}^{L\times L}$ defined by $P^{\mathrm{prev}}_{11} = 1, P^{\mathrm{prev}}_{i,i-1} = 1 \quad \text{for } i = 2,\dots,L$, with all remaining entries equal to zero. It has one nonzero entry per row and at most two per column, so it is a causal-compatible $(2,1)$-sparse pattern under Definition 3.1. This pattern is widely observed in real-world models (Clark et al., 2019; Cheng et al., 2026) and we also visualize it in Figure 5.

Proposition 3.2 (Norm lower bound for tied projections)**.** *Let $L > 1$ and $0 < \varepsilon < 1/2$. For every $d \ge 1$ and every $\boldsymbol{Z} \in \mathbb{R}^{L\times d}$ satisfying $\left\|\mathrm{Softmax}\big(\boldsymbol{Z}\boldsymbol{Z}^\top + \boldsymbol{M}_{\mathrm{MASK}}\big) - \boldsymbol{P}^{\mathrm{prev}}\right\|_2 < \varepsilon$, we have*

$$\|\boldsymbol{Z}\|_F > \sqrt{L(L-1)\log\frac{1-\varepsilon}{\varepsilon}}.$$

For fixed $\varepsilon < 1/2$, Proposition 3.2 gives an $\Omega(L)$ lower bound that holds independently of the head dimension. By contrast, Theorem 3.3, applied with $k = 2$ and $\gamma = 1$, gives an untied construction with norm $\mathcal{O}(\sqrt{L\log L})$ and head dimension $\mathcal{O}((\log L)^3)$. Together with Theorem 3.2, this shows that causal masking restores the qualitative realizability of tied projections but does not eliminate their quantitative cost. Since some positional encodings such as RoPE (Su et al., 2024) and absolute positional encoding (Vaswani et al., 2017) preserve the structure of the Gram matrix, the results in Propositions 3.1 and 3.2 also hold for models that employ these positional encodings. We provide experimental verification in Section 5.

4 LIMITATIONS OF TIED K AND V

We now characterize the exact representational cost of tying the key and value projections in a single-head self-attention layer. The key projection determines which input features are used for *addressing*, namely, for deciding where information should be retrieved from. In contrast, the value projection determines which features are transmitted as *content* once a token has been selected. A standard untied head may represent these two roles in different feature spaces, whereas a K=V head must encode them through a single shared representation. Our goal is to quantify the exact head dimension required by this shared representation.

Fix a sequence length $L \geq 2$, and let $\boldsymbol{M}_{\mathrm{MASK}}$ be either the zero mask or the standard causal mask. Let the parameters of a teacher self-attention head be $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d_{\mathrm{hid}} \times d}, \boldsymbol{W}_O \in \mathbb{R}^{d \times d_{\mathrm{hid}}}$, and define its effective score and value matrices by $\boldsymbol{B} := \boldsymbol{W}_Q \boldsymbol{W}_K^\top \in \mathbb{R}^{d_{\mathrm{hid}} \times d_{\mathrm{hid}}}, \boldsymbol{C} := \boldsymbol{W}_V \boldsymbol{W}_O \in \mathbb{R}^{d_{\mathrm{hid}} \times d_{\mathrm{hid}}}$. The resulting attention map can be written as

$$\mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}, \boldsymbol{C}) := \mathrm{Softmax}\big(\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^\top + \boldsymbol{M}_{\mathrm{MASK}}\big)\,\boldsymbol{X}\boldsymbol{C}.$$

The teacher and student always use the same mask. To identify the feature spaces used for addressing and content transmission, write the i-th row of $\boldsymbol{X}$ as $\boldsymbol{x}_i^\top$. The attention logit between positions i and j can be written as $\boldsymbol{x}_i^\top \boldsymbol{B}\boldsymbol{x}_j = \boldsymbol{x}_j^\top \boldsymbol{B}^\top \boldsymbol{x}_i$. Thus, for any fixed query, the key token is examined only through directions in $\mathrm{Col}(\boldsymbol{B}^\top)$. Similarly, every coordinate of the transformed value $\boldsymbol{x}_j^\top \boldsymbol{C}$ is obtained by taking an inner product between $\boldsymbol{x}_j$ and a column of $\boldsymbol{C}$. Motivated by these observations, we define the effective key and value subspaces as $\mathcal{S}_K := \mathrm{Col}(\boldsymbol{B}^\top), \mathcal{S}_V := \mathrm{Col}(\boldsymbol{C})$. The subspace $\mathcal{S}_K$ therefore contains the input directions that can affect the routing decisions, whereas $\mathcal{S}_V$ contains the input directions that can affect the transmitted content.

We collect these two sets of directions in the joint address–content matrix $\boldsymbol{M} := \begin{bmatrix}\boldsymbol{B}^\top & \boldsymbol{C}\end{bmatrix} \in \mathbb{R}^{d_{\mathrm{hid}} \times 2d_{\mathrm{hid}}}$. Its column space is precisely $\mathrm{Col}(\boldsymbol{M}) = \mathcal{S}_K + \mathcal{S}_V := \mathrm{Span}\{\mathcal{S}_K \cup \mathcal{S}_V\}$. Consequently, $\mathrm{rank}(\boldsymbol{M})$ is the number of independent directions required to support both addressing and content transmission. We define $r_K := \dim(\mathcal{S}_K) = \mathrm{rank}(\boldsymbol{B})$, $r_V := \dim(\mathcal{S}_V) = \mathrm{rank}(\boldsymbol{C})$, $r_\cap := \dim(\mathcal{S}_K \cap \mathcal{S}_V)$, and $r_M := \dim(\mathcal{S}_K + \mathcal{S}_V) = \mathrm{rank}(\boldsymbol{M})$.

Theorem 4.1. *Suppose that $\boldsymbol{C} \neq \boldsymbol{0}$. For a positive integer s, a student self-attention head satisfying $\boldsymbol{W}_K = \boldsymbol{W}_V$ is parametrized by $\widetilde{\boldsymbol{W}}_Q, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{d_{\mathrm{hid}} \times s}, \widetilde{\boldsymbol{W}}_O \in \mathbb{R}^{s \times d_{\mathrm{hid}}}$, where $\widetilde{\boldsymbol{W}}$ is the shared key–value projection. Its effective matrices are $\widetilde{\boldsymbol{B}} := \widetilde{\boldsymbol{W}}_Q \widetilde{\boldsymbol{W}}^\top, \widetilde{\boldsymbol{C}} := \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}_O$. Define the minimal head dimension required to reproduce the teacher exactly by*

$$d^* := \min\left\{ s \in \mathbb{N} \;\middle|\; \begin{array}{l} s \geq 1, \text{ and there exist } \widetilde{\boldsymbol{W}}_Q, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}}_O \text{ of the above dimensions} \\ \text{such that } \mathrm{Attn}(\boldsymbol{X}; \widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{C}}) = \mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}, \boldsymbol{C}) \text{ for every } \boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}} \end{array} \right\}.$$

Then $d^* = \mathrm{rank}(\boldsymbol{M}) = \dim\big(\mathcal{S}_K + \mathcal{S}_V\big)$.

Proof idea. The key step is functional identifiability: equality of the two attention functions for every input forces equality of both effective matrices. Once this is established, K=V tying requires $\boldsymbol{B}^\top$ and $\boldsymbol{C}$ to factor through a common s-dimensional space, giving the joint-rank lower bound and a matching construction. The identifiability argument also applies under a causal mask because the last query can attend to every token; see Appendix G.1.

Interpretation of Theorem 4.1. The larger head dimension required by K=V tying comes from forcing the addressing and content features to share the same representation space. As a result, the shared projection must preserve both subspaces simultaneously, and therefore its column space must contain their joint span $\mathcal{S}_K + \mathcal{S}_V$. Hence, $d^* = \dim(\mathcal{S}_K + \mathcal{S}_V) = r_K + r_V - r_\cap$. Thus, only the overlapping directions can be reused; non-overlapping addressing and content directions consume separate dimensions in the shared representation. In the extreme case of two disjoint subspaces of equal dimension r, an untied head needs dimension r, whereas a tied K=V head needs dimension $\min\{2r, d_{\mathrm{hid}}\}$.

Why effective subspaces matter. The theorem is formulated in terms of the effective subspaces $\mathcal{S}_K = \mathrm{Col}(\boldsymbol{B}^\top), \mathcal{S}_V = \mathrm{Col}(\boldsymbol{C})$, rather than the raw projection spaces $\mathrm{Col}(\boldsymbol{W}_K)$ and $\mathrm{Col}(\boldsymbol{W}_V)$.

Indeed, since $\boldsymbol{B}^\top = \boldsymbol{W}_K \boldsymbol{W}_Q^\top, \boldsymbol{C} = \boldsymbol{W}_V \boldsymbol{W}_O$, we always have $\mathcal{S}_K \subseteq \mathrm{Col}(\boldsymbol{W}_K), \mathcal{S}_V \subseteq \mathrm{Col}(\boldsymbol{W}_V)$. If $\boldsymbol{W}_Q$ is rank deficient, some directions represented by $\boldsymbol{W}_K$ never contribute to the attention scores. Similarly, if $\boldsymbol{W}_O$ is rank deficient, some directions represented by $\boldsymbol{W}_V$ never reach the final output. The raw projection spaces may therefore contain redundant directions that do not affect the function implemented by the attention layer, whereas the effective subspaces retain only the functionally relevant directions.

Implications for KV cache sharing. The causal version of Theorem 4.1 also clarifies the storage implications of exact tying. Suppose $r_K = r_V = r$. A minimal untied head with a conventional separate key/value cache stores $2r$ scalars per token, whereas an exactly equivalent tied head needs a shared representation of dimension $2r - r_\cap$. Relative to this separate-cache baseline, the saving is $r_\cap$ scalars per token, or a fraction $r_\cap/(2r)$. Thus, preserving the full attention function can offset the nominal cache benefit of tying when the effective subspaces have little overlap. This is a scalar-storage comparison within our attention model, without additional cache compression; it does not assert an implementation-level latency or throughput gain.

Exact versus approximate realization. Theorem 4.1 gives a sharp characterization of the head dimension required for exact functional equivalence: a tied K=V head must preserve the entire joint address–content subspace, and hence requires dimension $\mathrm{rank}([\boldsymbol{B}^\top, \boldsymbol{C}])$. For approximate realization on a bounded input domain, the same geometric picture naturally extends from algebraic rank to numerical rank. Specifically, Corollary G.1 shows that projecting the joint matrix $[\boldsymbol{B}^\top, \boldsymbol{C}]$ onto its leading s-dimensional subspace yields a tied student whose uniform approximation error is controlled by $\sigma_{s+1}([\boldsymbol{B}^\top, \boldsymbol{C}])$. Thus, exact realization is governed by the dimension of the joint address–content space, whereas approximate realization can exploit spectral decay in this space: when the singular values decay rapidly, a substantially smaller shared representation can suffice with small approximation error.

5 Experiments

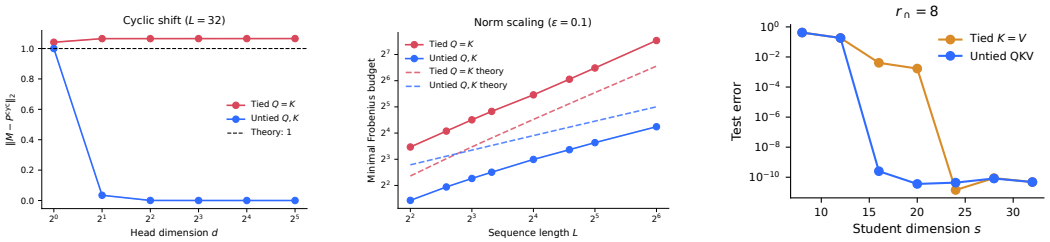
We use controlled synthetic experiments to test three predictions of our theory: the unmasked routing obstruction for Q=K, the norm separation under causal masking, and the head-dimension cost of K=V. Throughout, $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$ and $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$ denote the projected query and key features; under Q=K, we write $\boldsymbol{Z} = \boldsymbol{Q} = \boldsymbol{K}$. Extensive experiments in (Kayyam et al., 2026) have shown that different tied projections lead to varying degrees of task-specific performance degradation. We therefore focus primarily on validating our theoretical results in controlled settings.

5.1 Synthetic Experiments

Cyclic-shift obstruction. We first examine Proposition 3.1 using the cyclic-shift pattern $\boldsymbol{P}^{\mathrm{cyc}}$. For a representative sequence length, we vary the head dimension and directly optimize $\boldsymbol{M}_{\mathrm{untied}} = \mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top), \boldsymbol{M}_{\mathrm{Q=K}} = \mathrm{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top)$, by minimizing their cross-entropy with the target. Approximation quality is measured by $\|\boldsymbol{M} - \boldsymbol{P}^{\mathrm{cyc}}\|_2$. As shown in Figure 1a, the untied error decreases rapidly with the head dimension and approaches numerical precision, whereas the Q=K error remains above the dimension-independent lower bound. The learned attention matrices in Figure 2 further show that the untied head recovers both the shifted diagonal and the wrap-around interaction, while the tied head fails to realize the directed cycle. Increasing the head dimension cannot overcome the Gram-matrix constraint imposed by Q=K.

Norm separation under causal masking. We next consider the previous-token pattern $\boldsymbol{P}^{\mathrm{prev}}$, with $\boldsymbol{M}_{\mathrm{mask}}$ denoting the causal mask. For a fixed error tolerance ε, we try to find the minimum norm budgets that achieve an error less than ε, which are denoted by $R_{Q=K}(L)$ and $R_{QK}(L)$, respectively. We set the head dimension equal to the sequence length and estimate these budgets using projected optimization and bisection.

Figure 1b is consistent with Theorem 3.3 and Proposition 3.2. The tied budget grows approximately linearly with L, whereas the untied budget follows a substantially slower $\sqrt{L \log L}$ growth. Accordingly, as shown in Figure 3, $R_{Q=K}(L)/L$ and $R_{QK}(L)/\sqrt{L \log L}$ remain nearly constant. We further report the norms at different token position in Figure 3, showing that the tied representation



(a) Approximation error of $\boldsymbol{P}^{\mathrm{cyc}}$.

(b) The norm budget to represent $\boldsymbol{P}^{\mathrm{prev}}$.

(c) Head dimension budget under K=V.

Figure 1: Empirical verification of theoretical results.

concentrates much larger norms at earlier token positions, which aligns with our proof of Proposition 3.2.

Joint-subspace separation for K=V. We finally test Theorem 4.1. Recall the effective matrices $\boldsymbol{B} = \boldsymbol{W}_Q \boldsymbol{W}_K^\top, \boldsymbol{C} = \boldsymbol{W}_V \boldsymbol{W}_O$, and the associated subspaces $\mathcal{S}_K = \mathrm{Col}(\boldsymbol{B}^\top), \mathcal{S}_V = \mathrm{Col}(\boldsymbol{C})$. The experiments in this subsection use the zero mask. We construct teachers with prescribed ranks $r_K = \dim(\mathcal{S}_K) = 16$ and $r_V = \dim(\mathcal{S}_V) = 16$, vary the intersection dimension $r_\cap = \dim(\mathcal{S}_K \cap \mathcal{S}_V)$, and evaluate students of different head dimensions s. Setting $r_\cap = 8$, Figure 1c shows that the untied QKV student reaches numerical precision at $s = \max\{r_K, r_V\} = 16$, whereas the K=V student requires $s = r_K + r_V - r_\cap = \dim(\mathcal{S}_K + \mathcal{S}_V) = 24$.

5.2 Experiments on Real-world Models

Evidence in real-world models. We probe pre-trained BERT (Devlin et al., 2019) and LLaMA (Touvron et al., 2023) models to examine whether the structures predicted by our theory emerge in practice. Figures 4–5 reveal cyclic-shift-like attention patterns in BERT and previous-token attention patterns in LLaMA2-7B, indicating that our routing witnesses are not merely synthetic constructions. We further compute the 99%-energy ranks of $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{M}$ as empirical proxies for r_K, r_V, and r_M, respectively. As shown in Figure 6, r_M remains relatively small in the early layers but becomes substantially larger than the individual ranks in later layers.

6 Conclusion

We studied how projection tying changes the expressivity and resource requirements of self-attention. While unmasked Transformers with positional embeddings remain universal approximators even under full Q=K=V tying, sharper differences emerge at the head level. Under Q=K, the Gram structure of the logits restricts directional routing in unmasked attention; causal masking restores realizability but can still incur substantially larger projected-feature norms than untied attention. Under K=V, attention-pattern realizability is preserved, but addressing and content transmission must share a representation, so exact equivalence requires a head dimension determined by the joint span of the effective key and value subspaces. These results show that independent projections provide benefits not captured by model-level universality or parameter counts alone: separate Q and K support flexible routing, while separate K and V allow addressing and content to occupy different feature spaces.

Limitations: Although the universal approximation capability of Transformers with tied projections is established in Section 2, a lower bound on the approximation rate with regularity assumptions is still needed for a more complete comparison of the expressive power of standard Transformers and those with tied projections. Moreover, all of our upper bounds are obtained through constructive arguments, and standard optimization methods may therefore fail to recover the proposed solutions or may converge slowly. Similarly, we do not investigate how tied projections affect the training dynamics and inductive biases of self-attention. Finally, Sections 3 and 4 primarily focus on single-layer, single-head self-attention, leaving a gap between our controlled experimental settings and real-world models.

AI USE STATEMENT

We used generative AI tools primarily to assist with drafting descriptive sections, language refinement, proofreading, and coding-related tasks. The core ideas, methodology, analyses, and contributions presented in this work were independently developed by the authors. All AI-assisted content was carefully reviewed, validated, and, where necessary, revised by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, analyses, and other materials produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed proofs in the Appendix, while all assumptions and definitions are stated explicitly and comprehensively in the main text. A detailed description of the experimental setup is provided in Appendix C.2. Since all experiments are designed to verify our theoretical results, we do not rely on any specialized datasets or data-processing procedures. The code will be made publicly available upon publication.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 72–86. PMLR, 2023.

Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Exact sequence interpolation with transformers. *Mathematical Foundations of Machine Learning*, 2(1):2, 2026.

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=y9Xp9NozPR.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, 2020.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Ali Borji. Key-value transformer. *arXiv preprint arXiv:2305.19129*, 2023.

Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. A unified framework on the universal approximation of transformer-type architectures. *arXiv preprint arXiv:2506.23551*, 2025.

Yuan Cheng, Fengzhuo Zhang, Yunlong Hou, Cunxiao Du, Chao Du, Tianyu Pang, Aixin Sun, and Zhuoran Yang. Demystifying the slash pattern in attention: The role of rope. *arXiv preprint arXiv:2601.08297*, 2026.

David Chiang, Peter Cholak, and Anand Pillay. Tighter bounds on the expressivity of transformer encoders. In *International Conference on Machine Learning*, pp. 5544–5562. PMLR, 2023.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

Léo Dana, Muni Sreenivas Pydi, and Yann Chevaleyre. Memorization in attention-only transformers. *arXiv preprint arXiv:2411.10115*, 2024.

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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Yiding Hao, Dana Angluin, and Robert Frank. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022.

Alexander Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. *Advances in Neural Information Processing Systems*, 37:42162–42210, 2024.

Xin He, Yuling Jiao, Xiliang Lu, and Jerry Zhijian Yang. Approximation error upper and lower bounds for h\"{o} lder class with transformers. *arXiv preprint arXiv:2605.07463*, 2026.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. *arXiv preprint arXiv:2504.15956*, 2025.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for transformer networks with application to regression. *arXiv preprint arXiv:2504.12175*, 2025.

Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Transformers can overcome the curse of dimensionality: A theoretical study from an approximation perspective. *Journal of Machine Learning Research*, 27(50):1–34, 2026.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *arXiv preprint arXiv:2307.14023*, 2023.

Tokio Kajitsuka and Issei Sato. Optimal memorization capacity of transformers. *arXiv preprint arXiv:2409.17677*, 2024.

Marko Karbevski and Antonij Mijoski. Key and value weights are probably all you need: On the necessity of the query, key, value weight triplet in self-attention transformers. *arXiv preprint arXiv:2510.23912*, 2025.

Ali Kayyam, Anusha Madan Gopal, and M Anthony Lewis. Do transformers need three projections? systematic study of qkv variants. In *Forty-third International Conference on Machine Learning*, 2026.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Md Kowsher, Nusrat Jahan Prottasha, Chun-Nam Yu, Ozlem Garibay, and Niloofar Yousefi. Does self-attention need separate weights in transformers? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 535–543, 2025.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. *arXiv preprint arXiv:2502.20766*, 2025.

Yanming Lai and Defeng Sun. Standard transformers achieve the minimax rate in nonparametric regression with $c^{s,\lambda}$ targets, 2026. URL `https://arxiv.org/abs/2602.20555`.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. The depth-to-width interplay in self-attention. *arXiv preprint arXiv:2006.12467*, 2020.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*, 2021.

Hude Liu, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. Attention mechanism, max-affine partition, and universal approximation. *arXiv preprint arXiv:2504.19901*, 2025.

Liam Madden, Curtis Fox, and Christos Thrampoulidis. Next-token prediction capacity: general upper bounds and a lower bound for transformers. *IEEE Transactions on Information Theory*, 71 (9):7134–7148, 2025.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. *arXiv preprint arXiv:2306.02010*, 2023.

William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on learning theory*, pp. 3627–3661. PMLR, 2021.

Parikshit Ram, Kenneth Clarkson, Tim Klinger, Shashanka Ubaru, and Alexander Gray. Transformers learn faster with semantic focus. *Advances in Neural Information Processing Systems*, 38: 33438–33502, 2026.

Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *Advances in neural information processing systems*, 36:36677–36707, 2023.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Zhaiming Shen, Alex Havrilla, Rongjie Lai, Alexander Cloninger, and Wenjing Liao. Transformers for learning on noisy and task-level manifolds: Approximation and generalization insights. *arXiv preprint arXiv:2505.03205*, 2025.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pp. 33416–33447. PMLR, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Pavlo Vasylenko, Hugo Pitorro, Andre Martins, and Marcos V Treviso. Long-context generalization with sparse attention. In *International Conference on Learning Representations*, volume 2026, pp. 15256–15303, 2026.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Hengshuai Yao, Xing Chen, Ahmed Murtadha, and Guan Wang. Thin keys, full values: Reducing kv cache via low-dimensional attention selection. *arXiv preprint arXiv:2603.04427*, 2026.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on transformer approximation. In *International Conference on Learning Representations*, volume 2026, pp. 77573–77607, 2026.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O (n) connections are expressive enough: Universal approximability of sparse transformers. *Advances in Neural Information Processing Systems*, 33:13783–13794, 2020.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Magauiya Zhussip, Dmitriy Shopkhoev, Ammar Ali, and Stamatios Lefkimmiatis. Share your attention: Transformer weight sharing via matrix-based dictionary learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29260–29268, 2026.

Appendix

Table of Contents

A Notation Table **15**

B Related work **15**

C Additional Experimental Results and Detailed Experimental Setup **16**
C.1 Additional Experimental Results 16
C.2 Experimental Setup 18

D A Summary of the Proof Techniques and Their Novelty **19**

E Technical Details in Section 2 **21**
E.1 Proof of Theorem 2.1 21
E.2 Proof of Lemma E.1 27
E.3 Proof of Lemma E.2 30
E.4 Proof of Lemma E.3 35
E.5 A Discussion on Parameter Efficiency 37

F Technical details in Section 3 **41**
F.1 Proof of Theorem 3.1 41
F.2 Approximating Sparse Attention Patterns with K=V 43
F.3 Proof of Proposition 3.1 44
F.4 Proof of Theorem 3.2 45
F.5 Proof of Theorem 3.3 48
F.6 Proof of Proposition 3.2 50
F.7 Supporting Lemmas 51

G Technical Details in Section 4 **60**
G.1 Proof of Theorem 4.1 60
G.2 Approximate realization 62

A NOTATION TABLE

Symbol	Meaning
L, d_{hid}	Sequence length and hidden dimension.
d, s	Head dimension and student head dimension, respectively.
$\boldsymbol{X}, \boldsymbol{E}$	Input representation and positional embedding in $\mathbb{R}^{L\times d_{\text{hid}}}$.
$\boldsymbol{M}_{\text{MASK}}$	Mask used in the single-head separation results.
$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{Z}$	Projected queries, projected keys, and their common representation under Q=K.
(k, γ)	Row/column sparsity bound and within-row nonzero-weight ratio bound.
$\boldsymbol{P}^{\text{cyc}}, \boldsymbol{P}^{\text{prev}}$	Cyclic-shift and previous-token target patterns.
$\boldsymbol{B}, \boldsymbol{C}$	Effective matrices $\boldsymbol{W}_Q\boldsymbol{W}_K^\top$ and $\boldsymbol{W}_V\boldsymbol{W}_O$.
$\mathcal{S}_K, \mathcal{S}_V$	Effective key and value subspaces $\text{Col}(\boldsymbol{B}^\top)$ and $\text{Col}(\boldsymbol{C})$.
$\boldsymbol{M}$	Joint address–content matrix $[\boldsymbol{B}^\top, \boldsymbol{C}]$.
$r_K, r_V, r_M, r_\cap$	Dimensions of the effective subspaces, the joint space and their intersection.
$\|\cdot\|_2, \|\cdot\|_F$	Matrix spectral norm and Frobenius norm.
$\sigma_j(\boldsymbol{M})$	j-th largest singular value of the joint matrix.

B RELATED WORK

Parameter sharing in Transformers. The original Transformer (Vaswani et al., 2017) assigns independent projection matrices to queries, keys, and values. Prior work has explored several distinct forms of attention-parameter sharing, primarily through their empirical performance and efficiency. Kowsher et al. (2025) use a shared projection with separate diagonal modulations for queries, keys, and values, rather than strict Q=K=V tying. The basic formulation of Borji (2023) replaces queries with keys and uses logits proportional to $\boldsymbol{K}\boldsymbol{K}^\top$, corresponding to Q=K at the level of the attention operator, not K=V. Kayyam et al. (2026) compare projection-tying schemes empirically across synthetic, vision, and language tasks. Related approaches share matrix dictionaries across layers and projection types, as in MASA (Zhussip et al., 2026), or share keys and values across query heads, as in multi-query and grouped-query attention (Shazeer, 2019; Ainslie et al., 2023). Karbevski & Mijoski (2025) provide theoretical results on eliminating an attention projection through reparameterization under specified assumptions. These mechanisms differ from the exact within-head constraints studied here. Our focus is on their expressive consequences and the associated head-dimension and projected-feature-norm requirements.

Expressive power of Transformers. Transformers are universal approximators of permutation-equivariant functions, and positional encodings extend this property to general continuous sequence-to-sequence functions (Yun et al., 2019). Subsequent work has established universality under increasingly restricted architectures, including sparse connectivity, a single attention layer or head, efficient attention mechanisms, and models without position-wise feed-forward networks (Yun et al., 2020; Kajitsuka & Sato, 2023; Alberti et al., 2023; Cheng et al., 2025; Hu et al., 2025). Beyond qualitative universality, approximation and statistical rates have been derived for structured, low-dimensional, and smooth function classes (Takakura & Suzuki, 2023; Jiang & Li, 2024; Havrilla & Liao, 2024; Jiao et al., 2025; 2026). Most closely related to our parameter-efficiency discussion, He et al. (2026) establish block-complexity bounds for Softmax–ReLU Transformers approximating Hölder functions.

At the level of individual attention operators, Bhojanapalli et al. (2020) characterize the rank bottleneck induced by small head dimensions, while Likhosherstov et al. (2021) show that low-dimensional self-attention can approximate sparse stochastic attention patterns. Related work has also compared self-attention with convolutional, recurrent, and feed-forward architectures, revealing both depth advantages and structural limitations such as rank collapse (Cordonnier et al., 2019; Dong et al., 2021; Levine et al., 2020; Sanford et al., 2023).

Another line of work quantifies Transformer expressivity through finite-sample memorization. Kim et al. (2022) establish constructive memorization guarantees for Transformers, while Mahdavi et al. (2023) analyze how the capacity of multi-head attention depends on architectural resources such as the number of heads and context length. Subsequent studies show that memorization can already be achieved by shallow or low-rank attention architectures (Kajitsuka & Sato, 2023; Dana et al., 2024).

More recently, Kajitsuka & Sato (2024) and Madden et al. (2025) derive nearly tight parameter-capacity bounds for sequence memorization and next-token prediction. These results complement universal approximation theory by characterizing the finite resources required to fit prescribed input–output associations. Alcalde et al. (2026) establish exact interpolation of real-valued sequence-to-sequence datasets by deep single-head Softmax Transformers,

C Additional Experimental Results and Detailed Experimental Setup

C.1 Additional Experimental Results

A two-dimensional untied cyclic-shift construction. The following construction separates realizability from optimizer convergence in the cyclic-shift experiment. For $t > 0$, define

$$\boldsymbol{K}_{j,:} := \sqrt{t}\left(\cos\frac{2\pi j}{L}, \sin\frac{2\pi j}{L}\right), \qquad \boldsymbol{Q}_{i,:} := \sqrt{t}\left(\cos\frac{2\pi(i-1)}{L}, \sin\frac{2\pi(i-1)}{L}\right),$$

where indices are interpreted modulo L. The target position uniquely maximizes each score row. Let

$$\Delta_L := 1 - \cos(2\pi/L) > 0.$$

Every off-target score is at least $t\Delta_L$ below the target score, so the off-target probability mass in each row is at most $(L-1)e^{-t\Delta_L}$. Both the attention matrix and the target are circulant, and therefore

$$\left\|\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top) - \boldsymbol{P}^{\mathrm{cyc}}\right\|_2 \le 2(L-1)e^{-t\Delta_L}.$$

Thus, for every $\varepsilon \in (0,1)$, taking

$$t > \Delta_L^{-1}\log\frac{2(L-1)}{\varepsilon}$$

gives spectral error smaller than ε in head dimension two. In contrast, setting $\boldsymbol{Z} = \boldsymbol{0}$ gives uniform attention and

$$\left\|\frac{1}{L}\mathbf{1}_L\mathbf{1}_L^\top - \boldsymbol{P}^{\mathrm{cyc}}\right\|_2 = 1,$$

showing that the tied lower bound in Proposition 3.1 is sharp.

Learned cyclic attention patterns. Figure 2 shows that self-attention with untied Q,K can well implement the cyclic attention pattern, while tied self-attention presents an obscure pattern.

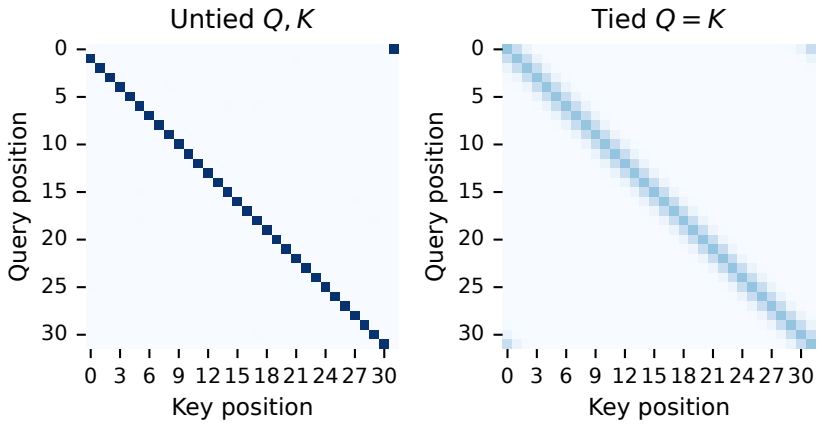


Figure 2: Learned attention patterns.

Interpreting the norm-budget search. The projected optimization and bisection procedure returns achieved feasible budgets rather than certified global minima. A candidate is counted as feasible only when it satisfies the masked spectral-error constraint. The rigorous tied reference is

$$\sqrt{L(L-1)\log\frac{1-\varepsilon}{\varepsilon}},$$

whereas the untied $\sqrt{L\log L}$ curve compares asymptotic growth. The proof of Proposition 3.2 also yields the row-wise envelope

$$\|\boldsymbol{z}_i\|_2^2 \geq 2(L-i)\log\frac{1-\varepsilon}{\varepsilon}, \qquad i \in [L],$$

which explains the accumulation visible in the learned tied representation.

Here, we provide the formal definition of $R_{Q=K}(L)$ and $R_{QK}(L)$, and Figure 3 shows the normalized norm budget.

$$R_{Q=K}(L) := \inf_{\|\operatorname{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top+\boldsymbol{M}_{\text{mask}})-\boldsymbol{P}^{\text{prev}}\|_2\leq\varepsilon} \|\boldsymbol{Z}\|_F,$$
$$R_{QK}(L) := \inf_{\|\operatorname{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top+\boldsymbol{M}_{\text{mask}})-\boldsymbol{P}^{\text{prev}}\|_2\leq\varepsilon} \max\{\|\boldsymbol{Q}\|_F, \|\boldsymbol{K}\|_F\}.$$

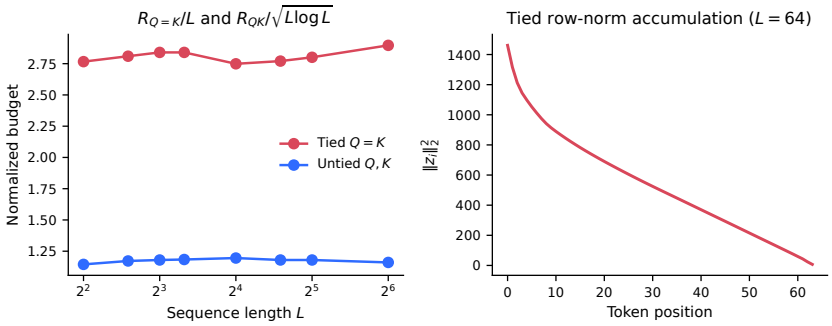


Figure 3: Normalized norm budget and the norms at different token position.

Observed attention patterns in real-world models. We plot the cyclic pattern observed in Bert and previous-token pattern in LLaMA2-7B in Figure 4 and Figure 5, respectively.

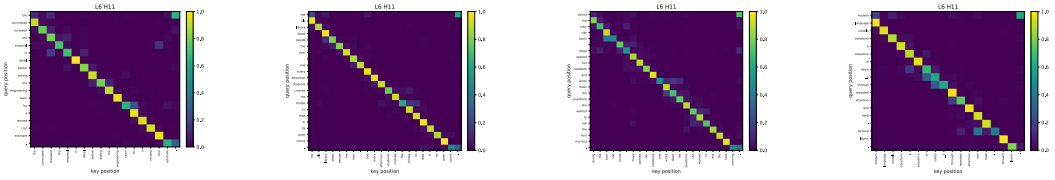


Figure 4: Cyclic patterns in BERT (Devlin et al., 2019).

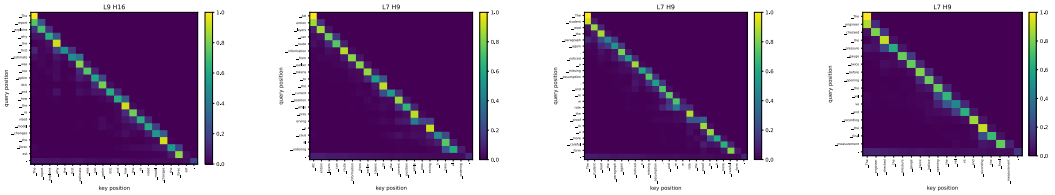
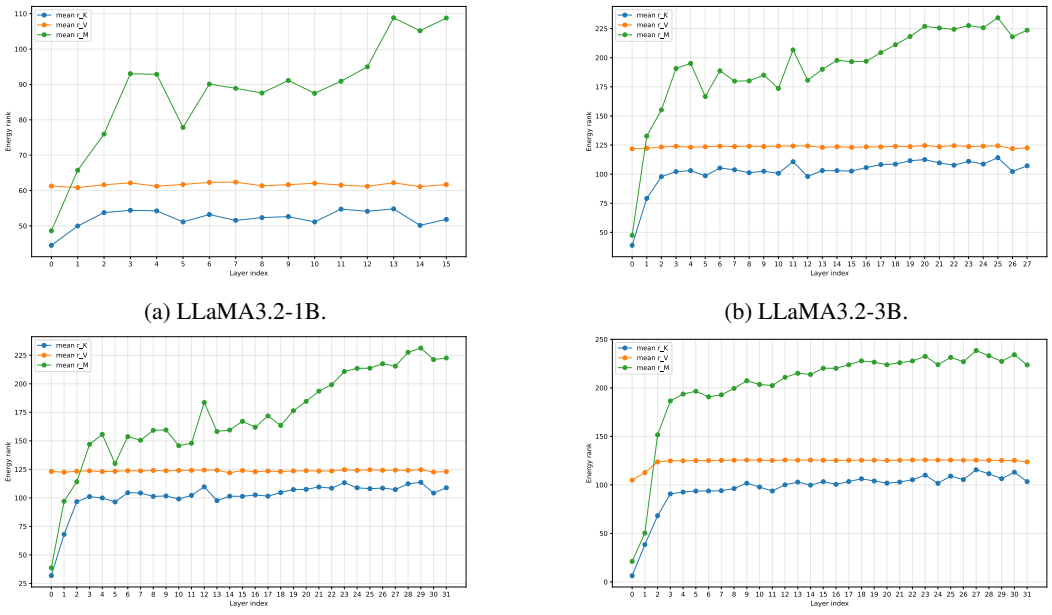


Figure 5: Previous-token patterns in LLaMA2-7B (Touvron et al., 2023).



(a) LLaMA3.2-1B.

(b) LLaMA3.2-3B.

(c) LLaMA3-8B.

(d) LLaMA2-7B.

Figure 6: Estimation of r_K, r_V and r_M.

Estimating r_K, r_V, and r_M. We plot the layer-wise r_K, r_V and r_M in Figure 6. It is obvious that in early layers, r_K and r_M are much smaller and they increase as layers go deeper, while r_V is nearly stable across layers.

C.2 Experimental Setup

We design the synthetic experiments to isolate the head-level phenomena predicted by the theory, rather than to evaluate downstream task performance. We refer readers to (Kayyam et al., 2026) for details about how tied projections affect performance on downstream tasks. Unless otherwise stated, all synthetic experiments are implemented in PyTorch, on a single NVIDIA A100 GPU and carried out in double precision (`float64`). We optimize the projected query/key representations or the effective attention matrices directly, thereby removing confounding effects from the rest of the Transformer architecture. Adam is used for all learned constructions, with early stopping based on the best achieved objective value. The primary approximation metric throughout Sections 5.1 is the spectral norm error, matching the norm used in the corresponding theoretical statements.

Cyclic-shift experiment. To test the unmasked routing obstruction for Q=K, we use the cyclic target $P^{\mathrm{cyc}}_{ij} = \mathbf{1}\{j = i - 1 \pmod{L}\}$. We consider sequence lengths $L \in \{8, 16, 32, 64, 128\}$ and head dimensions $d \in \{1, 2, 4, 8, 16, 32, 64\}$, excluding configurations with $d > L$. For the untied parameterization we optimize $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{L\times d}$, whereas for the tied parameterization we optimize $\boldsymbol{Z} \in \mathbb{R}^{L\times d}$ and use the Gram logits $\boldsymbol{Z}\boldsymbol{Z}^\top$. Parameters are trained by minimizing the cross-entropy with the target attention pattern. We use Adam with learning rate 3×10^{-2} for at most 1500 steps and independent restarts for each configuration, and retain the best solution. Approximation quality is reported as $\|\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top) - \boldsymbol{P}^{\mathrm{cyc}}\|_2$ or $\|\mathrm{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top) - \boldsymbol{P}^{\mathrm{cyc}}\|_2$, respectively.

Causal norm experiment. For the previous-token target $\boldsymbol{P}^{\mathrm{prev}}$, we estimate the smallest projected-feature budget required to reach spectral error at most $\epsilon = 0.1$ under a causal mask. We use $L \in \{4, 6, 8, 10, 16, 24, 32, 64\}$ and set the head dimension to $d = L$ for both parameterizations. For a candidate budget R, the tied representation is constrained by $\|\boldsymbol{Z}\|_F \le R$, whereas the untied representations satisfy $\|\boldsymbol{Q}\|_F, \|\boldsymbol{K}\|_F \le R$. We optimize under these hard Frobenius-ball constraints and use bisection over R to locate the smallest empirically feasible budget. Each constrained opti-

mization uses Adam with learning rate 5×10^{-3} for at most 20,000 steps. The optimization objective combines the spectral approximation error with target cross-entropy, while feasibility is determined solely by whether the resulting spectral error is at most ϵ. We compare the achieved budgets with the theoretical tied lower bound and the untied asymptotic reference derived in Section 3.

K=V joint-subspace experiment. To test the dimension characterization in Theorem 4.1, we construct teacher heads through their effective matrices $\boldsymbol{B} = \boldsymbol{W}_Q\boldsymbol{W}_K^\top$ and $\boldsymbol{C} = \boldsymbol{W}_V\boldsymbol{W}_O$. We set $d_{\mathrm{hid}} = 32$ and $r_K = r_V = 16$, and construct effective key and value subspaces with prescribed intersection dimensions $r_\cap \in \{0, 4, 8, 12, 16\}$. Student dimensions are varied over $s \in \{8, 12, 16, 20, 24, 28, 32\}$. We compare an untied QKV student with a K=V student. For each configuration we report both the analytic truncated-SVD optimum and a learned factorization of $(\boldsymbol{B}, \boldsymbol{C})$. The reconstruction objective is

$$E_{\mathrm{mat}} = \frac{\|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|_F^2 + \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F^2}{\|\boldsymbol{B}\|_F^2 + \|\boldsymbol{C}\|_F^2}.$$

The learned reconstruction is optimized with Adam for at most 1000 steps with learning rate 2×10^{-2}. We additionally perform functional distillation on independently sampled Gaussian inputs: each sequence has length 8, and we use 128 training and 128 held-out sequences. The functional students are optimized for at most 5000 steps with learning rate 5×10^{-3}. This experiment tests whether the empirical transition occurs at $s = \max\{r_K, r_V\}$ for QKV and $s = r_K + r_V - r_\cap$ for K=V, as predicted by Theorem 4.1.

Pretrained-model probes. We complement the controlled experiments with measurements on pretrained Transformers. For directional routing, we extract post-Softmax attention probabilities from `bert-base-uncased` and search all layers and heads for attention matrices close to P^{cyc}. For causal routing, we similarly examine `Llama-2-7B` and `Llama-3.2-1B` for the previous-token pattern P^{prev}. Special tokens are removed in the content-only analysis and the retained attention rows are renormalized. In addition to full-sequence measurements, we search contiguous content windows of length 16 for BERT and 32 for the causal models. Candidate patterns are ranked primarily by spectral error to the corresponding target pattern.

To examine the address–content geometry in pretrained LLMs, we analyze `Llama-2-7B`, `Llama-3.2-1B`, and `Llama-3-8B`. For every query head in every layer, we form the effective key, value, and joint operators corresponding to $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{M} = [\boldsymbol{B}^\top, \boldsymbol{C}]$. For grouped-query attention, each query head is paired with the KV head that it actually uses. Rather than materializing the hidden-dimension matrices, we recover their singular values from equivalent $d \times d$ and $2d \times 2d$ Gram matrices, using `float64` eigendecompositions. For a matrix A, we define its 99% energy rank as

$$r_{0.99}(A) := \min\left\{ r : \frac{\sum_{i=1}^{r} \sigma_i(A)^2}{\sum_i \sigma_i(A)^2} \geq 0.99 \right\}.$$

We report $r_{0.99}(B)$, $r_{0.99}(C)$, and $r_{0.99}(M)$ averaged over all query heads within each layer. These quantities are used as effective-dimension diagnostics. The analysis uses the raw projection subspaces and does not incorporate position-dependent rotary transformations, so it should be interpreted as a projection-space diagnostic rather than an exact reconstruction of the full pretrained attention operator.

D A Summary of the Proof Techniques and Their Novelty

This appendix summarizes the main proof ideas and clarifies their relation to existing techniques. The three phenomena studied in the paper require rather different arguments: a tied contextual-mapping construction for universality, geometric and random-projection arguments for Q=K, and an identifiability plus joint-factorization argument for K=V.

Universality under complete projection tying. The proof of Theorem 2.1 follows the standard quantization–contextual-mapping–value-mapping framework (Yun et al., 2019; Kajitsuka & Sato, 2023; Jiao et al., 2025; Cheng et al., 2025), but modifies each stage to operate in the fixed-width regime and, crucially, under complete projection tying.

- **Quantization.** Lemma E.1 gives an explicit depth–width trade-off: elementary quantization operations can be parallelized through width or serialized through depth. This unifies the fixed-width and fixed-depth constructions used in prior approximation results (He et al., 2026; Kajitsuka & Sato, 2023; Jiao et al., 2025).
- **Contextual mapping.** The main new ingredient is that a single shared rank-one projection

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V = \boldsymbol{v}\boldsymbol{u}^\top$$

  suffices to separate all token–context pairs in a finite token-wise separated vocabulary. Residual connections distinguish different tokens, while the Softmax-induced Boltzmann operator separates identical tokens appearing in different contexts. We note that this result is a natural extension of Theorem 2 in (Kajitsuka & Sato, 2023).
- **Value mapping.** Lemma E.3 first projects the finite set of contextual representations onto one dimension and then performs the required interpolation through sequential scalar corrections. This moves the dependence on the number of interpolation points from width to depth, yielding a width bound independent of the grid resolution. Our construction is different from that in (He et al., 2026).

Sparse routing with fixed untied projections. Theorem 3.1 constructs a low-rank score matrix for an arbitrary (k, γ)-sparse stochastic pattern. For each row, an interpolation polynomial matches the desired log-attention ratios on the support, while a nonnegative polynomial vanishing on the support suppresses all off-support entries. This extends the support-selection construction of Sanford et al. (2023) from approximately uniform weights to arbitrary sparse stochastic weights. Row and column sparsity imply a score rank of at most $2k + 1$, and Softmax row-shift invariance removes the constant direction, giving the dimension $\min\{2k, L - 1\}$. A fixed pair of coordinate projections then realizes the entire target class; only the input representation depends on the target. Corollary F.1 further shows that the key projection itself may be prescribed arbitrarily, provided it has full column rank, highlighting the routing flexibility retained under K=V.

A dimension-independent obstruction for unmasked Q=K. Proposition 3.1 uses the geometry of Gram logits rather than a rank argument. If $\boldsymbol{S} = \boldsymbol{Z}\boldsymbol{Z}^\top \succeq 0$, then $\exp^\circ(\boldsymbol{S})$ ($\exp^\circ$ is $\exp$ applied elementwisely) is also positive semidefinite by the Schur product theorem. Approximating the cyclic shift within spectral error strictly below one would force a matrix similar to a weighted cyclic shift to have all eigenvalues in the open right half-plane, contradicting the spectrum of the cyclic shift. This yields the sharp error floor 1 for every head dimension, showing that the obstruction is geometric and cannot be removed by increasing width.

Causal masking and the norm separation. Theorem 3.2 shows that causal masking removes this qualitative obstruction. The construction builds a positive-semidefinite score matrix whose Gram entries realize the required causal log-attention ratios, followed by a Softmax-invariant rank reduction to dimension at most $L - 1$.

Theorem 3.3 then gives a low-dimensional untied realization with controlled projected-feature norms. Building on the random-projection idea of Likhosherstov et al. (2021), our argument exploits sparsity directly: it preserves only score differences involving the target support, while simultaneously controlling the projected query and key norms. Row and column sparsity convert these local ratio guarantees into

$$\|\boldsymbol{M} - \boldsymbol{A}\|_2 \le \sqrt{\|\boldsymbol{M} - \boldsymbol{A}\|_1 \|\boldsymbol{M} - \boldsymbol{A}\|_\infty}.$$

This yields a joint dimension–norm–accuracy guarantee rather than only a low-dimensional realization. In contrast, Proposition 3.2 shows that for the previous-token pattern, Q=K forces an $\Omega(L)$ projected-feature norm: the required predecessor–diagonal logit gaps induce a recursive growth of token norms. Thus, causal masking restores realizability but not the representation efficiency of tied queries and keys.

The exact address–content cost of K=V. Theorem 4.1 combines an identifiability argument with a matrix factorization argument. Constant token inputs first identify the effective value matrix $\boldsymbol{C} = \boldsymbol{W}_V \boldsymbol{W}_O$; perturbing one token then reduces the attention weight to a strictly monotone scalar

function of $\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{w}$, identifying the effective score matrix $\boldsymbol{B} = \boldsymbol{W}_Q \boldsymbol{W}_K^\top$. Hence any exactly equivalent K=V head must factor both $\boldsymbol{B}^\top$ and $\boldsymbol{C}$ through the same shared projection, giving

$$d^\star \geq \operatorname{rank}\big([\boldsymbol{B}^\top, \boldsymbol{C}]\big).$$

A matching rank factorization attains equality, and therefore

$$d^\star = \operatorname{rank}\big([\boldsymbol{B}^\top, \boldsymbol{C}]\big) = \dim(\mathcal{S}_K + \mathcal{S}_V).$$

Thus, the exact cost of tying K and V is the dimension of the joint address–content feature space: overlapping directions can be reused, whereas complementary directions require additional head dimensions.

E Technical Details in Section 2

E.1 Proof of Theorem 2.1

The proof proceeds in five steps, which are nearly standard in existing works (Yun et al., 2019; Kajitsuka & Sato, 2023; Jiao et al., 2025; He et al., 2026). First, we partition the input domain into finitely many disjoint quantization cells, separated by narrow gaps, and approximate F by a function F' that is constant on each cell. Second, we construct a token-wise feed-forward module F_1 that maps every input in a given cell, after adding the positional encoding, exactly to the corresponding positional grid point. Third, a single self-attention layer F_2 satisfying $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$ maps all tokens on the finite positional grid to pairwise distinct context-dependent representations. Fourth, a token-wise feed-forward module F_3 interpolates these representations to the prescribed grid values of F. Consequently, the composition

$$T(\boldsymbol{X}) := F_3 \circ F_2 \circ F_1(\boldsymbol{X} + \boldsymbol{E})$$

agrees exactly with F' on the union of the quantization cells. Finally, after establishing a uniform bound on T, we choose the gaps sufficiently narrow that their total measure is small. This controls both $\|F - F'\|_{L^2}$ and $\|F' - T\|_{L^2}$, and the desired estimate follows from the triangle inequality.

Theorem 2.1 (Restatement). *There exists a universal constant $C > 0$ such that, for every $L, d_{\mathrm{hid}} \in \mathbb{N}$, there are positive integers H, d, W satisfying $\max\{H, d, W\} \leq CLd_{\mathrm{hid}}$ for which the following holds: for every $F \in \mathcal{C}_{L,d_{\mathrm{hid}}}$ and every $\varepsilon > 0$, there exists $T \in \mathfrak{T}_{Q=K=V}(H, d, W)$ such that*

$$\|F - T\|_{L^2} < \varepsilon.$$

Consequently, the same conclusion holds for $\mathfrak{T}_{Q=K}(H, d, W)$ and $\mathfrak{T}_{K=V}(H, d, W)$.

Proof of Theorem 2.1. Let's begin with the first step.

Step 1: Approximating F by a piecewise constant function F'. Let $D := Ld_{\mathrm{hid}}$. Since $F : [0,1]^{L\times d_{\mathrm{hid}}} \to \mathbb{R}^{L\times d_{\mathrm{hid}}}$ is continuous and $[0,1]^{L\times d_{\mathrm{hid}}}$ is compact, F is uniformly continuous. Moreover, we have

$$B_F := \max_{\boldsymbol{X}\in[0,1]^{L\times d_{\mathrm{hid}}}} \|F(\boldsymbol{X})\|_F < \infty.$$

By the uniform continuity of F, there exists $\rho > 0$ such that

$$\|\boldsymbol{X} - \boldsymbol{X}'\|_F < \rho \quad \Longrightarrow \quad \|F(\boldsymbol{X}) - F(\boldsymbol{X}')\|_{L^2} < \frac{\varepsilon}{4}, \tag{E.1}$$

for any $\boldsymbol{X}, \boldsymbol{X}' \in [0,1]^{L\times d_{\mathrm{hid}}}$. Choose an integer $M \geq 2$ sufficiently large such that $\frac{\sqrt{Ld_{\mathrm{hid}}}}{M} < \rho$, and define $h := \frac{1}{M}$. The parameter h determines the locations of the quantization grid points. We will later choose $\delta^* \in (0, h)$ sufficiently small and set $\delta := h - \delta^*$, so that $h = \delta + \delta^*$. Define the one-dimensional grid $\Gamma_{\delta,\delta^*} := \{0, h, 2h, \ldots, (M-1)h\}$, and the corresponding grid over $[0,1]^{L\times d_{\mathrm{hid}}}$

$$\begin{aligned}\mathcal{G}_{\delta,\delta^*} &:= \Gamma_{\delta,\delta^*}^{L\times d_{\mathrm{hid}}} \\ &= \big\{\boldsymbol{G} \in [0,1]^{L\times d_{\mathrm{hid}}} : \boldsymbol{G}_{i,j} \in \{0, h, \ldots, (M-1)h\} \text{ for all } i \in [L],\ j \in [d_{\mathrm{hid}}]\big\}.\end{aligned}$$

In particular, $|\mathcal{G}_{\delta,\delta^*}| = M^{Ld_{\mathrm{hid}}}$. For each $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$, define its associated quantization cell by

$$S_{\boldsymbol{G}} := \prod_{i\in[L]} \prod_{j\in[d_{\mathrm{hid}}]} [\boldsymbol{G}_{i,j}, \boldsymbol{G}_{i,j} + \delta].$$

Since $(M-1)h + \delta = 1 - \delta^* < 1$, we have $S_{\boldsymbol{G}} \subset [0,1]^{L\times d_{\mathrm{hid}}}$ for every $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$. Moreover, since $\delta < h$, the cells $\{S_{\boldsymbol{G}}\}_{\boldsymbol{G}\in\mathcal{G}_{\delta,\delta^*}}$ are pairwise disjoint. We now define the piecewise constant approximation F' by

$$F'(\boldsymbol{X}) := \sum_{\boldsymbol{G}\in\mathcal{G}_{\delta,\delta^*}} F(\boldsymbol{G})\, \mathbf{1}\{\boldsymbol{X} \in S_{\boldsymbol{G}}\}.$$

Notice that $F'(\boldsymbol{X}) = 0$ whenever $\boldsymbol{X}$ does not belong to any quantization cell. We next verify the approximation error inside the quantization cells. Fix any $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$ and $\boldsymbol{X} \in S_{\boldsymbol{G}}$. By the definition of $S_{\boldsymbol{G}}$,

$$|\boldsymbol{X}_{i,j} - \boldsymbol{G}_{i,j}| \le \delta$$

for every $i \in [L]$ and $j \in [d_{\mathrm{hid}}]$. Consequently,

$$\|\boldsymbol{X} - \boldsymbol{G}\|_F \le \sqrt{Ld_{\mathrm{hid}}}\,\delta < \sqrt{Ld_{\mathrm{hid}}}\,h = \frac{\sqrt{Ld_{\mathrm{hid}}}}{M} < \rho,$$

where the last inequality follows from equation E.1. Therefore, by equation E.1,

$$\|F(\boldsymbol{X}) - F'(\boldsymbol{X})\|_{L^2} = \|F(\boldsymbol{X}) - F(\boldsymbol{G})\|_{L^2} < \frac{\varepsilon}{4}, \qquad \boldsymbol{X} \in \bigcup_{\boldsymbol{G}\in\mathcal{G}_{\delta,\delta^*}} S_{\boldsymbol{G}}.$$

The cells do not cover the whole domain because of the gaps of width δ^*. Hence, the global approximation error between F and F' cannot be controlled by uniform continuity alone. The precise choice of δ^*, together with the estimate $\|F - F'\|_F < \frac{\varepsilon}{2}$, will be established in Step 5 by controlling the Lebesgue measure of the complement of the quantization cells.

Step 2: Quantization module F_1. We next construct a token-wise feed-forward module that maps every input inside a quantization cell to the corresponding positional grid point. Recall that the positional encoding is defined by

$$\boldsymbol{E} := [1, 2, \ldots, L]^\top \mathbf{1}_{d_{\mathrm{hid}}}^\top \in \mathbb{R}^{L\times d_{\mathrm{hid}}},$$

Define the positional grid

$$\mathcal{G}^p_{\delta,\delta^*} := \{\boldsymbol{G} + \boldsymbol{E} : \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}\}.$$

Although we retain the notation $\mathcal{G}^p_{\delta,\delta^*}$, the locations of its grid points depend only on

$$h = \delta + \delta^* = \frac{1}{M},$$

and hence are already fixed once M is fixed in Step 1. The following lemma gives a depth–width trade-off for the quantization module.

Lemma E.1 (Quantization). *Fix the integer M and $h = \frac{1}{M}$. For every $\delta^* \in (0, h)$, let $\delta := h - \delta^*$. Then, for every integer $s \in [L]$, there exists a quantization module $F_{1,s}^{(\delta^*)} : \mathbb{R}^{L\times d_{\mathrm{hid}}} \to \mathbb{R}^{L\times d_{\mathrm{hid}}}$, consisting of at most $D_{\mathrm{q}}(s) := M\left\lceil \frac{L}{s} \right\rceil$ feed-forward layers, each having hidden dimension at most $W_{\mathrm{q}}(s) := 3sd_{\mathrm{hid}}$, such that*

$$F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}, \qquad \forall \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}, \quad \forall \boldsymbol{X} \in S_{\boldsymbol{G}}.$$

Moreover, for every $\boldsymbol{X} \in [0,1]^{L\times d_{\mathrm{hid}}}$,

$$i \le \left[F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\right]_{i,j} \le i + 1, \qquad i \in [L], \quad j \in [d_{\mathrm{hid}}].$$

Consequently,

$$\left\|F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\right\|_F \le \left(d_{\mathrm{hid}} \sum_{i=1}^{L} (i+1)^2\right)^{1/2} \le (L+1)\sqrt{Ld_{\mathrm{hid}}}.$$

The proof of Lemma E.1 is provided in Appendix E.2. Lemma E.1 explicitly reveals a depth–width trade-off. At one extreme, choosing $s = L$ yields

$$D_{\mathrm{q}}(L) = M, \qquad W_{\mathrm{q}}(L) = 3Ld_{\mathrm{hid}},$$

which fully parallelizes all positional corrections at each quantization level. At the other extreme, choosing $s = 1$ yields

$$D_{\mathrm{q}}(1) = LM, \qquad W_{\mathrm{q}}(1) = 3d_{\mathrm{hid}}.$$

In particular, the latter construction has a feed-forward width that is a fixed constant multiple of the hidden dimension and is independent of both the number of quantization levels M and the approximation accuracy. In the remainder of the proof of Theorem 2.1, we use the constant-width choice $s = 1$, which aligns with the practical setting and write

$$F_1^{(\delta^*)} := F_{1,1}^{(\delta^*)}.$$

Thus $F_1^{(\delta^*)}$ consists of at most LM shared token-wise residual ReLU feed-forward layers, each with hidden dimension at most $3d_{\mathrm{hid}}$.

As before, the superscript (δ^*) emphasizes that the weights of the quantization module depend on δ^*. We do not fix δ^* at this stage. The contextual mapping F_2 and the value mapping F_3 depend only on the positional grid points and therefore only on $h = 1/M$. After constructing F_2 and F_3 and obtaining a uniform bound for the whole composition, we will choose $\delta^* > 0$ sufficiently small in Step 5 and then fix the corresponding module $F_1^{(\delta^*)}$.

Since for every $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$ and every $\boldsymbol{X} \in S_{\boldsymbol{G}}$, $F_1^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}$. Therefore, it remains to map each positional grid point $\boldsymbol{G} + \boldsymbol{E}$ to its prescribed value $F(\boldsymbol{G})$, which will be accomplished in Steps 3–4.

Step 3: Contextual mapping F_2. We next construct a single self-attention layer that maps all token vectors appearing in the positional grid to pairwise distinct context-dependent representations.

We first introduce some notation. Let $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be a finite collection of input sequences. For every $n \in [N]$, define its vocabulary set by $\mathcal{V}^{(n)} := \left\{ \boldsymbol{X}_{i,:}^{(n)} : i \in [L] \right\} \subset \mathbb{R}^{d_{\mathrm{hid}}}$, and define the whole vocabulary by $\mathcal{V} := \bigcup_{n=1}^{N} \mathcal{V}^{(n)}$.

Definition E.1 (Token-wise separatedness)**.** Let $r_{\min}, r_{\max}, \eta > 0$. The input sequences $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ are called token-wise $(r_{\min}, r_{\max}, \eta)$-separated if

1. for every $n \in [N]$ and $i \in [L]$, $r_{\min} \le \|\boldsymbol{X}_{i,:}^{(n)}\|_2 \le r_{\max}$;
2. for every $n, m \in [N]$ and $i, j \in [L]$ such that $\boldsymbol{X}_{i,:}^{(n)} \neq \boldsymbol{X}_{j,:}^{(m)}$, $\|\boldsymbol{X}_{i,:}^{(n)} - \boldsymbol{X}_{j,:}^{(m)}\|_2 \ge \eta$.

Definition E.2 (Contextual mapping)**.** Let $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$. A map $F : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}$ is called a (κ_1, κ_2)-contextual mapping for $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(N)}$ if

1. for every $n \in [N]$ and $i \in [L]$, $\|F(\boldsymbol{X}^{(n)})_{i,:}\|_2 \le \kappa_1$;
2. for every $n, m \in [N]$ and $i, j \in [L]$ satisfying $\mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}$ or $\boldsymbol{X}_{i,:}^{(n)} \neq \boldsymbol{X}_{j,:}^{(m)}$, we have $\|F(\boldsymbol{X}^{(n)})_{i,:} - F(\boldsymbol{X}^{(m)})_{j,:}\|_2 \ge \kappa_2$.

We can now state the contextual-mapping result needed in the proof.

Lemma E.2 (Contextual mapping)**.** *Let $\mathcal{G}_{\delta,\delta^*}^{p} = \{\boldsymbol{G} + \boldsymbol{E} : \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}\}$ be the positional grid constructed in Steps 1–2. Then there exist constants $\kappa_1, \kappa_2 > 0$ and a single one-head self-attention layer $F_2 : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}$, where $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V \in \mathbb{R}^{d_{\mathrm{hid}} \times d_{\mathrm{hid}}}$, such that F_2 is a (κ_1, κ_2)-contextual mapping for $\mathcal{G}_{\delta,\delta^*}^{p}$. Moreover, F_2, κ_1, and κ_2 depend on the positional grid only through $h = 1/M$ and hence are independent of the particular choice of $\delta^* \in (0, h)$.*

The proof of Lemma E.2 is provided in Appendix E.3. We now apply Lemma E.2. Since $|\mathcal{G}_{\delta,\delta^*}| = M^{Ld_{\text{hid}}}$, let $N_G := M^{Ld_{\text{hid}}}$, and enumerate the positional grid as

$$\mathcal{G}^p_{\delta,\delta^*} = \left\{\widetilde{\boldsymbol{G}}^{(1)}, \ldots, \widetilde{\boldsymbol{G}}^{(N_G)}\right\}, \qquad \widetilde{\boldsymbol{G}}^{(n)} = \boldsymbol{G}^{(n)} + \boldsymbol{E},$$

where $\boldsymbol{G}^{(n)} \in \mathcal{G}_{\delta,\delta^*}$. For every $n \in [N_G]$, define $\mathcal{V}^{(n)} := \left\{\widetilde{\boldsymbol{G}}^{(n)}_{i,:} : i \in [L]\right\}$. By Lemma E.2, for every $n \in [N_G]$ and $i \in [L]$, $\|F_2(\widetilde{\boldsymbol{G}}^{(n)})_{i,:}\|_2 \le \kappa_1$, and whenever $\mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}$ or $\widetilde{\boldsymbol{G}}^{(n)}_{i,:} \neq \widetilde{\boldsymbol{G}}^{(m)}_{j,:}$, we have $\|F_2(\widetilde{\boldsymbol{G}}^{(n)})_{i,:} - F_2(\widetilde{\boldsymbol{G}}^{(m)})_{j,:}\|_2 \ge \kappa_2$. We next show that the collection

$$\left\{F_2(\widetilde{\boldsymbol{G}}^{(n)})_{i,:} : n \in [N_G],\ i \in [L]\right\} \tag{E.2}$$

is pairwise distinct. Indeed, for every position $i \in [L]$, all possible token vectors at that position belong to the region $\mathcal{R}_i := [i, i+1-h]^{d_{\text{hid}}}$. The sets $\mathcal{R}_1, \ldots, \mathcal{R}_L$ are pairwise disjoint. In particular, every positional-grid sequence contains exactly one token in each $\mathcal{R}_i$. Consequently, if $\mathcal{V}^{(n)} = \mathcal{V}^{(m)}$, then the unique token contained in $\mathcal{R}_i$ must agree for every $i \in [L]$. Hence

$$\widetilde{\boldsymbol{G}}^{(n)}_{i,:} = \widetilde{\boldsymbol{G}}^{(m)}_{i,:}, \qquad i \in [L],$$

which implies $\widetilde{\boldsymbol{G}}^{(n)} = \widetilde{\boldsymbol{G}}^{(m)}$ and therefore $n = m$. Thus, if $n \neq m$, necessarily $\mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}$. On the other hand, when $n = m$ but $i \neq j$, the two token vectors belong to the disjoint sets $\mathcal{R}_i$ and $\mathcal{R}_j$, respectively, and hence

$$\widetilde{\boldsymbol{G}}^{(n)}_{i,:} \neq \widetilde{\boldsymbol{G}}^{(n)}_{j,:}.$$

Therefore, for every $(n, i) \neq (m, j)$, at least one of the two conditions in Lemma E.2 holds. By Lemma E.2,

$$\|F_2(\widetilde{\boldsymbol{G}}^{(n)})_{i,:} - F_2(\widetilde{\boldsymbol{G}}^{(m)})_{j,:}\|_2 \ge \kappa_2 > 0. \tag{E.3}$$

Hence all token vectors in equation E.2 are pairwise distinct. This property will allow us to construct the value mapping F_3 in the next step.

Step 4: Value mapping F_3. Recall from Step 3 that $N_G = M^{Ld_{\text{hid}}}$, and that the positional grid is enumerated as

$$\mathcal{G}^p_{\delta,\delta^*} = \left\{\widetilde{\boldsymbol{G}}^{(1)}, \ldots, \widetilde{\boldsymbol{G}}^{(N_G)}\right\}, \qquad \widetilde{\boldsymbol{G}}^{(n)} = \boldsymbol{G}^{(n)} + \boldsymbol{E},$$

where $\boldsymbol{G}^{(n)} \in \mathcal{G}_{\delta,\delta^*}$.

For every $n \in [N_G]$, define

$$\boldsymbol{Z}^{(n)} := F_2(\widetilde{\boldsymbol{G}}^{(n)}) = F_2(\boldsymbol{G}^{(n)} + \boldsymbol{E}),$$
$$\boldsymbol{Y}^{(n)} := F(\boldsymbol{G}^{(n)}).$$

By equation E.3, all token vectors in the collection

$$\left\{\boldsymbol{Z}^{(n)}_{i,:} : n \in [N_G],\ i \in [L]\right\}$$

are pairwise distinct. Moreover, by Lemma E.2,

$$B_Z := \max_{\substack{n \in [N_G] \\ i \in [L]}} \|\boldsymbol{Z}^{(n)}_{i,:}\|_2 \le \kappa_1. \tag{E.4}$$

Recall from Step 1 that

$$B_F := \max_{\boldsymbol{X} \in [0,1]^{L \times d_{\text{hid}}}} \|F(\boldsymbol{X})\|_F < \infty.$$

Hence

$$B_Y := \max_{\substack{n \in [N_G] \\ i \in [L]}} \|\boldsymbol{Y}^{(n)}_{i,:}\|_2 = \max_{\substack{n \in [N_G] \\ i \in [L]}} \|F(\boldsymbol{G}^{(n)})_{i,:}\|_2 \le B_F. \tag{E.5}$$

We use the following value-mapping lemma.

Lemma E.3 (Value mapping). *Let $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be a finite collection of sequences such that all their token vectors are pairwise distinct, i.e., $\boldsymbol{Z}^{(n)}_{i,:} \neq \boldsymbol{Z}^{(m)}_{j,:}$ whenever $(n,i) \neq (m,j)$. Let $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be arbitrary target sequences, and define*

$$B_Z := \max_{\substack{n \in [N] \\ i \in [L]}} \|\boldsymbol{Z}^{(n)}_{i,:}\|_2, \quad B_Y := \max_{\substack{n \in [N] \\ i \in [L]}} \|\boldsymbol{Y}^{(n)}_{i,:}\|_2.$$

Then there exists a value mapping module $F_3 : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}$, consisting of at most $NL + 1$ feed-forward layers, such that

$$F_3(\boldsymbol{Z}^{(n)}) = \boldsymbol{Y}^{(n)}, \qquad n \in [N].$$

Moreover, for every $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$,

$$\|F_3(\boldsymbol{X})\|_F \leq \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_Y + \frac{5}{3} B_Z\right). \tag{E.6}$$

The proof of Lemma E.3 is provided in Appendix E.4. Applying Lemma E.3 to the pairs

$$\left(\boldsymbol{Z}^{(n)}, \boldsymbol{Y}^{(n)}\right) = \left(F_2(\boldsymbol{G}^{(n)} + \boldsymbol{E}), F(\boldsymbol{G}^{(n)})\right), \qquad n \in [N_G],$$

we obtain a shared token-wise feed-forward module F_3 consisting of at most

$$N_G L + 1 = L M^{L d_{\mathrm{hid}}} + 1$$

feed-forward layers such that

$$F_3\left(F_2(\boldsymbol{G}^{(n)} + \boldsymbol{E})\right) = F(\boldsymbol{G}^{(n)}), \qquad n \in [N_G]. \tag{E.7}$$

Furthermore, combining equation E.6, equation E.4, and equation E.5, we obtain the global estimate

$$\begin{aligned} \|F_3(\boldsymbol{X})\|_F &\leq \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_Y + \frac{5}{3} B_Z\right) \\ &\leq \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_F + \frac{5}{3}\kappa_1\right), \quad \boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}. \end{aligned} \tag{E.8}$$

Finally, observe that both the interpolation inputs $F_2(\boldsymbol{G}^{(n)} + \boldsymbol{E})$ and the interpolation targets $F(\boldsymbol{G}^{(n)})$ depend only on the grid points

$$\boldsymbol{G}^{(n)} \in \{0, h, \dots, (M-1)h\}^{L \times d_{\mathrm{hid}}},$$

and hence only on $h = 1/M$. Therefore, the module F_3 can be fixed before δ^* is chosen and is independent of the particular choice of $\delta^* \in (0, h)$.

This completes the construction of the value mapping.

Step 5: Putting everything together. We now combine the constructions in Steps 1–4 and complete the approximation argument.

Recall that $D := L d_{\mathrm{hid}}, h = \frac{1}{M}$, where M was chosen in Step 1 so that $\sqrt{L d_{\mathrm{hid}}}\, h < \rho$. The positional grid $\mathcal{G}_{\delta,\delta^*} = \{0, h, \dots, (M-1)h\}^{L \times d_{\mathrm{hid}}}$ is therefore already fixed and depends only on h, and not on the particular decomposition $h = \delta + \delta^*$.

Recall also that the contextual mapping F_2 constructed in Step 3 and the value mapping F_3 constructed in Step 4 depend only on the grid points and hence only on h. In particular, both F_2 and F_3 are independent of the value of δ^* that will be chosen below.

We first derive a uniform bound for the whole composition outside the quantization cells. By Lemma E.1, for every $\delta^* \in (0, h)$ and every $\boldsymbol{X} \in [0,1]^{L \times d_{\mathrm{hid}}}$,

$$i \leq F_1^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})_{i,j} \leq i + 1, \qquad i \in [L], \quad j \in [d_{\mathrm{hid}}].$$

Define the compact set

$$\mathcal{K}_1 := \left\{ \boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}} : i \le \boldsymbol{X}_{i,j} \le i+1,\ i \in [L],\ j \in [d_{\text{hid}}] \right\}.$$

Then $F_1^{(\delta^*)}(\boldsymbol{X}+\boldsymbol{E}) \in \mathcal{K}_1$ for every $\boldsymbol{X} \in [0,1]^{L \times d_{\text{hid}}}$ and every $\delta^* \in (0,h)$. Since F_2 is continuous and $\mathcal{K}_1$ is compact, the constant

$$B_2 := \max_{\boldsymbol{X} \in \mathcal{K}_1} \|F_2(\boldsymbol{X})\|_F < \infty$$

is well defined. Notice that B_2 is independent of δ^*. Furthermore, by equation E.8, for every $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}}$,

$$\|F_3(\boldsymbol{X})\|_F \le \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_F + \frac{5}{3}\kappa_1\right). \tag{E.9}$$

Define

$$B_g := B_2 + \sqrt{L}\left(B_F + \frac{5}{3}\kappa_1\right).$$

Both B_2 and B_g are finite and independent of δ^*. We now choose δ^*. Take

$$0 < \delta^* < \min\left\{ h,\ \frac{3\varepsilon^2}{16DM(1+B_F)^2},\ \frac{\varepsilon^2}{4DM(1+B_g)^2} \right\}, \tag{E.10}$$

For this choice of δ^*, fix the corresponding quantization module

$$F_1 := F_1^{(\delta^*)}$$

constructed in Step 2, and let F' denote the piecewise constant function defined in Step 1 with this choice of δ and δ^*. Define

$$\Omega_{\text{in}} := \bigcup_{\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}} S_G, \quad \Omega_{\text{out}} := [0,1]^{L \times d_{\text{hid}}} \setminus \Omega_{\text{in}}.$$

We refer to Ω_{in} as the quantized region and Ω_{out} as the gap region. Since $\delta < h$, the quantization cells $\{S_G\}_{\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}}$ are pairwise disjoint. Each cell has Lebesgue measure $\mu(S_{\boldsymbol{G}}) = \delta^D$, and there are $|\mathcal{G}_{\delta,\delta^*}| = M^D$ such cells. Hence

$$\mu(\Omega_{\text{in}}) = M^D \delta^D = (M\delta)^D = (1 - M\delta^*)^D,$$

where we used the fact that $Mh = 1$ and $\delta = h - \delta^*$. Consequently,

$$\mu(\Omega_{\text{out}}) = 1 - (1 - M\delta^*)^D \le DM\delta^*,$$

where the last inequality follows from $1 - (1-x)^D \le Dx,\ x \in [0,1]$. By the choice of δ^* in equation E.10, we therefore have

$$\mu(\Omega_{\text{out}}) < \min\left\{ \frac{3\varepsilon^2}{16(1+B_F)^2},\ \frac{\varepsilon^2}{4(1+B_g)^2} \right\}.$$

We first bound the approximation error between F and F'. Recall from Step 1 that, for every $\boldsymbol{X} \in \Omega_{\text{in}}$, $\|F(\boldsymbol{X}) - F'(\boldsymbol{X})\|_F < \frac{\varepsilon}{4}$. On the other hand, by the definition of F', $F'(\boldsymbol{X}) = 0, \boldsymbol{X} \in \Omega_{\text{out}}$, and hence

$$\|F(\boldsymbol{X}) - F'(\boldsymbol{X})\|_F = \|F(\boldsymbol{X})\|_F \le B_F, \qquad \boldsymbol{X} \in \Omega_{\text{out}}.$$

Therefore,

$$\begin{aligned}
\|F - F'\|_{L^2}^2 &\le \int_{\Omega_{\text{in}}} \|F(\boldsymbol{X}) - F'(\boldsymbol{X})\|_F^2 \, dX + \int_{\Omega_{\text{out}}} \|F(\boldsymbol{X}) - F'(\boldsymbol{X})\|_F^2 \, dX \\
&< \frac{\varepsilon^2}{16}\mu(\Omega_{\text{in}}) + B_F^2 \mu(\Omega_{\text{out}}) \le \frac{\varepsilon^2}{16} + B_F^2 \frac{3\varepsilon^2}{16(1+B_F)^2} < \frac{\varepsilon^2}{16} + \frac{3\varepsilon^2}{16} = \frac{\varepsilon^2}{4}.
\end{aligned}$$

Thus, $\|F - F'\|_{L^2} < \frac{\varepsilon}{2}$. We next compare F' with the Transformer constructed in Steps 2–4. Define

$$T(\boldsymbol{X}) := F_3 \circ F_2 \circ F_1(\boldsymbol{X} + \boldsymbol{E}).$$

Fix any $\boldsymbol{X} \in \Omega_{\text{in}}$. Then there exists a unique $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$ such that $\boldsymbol{X} \in S_G$. By the quantization property in Step 2, $F_1(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}$. By the interpolation property equation E.7 established in Step 4, $F_3(F_2(\boldsymbol{G} + \boldsymbol{E})) = F(\boldsymbol{G})$. Consequently,

$$T(\boldsymbol{X}) = F(\boldsymbol{G}) = F'(\boldsymbol{X}), \qquad \boldsymbol{X} \in \Omega_{\text{in}}.$$

It remains to control the output of T on Ω_{out}. For every $\boldsymbol{X} \in [0,1]^{L \times d_{\text{hid}}}$, we have $F_1(\boldsymbol{X} + \boldsymbol{E}) \in \mathcal{K}_1$. Therefore, by the definition of B_2, $\|F_2(F_1(\boldsymbol{X} + \boldsymbol{E}))\|_F \le B_2$. Using equation E.9, we obtain

$$\begin{aligned}\|T(\boldsymbol{X})\|_F &= \|F_3\left(F_2(F_1(\boldsymbol{X} + \boldsymbol{E}))\right)\|_F \\ &\le \|F_2(F_1(\boldsymbol{X} + \boldsymbol{E}))\|_F + \sqrt{L}\left(B_F + \frac{5}{3}\kappa_1\right) \le B_g.\end{aligned}$$

Since $F' = T$ on Ω_{in} and $F' = 0$ on Ω_{out}, we obtain

$$\|F' - T\|_{L^2}^2 \le \int_{\Omega_{\text{out}}} \|T(\boldsymbol{X})\|_F^2 \, dX \le B_g^2 \mu(\Omega_{\text{out}}) < B_g^2 \frac{\varepsilon^2}{4(1+B_g)^2} < \frac{\varepsilon^2}{4}.$$

Hence $\|F' - T\|_{L^2} < \frac{\varepsilon}{2}$. Finally, by the triangle inequality,

$$\|F - T\|_{L^2} \le \|F - F'\|_{L^2} + \|F' - T\|_{L^2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

By construction, T consists of a positional encoding, a finite shared token-wise feed-forward module F_1, a single one-head self-attention layer F_2 satisfying

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V,$$

and a finite shared token-wise feed-forward module F_3. Therefore,

$$T \in \mathcal{T}_{Q=K=V},$$

which completes the proof of Theorem 2.1.

□

E.2 Proof of Lemma E.1

Lemma E.1 (Quantization; Restatement). *Fix the integer M and $h = \frac{1}{M}$. For every $\delta^* \in (0, h)$, let $\delta := h - \delta^*$. Then, for every integer $s \in [L]$, there exists a quantization module $F_{1,s}^{(\delta^*)} : \mathbb{R}^{L \times d_{\text{hid}}} \to \mathbb{R}^{L \times d_{\text{hid}}}$, consisting of at most $D_{\text{q}}(s) := M \left\lceil \frac{L}{s} \right\rceil$ feed-forward layers, each having hidden dimension at most $W_{\text{q}}(s) := 3sd_{\text{hid}}$, such that*

$$F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}, \qquad \forall \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}, \quad \forall \boldsymbol{X} \in S_{\boldsymbol{G}}.$$

Moreover, for every $\boldsymbol{X} \in [0,1]^{L \times d_{\text{hid}}}$,

$$i \le \left[F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\right]_{i,j} \le i + 1, \qquad i \in [L], \quad j \in [d_{\text{hid}}].$$

Consequently,

$$\begin{aligned}\left\|F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\right\|_F &\le \left(d_{\text{hid}} \sum_{i=1}^{L} (i+1)^2\right)^{1/2} \\ &\le (L+1)\sqrt{Ld_{\text{hid}}}.\end{aligned}$$

Proof of Lemma E.1. Fix any $s \in [L], \delta^* \in (0, h)$, and let $\delta := h - \delta^*$. Define

$$\psi(t) := -\sigma_{\mathrm{R}}(t) + \frac{h}{\delta^*}\sigma(t-\delta) - \frac{\delta}{\delta^*}\sigma(t-h). \tag{E.11}$$

Since $h = \delta + \delta^*$, a direct calculation gives

$$\psi(t) = \begin{cases} 0, & t \le 0, \\ -t, & 0 \le t \le \delta, \\ \dfrac{\delta}{\delta^*}(t-h), & \delta \le t \le h, \\ 0, & t \ge h. \end{cases}$$

In particular, we have $\mathrm{supp}(\psi) \subseteq [0, h]$. For any $a \in \mathbb{R}$, define

$$R_a(t) := t + \psi(t-a).$$

Then $R_a(t) = a,\ t \in [a, a+\delta]$, while $R_a(t) = t, t \notin [a, a+h]$, and importantly, we have

$$R_a([a, a+h]) \subseteq [a, a+h].$$

For later reference, we record the latter property as

$$R_a([a, a+h]) \subseteq [a, a+h]. \tag{E.12}$$

We next divide the L token positions into groups of size at most s. Let $R_s := \left\lceil \frac{L}{s} \right\rceil$, and define

$$P_r := \{(r-1)s + 1, \dots, \min\{rs, L\}\}, \qquad r \in [R_s].$$

Then $[L] = \bigcup_{r=1}^{R_s} P_r$, the sets $P_1, \dots, P_{R_s}$ are pairwise disjoint, and $|P_r| \le s$. For every quantization level $k \in \{0, \dots, M-1\}$ and every position group $r \in [R_s]$, define

$$q_{k,r}(t) := t + \sum_{p \in P_r} \psi(t - (p + kh)). \tag{E.13}$$

Fix any token position $i \in [L]$ and let $r(i)$ denote the unique group such that $i \in P_{r(i)}$. For $t \in [i, i+1]$, if $r \ne r(i)$, then every correction term in equation E.13 vanishes, and hence

$$q_{k,r}(t) = t. \tag{E.14}$$

On the other hand, when $r = r(i)$, all terms vanish except possibly the term indexed by $p = i$, so that $q_{k,r(i)}(t) = R_{i+kh}(t),\ t \in [i, i+1]$. Consequently, by equation E.12,

$$q_{k,r}([i, i+1]) \subseteq [i, i+1] \tag{E.15}$$

for every $i \in [L]$, $k \in \{0, \dots, M-1\}$, and $r \in [R_s]$. We next lift $q_{k,r}$ coordinate-wise to a token-wise map

$$\phi_{k,r} : \mathbb{R}^{d_{\mathrm{hid}}} \to \mathbb{R}^{d_{\mathrm{hid}}},$$

defined by

$$\phi_{k,r}(\boldsymbol{z})_j := q_{k,r}(\boldsymbol{z}_j), \qquad j \in [d_{\mathrm{hid}}].$$

Using equation E.11, we may write

$$\begin{aligned} &\phi_{k,r}(\boldsymbol{z}) \\ &= z + \sum_{j=1}^{d_{\mathrm{hid}}} \sum_{p \in P_r} e_j \Bigg[\quad -\sigma\left(\boldsymbol{z}_j - p - kh\right) + \frac{h}{\delta^*}\sigma\left(\boldsymbol{z}_j - p - kh - \delta\right) - \frac{\delta}{\delta^*}\sigma\left(\boldsymbol{z}_j - p - (k+1)h\right) \Bigg]. \end{aligned}$$

Thus $\phi_{k,r}$ is exactly representable by a residual ReLU feed-forward layer with at most

$$3|P_r| d_{\mathrm{hid}} \le 3 s d_{\mathrm{hid}}$$

hidden neurons. Applying $\phi_{k,r}$ row-wise defines a token-wise feed-forward layer

$$\Phi_{k,r} : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}},$$

where

$$\Phi_{k,r}(\boldsymbol{Z})_{i,:} = \phi_{k,r}(\boldsymbol{Z}_{i,:}), \qquad i \in [L].$$

Define the complete quantization module by processing all position groups at each quantization level:

$$\begin{aligned} F_{1,s}^{(\delta^*)} :=& \Phi_{M-1,R_s} \circ \cdots \circ \Phi_{M-1,1} \circ \Phi_{M-2,R_s} \circ \cdots \circ \Phi_{M-2,1} \\ & \circ \cdots \circ \Phi_{0,R_s} \circ \cdots \circ \Phi_{0,1}. \end{aligned}$$

This module contains at most $MR_s = M\left\lceil \frac{L}{s} \right\rceil$ feed-forward layers, each having hidden dimension at most $3sd_{\text{hid}}$. We now prove the exact quantization property. Fix $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$, $\boldsymbol{X} \in S_G$, and set $Z := \boldsymbol{X} + \boldsymbol{E}$. For every $i \in [L]$ and $j \in [d_{\text{hid}}]$, there exists a unique

$$m_{i,j} \in \{0, \ldots, M-1\}$$

such that $G_{i,j} = m_{i,j}h$. Since $\boldsymbol{X} \in S_{\boldsymbol{G}}$,

$$\boldsymbol{Z}_{i,j} \in [i + m_{i,j}h,\ i + m_{i,j}h + \delta]. \tag{E.16}$$

Let $r(i)$ be the unique group containing position i. For every $k < m_{i,j}$, the layer $\Phi_{k,r(i)}$ leaves the coordinate $Z_{i,j}$ unchanged because its value lies to the right of

$$[i + kh, i + (k+1)h],$$

while all layers $\Phi_{k,r}$ with $r \neq r(i)$ act as the identity on position i by equation E.14.

At $k = m_{i,j}$, the layer $\Phi_{m_{i,j},r(i)}$ maps the interval in equation E.16 exactly to its left endpoint:

$$q_{m_{i,j},r(i)}(\boldsymbol{Z}_{i,j}) = i + m_{i,j}h.$$

For every $k > m_{i,j}$, the value $i + m_{i,j}h$ lies to the left of the support of the k-th correction and is therefore never modified again. Hence

$$[F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})]_{i,j} = i + m_{i,j}h.$$

Since $(\boldsymbol{G} + \boldsymbol{E})_{i,j} = G_{i,j} + i = m_{i,j}h + i$, we conclude that

$$F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}.$$

Finally, we establish the uniform boundedness. For every $\boldsymbol{X} \in [0,1]^{L \times d_{\text{hid}}}$,

$$(\boldsymbol{X} + \boldsymbol{E})_{i,j} \in [i, i+1].$$

By equation E.15, every layer $\Phi_{k,r}$ preserves these intervals. Therefore,

$$i \leq [F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})]_{i,j} \leq i + 1$$

for every $i \in [L]$ and $j \in [d_{\text{hid}}]$. Consequently,

$$\|F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\|_F^2 \leq d_{\text{hid}} \sum_{i=1}^{L} (i+1)^2 \leq L d_{\text{hid}} (L+1)^2,$$

and hence

$$\|F_{1,s}^{(\delta^*)}(\boldsymbol{X} + \boldsymbol{E})\|_F \leq (L+1)\sqrt{L d_{\text{hid}}}.$$

This completes the proof. □

E.3 Proof of Lemma E.2

Throughout this subsection, all standalone vectors are regarded as column vectors. In particular, every token belongs to $\mathbb{R}^{d_{\mathrm{hid}}}$. For a sequence matrix $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$, we write

$$x_i(\boldsymbol{X}) := (\boldsymbol{X}_{i,:})^\top \in \mathbb{R}^{d_{\mathrm{hid}}}, \qquad i \in [L].$$

Thus the i-th row of $\boldsymbol{X}$ is $x_i(\boldsymbol{X})^\top$; the matrix representation stores tokens row-wise, while the tokens themselves are always column vectors. Whenever a row vector is required by a matrix operation, we write the transpose explicitly.

We first introduce several auxiliary lemmas.

Lemma E.4 ((Park et al., 2021)). *Let $\mathcal{S} \subset \mathbb{R}^{d_{\mathrm{hid}}}$ be a finite set with $|\mathcal{S}| = K$. Then there exists a unit vector $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{hid}}}$ such that, for every distinct $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$,*

$$\frac{1}{K^2}\sqrt{\frac{8}{\pi d_{\mathrm{hid}}}}\|\boldsymbol{x}-\boldsymbol{y}\|_2 \le |\boldsymbol{v}^\top(\boldsymbol{x}-\boldsymbol{y})| \le \|\boldsymbol{x}-\boldsymbol{y}\|_2.$$

Lemma E.5 (Modified from (Kajitsuka & Sato, 2023)). *Let $r_{\min}, r_{\max}, \eta > 0$. Given any $(r_{\min}, r_{\max}, \eta)$-separated vocabulary $\mathcal{V} \subset \mathbb{R}^d$ with $|\mathcal{V}| < \infty$. Then, for every $\Delta_0 > 0$, there exist a unit vector $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{hid}}}$ and a vector $\boldsymbol{u} \in \mathbb{R}^{d_{\mathrm{hid}}}$ such that, with*

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{v}\boldsymbol{u}^\top,$$

we have

$$\left|(\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_a^\top \boldsymbol{W}_K)^\top - (\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_b^\top \boldsymbol{W}_K)^\top\right| \ge \Delta_0 \tag{E.17}$$

for every $\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c \in \mathcal{V}$ with $\boldsymbol{x}_a \neq \boldsymbol{x}_b$. Moreover, the same vector $\boldsymbol{v}$ satisfies

$$\frac{\|\boldsymbol{x}\|_2}{(|\mathcal{V}|+1)^2\sqrt{d_{\mathrm{hid}}}} \le |\boldsymbol{v}^\top \boldsymbol{x}| \le \|\boldsymbol{x}\|_2, \qquad \boldsymbol{x} \in \mathcal{V}. \tag{E.18}$$

One may choose $\boldsymbol{u}$ such that

$$\|\boldsymbol{u}\|_2^2 = \frac{(|\mathcal{V}|+1)^4 d_{\mathrm{hid}} \Delta_0}{\eta r_{\min}}. \tag{E.19}$$

Proof of Lemma E.5. Apply Lemma E.4 to $\mathcal{V} \cup \{0\}$. Since $r_{\min} > 0$, we have $0 \notin \mathcal{V}$, and hence $|\mathcal{V} \cup \{0\}| = |\mathcal{V}| + 1$. Therefore, there exists a unit vector $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{hid}}}$ such that

$$|\boldsymbol{v}^\top(\boldsymbol{x}-\boldsymbol{y})| \ge \frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{(|\mathcal{V}|+1)^2\sqrt{d_{\mathrm{hid}}}}$$

for every distinct $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} \cup \{0\}$. Taking $y = 0$ gives equation E.18. Choose u according to equation E.19 and set

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{v}\boldsymbol{u}^\top.$$

For any $\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c \in \mathcal{V}$ with $\boldsymbol{x}_a \neq \boldsymbol{x}_b$,

$$\begin{aligned}
&\left|(\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_a^\top \boldsymbol{W}_K)^\top - (\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_b^\top \boldsymbol{W}_K)^\top\right| \\
&= |\boldsymbol{v}^\top \boldsymbol{x}_c|\,|\boldsymbol{v}^\top(\boldsymbol{x}_a - \boldsymbol{x}_b)|\,\|\boldsymbol{u}\|_2^2 \\
&\ge \frac{r_{\min}}{(|\mathcal{V}|+1)^2\sqrt{d_{\mathrm{hid}}}} \cdot \frac{\eta}{(|\mathcal{V}|+1)^2\sqrt{d_{\mathrm{hid}}}} \cdot \frac{(|\mathcal{V}|+1)^4 d_{\mathrm{hid}} \Delta_0}{\eta r_{\min}} \\
&= \Delta_0.
\end{aligned}$$

This proves equation E.17.

□

We next use the following separation property of the Boltzmann operator. For $a \in \mathbb{R}^L$, define

$$\mathrm{Boltz}(\boldsymbol{a}) := \mathrm{Softmax}(\boldsymbol{a}^\top)\boldsymbol{a}.$$

Lemma E.6 ((Kajitsuka & Sato, 2023)). *Let $L \geq 2$ and let $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(K)} \in \mathbb{R}^L$. Suppose that, for some $R > 0$, $|\boldsymbol{a}_i^{(n)}| < R, n \in [K],\ i \in [L]$, that every $a^{(n)}$ has no duplicate entries, and that any two unequal scalar entries appearing in the collection satisfy $a_i^{(n)} \neq a_j^{(m)} \Rightarrow |a_i^{(n)} - a_j^{(m)}| \geq 2\log L + 3$. Then, whenever $\left\{a_i^{(n)} : i \in [L]\right\} \neq \left\{a_i^{(m)} : i \in [L]\right\}$, we have*

$$\left|\mathrm{Boltz}(a^{(n)}) - \mathrm{Boltz}(a^{(m)})\right| > (\log L)^2 e^{-2R}.$$

The above formulation makes explicit the permutation invariance of the Boltzmann operator: two score vectors that differ only by a permutation have the same Boltzmann output, and hence separation is asserted only when their sets of score values are different.

We are now ready to prove the general contextual-mapping result used for the positional grid.

Lemma E.7 (One-head contextual mapping). *Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be token-wise $(r_{\min}, r_{\max}, \eta)$-separated input sequences. For each $n \in [N]$, write $x_i^{(n)} := x_i(\boldsymbol{X}^{(n)}) = (\boldsymbol{X}_{i,:}^{(n)})^\top,\ i \in [L]$, and define $\mathcal{V}^{(n)} := \{x_i^{(n)} : i \in [L]\},\ \mathcal{V} := \bigcup_{n=1}^N \mathcal{V}^{(n)}$. Assume additionally that each sequence contains no duplicate token, i.e., $x_i^{(n)} \neq x_j^{(n)},\ n \in [N],\ i \neq j$. Then there exist $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O \in \mathbb{R}^{d_{\mathrm{hid}} \times d_{\mathrm{hid}}}$ with*

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$$

such that the one-head self-attention layer

$$F_2(\boldsymbol{X}) := \boldsymbol{X} + \mathrm{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\right)(\boldsymbol{X}\boldsymbol{W}_V)\boldsymbol{W}_O$$

is a (κ_1, κ_2)-contextual mapping for $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}$ for some $\kappa_1, \kappa_2 > 0$.

More precisely, when $L \geq 2$, define

$$\Delta_0 := 2\log L + 3, \quad U^2 := \frac{(|\mathcal{V}|+1)^4 d_{\mathrm{hid}} \Delta_0}{\eta r_{\min}},$$
$$C := U^2 r_{\max}^2, \quad R := C + 1. \tag{E.20}$$

Then we may take

$$\kappa_1 := r_{\max} + \frac{\eta}{4}, \quad \kappa_{\mathrm{ctx}} := \frac{(\log L)^2 e^{-2R}\eta}{4C}, \quad \kappa_2 := \min\left\{\frac{\eta}{2}, \kappa_{\mathrm{ctx}}\right\} > 0. \tag{E.21}$$

Proof. When $L = 1$, the claim is immediate: take $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V = \boldsymbol{W}_O = \boldsymbol{0}$, so that $F_2(\boldsymbol{X}) = \boldsymbol{X}$. Since a one-token vocabulary is different if and only if its unique token is different, the token-wise separatedness assumption directly gives a contextual mapping. Hence, in the remainder of the proof we assume $L \geq 2$.

Let Δ_0, U, C, and R be defined by equation E.20. Applying Lemma E.5 to the whole vocabulary $\mathcal{V}$ with separation level Δ_0, there exist a unit vector $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{hid}}}$ and $\boldsymbol{u} \in \mathbb{R}^{d_{\mathrm{hid}}}$ with $\|\boldsymbol{u}\|_2^2 = U^2$ such that $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{v}\boldsymbol{u}^\top$ satisfies

$$\left|(\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_a^\top \boldsymbol{W}_K)^\top - (\boldsymbol{x}_c^\top \boldsymbol{W}_Q)(\boldsymbol{x}_b^\top \boldsymbol{W}_K)^\top\right| \geq \Delta_0 \tag{E.22}$$

for every $\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c \in \mathcal{V}$ with $\boldsymbol{x}_a \neq \boldsymbol{x}_b$. Set

$$\boldsymbol{W}_V := \boldsymbol{v}\boldsymbol{u}^\top,$$

so that

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V.$$

Choose any vector $\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{hid}}}$ satisfying

$$\|\boldsymbol{w}\|_2 = \frac{\eta}{4r_{\max}},$$

and define

$$\boldsymbol{W}_O := \frac{\boldsymbol{u}\boldsymbol{w}^\top}{\|\boldsymbol{u}\|_2^2}.$$

Then

$$\boldsymbol{W}_O^\top \boldsymbol{u} = \boldsymbol{w}, \qquad \|\boldsymbol{W}_O^\top \boldsymbol{u}\|_2 = \frac{\eta}{4r_{\max}}.$$

For convenience, denote the self-attention contribution without the skip connection by

$$\mathrm{Attn}(\boldsymbol{X}) := \mathrm{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\right)(\boldsymbol{X}\boldsymbol{W}_V)\boldsymbol{W}_O,$$

so that

$$F_2(\boldsymbol{X}) = \boldsymbol{X} + \mathrm{Attn}(\boldsymbol{X}).$$

For $n \in [N]$ and $i \in [L]$, define the corresponding column-vector attention contribution and output token by

$$\boldsymbol{c}_i^{(n)} := \mathrm{Attn}(\boldsymbol{X}^{(n)})_{i,:}^\top, \qquad \boldsymbol{f}_i^{(n)} := F_2(\boldsymbol{X}^{(n)})_{i,:}^\top = \boldsymbol{x}_i^{(n)} + \boldsymbol{c}_i^{(n)}.$$

We first obtain a uniform bound on the attention contribution. Fix $n \in [N]$ and $i \in [L]$, and let $\boldsymbol{p}^{(n,i)} \in \mathbb{R}^L$ be defined so that $(\boldsymbol{p}^{(n,i)})^\top$ is the i-th row of the attention weight matrix. Since $\boldsymbol{p}^{(n,i)}$ is a probability vector,

$$\begin{aligned}
\|\boldsymbol{c}_i^{(n)}\|_2 &= \left\| \sum_{s=1}^{L} \boldsymbol{p}_s^{(n,i)} (\boldsymbol{v}^\top \boldsymbol{x}_s^{(n)}) \boldsymbol{W}_O^\top \boldsymbol{u} \right\|_2 \\
&\le \sum_{s=1}^{L} \boldsymbol{p}_s^{(n,i)} |\boldsymbol{v}^\top \boldsymbol{x}_s^{(n)}| \|\boldsymbol{W}_O^\top \boldsymbol{u}\|_2 \\
&\le r_{\max} \frac{\eta}{4r_{\max}} = \frac{\eta}{4}.
\end{aligned} \tag{E.23}$$

Consequently,

$$\begin{aligned}
\|\boldsymbol{f}_i^{(n)}\|_2 &\le \|\boldsymbol{x}_i^{(n)}\|_2 + \|\boldsymbol{c}_i^{(n)}\|_2 \\
&\le r_{\max} + \frac{\eta}{4} = \kappa_1.
\end{aligned} \tag{E.24}$$

It remains to establish separation. We consider two cases.

Case 1: The query tokens are different. Suppose

$$\boldsymbol{x}_i^{(n)} \neq \boldsymbol{x}_j^{(m)}.$$

By token-wise separatedness and equation E.23,

$$\begin{aligned}
&\|\boldsymbol{f}_i^{(n)} - \boldsymbol{f}_j^{(m)}\|_2 \\
&\ge \|\boldsymbol{x}_i^{(n)} - \boldsymbol{x}_j^{(m)}\|_2 - \|\boldsymbol{c}_i^{(n)}\|_2 - \|\boldsymbol{c}_j^{(m)}\|_2 \\
&\ge \eta - \frac{\eta}{4} - \frac{\eta}{4} = \frac{\eta}{2}.
\end{aligned} \tag{E.25}$$

Case 2: The query tokens are identical but the contexts are different. Suppose

$$\boldsymbol{x}_i^{(n)} = \boldsymbol{x}_j^{(m)} =: \boldsymbol{x}, \qquad \mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}. \tag{E.26}$$

Define the corresponding score vectors $\boldsymbol{a}^{(n)}, \boldsymbol{a}^{(m)} \in \mathbb{R}^L$ by

$$\begin{aligned}
\boldsymbol{a}_s^{(n)} &:= (\boldsymbol{x}^\top \boldsymbol{W}_Q)((\boldsymbol{x}_s^{(n)})^\top \boldsymbol{W}_K)^\top, \qquad s \in [L], \\
\boldsymbol{a}_s^{(m)} &:= (\boldsymbol{x}^\top \boldsymbol{W}_Q)((\boldsymbol{x}_s^{(m)})^\top \boldsymbol{W}_K)^\top, \qquad s \in [L].
\end{aligned}$$

Since

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{v}\boldsymbol{u}^\top,$$

we have

$$\boldsymbol{a}_s^{(n)} = (\boldsymbol{v}^\top \boldsymbol{x})(\boldsymbol{v}^\top \boldsymbol{x}_s^{(n)})U^2, \tag{E.27}$$

and similarly for $\boldsymbol{a}_s^{(m)}$. By equation E.22, whenever two key tokens are different, their corresponding score values differ by at least Δ_0. Since

$$\Delta_0 = 2\log L + 3 \geq 2\log L + 3,$$

any two unequal score values appearing in $\boldsymbol{a}^{(n)}$ and $\boldsymbol{a}^{(m)}$ satisfy the separation condition in Lemma E.6. Moreover, because each input sequence contains no duplicate tokens, both score vectors contain no duplicate entries.

We also have

$$\begin{aligned} |\boldsymbol{a}_s^{(n)}| &\leq |\boldsymbol{v}^\top \boldsymbol{x}||\boldsymbol{v}^\top \boldsymbol{x}_s^{(n)}|U^2 \\ &\leq r_{\max}^2 U^2 = C < R, \end{aligned}$$

and the same estimate holds for $\boldsymbol{a}^{(m)}$. Next, the map

$$\begin{aligned} T_x : \mathcal{V} &\to \mathbb{R}, \\ \boldsymbol{y} &\mapsto (\boldsymbol{x}^\top \boldsymbol{W}_Q)(\boldsymbol{y}^\top \boldsymbol{W}_K)^\top \end{aligned}$$

is injective on $\mathcal{V}$ by equation E.22. Therefore,

$$\mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}$$

implies

$$\left\{\boldsymbol{a}_s^{(n)} : s \in [L]\right\} \neq \left\{\boldsymbol{a}_s^{(m)} : s \in [L]\right\}.$$

Lemma E.6 thus yields

$$\left|\mathrm{Boltz}(\boldsymbol{a}^{(n)}) - \mathrm{Boltz}(\boldsymbol{a}^{(m)})\right| > (\log L)^2 e^{-2R}.$$

Define

$$\begin{aligned} b_n &:= \mathrm{Softmax}((a^{(n)})^\top)\boldsymbol{X}^{(n)}\boldsymbol{v}, \\ b_m &:= \mathrm{Softmax}((a^{(m)})^\top)\boldsymbol{X}^{(m)}\boldsymbol{v}. \end{aligned}$$

Using equation E.27,

$$\begin{aligned} \mathrm{Boltz}(\boldsymbol{a}^{(n)}) &= (\boldsymbol{v}^\top \boldsymbol{x})U^2 b_n, \\ \mathrm{Boltz}(\boldsymbol{a}^{(m)}) &= (\boldsymbol{v}^\top \boldsymbol{x})U^2 b_m. \end{aligned}$$

Consequently,

$$\begin{aligned} |b_n - b_m| &= \frac{\left|\mathrm{Boltz}(\boldsymbol{a}^{(n)}) - \mathrm{Boltz}(\boldsymbol{a}^{(m)})\right|}{|\boldsymbol{v}^\top \boldsymbol{x}|U^2} \\ &> \frac{(\log L)^2 e^{-2R}}{r_{\max} U^2}, \end{aligned}$$

where we used $|\boldsymbol{v}^\top \boldsymbol{x}| \leq \|\boldsymbol{x}\|_2 \leq r_{\max}$. Since the residual terms cancel in equation E.26,

$$\begin{aligned} \|\boldsymbol{f}_i^{(n)} - \boldsymbol{f}_j^{(m)}\|_2 &= \|\boldsymbol{c}_i^{(n)} - \boldsymbol{c}_j^{(m)}\|_2 \\ &= |\boldsymbol{b}_n - \boldsymbol{b}_m|\|\boldsymbol{W}_O^\top \boldsymbol{u}\|_2 > \frac{(\log L)^2 e^{-2R}}{r_{\max} U^2} \cdot \frac{\eta}{4 r_{\max}} = \frac{(\log L)^2 e^{-2R}\eta}{4U^2 r_{\max}^2} \\ &= \kappa_{\mathrm{ctx}}. \end{aligned} \tag{E.28}$$

Combining equation E.25 and equation E.28, we obtain

$$\|\boldsymbol{f}_i^{(n)} - \boldsymbol{f}_j^{(m)}\|_2 \geq \kappa_2$$

whenever

$$\boldsymbol{x}_i^{(n)} \neq \boldsymbol{x}_j^{(m)} \qquad \text{or} \qquad \mathcal{V}^{(n)} \neq \mathcal{V}^{(m)}.$$

Together with equation E.24, this proves that F_2 is a (κ_1, κ_2)-contextual mapping. □

We can now prove Lemma E.2.

Proof of Lemma E.2. Recall that $h = \frac{1}{M}$ and $\mathcal{G}^p_{\delta,\delta^*} = \{\boldsymbol{G} + \boldsymbol{E} : \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}\}$. Since the locations of the grid points are $\{0, h, \dots, (M-1)h\}$, the positional grid depends only on h, and not on the particular decomposition $h = \delta + \delta^*$. The number of positional-grid sequences is $N_{\boldsymbol{G}} = |\mathcal{G}^p_{\delta,\delta^*}| = M^{Ld_{\text{hid}}}$. We next characterize its whole token vocabulary. For any $\widetilde{\boldsymbol{G}} \in \mathcal{G}^p_{\delta,\delta^*}$ and any position $i \in [L]$, write $\widetilde{\boldsymbol{g}}_i := (\widetilde{\boldsymbol{G}}_{i,:})^\top$. Then $\widetilde{g}_i$ has the form

$$\widetilde{\boldsymbol{g}}_i = i\mathbf{1}_{d_{\text{hid}}} + \boldsymbol{z}, \qquad \boldsymbol{z} \in \{0, h, \dots, (M-1)h\}^{d_{\text{hid}}}.$$

Hence there are exactly $M^{d_{\text{hid}}}$ possible token vectors at that position. Moreover, token sets corresponding to different positions are disjoint. Therefore, the whole vocabulary satisfies $|\mathcal{V}| = LM^{d_{\text{hid}}}$.

We now verify token-wise separatedness. For every position $i \in [L]$ and coordinate $j \in [d_{\text{hid}}]$,

$$i \le (\widetilde{\boldsymbol{g}}_i)_j \le i + 1 - h. \tag{E.29}$$

Hence, $\sqrt{d_{\text{hid}}} \le \|\widetilde{\boldsymbol{g}}_i\|_2 < \sqrt{d_{\text{hid}}}(L+1)$. Thus we may choose

$$r_{\min} := \sqrt{d_{\text{hid}}}, \qquad r_{\max} := \sqrt{d_{\text{hid}}}(L+1).$$

Consider now two positional-grid sequences $\widetilde{\boldsymbol{G}}^{(n)}, \widetilde{\boldsymbol{G}}^{(m)}$ and write

$$\widetilde{\boldsymbol{g}}^{(n)}_i := (\widetilde{\boldsymbol{G}}^{(n)}_{i,:})^\top, \qquad \widetilde{\boldsymbol{g}}^{(m)}_j := (\widetilde{\boldsymbol{G}}^{(m)}_{j,:})^\top.$$

Consider two such token vectors that are distinct. If $i = j$, then the two underlying grid vectors differ in at least one coordinate by a nonzero integer multiple of h. Therefore,

$$\|\widetilde{\boldsymbol{g}}^{(n)}_i - \widetilde{\boldsymbol{g}}^{(m)}_i\|_2 \ge h.$$

If $i \ne j$, assume without loss of generality $i < j$. By equation E.29, for every coordinate s,

$$\begin{aligned}(\widetilde{\boldsymbol{g}}^{(m)}_j)_s - (\widetilde{\boldsymbol{g}}^{(n)}_i)_s &\ge j - (i + 1 - h)\\ &\ge h.\end{aligned}$$

Thus, in particular,

$$\|\widetilde{\boldsymbol{g}}^{(n)}_i - \widetilde{\boldsymbol{g}}^{(m)}_j\|_2 \ge h.$$

Consequently, the entire positional grid is token-wise

$$\left(\sqrt{d_{\text{hid}}}, \sqrt{d_{\text{hid}}}(L+1), h\right)$$

-separated. Furthermore, every individual positional-grid sequence contains no duplicate tokens, since tokens at different positions belong to disjoint regions $[i, i+1-h]^{d_{\text{hid}}}$. We may therefore apply Lemma E.7 with

$$\begin{aligned}r_{\min} &= \sqrt{d_{\text{hid}}},\\ r_{\max} &= \sqrt{d_{\text{hid}}}(L+1),\\ \eta &= h,\\ |\mathcal{V}| &= LM^{d_{\text{hid}}}.\end{aligned}$$

It follows that there exists a single one-head self-attention layer F_2 with

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$$

which is a (κ_1, κ_2)-contextual mapping for the positional grid, where

$$\kappa_1 = \sqrt{d_{\text{hid}}}(L+1) + \frac{h}{4} < \infty,$$

and $\kappa_2 > 0$ is given by equation E.21.

Since all quantities entering the construction depend only on L, d_{hid}, M, and $h = 1/M$, the resulting F_2, κ_1, and κ_2 are independent of δ^*. This completes the proof.

□

E.4 Proof of Lemma E.3

Lemma E.3 (Value Mapping; Restatement). *Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be a collection of input sequences such that all their token vectors are pairwise distinct, i.e., $\boldsymbol{X}^{(i)}_{k,:} \neq \boldsymbol{X}^{(j)}_{\ell,:}$ whenever $(i,k) \neq (j,\ell)$. Let $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(N)} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ be arbitrary target sequences. Define*

$$B_X := \max_{i \in [N], k \in [L]} \|\boldsymbol{X}^{(i)}_{k,:}\|_2, \qquad B_Y := \max_{i \in [N], k \in [L]} \|\boldsymbol{Y}^{(i)}_{k,:}\|_2.$$

Then there exists a sequence of at most $NL + 1$ feed-forward layers whose composition $F_3 : \mathbb{R}^{L \times d_{\mathrm{hid}}} \to \mathbb{R}^{L \times d_{\mathrm{hid}}}$ satisfies

$$F_3(\boldsymbol{X}^{(i)}) = \boldsymbol{Y}^{(i)}, \qquad i \in [N].$$

Moreover, for every $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$, we have

$$\|F_3(\boldsymbol{X})\|_F \leq \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_Y + \frac{5}{3} B_X\right).$$

Proof of Lemma E.3. Set $K := NL$. Enumerate all token vectors appearing in the input sequences as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^{d_{\mathrm{hid}}}$, and enumerate the corresponding target token vectors as $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{R}^{d_{\mathrm{hid}}}$, so that each pair $(\boldsymbol{z}_m, \boldsymbol{y}_m)$ is of the form $\boldsymbol{z}_m = \boldsymbol{X}^{(i)}_{k,:}$, $\boldsymbol{y}_m = \boldsymbol{Y}^{(i)}_{k,:}$ for some $i \in [N]$ and $k \in [L]$. By assumption, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ are pairwise distinct. Using Lemma E.4, we know that there exists a unit vector $\boldsymbol{a} \in \mathbb{R}^{d_{\mathrm{hid}}}$ such that

$$\boldsymbol{a}^\top \boldsymbol{x}_m \neq \boldsymbol{a}^\top \boldsymbol{x}_{m'} \qquad \text{for all } m \neq m'.$$

Let $s_m := \boldsymbol{a}^\top \boldsymbol{x}_m$. After relabeling the pairs $(\boldsymbol{x}_m, \boldsymbol{y}_m)$, we may assume $s_1 < s_2 < \cdots < s_M$. Define

$$\Delta := \min_{m \neq m'} |s_m - s_{m'}| > 0, \qquad \rho := \frac{\Delta}{3}.$$

For every $m \in [M]$, define

$$q_m(t) := \frac{1}{\rho}\left[\sigma_{\mathrm{R}}(t - s_m + \rho) - 2\sigma_{\mathrm{R}}(t - s_m) + \sigma_{\mathrm{R}}(t - s_m - \rho)\right].$$

Then $0 \leq q_m(t) \leq 1$, $\mathrm{supp}(q_m) \subseteq [s_m - \rho, s_m + \rho]$, and $q_m(s_{m'}) = \mathbf{1}\{m = m'\}$. Since $|s_m - s_{m'}| \geq 3\rho$, the supports of $q_1, \ldots, q_M$ are pairwise disjoint. We next introduce a common translation that moves every target vector to the left of all detector regions. Let

$$s_{\min} := s_1, \qquad y_{\max} := \max_{m \in [M]} \boldsymbol{a}^\top \boldsymbol{y}_m,$$

and define

$$\tau := \big(y_{\max} - s_{\min} + \rho\big)_+, \qquad \boldsymbol{c} := -\tau \boldsymbol{a},$$

where $(t)_+ := \max\{t, 0\}$. Then, for every $m \in [M]$,

$$a^\top(\boldsymbol{y}_m + \boldsymbol{c}) = a^\top \boldsymbol{y}_m - \tau \leq s_{\min} - \rho.$$

Thus all translated target vectors $\boldsymbol{y}_m + \boldsymbol{c}$ lie to the left of every detector support in the a-direction.

For each $m \in [M]$, define the token-wise residual map

$$\phi_m(\boldsymbol{x}) := \boldsymbol{x} + (\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{x}_m) q_m(\boldsymbol{a}^\top \boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}}.$$

Using the definition of q_m,

$$\phi_m(\boldsymbol{x}) = \boldsymbol{x} + \frac{\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{x}_m}{\rho}\Big[\sigma_{\mathrm{R}}(\boldsymbol{a}^\top \boldsymbol{x} - s_m + \rho) - 2\sigma_{\mathrm{R}}(\boldsymbol{a}^\top \boldsymbol{x} - s_m) + \sigma_{\mathrm{R}}(\boldsymbol{a}^\top \boldsymbol{x} - s_m - \rho)\Big].$$

Hence ϕ_m is exactly representable by a feed-forward layer with three hidden neurons.

Applying the same token-wise map to every row defines a shared feed-forward layer

$$F_m : \mathbb{R}^{L\times d_{\text{hid}}} \to \mathbb{R}^{L\times d_{\text{hid}}}, \qquad F_m(\boldsymbol{X})_{k,:} = \phi_m(\boldsymbol{X}_{k,:}^\top).$$

Therefore, F_m has precisely the usual shared token-wise feed-forward structure. We now show that once a token vector is modified by one layer, it cannot be modified by any subsequent layer. Suppose that $q_m(\boldsymbol{a}^\top \boldsymbol{x}) > 0$. Then

$$\boldsymbol{a}^\top \boldsymbol{x} \in (s_m - \rho, s_m + \rho).$$

Furthermore,

$$\begin{aligned}\boldsymbol{a}^\top(\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{x}_m) &= \boldsymbol{a}^\top(\boldsymbol{y}_m + \boldsymbol{c}) - s_m \\ &\le s_{\min} - \rho - s_m \\ &\le -\rho.\end{aligned}$$

Consequently,

$$\boldsymbol{a}^\top \phi_m(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + q_m(\boldsymbol{a}^\top \boldsymbol{x})\boldsymbol{a}^\top(\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{x}_m) \le \boldsymbol{a}^\top \boldsymbol{x} \le s_m + \rho.$$

For every $m' > m$,

$$s_{m'} - \rho \ge s_m + \Delta - \rho = s_m + 2\rho.$$

Hence

$$\boldsymbol{a}^\top \phi_m(\boldsymbol{x}) < s_{m'} - \rho,$$

and therefore

$$q_{m'}(\boldsymbol{a}^\top \phi_m(\boldsymbol{x})) = 0.$$

Thus, along the trajectory of any token vector, at most one of the maps $\phi_1, \dots, \phi_M$ acts nontrivially.

We next verify the desired interpolation property. Fix $m \in [M]$. For every $m' < m$,

$$q_{m'}(\boldsymbol{a}^\top \boldsymbol{x}_m) = q_{m'}(s_m) = 0,$$

so the first $m-1$ layers leave $\boldsymbol{x}_m$ unchanged. At the m-th layer,

$$q_m(\boldsymbol{a}^\top \boldsymbol{x}_m) = q_m(s_m) = 1,$$

and therefore

$$\phi_m(\boldsymbol{x}_m) = \boldsymbol{y}_m + \boldsymbol{c}.$$

Since

$$\boldsymbol{a}^\top(\boldsymbol{y}_m + \boldsymbol{c}) \le s_{\min} - \rho,$$

all subsequent detectors vanish. Hence

$$\phi_M \circ \cdots \circ \phi_1(\boldsymbol{x}_m) = \boldsymbol{y}_m + \boldsymbol{c}.$$

Define the final residual affine map

$$\phi_{M+1}(\boldsymbol{x}) := \boldsymbol{x} - \boldsymbol{c}.$$

Its shared sequence-level version is given by

$$F_{M+1}(\boldsymbol{X})_{k,:} = \phi_{M+1}(\boldsymbol{X}_{k,:}^\top).$$

Therefore, setting

$$F_3 := F_{M+1} \circ F_M \circ \cdots \circ F_1,$$

we obtain

$$F_3(\boldsymbol{X}^{(i)}) = \boldsymbol{Y}^{(i)} \qquad \text{for every } i \in [N].$$

It remains to establish the uniform norm bound. First observe that

$$y_{\max} \le B_Y, \qquad -s_{\min} \le B_X.$$

Moreover,

$$\Delta \le \max_{m,m'} \|\boldsymbol{x}_m - \boldsymbol{x}_{m'}\|_2 \le 2B_X,$$

and hence $\rho \le \frac{2}{3}B_X$. Therefore,

$$\tau = \|\boldsymbol{c}\|_2 \le B_Y + \frac{5}{3}B_X.$$

Consider an arbitrary token vector $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}}$. If none of the detectors is activated, then

$$\phi_{M+1} \circ \cdots \circ \phi_1(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{c} = \boldsymbol{x} + \tau\boldsymbol{a},$$

and hence

$$\|\phi_{M+1} \circ \cdots \circ \phi_1(\boldsymbol{x})\|_2 \le \|\boldsymbol{x}\|_2 + \tau.$$

Otherwise, exactly one detector, say q_m, is activated. Let

$$\lambda := q_m(\boldsymbol{a}^\top \boldsymbol{x}) \in (0, 1].$$

Since all other layers act as the identity along this trajectory,

$$\phi_M \circ \cdots \circ \phi_1(\boldsymbol{x}) = \boldsymbol{x} + \lambda(\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{z}_m).$$

Applying the final translation and using $\boldsymbol{c} = -\tau\boldsymbol{a}$ gives

$$\begin{aligned}\phi_{M+1} \circ \cdots \circ \phi_1(\boldsymbol{x}) &= \boldsymbol{x} + \lambda(\boldsymbol{y}_m + \boldsymbol{c} - \boldsymbol{x}_m) - \boldsymbol{c} \\ &= \boldsymbol{x} + \lambda(\boldsymbol{y}_m - \boldsymbol{x}_m) + (1-\lambda)\tau\boldsymbol{a}.\end{aligned}$$

Thus

$$\begin{aligned}\|\phi_{M+1} \circ \cdots \circ \phi_1(\boldsymbol{x})\|_2 &\le \|\boldsymbol{x}\|_2 + \lambda\|\boldsymbol{y}_m - \boldsymbol{x}_m\|_2 + (1-\lambda)\tau \\ &\le \|\boldsymbol{x}\|_2 + \lambda(B_Y + B_X) + (1-\lambda)\tau \\ &\le \|\boldsymbol{x}\|_2 + \max\{B_Y + B_X, \tau\} \\ &\le \|\boldsymbol{x}\|_2 + B_Y + \frac{5}{3}B_X.\end{aligned}$$

The same estimate holds when no detector is activated. Therefore, for every $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}}$,

$$\|\phi_{M+1} \circ \cdots \circ \phi_1(\boldsymbol{x})\|_2 \le \|\boldsymbol{x}\|_2 + B_Y + \frac{5}{3}B_X.$$

Finally, by Minkowski's inequality,

$$\begin{aligned}\|F_3(\boldsymbol{X})\|_F &= \left(\sum_{k=1}^{L} \|[F_3(\boldsymbol{X})]_{k,:}\|_2^2\right)^{1/2} \\ &\le \left(\sum_{k=1}^{L} \left(\|\boldsymbol{X}_{k,:}\|_2 + B_Y + \frac{5}{3}B_X\right)^2\right)^{1/2} \\ &\le \|\boldsymbol{X}\|_F + \sqrt{L}\left(B_Y + \frac{5}{3}B_X\right).\end{aligned}$$

This completes the proof. □

E.5 A Discussion on Parameter Efficiency

Theorem 2.1 is qualitative because the grid resolution in its proof is chosen only through uniform continuity. For targets with a prescribed modulus of continuity, the same construction yields an explicit bound on the number of Transformer blocks. We specialize to the Hölder setting of (He et al., 2026) and show that tying the query, key, and value projections preserves the known dependence on the approximation accuracy.

Let

$$d_0 := Ld_{\mathrm{hid}}.$$

After identifying $\mathbb{R}^{L \times d_{\mathrm{hid}}}$ with $\mathbb{R}^{d_0}$ through vectorization, write

$$\mathrm{vec}(F(\boldsymbol{X})) = \big(F_1(\boldsymbol{X}), \ldots, F_{d_0}(\boldsymbol{X})\big)^\top.$$

For $\alpha \in (0,1]$ and $K \geq 1$, define

$$\mathcal{H}^{\alpha}_{L,d_{\mathrm{hid}}}(K) := \Bigg\{ F : [0,1]^{L\times d_{\mathrm{hid}}} \to \mathbb{R}^{L\times d_{\mathrm{hid}}} : \max_{r\in[d_0]} \sup_{\boldsymbol{X}\in[0,1]^{L\times d_{\mathrm{hid}}}} |F_r(\boldsymbol{X})| \leq K, \\ \max_{r\in[d_0]} \sup_{\substack{\boldsymbol{X},\boldsymbol{Y}\in[0,1]^{L\times d_{\mathrm{hid}}} \\ \boldsymbol{X}\neq\boldsymbol{Y}}} \frac{|F_r(\boldsymbol{X}) - F_r(\boldsymbol{Y})|}{\|\boldsymbol{X}-\boldsymbol{Y}\|_F^{\alpha}} \leq K \Bigg\}. \tag{E.30}$$

Because vectorization preserves the Euclidean/Frobenius norm, this is the sequence-to-sequence form of $\mathcal{H}^{\alpha}_{d_0}([0,1]^{d_0}, K)$ considered by He et al. (2026).

Corollary E.1 (Hölder approximation with tied projections)**.** *Let $L, d_{\mathrm{hid}} \in \mathbb{N}$, set $d_0 = Ld_{\mathrm{hid}}$, and suppose that $\alpha \in (0,1]$, $d_0 > 2\alpha$, and $K \geq 1$. For every*

$$F \in \mathcal{H}^{\alpha}_{L,d_{\mathrm{hid}}}(K) \qquad \textit{and} \qquad 0 < \varepsilon < 1,$$

define

$$M_\varepsilon := \max\left\{2, \left\lceil \left(\frac{4Kd_0^{(1+\alpha)/2}}{\varepsilon}\right)^{1/\alpha} \right\rceil \right\}. \tag{E.31}$$

Then, for every $s \in [L]$, there exists an unmasked Transformer

$$T \in \mathcal{T}_{Q=K=V}\big(N_s, 1, d_{\mathrm{hid}}, W_s\big), \qquad W_s := 3sd_{\mathrm{hid}},$$

such that

$$\|F - T\|_{L^2} < \varepsilon$$

and

$$N_s \leq M_\varepsilon \left\lceil \frac{L}{s} \right\rceil + LM_\varepsilon^{d_0} + 1. \tag{E.32}$$

In particular, taking $s = L$ gives

$$N_L \leq M_\varepsilon + LM_\varepsilon^{d_0} + 1 = \mathcal{O}_{L,d_{\mathrm{hid}},K,\alpha}\left(\varepsilon^{-d_0/\alpha}\right). \tag{E.33}$$

Thus, a fixed-width Transformer with $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$ achieves the same order of dependence on ε as the standard untied Transformer upper bound of He et al. (2026).

Moreover, if all parameters of the ambient architecture are counted, including a positional embedding, then

$$P_s \leq d_0 + N_s\left(2d_{\mathrm{hid}}^2 + 2W_s d_{\mathrm{hid}} + W_s + d_{\mathrm{hid}}\right). \tag{E.34}$$

Consequently,

$$P_s = \mathcal{O}_{L,d_{\mathrm{hid}},K,\alpha}\left(\varepsilon^{-d_0/\alpha}\right)$$

for every fixed $s \in [L]$.

Proof. We apply the construction in the proof of Theorem 2.1 while tracking its dependence on the Hölder modulus.

Let $M := M_\varepsilon$, $h := 1/M$, choose $\delta^* \in (0,h)$, and set

$$\delta := h - \delta^*.$$

Define

$$\mathcal{G}_{\delta,\delta^*} := \{0, h, \ldots, (M-1)h\}^{L\times d_{\mathrm{hid}}},$$

and, for each $\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}$,

$$S_{\boldsymbol{G}} := \prod_{i=1}^{L}\prod_{j=1}^{d_{\mathrm{hid}}} [G_{ij}, G_{ij} + \delta].$$

If $\boldsymbol{X} \in S_{\boldsymbol{G}}$, then

$$\|\boldsymbol{X} - \boldsymbol{G}\|_F < \frac{\sqrt{d_0}}{M}.$$

Since every coordinate F_r is (K, α)-Hölder continuous,

$$\|F(\boldsymbol{X}) - F(\boldsymbol{G})\|_F \le K d_0^{(1+\alpha)/2} M^{-\alpha} \le \frac{\varepsilon}{4}. \tag{E.35}$$

Hence the required grid resolution satisfies $M = \mathcal{O}(\varepsilon^{-1/\alpha})$.

Define the piecewise-constant approximation

$$F'(\boldsymbol{X}) := \sum_{\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}} F(\boldsymbol{G}) \mathbf{1}\{\boldsymbol{X} \in S_{\boldsymbol{G}}\}.$$

We next apply the three modules used in the proof of Theorem 2.1. By Lemma E.1, for every $s \in [L]$, there exists a quantization module $F_{1,s}$ with at most

$$M \left\lceil \frac{L}{s} \right\rceil$$

residual ReLU feed-forward layers and hidden width

$$W_s = 3 s d_{\mathrm{hid}}.$$

With positional embedding

$$\boldsymbol{E} = [1, 2, \ldots, L]^\top \mathbf{1}_{d_{\mathrm{hid}}}^\top,$$

it satisfies

$$F_{1,s}(\boldsymbol{X} + \boldsymbol{E}) = \boldsymbol{G} + \boldsymbol{E}, \qquad \boldsymbol{X} \in S_{\boldsymbol{G}}.$$

Since

$$|\mathcal{G}_{\delta,\delta^*}| = M^{d_0},$$

Lemma E.2 provides a single one-head self-attention layer F_2 satisfying

$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$$

such that

$$\{F_2(\boldsymbol{G} + \boldsymbol{E})_{i,:} : \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*},\ i \in [L]\}$$

are pairwise distinct. Lemma E.3 then gives a position-wise value-mapping module F_3, composed of at most

$$L M^{d_0} + 1$$

residual ReLU feed-forward layers, such that

$$F_3\big(F_2(\boldsymbol{G} + \boldsymbol{E})\big) = F(\boldsymbol{G}), \qquad \boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}. \tag{E.36}$$

It remains to control the gap between the quantization cells. Define

$$\Omega_{\mathrm{in}} := \bigcup_{\boldsymbol{G} \in \mathcal{G}_{\delta,\delta^*}} S_{\boldsymbol{G}}, \qquad \Omega_{\mathrm{out}} := [0,1]^{L \times d_{\mathrm{hid}}} \setminus \Omega_{\mathrm{in}}.$$

Since there are M^{d_0} cells,

$$\mu(\Omega_{\mathrm{out}}) = 1 - (M\delta)^{d_0} \tag{E.37}$$

$$= 1 - (1 - M\delta^*)^{d_0} \tag{E.38}$$

$$\le d_0 M \delta^*. \tag{E.39}$$

Let

$$B_F := \sup_{\boldsymbol{X} \in [0,1]^{L \times d_{\mathrm{hid}}}} \|F(\boldsymbol{X})\|_F.$$

By equation E.30, $B_F \le K\sqrt{d_0}$.

Moreover, the output of $F_{1,s}$ lies in the compact set

$$\mathcal{K}_1 := \left\{ \boldsymbol{Z} \in \mathbb{R}^{L \times d_{\mathrm{hid}}} : i \le Z_{ij} \le i+1 \text{ for every } i, j \right\}.$$

Since $F_3 \circ F_2$ is continuous,

$$B_T := \max_{\boldsymbol{Z} \in \mathcal{K}_1} \|F_3(F_2(\boldsymbol{Z}))\|_F < \infty.$$

The quantity B_T is independent of δ^*. Therefore, choosing $\delta^* > 0$ sufficiently small gives

$$\|F - F'\|_{L^2} < \frac{\varepsilon}{2}.$$

Now define

$$T := F_3 \circ F_2 \circ F_{1,s} \circ \mathrm{PE}_{\boldsymbol{E}}.$$

By the quantization property and equation E.36,

$$T(\boldsymbol{X}) = F'(\boldsymbol{X}), \qquad \boldsymbol{X} \in \Omega_{\mathrm{in}}.$$

Since $\|T(\boldsymbol{X})\|_F \le B_T$ on Ω_{out}, decreasing δ^* further if necessary gives

$$\|F' - T\|_{L^2} < \frac{\varepsilon}{2}.$$

Hence

$$\|F - T\|_{L^2} \le \|F - F'\|_{L^2} + \|F' - T\|_{L^2} < \varepsilon.$$

Finally, the quantization module contributes at most

$$M \left\lceil \frac{L}{s} \right\rceil$$

Transformer blocks, while the contextual and value-mapping modules contribute at most

$$LM^{d_0} + 1$$

additional blocks. Therefore,

$$N_s \le M \left\lceil \frac{L}{s} \right\rceil + LM^{d_0} + 1,$$

which proves equation E.32. Since

$$M_\varepsilon = \mathcal{O}_{d_0, K, \alpha}\left(\varepsilon^{-1/\alpha}\right),$$

equation E.33 follows.

A one-head block with head dimension d_{hid}, feed-forward width W_s, and tied query–key–value projection has at most

$$2d_{\mathrm{hid}}^2 + 2W_s d_{\mathrm{hid}} + W_s + d_{\mathrm{hid}}$$

parameters. Adding at most d_0 parameters for the positional embedding proves equation E.34. □

Comparison with untied Transformers. For the same Hölder class, He et al. (2026) construct a standard Transformer with

$$d_0 M + 2LM^{d_0} + 1$$

blocks at grid resolution $M \asymp \varepsilon^{-1/\alpha}$. Our construction instead uses

$$M \left\lceil \frac{L}{s} \right\rceil + LM^{d_0} + 1$$

blocks. In both constructions, the dominant term is the value mapping over M^{d_0} quantized sequences. Therefore,

$$N_{\mathrm{tied}}(\varepsilon) = \mathcal{O}\left(\varepsilon^{-d_0/\alpha}\right),$$

which has the same approximation-rate exponent as the untied construction. Thus, within the quantization–contextual-mapping–value-mapping framework, imposing $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}_V$ does not worsen the known dependence on the target accuracy. The dominant complexity instead comes from exhaustive value interpolation over the d_0-dimensional quantization grid.

F TECHNICAL DETAILS IN SECTION 3

F.1 PROOF OF THEOREM 3.1

Theorem 3.1 (Restatement). *Fix $L \geq 2$ and $1 \leq k \leq L$, and set $d := \min\{2k, L-1\}$, $d_{\mathrm{hid}} := 2d$. There exist fixed matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}} \times d}$, depending only on k and L, such that the following holds. For every $\gamma \geq 1$, every $\varepsilon > 0$, and every (k,γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L\times L}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}}$ satisfying*

$$\| \operatorname{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\big) - \boldsymbol{A}\|_2 < \varepsilon.$$

Proof of Theorem 3.1. Let

$$\bar{\varepsilon} = \min\{\varepsilon, 1\}, \quad \alpha = \frac{\bar{\varepsilon}}{4L\sqrt{k}}, \quad \Gamma = \log(1/\alpha).$$

Let $x_1, \cdots, x_L \in \mathbb{R}$ be mutually different. For each row index i define

$$S_i = \{j : \boldsymbol{A}_{i,j} > 0\}, \quad r_i = |S_i| \leq k.$$

Since A is right stochastic, then $S_i \neq \varnothing$.

We can construct an interpolation polynomial with $\deg(h_i) \leq r_i - 1$ such that

$$h_i(x_j) = \log \boldsymbol{A}_{i,j}, \quad j \in S_i.$$

Define

$$g_i(x) = \prod_{s\in S_i} (x - x_s)^2.$$

Then

$$\begin{aligned} g_i(x_s) &= 0, \quad s \in S_i, \\ g_i(x_j) &> 0, \quad j \notin S_i. \end{aligned}$$

Let $m_i := \min_{s\in S_i} \log \boldsymbol{A}_{i,s}$. We can select a large enough $C_i \geq 0$, such that

$$h_i(x_j) - C_i g_i(x_j) \leq m_i - \Gamma, \quad j \notin S_i.$$

For example, $C_i = \max_{j\notin S_i} \frac{(h_i(x_j) - m_i + \Gamma)_+}{g_i(x_j)}$. Define

$$p_i(x) = h_i(x) - C_i g_i(x),$$

which is a polynomial with $\deg(p_i) \leq 2k$. As a result, we obtain

$$p_i(x_j) = \log \boldsymbol{A}_{i,j}, \quad j \in S_i$$

and

$$p_i(x_j) \leq m_i - \Gamma, \quad j \notin S_i.$$

We define the matrix $\boldsymbol{P} \in \mathbb{R}^{L\times L}$ with $\boldsymbol{P}_{i,j} = p_i(x_j)$. As the order of p_i is less than $2k$, we know that $\operatorname{span}(\mathbf{1}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(2k)})$ where $\boldsymbol{x}^{(i)} = \begin{pmatrix} x_1^i & \cdots & x_L^i \end{pmatrix}^\top$. So, $\operatorname{rank}(\boldsymbol{P}) \leq 2k+1$.

Let $\boldsymbol{H}_L = I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top$ and define $\boldsymbol{S} = \boldsymbol{P}\boldsymbol{H}_L$. It is straightforward to verify that

$$\operatorname{Softmax}(\boldsymbol{S}) = \operatorname{Softmax}(\boldsymbol{P}).$$

Since $\mathbf{1}^\top \boldsymbol{H}_L = \mathbf{0}$, then

$$\operatorname{rank}(\boldsymbol{S}) \leq \min\{2k, L-1\}.$$

Applying singular-value decomposition to $\boldsymbol{S}$, we have

$$\boldsymbol{S} = \boldsymbol{U}_r \boldsymbol{\Sigma} \boldsymbol{V}_r^\top, \quad r = \operatorname{rank}(\boldsymbol{S}).$$

Define

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{U}_r \boldsymbol{\Sigma}_r^{1/2} & \mathbf{0}_{L\times(d-r)} \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{V}_r \boldsymbol{\Sigma}_r^{1/2} & \mathbf{0}_{L\times(d-r)} \end{pmatrix},$$

and

$$\boldsymbol{W}_Q = \begin{pmatrix} \boldsymbol{I}_d \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{d_{hid}\times d}, \quad \boldsymbol{W}_K = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{I}_d \end{pmatrix} \in \mathbb{R}^{d_{hid}\times d}.$$

Let

$$\boldsymbol{X} = [\boldsymbol{Q} \quad \boldsymbol{K}] \in \mathbb{R}^{L\times d_{hid}}.$$

Then

$$\mathrm{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\right) = \mathrm{Softmax}(\boldsymbol{S}).$$

To control the error, let $\boldsymbol{M} = \mathrm{Softmax}(\boldsymbol{S}) = \mathrm{Softmax}(\boldsymbol{P})$. For $j_1, j_2 \in S_i$, we have

$$\boldsymbol{P}_{i,j_1} - \boldsymbol{P}_{i,j_2} = \log \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}},$$

so

$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} = \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}}.$$

If $j_1 \notin S_i$ and $j_2 \in S_i$, then

$$\boldsymbol{P}_{i,j_1} - \boldsymbol{P}_{i,j_2} \le -\Gamma.$$

Consequently

$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} \le e^{-\Gamma} = \alpha.$$

Particularly, we have $\boldsymbol{M}_{i,j} \le \alpha$ if $j \notin S_i$. Define

$$\eta_i = \sum_{j\notin S_i} \boldsymbol{M}_{i,j}.$$

Then $\eta_i \le L\alpha$, and importantly

$$\begin{aligned} \sum_{\ell\in S_i} \frac{\boldsymbol{M}_{i,j}}{\boldsymbol{M}_{i,\ell}} = \sum_{\ell\in S_i} \frac{\boldsymbol{A}_{i,j}}{\boldsymbol{A}_{i,\ell}} &\Rightarrow \boldsymbol{M}_{i,j} \sum_{\ell\in S_i} \boldsymbol{A}_{i,\ell} = \boldsymbol{A}_{i,j} \sum_{\ell\in S_i} \boldsymbol{M}_{i,j} \\ &\Rightarrow \boldsymbol{M}_{i,j}\cdot 1 = \boldsymbol{A}_{i,j}\cdot(1-\eta_i), \end{aligned}$$

which means that

$$\boldsymbol{M}_{i,j} = (1-\eta_i)\boldsymbol{A}_{i,j}, \quad j \in S_i.$$

Consequently

$$\begin{aligned} \sum_{j=1}^{L} |\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| &= \sum_{j\in S_i} \eta_i \boldsymbol{A}_{i,j} + \sum_{j\notin S_i} \boldsymbol{M}_{i,j} \\ &= 2\eta_i \le 2L\alpha. \end{aligned}$$

As a result, we obtain

$$\|\boldsymbol{M} - \boldsymbol{A}\|_\infty \le 2L\alpha.$$

We now fix any column index j and let

$$I_j = \{i : \boldsymbol{A}_{i,j} > 0\}.$$

According to the sparsity assumption, we know that $|I_j| \le k$. For any $i \in I_j$, we have

$$|\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| = \eta_i \boldsymbol{A}_{i,j} \le L\alpha.$$

Then

$$\sum_{i \in I_j} |\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| \leq kL\alpha.$$

For $i \notin I_j$, similarly we have $\boldsymbol{M}_{i,j} \leq \alpha$, then $\sum_{i \notin I_j} \boldsymbol{M}_{i,j} \leq L\alpha$. As a result

$$\|\boldsymbol{M} - \boldsymbol{A}\|_1 \leq (k+1)L\alpha \leq 2kL\alpha.$$

Finally, we can bound the spectral norm

$$\begin{aligned}
\|\boldsymbol{M} - \boldsymbol{A}\|_2 &\leq \sqrt{\|\boldsymbol{M} - \boldsymbol{A}\|_1 \|\boldsymbol{M} - \boldsymbol{A}\|_\infty} \\
&\leq \sqrt{(2kL\alpha)(2L\alpha)} \\
&= 2L\sqrt{k}\alpha \\
&= \frac{\bar{\varepsilon}}{2} \\
&< \varepsilon,
\end{aligned}$$

which completes the proof. □

F.2 Approximating Sparse Attention Patterns with K=V

Corollary F.1. *Fix $L \geq 2$ and $1 \leq k \leq L$, and set $d := \min\{2k, L-1\}, d_{\text{hid}} := 2d$. For every fixed matrix $\boldsymbol{W}_K \in \mathbb{R}^{d_{\text{hid}} \times d}$ with full column rank, there exists $\boldsymbol{W}_Q \in \mathbb{R}^{d_{\text{hid}} \times d}$ such that, for every $\gamma \geq 1$, every $\varepsilon > 0$, and every (k, γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L \times L}$, one can find $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}}$ satisfying*

$$\left\|\text{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\right) - \boldsymbol{A}\right\|_2 < \varepsilon.$$

Proof of Corollary F.1. Fix a full-column-rank matrix $\boldsymbol{W}_K \in \mathbb{R}^{d_{\text{hid}} \times d}$. Since

$$\text{rank}(\boldsymbol{W}_K) = d \qquad \text{and} \qquad d_{\text{hid}} = 2d,$$

Lemma F.1, applied with $m = d$, yields a matrix $\boldsymbol{W}_Q \in \mathbb{R}^{d_{\text{hid}} \times d}$ such that, for every $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{L \times d}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}}$ satisfying

$$(\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top = \boldsymbol{U}\boldsymbol{V}^\top.$$

Importantly, $\boldsymbol{W}_Q$ depends only on the prescribed matrix $\boldsymbol{W}_K$.

Now fix $\gamma \geq 1$, $\varepsilon > 0$, and a (k, γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L \times L}$. By Theorem 3.1, there exist fixed matrices

$$\widehat{\boldsymbol{W}}_Q, \widehat{\boldsymbol{W}}_K \in \mathbb{R}^{d_{\text{hid}} \times d}$$

and a matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{L \times d_{\text{hid}}}$ such that

$$\left\|\text{Softmax}\left((\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_Q)(\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_K)^\top\right) - \boldsymbol{A}\right\|_2 < \varepsilon.$$

Define

$$\boldsymbol{U} := \widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_Q, \qquad \boldsymbol{V} := \widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_K.$$

Then $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{L \times d}$. By the property obtained from Lemma F.1, there exists $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}}$ such that

$$(\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top = \boldsymbol{U}\boldsymbol{V}^\top = (\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_Q)(\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_K)^\top.$$

Therefore,

$$\begin{aligned}
&\left\|\text{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\right) - \boldsymbol{A}\right\|_2 \\
&\quad = \left\|\text{Softmax}\left((\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_Q)(\widehat{\boldsymbol{X}}\widehat{\boldsymbol{W}}_K)^\top\right) - \boldsymbol{A}\right\|_2 < \varepsilon,
\end{aligned}$$

which proves the result. □

F.3 Proof of Proposition 3.1

Proposition 3.1 (Restatement). *Let $L \geq 2$, and define the cyclic-shift pattern $\boldsymbol{P}^{\mathrm{cyc}} \in \mathbb{R}^{L\times L}$ by $P^{\mathrm{cyc}}_{ij} := \mathbf{1}\{j \equiv i-1 \pmod L\}$. Then, for every $d \geq 1$ and every $\boldsymbol{Z} \in \mathbb{R}^{L\times d}$,*

$$\left\|\mathrm{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top) - \boldsymbol{P}^{\mathrm{cyc}}\right\|_2 \geq 1.$$

Proof of Proposition 3.1. Let

$$\boldsymbol{S} := \boldsymbol{Z}\boldsymbol{Z}^\top, \qquad \boldsymbol{K} := \exp^\circ(\boldsymbol{S}), \qquad \boldsymbol{D} := \mathrm{diag}(\boldsymbol{K}\mathbf{1}),$$

where $\exp^\circ$ denotes the entrywise exponential. Then

$$\boldsymbol{M} := \mathrm{Softmax}(\boldsymbol{Z}\boldsymbol{Z}^\top) = \boldsymbol{D}^{-1}\boldsymbol{K}.$$

Since $\boldsymbol{S} = \boldsymbol{Z}\boldsymbol{Z}^\top$ is positive semidefinite, the Schur product theorem gives

$$\boldsymbol{S}^{\circ n} \succeq 0 \qquad \text{for every } n \geq 0.$$

Therefore,

$$\boldsymbol{K} = \exp^\circ(\boldsymbol{S}) = \sum_{n=0}^{\infty} \frac{1}{n!}\boldsymbol{S}^{\circ n} \succeq 0.$$

Also, every entry of $\boldsymbol{K}$ is strictly positive, so $\boldsymbol{D}$ is a positive diagonal matrix.

Write $\boldsymbol{P} := \boldsymbol{P}^{\mathrm{cyc}}$. Suppose, for the sake of contradiction, that

$$\|\boldsymbol{M} - \boldsymbol{P}\|_2 < 1.$$

Because $\boldsymbol{P}$ is orthogonal,

$$\boldsymbol{M} = \boldsymbol{P}\big(\boldsymbol{I}_L + \boldsymbol{P}^\top(\boldsymbol{M} - \boldsymbol{P})\big),$$

and

$$\|\boldsymbol{P}^\top(\boldsymbol{M} - \boldsymbol{P})\|_2 < 1.$$

Hence $\boldsymbol{M}$ is invertible. Since $\boldsymbol{K} = \boldsymbol{D}\boldsymbol{M}$, the matrix $\boldsymbol{K}$ is also invertible. Together with $\boldsymbol{K} \succeq 0$, this implies

$$\boldsymbol{K} \succ 0.$$

For every nonzero $\boldsymbol{x} \in \mathbb{R}^L$, the assumed strict inequality gives

$$\|(\boldsymbol{M} - \boldsymbol{P})\boldsymbol{x}\|_2 < \|\boldsymbol{x}\|_2 = \|\boldsymbol{P}\boldsymbol{x}\|_2.$$

Expanding the squares, we obtain

$$\|\boldsymbol{M}\boldsymbol{x}\|_2^2 + \|\boldsymbol{P}\boldsymbol{x}\|_2^2 - 2\boldsymbol{x}^\top\boldsymbol{M}^\top\boldsymbol{P}\boldsymbol{x} < \|\boldsymbol{P}\boldsymbol{x}\|_2^2,$$

and therefore

$$\boldsymbol{x}^\top\boldsymbol{M}^\top\boldsymbol{P}\boldsymbol{x} > \frac{1}{2}\|\boldsymbol{M}\boldsymbol{x}\|_2^2 > 0.$$

Consequently, the symmetric matrix

$$\boldsymbol{H} := \frac{1}{2}\big(\boldsymbol{M}^\top\boldsymbol{P} + \boldsymbol{P}^\top\boldsymbol{M}\big)$$

is positive definite.

Since $\boldsymbol{M} = \boldsymbol{D}^{-1}\boldsymbol{K}$ and $\boldsymbol{K}$ is symmetric,

$$\boldsymbol{M}^\top\boldsymbol{P} = \boldsymbol{K}\boldsymbol{D}^{-1}\boldsymbol{P}.$$

Define

$$\boldsymbol{C} := \boldsymbol{D}^{-1}\boldsymbol{P}, \qquad \boldsymbol{B} := \boldsymbol{K}^{1/2}\boldsymbol{C}\boldsymbol{K}^{-1/2}.$$

A congruence transformation gives

$$\begin{aligned}\boldsymbol{K}^{-1/2}\boldsymbol{H}\boldsymbol{K}^{-1/2} &= \frac{1}{2}\left(\boldsymbol{K}^{1/2}\boldsymbol{C}\boldsymbol{K}^{-1/2} + \boldsymbol{K}^{-1/2}\boldsymbol{C}^\top\boldsymbol{K}^{1/2}\right)\\ &= \frac{1}{2}(\boldsymbol{B} + \boldsymbol{B}^\top).\end{aligned}$$

Thus

$$\frac{1}{2}(\boldsymbol{B}+\boldsymbol{B}^\top) \succ 0.$$

It follows that every eigenvalue of $\boldsymbol{B}$ has strictly positive real part. Indeed, if $\boldsymbol{B}\boldsymbol{u} = \lambda\boldsymbol{u}$ for some nonzero $\boldsymbol{u} \in \boldsymbol{C}^L$, then

$$\mathrm{Re}(\lambda)\|\boldsymbol{u}\|_2^2 = \boldsymbol{u}^* \frac{\boldsymbol{B}+\boldsymbol{B}^\top}{2}\boldsymbol{u} > 0.$$

On the other hand, $\boldsymbol{B}$ is similar to $\boldsymbol{C}$, and hence

$$\sigma(\boldsymbol{B}) = \sigma(\boldsymbol{C}).$$

Write

$$\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_L), \qquad d_i > 0,$$

and define

$$\rho := \left(\prod_{i=1}^{L} d_i^{-1}\right)^{1/L} > 0.$$

Because $\boldsymbol{C} = \boldsymbol{D}^{-1}\boldsymbol{P}$ is a weighted cyclic-permutation matrix,

$$\boldsymbol{C}^L = \rho^L \boldsymbol{I}_L,$$

and its eigenvalues are

$$\sigma(\boldsymbol{C}) = \left\{\rho \exp\left(\frac{2\pi \mathrm{i} k}{L}\right) : k = 0, \dots, L-1\right\}.$$

For every $L \geq 2$, this set contains an eigenvalue with negative real part. Specifically, when L is even, take $k = L/2$, while when L is odd, take $k = (L-1)/2$. This contradicts the fact that every eigenvalue of $\boldsymbol{B}$ has strictly positive real part. Therefore,

$$\|\boldsymbol{M} - \boldsymbol{P}^{\mathrm{cyc}}\|_2 \geq 1.$$

Finally, the bound is sharp. If $\boldsymbol{Z} = \boldsymbol{0}$, then

$$\boldsymbol{M} = \frac{1}{L}\boldsymbol{1}\boldsymbol{1}^\top =: \boldsymbol{P}_i.$$

Since $\boldsymbol{P}_i$ is the orthogonal projection onto $\mathrm{span}\{\boldsymbol{1}\}$ and $\boldsymbol{P}$ acts orthogonally on its orthogonal complement,

$$(\boldsymbol{P}_i - \boldsymbol{P})^\top(\boldsymbol{P}_i - \boldsymbol{P}) = \boldsymbol{I}_L - \boldsymbol{P}_i.$$

Hence

$$\|\boldsymbol{P}_i - \boldsymbol{P}\|_2 = 1.$$

Thus the constant 1 cannot be further improved. □

F.4 Proof of Theorem 3.2

Theorem 3.2 (Restatement). *Let $L \geq 2$ and set $d = d_{\mathrm{hid}} = L-1$ and $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{I}_d$. For every causal-compatible row-stochastic matrix $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ and every $\varepsilon > 0$, there exists $\boldsymbol{X} \in \mathbb{R}^{L\times d}$ for which $\boldsymbol{M} := \mathrm{Softmax}\big(\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{M}_{\mathrm{MASK}}\big)$ satisfies*

$$\|\boldsymbol{M} - \boldsymbol{A}\|_2 < \varepsilon.$$

Proof of Theorem 3.2. Set $k := L$ and define

$$\bar{\varepsilon} := \min\{\varepsilon, 1\}, \quad \alpha := \frac{\bar{\varepsilon}}{4L\sqrt{k}}, \quad \Gamma := \log(1/\alpha).$$

Construct $L(L+1)/2$ mutually orthogonal coordinate:

$$\{\boldsymbol{p}_i : 1 \leq i \leq L\} \cup \{\boldsymbol{e}_{i,j} : 1 \leq j < i \leq L\},$$

where $\boldsymbol{p}_i$ is designed solely for position i and, $\boldsymbol{e}_{i,j}$ is designed for causal pair (i,j). For each i, define

$$S_i = \{j \le i : \boldsymbol{A}_{i,j} > 0\}, \quad r_i = |S_i \cap \{1, \cdots, i-1\}|.$$

In the following, we construct $\boldsymbol{z}_i$ from $i = L$ to $i = 1$. Suppose that we have already finished the construction of $\ell > i$ and let $K_i = \sum_{\ell > i, \boldsymbol{A}_{\ell,i} > 0} \boldsymbol{b}_{\ell,i}^2$, where $\boldsymbol{b}_{\ell,i}$ is the coordinate of $\boldsymbol{z}_i$ on $\boldsymbol{e}_{\ell,i}$.

We choose a constant λ_i such that

$$\lambda_i + \log \boldsymbol{A}_{i,j} \ge K_i + \Gamma + 2, \quad j \in S_i. \tag{F.1}$$

For each $j \in S_i$, define

$$t_{i,j} = \lambda_i + \log \boldsymbol{A}_{i,j}.$$

If $r_i > 0$, for each $j < i$ and $\boldsymbol{A}_{i,j} > 0$, let the coefficient of z_i on $\boldsymbol{e}_{i,j}$ be

$$\boldsymbol{z}_i[\boldsymbol{e}_{i,j}] = a_{i,j} = 1/\sqrt{r_i},$$

and let

$$\boldsymbol{z}_j[\boldsymbol{e}_{i,j}] = b_{i,j} = \sqrt{r_i} t_{i,j}.$$

Thus

$$a_{i,j} b_{i,j} = t_{i,j}, \quad \sum_{j<i, \boldsymbol{A}_{i,j}>0} a_{i,j}^2 = \begin{cases} 1 & r_i > 0, \\ 0 & r_i = 0. \end{cases}$$

If $\boldsymbol{A}_{i,i} > 0$, we define the coefficient of private coordinate as

$$\boldsymbol{z}_i[\boldsymbol{p}_i] = \sqrt{t_{i,i} - K_i - \sum_{j<i, \boldsymbol{A}_{i,j}>0} a_{i,j}^2},$$

which is positive as $t_{i,i} \ge K_i + \Gamma + 2 > K_i + 1$. If $\boldsymbol{A}_{i,i} = 0$, let

$$\boldsymbol{z}_i[\boldsymbol{p}_i] = 0.$$

For any $j < i$, the coordinates of $\boldsymbol{z}_i, \boldsymbol{z}_j$ are mutually non-zero only on $\boldsymbol{e}_{i,j}$. Thus

$$\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle = \begin{cases} a_{i,j} b_{i,j} = t_{i,j} = \lambda_i + \log \boldsymbol{A}_{i,j} & \boldsymbol{A}_{i,j} > 0, \\ 0 & \boldsymbol{A}_{i,j} = 0. \end{cases}$$

If $\boldsymbol{A}_{i,i} > 0$, then

$$\|\boldsymbol{z}_i\|^2 = K_i + \sum_{j<i, \boldsymbol{A}_{i,j}>0} a_{i,j}^2 + \boldsymbol{z}_i[\boldsymbol{p}_i]^2 = t_{i,i} = \lambda_i + \log \boldsymbol{A}_{i,i},$$

Then for any $j \in S_i$ including $j = i$, we have

$$\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle = \lambda_i + \log \boldsymbol{A}_{i,j}.$$

If $\boldsymbol{A}_{i,i} = 0$, then $r_i > 0$, and we obtain

$$\|\boldsymbol{z}_i\|^2 = K_i + \sum_{j<i, \boldsymbol{A}_{i,j}>0} a_{i,j}^2 = K_i + 1$$

By Eq.(F.1), for any $j \in S_i$, we have

$$\lambda_i + \log \boldsymbol{A}_{i,j} \ge K_i + \Gamma + 2 \ge \|\boldsymbol{z}_i\|^2 + \Gamma.$$

Let $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}^\top + \boldsymbol{M}_{\mathrm{MASK}}$, $\boldsymbol{M} = \mathrm{Softmax}(\boldsymbol{G})$. Overall, for each row index i, we obtain

- For any $j_1, j_2 \in S_i$, then $\boldsymbol{A}_{i,j_1}, \boldsymbol{A}_{i,j_2} > 0$, we obtain

$$\boldsymbol{G}_{i,j_1} - \boldsymbol{G}_{i,j_2} = \langle \boldsymbol{z}_i, \boldsymbol{z}_{j_1} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{z}_{j_2} \rangle = \log \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}},$$

  then

$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} = \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}}.$$

- On the other hand, if $j_1 \le i$, $\boldsymbol{A}_{i,j_1} = 0$, while $j_2 \in S_i$ and $\boldsymbol{A}_{i,j_2} > 0$, then we consider two cases
  - If $j_1 < i$, then $\boldsymbol{G}_{i,j_1} = 0$, and $\boldsymbol{G}_{i,j_2} \ge \Gamma + 2$.
  - If $j_1 = i$, then $\boldsymbol{A}_{i,i} = 0$, and $\boldsymbol{G}_{i,i} = \|\boldsymbol{z}_i\|^2 \le \boldsymbol{G}_{i,j_2} - \Gamma$.

  As a result, we have
  $$\boldsymbol{G}_{i,j_1} - \boldsymbol{G}_{i,j_2} = \langle \boldsymbol{z}_i, \boldsymbol{z}_{j_1} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{z}_{j_2} \rangle \le -\Gamma,$$
  and
  $$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} \le e^{-\Gamma} = \alpha.$$

Let $\eta_i = \sum_{j \le i, \boldsymbol{A}_{i,j}=0} \boldsymbol{M}_{i,j}$. For any position with zero value, we have $\boldsymbol{M}_{i,j} \le \alpha$, so
$$\eta_i \le L\alpha.$$
Moreover, if $\boldsymbol{M}_{i,j} \ne 0$, then $\boldsymbol{M}_{i,j} = c_i \boldsymbol{A}_{i,j}$, $j \in S_i$, where $c_i = 1 - \eta_i$. Then
$$\sum_{j=1}^{L} |\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| = 2\eta_i \le 2L\alpha.$$
As a result, we have
$$\|\boldsymbol{M} - \boldsymbol{A}\|_\infty \le 2L\alpha.$$
Fix any column index j, for j such that $\boldsymbol{A}_{i,j} > 0$, we have
$$|\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| = \eta_i \boldsymbol{A}_{i,j} \le L\alpha \boldsymbol{A}_{i,j}.$$
Then
$$\sum_{i=1}^{L} |\boldsymbol{M}_{i,j} - \boldsymbol{A}_{i,j}| \le (k+1)L\alpha \le 2kL\alpha,$$
then
$$\|\boldsymbol{M} - \boldsymbol{A}\|_1 \le 2kL\alpha.$$
Finally, by using the standard inequality about matrix norm, we obtain
$$\begin{aligned} \|\boldsymbol{M} - \boldsymbol{A}\|_2 &\le \sqrt{\|\boldsymbol{M} - \boldsymbol{A}\|_1 \|\boldsymbol{M} - \boldsymbol{A}\|_\infty} \\ &\le 2L\sqrt{k}\alpha \\ &= \bar{\varepsilon}/2 < \varepsilon. \end{aligned}$$
Let $\boldsymbol{H} := \boldsymbol{Z}\boldsymbol{Z}^\top \succeq \boldsymbol{0}$. If $\mathrm{rank}(\boldsymbol{H}) \le L - 1$, we directly let $\tilde{\boldsymbol{H}} = \boldsymbol{H}$. otherwise, $\boldsymbol{H}$ is full rank, hence $\boldsymbol{H}^{-1}$ exists. Define $c := 1/(\boldsymbol{1}^\top \boldsymbol{H}^{-1} \boldsymbol{1})$ and let $\tilde{\boldsymbol{H}} := \boldsymbol{H} - c\boldsymbol{1}\boldsymbol{1}^\top$, and $\boldsymbol{v} := \boldsymbol{H}^{-1/2}\boldsymbol{1}$. Since $c = 1/\|\boldsymbol{v}\|^2$, we have
$$\tilde{\boldsymbol{H}} = \boldsymbol{H}^{1/2} \left(\boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|^2} \right) \boldsymbol{H}^{1/2}.$$
Importantly, we have $\mathrm{rank}(\tilde{\boldsymbol{H}}) = L - 1$, and $\tilde{\boldsymbol{H}}_{i,j} = \boldsymbol{H}_{i,j} - c$ for some constant c. Since Softmax function is invariant under row-wise additive shift, we obtain
$$\mathrm{Softmax}(\tilde{\boldsymbol{H}} + \boldsymbol{M}_{\mathrm{MASK}}) = \mathrm{Softmax}(\boldsymbol{H} + \boldsymbol{M}_{\mathrm{MASK}}).$$
Now, since $\tilde{\boldsymbol{H}} \succeq \boldsymbol{0}$ and $\mathrm{rank}(\tilde{\boldsymbol{H}}) \le L - 1$. There exist $\tilde{\boldsymbol{X}} \in \mathbb{R}^{L \times (L-1)}$ such that
$$\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^\top.$$
Let $\boldsymbol{X} = \tilde{\boldsymbol{X}}$, and define
$$\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{I}_d \in \mathbb{R}^{d \times d}.$$
It is straightforward to verify that
$$\mathrm{Softmax}\left((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top + \boldsymbol{M}_{\mathrm{MASK}} \right) = \mathrm{Softmax}(\boldsymbol{G}) = \boldsymbol{M},$$
which completes the proof. □

F.5 Proof of Theorem 3.3

Theorem 3.3 (Restatement). *Let $L > 1$, $1 \le k \le L$, $\gamma \ge 1$, $\varepsilon > 0$, d even, and $d_{\mathrm{hid}} \ge d$. Define*

$$\bar{\varepsilon} := \min\{\varepsilon, 1\}, \quad \Lambda := \max\left\{\log\gamma + \log\frac{64L\sqrt{k}}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{\sqrt{k}}, 1\right\}.$$

Suppose that $d \ge \max\left\{\frac{32k^2}{\bar{\varepsilon}^2}\Lambda^2\log\big(8kL(L-1)\big),\, 64\log(16L)\right\}$. Then there exist fixed matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}}\times d}$, depending only on d_{hid} and d, with the following property. For every causal-compatible (k,γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L\times L}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L\times d_{\mathrm{hid}}}$ such that

$$\big\|\mathrm{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top + \boldsymbol{M}_{\mathrm{MASK}}\big) - \boldsymbol{A}\big\|_2 < \varepsilon,$$

where $\max\{\|\boldsymbol{X}\boldsymbol{W}_Q\|_F, \|\boldsymbol{X}\boldsymbol{W}_K\|_F\} \le \sqrt{\frac{3}{2}}\sqrt{L}\,k^{1/4}\sqrt{\Lambda}$.

Proof of Theorem 3.3. We present the proof for unmasked attention. For a causal-compatible target, causal masking preserves ratios between retained attention entries and sets masked zero-target entries to zero. Hence the same ratio bounds and subsequent spectral-norm estimates remain valid, while the query and key representations, and therefore their norm bounds, are unchanged. Let

$$\bar{\varepsilon} := \min\{\varepsilon, 1\}, \qquad \alpha := \frac{\bar{\varepsilon}}{64L\sqrt{k}}, \qquad \beta := \frac{\bar{\varepsilon}}{\sqrt{k}}.$$

Then $0 < \alpha < 1$ and $0 < \beta < \sqrt{2}$. Moreover,

$$\max\{\log\gamma - \log\alpha + \beta, 1\} = \max\left\{\log\gamma + \log\frac{64L\sqrt{k}}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{\sqrt{k}}, 1\right\} = \Lambda,$$

and

$$32\beta^{-2}k\Lambda^2\log\big(8kL(L-1)\big) = \frac{32k^2\Lambda^2}{\bar{\varepsilon}^2}\log\big(8kL(L-1)\big).$$

Therefore, the assumed lower bound on d allows us to apply Lemma F.2 with $\varepsilon_1 = \alpha$ and $\varepsilon_2 = \beta$.

The conclusion of Lemma F.2 is deterministic. In particular, there exist fixed matrices

$$\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}}\times d},$$

depending only on d_{hid} and d, such that, for every (k,γ)-sparse attention pattern $\boldsymbol{A} \in \mathbb{R}^{L\times L}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L\times d_{\mathrm{hid}}}$ for which

$$\max\big\{\|\boldsymbol{X}\boldsymbol{W}_Q\|_{\mathrm{F}}, \|\boldsymbol{X}\boldsymbol{W}_K\|_{\mathrm{F}}\big\} \le \sqrt{\frac{3}{2}}\,L^{1/2}k^{1/4}\Lambda^{1/2},$$

and the matrix

$$\boldsymbol{M} := \mathrm{Softmax}\big((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top\big)$$

satisfies the following two ratio properties:

$$\frac{M_{i,j_1}}{M_{i,j_2}} < \alpha$$

whenever $A_{i,j_1} = 0$ and $A_{i,j_2} > 0$, and

$$e^{-\beta}\frac{A_{i,j_1}}{A_{i,j_2}} < \frac{M_{i,j_1}}{M_{i,j_2}} < e^{\beta}\frac{A_{i,j_1}}{A_{i,j_2}}$$

whenever $A_{i,j_1}, A_{i,j_2} > 0$.

It remains to convert these ratio bounds into a spectral-norm bound. Let

$$\boldsymbol{E} := \boldsymbol{M} - \boldsymbol{A}.$$

For each $i \in [L]$, define

$$S_i := \{j \in [L] : A_{i,j} > 0\}, \qquad \eta_i := \sum_{j\notin S_i} M_{i,j}.$$

Because $\boldsymbol{A}$ is row-stochastic, $S_i \neq \varnothing$. For any $j \notin S_i$, choosing an arbitrary $\ell \in S_i$ gives

$$M_{i,j} < \alpha M_{i,\ell} \leq \alpha.$$

Consequently,

$$\eta_i \leq L\alpha.$$

For $j \in S_i$, define

$$q_{i,j} := \frac{M_{i,j}}{1-\eta_i}, \qquad r_{i,j} := \frac{q_{i,j}}{A_{i,j}}.$$

Then

$$\sum_{j \in S_i} q_{i,j} = 1, \qquad \sum_{j \in S_i} A_{i,j} r_{i,j} = 1.$$

Furthermore, for every $j_1, j_2 \in S_i$,

$$e^{-\beta} < \frac{r_{i,j_1}}{r_{i,j_2}} < e^{\beta}.$$

Applying Lemma F.6 with $p_j = A_{i,j}$ and $\kappa = e^{\beta}$ yields

$$\sum_{j \in S_i} A_{i,j} |r_{i,j} - 1| \leq \frac{\beta}{2}.$$

Since

$$M_{i,j} = (1-\eta_i) A_{i,j} r_{i,j}, \qquad j \in S_i,$$

we obtain

$$\begin{aligned} \sum_{j \in S_i} |M_{i,j} - A_{i,j}| &\leq \sum_{j \in S_i} A_{i,j} |r_{i,j} - 1| + \eta_i \sum_{j \in S_i} A_{i,j} r_{i,j} \\ &\leq \frac{\beta}{2} + \eta_i. \end{aligned}$$

On the complement of S_i,

$$\sum_{j \notin S_i} |M_{i,j} - A_{i,j}| = \sum_{j \notin S_i} M_{i,j} = \eta_i.$$

Therefore,

$$\sum_{j=1}^{L} |E_{i,j}| \leq \frac{\beta}{2} + 2\eta_i \leq \frac{\beta}{2} + 2L\alpha,$$

and hence

$$\|\boldsymbol{E}\|_\infty \leq \frac{\beta}{2} + 2L\alpha.$$

We next control the maximum column sum. Fix $j \in [L]$ and define

$$I_j := \{i \in [L] : A_{i,j} > 0\}.$$

By the column-sparsity assumption, $|I_j| \leq k$. If $i \notin I_j$, then $A_{i,j} = 0$, and the first ratio property gives $M_{i,j} < \alpha$. Thus,

$$\sum_{i \notin I_j} |E_{i,j}| = \sum_{i \notin I_j} M_{i,j} \leq L\alpha.$$

If $i \in I_j$, then

$$\begin{aligned} |M_{i,j} - A_{i,j}| &= \left|(1-\eta_i) q_{i,j} - A_{i,j}\right| \\ &\leq |q_{i,j} - A_{i,j}| + \eta_i q_{i,j} \\ &= A_{i,j} |r_{i,j} - 1| + \eta_i q_{i,j}. \end{aligned}$$

The ratio bound and the identity $\sum_{\ell \in S_i} A_{i,\ell} r_{i,\ell} = 1$ imply

$$e^{-\beta} \leq r_{i,j} \leq e^{\beta}.$$

Since $q_{i,j} \le 1$ and $\eta_i \le L\alpha$, it follows that

$$|M_{i,j} - A_{i,j}| \le A_{i,j}(e^\beta - 1) + L\alpha.$$

Consequently,

$$\begin{aligned}\sum_{i=1}^{L} |E_{i,j}| &\le L\alpha + (e^\beta - 1)\sum_{i \in I_j} A_{i,j} + |I_j| L\alpha \\ &\le L\alpha + k(e^\beta - 1 + L\alpha).\end{aligned}$$

Taking the maximum over j gives

$$\|\boldsymbol{E}\|_1 \le L\alpha + k(e^\beta - 1 + L\alpha).$$

Substituting the definitions of α and β, we obtain

$$\|\boldsymbol{E}\|_\infty \le \frac{\bar{\varepsilon}}{2\sqrt{k}} + \frac{\bar{\varepsilon}}{32\sqrt{k}} = \frac{17\bar{\varepsilon}}{32\sqrt{k}}.$$

Because $0 < \beta \le 1$, convexity of the exponential function gives

$$e^\beta - 1 \le (e-1)\beta.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{E}\|_1 &\le \frac{\bar{\varepsilon}}{64\sqrt{k}} + k\left((e-1)\frac{\bar{\varepsilon}}{\sqrt{k}} + \frac{\bar{\varepsilon}}{64\sqrt{k}}\right) \\ &\le \left(e - 1 + \frac{1}{32}\right)\bar{\varepsilon}\sqrt{k}.\end{aligned}$$

Finally, the standard inequality $\|\boldsymbol{E}\|_2 \le \sqrt{\|\boldsymbol{E}\|_1 \|\boldsymbol{E}\|_\infty}$ yields

$$\begin{aligned}\|\boldsymbol{M} - \boldsymbol{A}\|_2 &\le \bar{\varepsilon}\sqrt{\frac{17}{32}\left(e - 1 + \frac{1}{32}\right)} \\ &< \bar{\varepsilon} \le \varepsilon.\end{aligned}$$

This proves the result. □

F.6 Proof of Proposition 3.2

Proposition 3.2 (Restatement). *Let $L > 1$ and $0 < \varepsilon < 1/2$. Let $\boldsymbol{M}_{\mathrm{MASK}}$ be either the zero mask or the causal mask. For every $d \ge 1$ and every $\boldsymbol{Z} \in \mathbb{R}^{L \times d}$ satisfying $\left\|\mathrm{Softmax}\left(\boldsymbol{Z}\boldsymbol{Z}^\top + \boldsymbol{M}_{\mathrm{MASK}}\right) - \boldsymbol{P}^{\mathrm{prev}}\right\|_2 < \varepsilon$, we have*

$$\|\boldsymbol{Z}\|_F > \sqrt{L(L-1)\log\frac{1-\varepsilon}{\varepsilon}}.$$

Proof of Proposition 3.2. Let

$$\boldsymbol{M} := \mathrm{Softmax}\left(\boldsymbol{Z}\boldsymbol{Z}^\top + \boldsymbol{M}_{\mathrm{MASK}}\right), \qquad \gamma_\varepsilon := \log\frac{1-\varepsilon}{\varepsilon} > 0.$$

Since each matrix entry is bounded in absolute value by the spectral norm, the error assumption gives, for every $i = 2, \dots, L$,

$$\boldsymbol{M}_{i,i-1} > 1 - \varepsilon, \qquad \boldsymbol{M}_{i,i} < \varepsilon.$$

Both positions $i-1$ and i are allowed under either mask. Therefore,

$$\frac{\boldsymbol{M}_{i,i-1}}{\boldsymbol{M}_{i,i}} = \exp\left(\langle \boldsymbol{z}_i, \boldsymbol{z}_{i-1}\rangle - \|\boldsymbol{z}_i\|_2^2\right) > \frac{1-\varepsilon}{\varepsilon},$$

where $\boldsymbol{z}_i^\top$ denotes the i-th row of $\boldsymbol{Z}$. Hence

$$\langle \boldsymbol{z}_i, \boldsymbol{z}_{i-1} \rangle - \|\boldsymbol{z}_i\|_2^2 > \gamma_\varepsilon.$$

Writing $r_i := \|\boldsymbol{z}_i\|_2$, the identity

$$2\left(\langle \boldsymbol{z}_i, \boldsymbol{z}_{i-1} \rangle - \|\boldsymbol{z}_i\|_2^2\right) = r_{i-1}^2 - r_i^2 - \|\boldsymbol{z}_{i-1} - \boldsymbol{z}_i\|_2^2$$

implies

$$r_{i-1}^2 > r_i^2 + 2\gamma_\varepsilon, \qquad i = 2, \ldots, L.$$

Iterating gives $r_i^2 > r_L^2 + 2(L-i)\gamma_\varepsilon$ for $i < L$. Consequently,

$$\begin{aligned} \|\boldsymbol{Z}\|_F^2 &= \sum_{i=1}^{L} r_i^2 \\ &> L r_L^2 + 2\gamma_\varepsilon \sum_{i=1}^{L-1} (L-i) \\ &\geq L(L-1)\gamma_\varepsilon. \end{aligned}$$

Taking square roots proves the claim. □

F.7 Supporting Lemmas

Lemma F.1. *Let $1 \leq m \leq d$, let $\boldsymbol{W}_K \in \mathbb{R}^{d_{\text{hid}} \times d}$ satisfy $\operatorname{rank}(\boldsymbol{W}_K) \geq m$, and suppose that $d_{\text{hid}} \geq 2m$. Then there exists $\boldsymbol{W}_Q \in \mathbb{R}^{d_{\text{hid}} \times d}$ such that, for every $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{L \times m}$, there exists $\boldsymbol{X} \in \mathbb{R}^{L \times d_{\text{hid}}}$ satisfying*

$$(\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top = \boldsymbol{U}\boldsymbol{V}^\top.$$

Proof of Lemma F.1. Since $\operatorname{rank}(\boldsymbol{W}_K) \geq m$, we can choose a column orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{d \times m}$, $\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}$, such that $\boldsymbol{C} := \boldsymbol{W}_K \boldsymbol{R} \in \mathbb{R}^{d_{hid} \times m}$ is full column rank, that is, $\operatorname{rank}(\boldsymbol{C}) = m$. Since $d_{hid} \geq m$, the dimension of space $\operatorname{col}(\boldsymbol{C})^\perp$ is at least m. Consequently, we can construct $\boldsymbol{G} \in \mathbb{R}^{d_{hid} \times m}$ such that $\operatorname{rank}[\boldsymbol{G}, \boldsymbol{C}] = 2m$. Define

$$\boldsymbol{W}_Q := \boldsymbol{G}\boldsymbol{R}^\top.$$

Then, for the previously fixed $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{L \times m}$, as $\boldsymbol{H} := [\boldsymbol{G}, \boldsymbol{C}] \in \mathbb{R}^{d_{hid} \times 2m}$ is full column rank, we define

$$\boldsymbol{X} = [\boldsymbol{U}, \boldsymbol{V}](\boldsymbol{H}^\top \boldsymbol{H})^{-1} \boldsymbol{H}^\top.$$

It is straightforward to verify that

$$\boldsymbol{X}\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{G}, \boldsymbol{C}] = [\boldsymbol{X}\boldsymbol{G}, \boldsymbol{X}\boldsymbol{C}] = [\boldsymbol{U}, \boldsymbol{V}].$$

Then, we obtain

$$\boldsymbol{X}\boldsymbol{G} = \boldsymbol{U}, \quad \boldsymbol{X}\boldsymbol{C} = \boldsymbol{V}.$$

On the other hand,

$$\boldsymbol{X}\boldsymbol{W}_Q = \boldsymbol{X}\boldsymbol{G}\boldsymbol{R}^\top = \boldsymbol{U}\boldsymbol{R}^\top,$$

and

$$\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{R} = \boldsymbol{X}\boldsymbol{C} = \boldsymbol{V}.$$

Consequently

$$\begin{aligned} (\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top &= \boldsymbol{X}\boldsymbol{G}\boldsymbol{R}^\top \boldsymbol{W}_K^\top \boldsymbol{X}^\top \\ &= (\boldsymbol{X}\boldsymbol{G})(\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{R})^\top \\ &= \boldsymbol{U}\boldsymbol{V}^\top, \end{aligned}$$

which completes the proof.

□

Lemma F.2. *Let $L > 1$, $1 \le k \le L$, and $d_{\mathrm{hid}} \ge d$ be positive integers, where d is even. Let*

$$0 < \varepsilon_1 < 1, \qquad 0 < \varepsilon_2 < \sqrt{2}, \qquad \gamma \ge 1,$$

and define

$$\Lambda := \max\left\{\log\gamma - \log\varepsilon_1 + \varepsilon_2,\, 1\right\}.$$

Suppose that

$$d \ge \max\left\{32\varepsilon_2^{-2}k\Lambda^2\log\left(8kL(L-1)\right),\; 64\log(16L)\right\}. \tag{F.2}$$

Then there exist fixed matrices

$$\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_{\mathrm{hid}}\times d},$$

depending only on d_{hid} and d, such that the following holds.

For every (k,γ)-sparse attention pattern $\boldsymbol{A}$. there exists a matrix $\boldsymbol{X} \in \mathbb{R}^{L\times d_{\mathrm{hid}}}$ satisfying

$$\max\left\{\|\boldsymbol{X}\boldsymbol{W}_Q\|_F,\, \|\boldsymbol{X}\boldsymbol{W}_K\|_F\right\} \le \sqrt{\frac{3}{2}}\,\sqrt{L}\,k^{1/4}\sqrt{\Lambda}, \tag{F.3}$$

such that

$$\boldsymbol{M} = \mathrm{Softmax}((\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top)$$

satisfies

- *for every $i, j_1, j_2 \in [L]$ such that $\boldsymbol{A}_{i,j_1} = 0$ and $\boldsymbol{A}_{i,j_2} > 0$,*
$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} < \varepsilon_1;$$
- *for every $i, j_1, j_2 \in [L]$ such that $\boldsymbol{A}_{i,j_1} > 0$ and $\boldsymbol{A}_{i,j_2} > 0$,*
$$e^{-\varepsilon_2}\frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}} < \frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} < e^{\varepsilon_2}\frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}}.$$

Proof of Lemma F.2. Fix an admissible matrix $\boldsymbol{A}$. For each $i \in [L]$, define

$$A_i^{\min} := \min_{j:\,\boldsymbol{A}_{i,j}>0} \boldsymbol{A}_{i,j}.$$

Since $\boldsymbol{A}$ is right-stochastic, every row of $\boldsymbol{A}$ contains at least one positive entry, and hence $A_i^{\min}$ is well-defined.

Define $\boldsymbol{B} \in \mathbb{R}^{L\times L}$ by

$$\boldsymbol{B}_{i,j} = \begin{cases} 0, & \boldsymbol{A}_{i,j} = 0, \\ \log\dfrac{\boldsymbol{A}_{i,j}}{A_i^{\min}} - \log\varepsilon_1 + \varepsilon_2, & \boldsymbol{A}_{i,j} > 0. \end{cases} \tag{F.4}$$

By the γ-variation-bounded assumption,

$$0 \le \log\frac{\boldsymbol{A}_{i,j}}{A_i^{\min}} \le \log\gamma$$

whenever $\boldsymbol{A}_{i,j} > 0$. Therefore,

$$0 \le \boldsymbol{B}_{i,j} \le \Lambda. \tag{F.5}$$

Moreover, each row of $\boldsymbol{B}$ contains at most k nonzero entries. Consequently,

$$\|\boldsymbol{B}_{i,:}\|_2 \le \sqrt{k}\Lambda, \qquad i \in [L]. \tag{F.6}$$

We next show the existence of a suitable low-dimensional projection. Let

$$m := \frac{d}{2},$$

and consider a random matrix $\boldsymbol{R} \in \mathbb{R}^{m \times L}$ whose entries are independently distributed according to

$$\boldsymbol{R}_{a,b} \overset{\text{i.i.d.}}{\sim} N(0, 1/m).$$

The randomness of $\boldsymbol{R}$ will only be used to establish existence; a deterministic realization will be fixed below.

For distinct $j_1, j_2 \in [L]$, let

$$\boldsymbol{h}_{j_1,j_2} := \boldsymbol{e}_{j_1} - \boldsymbol{e}_{j_2},$$

where $\boldsymbol{e}_j$ is the j-th standard basis vector in $\mathbb{R}^L$. Then

$$\|\boldsymbol{h}_{j_1,j_2}\|_2 = \sqrt{2},$$

and

$$\left\langle \boldsymbol{B}_{i,:}^\top, \boldsymbol{h}_{j_1,j_2} \right\rangle = \boldsymbol{B}_{i,j_1} - \boldsymbol{B}_{i,j_2}. \tag{F.7}$$

Set

$$t := \frac{\varepsilon_2}{\sqrt{2k}\Lambda}.$$

Since $0 < \varepsilon_2 < \sqrt{2}$, $k \geq 1$, and $\Lambda \geq 1$, we have $0 < t < 1$. Applying Lemma F.3 with

$$\boldsymbol{u} = \boldsymbol{B}_{i,:}^\top, \qquad \boldsymbol{v} = \boldsymbol{h}_{j_1,j_2},$$

and using equation F.6, we obtain

$$\begin{aligned} &\mathbb{P}\left(\left|\left\langle \boldsymbol{R}\boldsymbol{B}_{i,:}^\top, \boldsymbol{R}\boldsymbol{h}_{j_1,j_2} \right\rangle - \boldsymbol{B}_{i,j_1} + \boldsymbol{B}_{i,j_2}\right| \geq \varepsilon_2\right) \\ &\qquad \leq 2\exp\left(-\frac{m\varepsilon_2^2}{16k\Lambda^2}\right) = 2\exp\left(-\frac{d\varepsilon_2^2}{32k\Lambda^2}\right). \end{aligned} \tag{F.8}$$

For each $i \in [L]$, let

$$s_i := \left|\{j \in [L] : \boldsymbol{A}_{i,j} > 0\}\right|.$$

Then $s_i \leq k$. We only need to control two types of ordered pairs:

$$\begin{aligned} &\boldsymbol{A}_{i,j_1} = 0, \quad \boldsymbol{A}_{i,j_2} > 0, \\ &\boldsymbol{A}_{i,j_1} > 0, \quad \boldsymbol{A}_{i,j_2} > 0, \quad j_1 \neq j_2. \end{aligned}$$

Their total number in row i is

$$s_i(L - s_i) + s_i(s_i - 1) = s_i(L - 1) \leq k(L - 1).$$

Hence, by a union bound over all rows and all relevant ordered pairs,

$$\mathbb{P}(\mathcal{E}_1^c) \leq 2kL(L-1)\exp\left(-\frac{d\varepsilon_2^2}{32k\Lambda^2}\right), \tag{F.9}$$

where $\mathcal{E}_1$ denotes the event that

$$\left|\left\langle \boldsymbol{R}\boldsymbol{B}_{i,:}^\top, \boldsymbol{R}\boldsymbol{h}_{j_1,j_2} \right\rangle - \boldsymbol{B}_{i,j_1} + \boldsymbol{B}_{i,j_2}\right| < \varepsilon_2 \tag{F.10}$$

holds simultaneously for all relevant triples (i, j_1, j_2).

By equation F.2,

$$\frac{d\varepsilon_2^2}{32k\Lambda^2} \geq \log\left(8kL(L-1)\right),$$

and therefore

$$\mathbb{P}(\mathcal{E}_1^c) \le \frac{1}{4}. \tag{F.11}$$

We next control the norms of the projected vectors. For any fixed $\boldsymbol{u} \in \mathbb{R}^L$, Lemma F.3 applied with $\boldsymbol{u} = \boldsymbol{v}$ and $t = 1/2$ gives

$$\mathbb{P}\left(\left|\|\boldsymbol{R}\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2\right| \ge \frac{1}{2}\|\boldsymbol{u}\|_2^2\right) \le 2\exp(-m/32). \tag{F.12}$$

Apply equation F.12 to the $2L$ vectors

$$\boldsymbol{B}_{1,:}^\top, \dots, \boldsymbol{B}_{L,:}^\top, \qquad \boldsymbol{e}_1, \dots, \boldsymbol{e}_L.$$

A union bound yields an event $\mathcal{E}_2$ satisfying

$$\mathbb{P}(\mathcal{E}_2^c) \le 4L\exp(-m/32) = 4L\exp(-d/64), \tag{F.13}$$

such that on $\mathcal{E}_2$,

$$\|\boldsymbol{R}\boldsymbol{B}_{i,:}^\top\|_2^2 \le \frac{3}{2}\|\boldsymbol{B}_{i,:}\|_2^2, \qquad i \in [L], \tag{F.14}$$

and

$$\|\boldsymbol{R}\boldsymbol{e}_j\|_2^2 \le \frac{3}{2}, \qquad j \in [L]. \tag{F.15}$$

Again by equation F.2,

$$d \ge 64\log(16L),$$

so

$$\mathbb{P}(\mathcal{E}_2^c) \le \frac{1}{4}. \tag{F.16}$$

Combining equation F.11 and equation F.16,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) &\ge 1 - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c)\\ &\ge \frac{1}{2} > 0.\end{aligned}$$

Therefore, there exists at least one deterministic matrix

$$\boldsymbol{R}_\star \in \mathbb{R}^{m\times L}$$

for which both $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Fix such an $\boldsymbol{R}_\star$ for the remainder of the proof.

Define the balancing factor

$$c := k^{-1/4}\Lambda^{-1/2}, \tag{F.17}$$

and set

$$\begin{aligned}\boldsymbol{X}^{(1)} &:= c\,\boldsymbol{B}\boldsymbol{R}_\star^\top \in \mathbb{R}^{L\times m},\\ \boldsymbol{X}^{(2)} &:= c^{-1}\boldsymbol{R}_\star^\top \in \mathbb{R}^{L\times m}.\end{aligned}$$

We then define

$$\boldsymbol{X} := \begin{bmatrix}\boldsymbol{X}^{(1)} & \boldsymbol{X}^{(2)} & \boldsymbol{0}_{L\times(d_{\mathrm{hid}}-d)}\end{bmatrix} \in \mathbb{R}^{L\times d_{\mathrm{hid}}}. \tag{F.18}$$

Choose the fixed matrices

$$\boldsymbol{W}_Q = \begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}_{m\times m}\\ \boldsymbol{0}_{m\times m} & \boldsymbol{0}_{m\times m}\\ \boldsymbol{0}_{(d_{\mathrm{hid}}-d)\times m} & \boldsymbol{0}_{(d_{\mathrm{hid}}-d)\times m}\end{bmatrix}, \tag{F.19}$$

$$\boldsymbol{W}_K = \begin{bmatrix}\boldsymbol{0}_{m\times m} & \boldsymbol{0}_{m\times m}\\ \boldsymbol{I}_m & \boldsymbol{0}_{m\times m}\\ \boldsymbol{0}_{(d_{\mathrm{hid}}-d)\times m} & \boldsymbol{0}_{(d_{\mathrm{hid}}-d)\times m}\end{bmatrix}. \tag{F.20}$$

These matrices depend only on d_{hid} and d, and in particular are independent of $\boldsymbol{A}$.

By construction,

$$\boldsymbol{X}\boldsymbol{W}_Q = \begin{bmatrix}\boldsymbol{X}^{(1)} & \boldsymbol{0}_{L\times m}\end{bmatrix},$$
$$\boldsymbol{X}\boldsymbol{W}_K = \begin{bmatrix}\boldsymbol{X}^{(2)} & \boldsymbol{0}_{L\times m}\end{bmatrix}.$$

Hence the score matrix satisfies

$$\begin{aligned}\boldsymbol{S} &:= (\boldsymbol{X}\boldsymbol{W}_Q)(\boldsymbol{X}\boldsymbol{W}_K)^\top \\ &= \boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)\top} \\ &= \boldsymbol{B}\boldsymbol{R}_\star^\top \boldsymbol{R}_\star. \end{aligned} \tag{F.21}$$

In particular, the balancing factor c cancels from the score matrix.

Since $\boldsymbol{R}_\star \in \mathcal{E}_1$, for every relevant triple (i, j_1, j_2),

$$|\boldsymbol{S}_{i,j_1} - \boldsymbol{S}_{i,j_2} - \boldsymbol{B}_{i,j_1} + \boldsymbol{B}_{i,j_2}| < \varepsilon_2. \tag{F.22}$$

Let

$$\boldsymbol{M}_{i,j} = \frac{\exp(\boldsymbol{S}_{i,j})}{\sum_{\ell=1}^{L}\exp(\boldsymbol{S}_{i,\ell})}.$$

For every $i, j_1, j_2 \in [L]$, the normalization factor cancels in the ratio, and therefore

$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} = \exp\left(\boldsymbol{S}_{i,j_1} - \boldsymbol{S}_{i,j_2}\right). \tag{F.23}$$

We now verify the two desired ratio properties.

First, suppose that

$$\boldsymbol{A}_{i,j_1} = 0, \qquad \boldsymbol{A}_{i,j_2} > 0.$$

Then

$$\boldsymbol{B}_{i,j_1} = 0$$

and

$$\boldsymbol{B}_{i,j_2} = \log\frac{\boldsymbol{A}_{i,j_2}}{A_i^{\min}} - \log\varepsilon_1 + \varepsilon_2.$$

It follows from equation F.22 that

$$\begin{aligned}\boldsymbol{S}_{i,j_1} - \boldsymbol{S}_{i,j_2} &< \boldsymbol{B}_{i,j_1} - \boldsymbol{B}_{i,j_2} + \varepsilon_2 \\ &= -\log\frac{\boldsymbol{A}_{i,j_2}}{A_i^{\min}} + \log\varepsilon_1 \\ &\le \log\varepsilon_1,\end{aligned}$$

where the last inequality follows from $\boldsymbol{A}_{i,j_2} \ge A_i^{\min}$. Thus,

$$\frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} < \varepsilon_1.$$

Next, suppose that

$$\boldsymbol{A}_{i,j_1} > 0, \qquad \boldsymbol{A}_{i,j_2} > 0.$$

The row-dependent terms in equation F.4 cancel, giving

$$\boldsymbol{B}_{i,j_1} - \boldsymbol{B}_{i,j_2} = \log\frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}}.$$

Therefore, equation F.22 implies

$$-\varepsilon_2 < \boldsymbol{S}_{i,j_1} - \boldsymbol{S}_{i,j_2} - \log \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}} < \varepsilon_2.$$

Exponentiating yields

$$e^{-\varepsilon_2} \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}} < \frac{\boldsymbol{M}_{i,j_1}}{\boldsymbol{M}_{i,j_2}} < e^{\varepsilon_2} \frac{\boldsymbol{A}_{i,j_1}}{\boldsymbol{A}_{i,j_2}}.$$

It remains to verify the norm bound. Since $\boldsymbol{R}_\star \in \mathcal{E}_2$, we have

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{W}_Q\|_F^2 = \|\boldsymbol{X}^{(1)}\|_F^2 = c^2 \sum_{i=1}^{L} \|\boldsymbol{R}_\star \boldsymbol{B}_{i,:}^\top\|_2^2 \le \frac{3}{2} c^2 \sum_{i=1}^{L} \|\boldsymbol{B}_{i,:}\|_2^2 \\ = \frac{3}{2} c^2 \|\boldsymbol{B}\|_F^2 \le \frac{3}{2} c^2 L k \Lambda^2 = \frac{3}{2} L \sqrt{k} \Lambda.\end{aligned}$$

Similarly,

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{W}_K\|_F^2 = \|\boldsymbol{X}^{(2)}\|_F^2 = c^{-2} \|\boldsymbol{R}_\star^\top\|_F^2 = c^{-2} \sum_{j=1}^{L} \|\boldsymbol{R}_\star \boldsymbol{e}_j\|_2^2 \\ \le \frac{3}{2} c^{-2} L = \frac{3}{2} L \sqrt{k} \Lambda.\end{aligned}$$

Taking square roots gives

$$\max\left\{\|\boldsymbol{X}\boldsymbol{W}_Q\|_F,\, \|\boldsymbol{X}\boldsymbol{W}_K\|_F\right\} \le \sqrt{\frac{3}{2}}\, \sqrt{L}\, k^{1/4} \sqrt{\Lambda}.$$

This completes the proof. □

Lemma F.3. *Let $\mathbf{R} \in \mathbb{R}^{m \times n}$ and $\mathbf{R}_{i,j} \sim N(0, 1/m)$. Then for any fixed vector $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and any $t \in (0, 1)$, we have*

$$\mathbb{P}\left(|\langle \mathbf{R}\boldsymbol{u}, \mathbf{R}\boldsymbol{v}\rangle - \langle \boldsymbol{u}, \boldsymbol{v}\rangle| \ge t\|\boldsymbol{u}\|\|\boldsymbol{v}\|\right) \le 2\exp\left(-\frac{mt^2}{8}\right).$$

Proof of Lemma F.3. If $\boldsymbol{u} = \mathbf{0}$ or $\boldsymbol{v} = \mathbf{0}$, the conclusion is trivial. Let

$$\bar{\boldsymbol{u}} = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}, \quad \bar{\boldsymbol{v}} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}, \quad \rho = \langle \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}\rangle.$$

Denote the r-th row of $\sqrt{m}\mathbf{R}$ as $\mathbf{g}_r^\top$, and define

$$\mathrm{x}_r = \langle \mathbf{g}_r, \bar{\boldsymbol{u}}\rangle, \quad \mathrm{y}_r = \langle \mathbf{g}_r, \bar{\boldsymbol{v}}\rangle.$$

Then it is clear that $(\mathrm{x}_r, \mathrm{y}_r)$ is a standard two dimensional Gaussian random variable with a correlation coefficient of ρ, and further

$$\langle \mathbf{R}\bar{\boldsymbol{u}}, \mathbf{R}\bar{\boldsymbol{v}}\rangle - \rho = \frac{1}{m} \sum_{r=1}^{m} \left(\mathrm{x}_r \mathrm{y}_r - \rho\right).$$

We let

$$a = \frac{1+\rho}{2}, \quad b = \frac{1-\rho}{2}.$$

Using Lemma F.4, there exists independent random variables $\mathrm{z}_1, \mathrm{z}_2 \sim N(0, 1)$ such that

$$\mathrm{x}_r \mathrm{y}_r - \rho \overset{d}{=} a(\mathrm{z}_1^2 - 1) - b(\mathrm{z}_2^2 - 1).$$

As a result, for any $\lambda \in \mathbb{R}$ with $|\lambda| \le \frac{1}{4}$, we have

$$\begin{aligned}\mathbb{E}\exp(\lambda(x_r y_r - \rho)) &= \mathbb{E}\exp(a\lambda(\mathrm{z}_1^2 - 1) - b\lambda(\mathrm{z}_2^2 - 1)) \\ &= \mathbb{E}\exp(a\lambda(\mathrm{z}_1^2 - 1)) \cdot \mathbb{E}\exp(-b\lambda(\mathrm{z}_2^2 - 1)),\end{aligned}$$

which implies

$$\log \mathbb{E}\exp(\lambda(x_r y_r - \rho)) = \log \mathbb{E}\exp(a\lambda(\mathrm{z}_1^2 - 1)) + \log \mathbb{E}\exp(-b\lambda(\mathrm{z}_2^2 - 1)).$$

According to Lemma F.5, we know that for any $\mathrm{z} \sim N(0,1)$ and $s \in \mathbb{R}$ with $|s| \le \frac{1}{4}$, the following holds

$$\log \mathbb{E}\exp(s(\mathrm{z}^2 - 1)) = -s - \frac{1}{2}\log(1 - 2s) \le 2s^2.$$

Notice that when $|\lambda| \le 1/4$, we have $|a\lambda| \le 1/4$ and $|-b\lambda| \le 1/4$. Hence, we have

$$\log \mathbb{E}\exp(\lambda(\mathrm{x}_r \mathrm{y}_r - \rho)) \le 2a^2\lambda^2 + 2b^2\lambda^2 \le 2\lambda^2, \tag{F.24}$$

where in the last inequality we use the fact that if $a+b=1$ and $a, b \ge 0$, then $a^2+b^2 \le (a+b)^2 = 1$. It remains to derive the probability bound. Let $\mathrm{w}_r = \mathrm{x}_r \mathrm{y}_r - \rho$, which are mutually independet. For any $\lambda > 0$, using Markov inequality, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r \ge t\right) &= \mathbb{P}\left(e^{\lambda\sum_{r=1}^m \mathrm{w}_r} \ge e^{\lambda m t}\right) \\ &\le e^{-\lambda m t}\mathbb{E}e^{\lambda\sum_{r=1}^m \mathrm{w}_r}.\end{aligned}$$

By independence,

$$\mathbb{E}e^{\lambda\sum_{r=1}^m \mathrm{w}_r} = \prod_{r=1}^m \mathbb{E}e^{\lambda \mathrm{w}_r}.$$

Applying Eq.F.24, for any $0 < \lambda \le 1/4$, we have

$$\mathbb{E}e^{\lambda \mathrm{w}_r} \le e^{2\lambda^2}.$$

Thus

$$\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r \ge t\right) \le e^{-\lambda m t}\left(e^{2\lambda^2}\right)^m = \exp(m(-\lambda t + 2\lambda^2)).$$

Let $\lambda = t/4$. As $0 < t < 1$, it is clear that $0 < \lambda < 1/4$, and $-\lambda t + 2\lambda^2 = -t^2/8$. Therefore,

$$\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r \ge t\right) \le e^{-mt^2/8}.$$

On the other side,

$$\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r \le -t\right) = \mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m (-\mathrm{w}_r) \ge t\right).$$

Since Eq.F.24 holds for any $|\lambda| \le 1/4$, the following holds naturally

$$\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r \le -t\right) \le e^{-mt^2/8}.$$

Combing the two cases, we have

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^m \mathrm{w}_r\right| \ge t\right) \le 2e^{-mt^2/8},$$

which completes the proof. □

Lemma F.4. *Let* $(\mathrm{x},\mathrm{y}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$ *with* $\rho \in [-1,1]$. *Define* $a = \frac{1+\rho}{2}$, $b = \frac{1-\rho}{2}$. *Then*

$$\mathrm{xy} - \rho \overset{d}{=} a\left(\mathrm{z}_1^2 - 1\right) - b\left(\mathrm{z}_2^2 - 1\right),$$

where $\mathrm{z}_1, \mathrm{z}_2 \overset{\text{i.i.d}}{\sim} N(0,1)$.

Proof of Lemma F.4. Define

$$\mathrm{x}' = \sqrt{a}\mathrm{z}_1 + \sqrt{b}\mathrm{z}_2, \quad \mathrm{y}' = \sqrt{a}\mathrm{z}_1 - \sqrt{b}\mathrm{z}_2.$$

We first verify that $(\mathrm{x}', \mathrm{y}')$ has the same Gaussian distribution as (x, y). Specifically,

$$\mathrm{Var}(\mathrm{x}\prime) = \mathrm{Var}(\mathrm{y}') = a + b = 1,$$

and

$$\begin{aligned}\mathrm{Cov}(\mathrm{x}', \mathrm{y}') &= \mathbb{E}\left[(\sqrt{a}\mathrm{z}_1 + \sqrt{b}\mathrm{z}_2)(\sqrt{a}\mathrm{z}_1 - \sqrt{b}\mathrm{z}_2)\right]\\ &= \mathbb{E}a\mathrm{z}_1^2 - \mathbb{E}b\mathrm{z}_2^2\\ &= a - b\\ &= \rho.\end{aligned}$$

Therefore, since both (x, y) and $(\mathrm{x}', \mathrm{y}')$ are zero-mean jointly Gaussian vectors with the same covariance matrix, we have

$$(\mathrm{x}, \mathrm{y}) \overset{d}{=} (\mathrm{x}', \mathrm{y}').$$

It follows that

$$\begin{aligned}\mathrm{xy} &\overset{d}{=} (\sqrt{a}\mathrm{z}_1 + \sqrt{b}\mathrm{z}_2)(\sqrt{a}\mathrm{z}_1 - \sqrt{b}\mathrm{z}_2)\\ &= a\mathrm{z}_1^2 - b\mathrm{z}_2^2.\end{aligned}$$

Using $\rho = a - b$, we further obtain

$$\begin{aligned}\mathrm{xy} - \rho &\overset{d}{=} a\mathrm{z}_1^2 - b\mathrm{z}_2^2 - (a-b)\\ &= a(\mathrm{z}_1^2 - 1) - b(\mathrm{z}_2^2 - 1),\end{aligned}$$

which completes the proof.

□

Lemma F.5. *For any* $\mathrm{z} \sim N(0,1)$ *and* $s \in \mathbb{R}$ *with* $|s| \le \frac{1}{4}$, *we have*

$$\log \mathbb{E}\exp(s(\mathrm{z}^2 - 1)) = -s - \frac{1}{2}\log(1-2s) \le 2s^2.$$

Proof of Lemma F.5. By separating the constant term, it is clear that $\mathbb{E}e^{s(\mathrm{z}^2-1)} = e^{-s}\mathbb{E}e^{s\mathrm{z}^2}$. To derive $\mathbb{E}e^{s\mathrm{z}^2}$, which is the moment generating function of z^2, we have

$$\begin{aligned}\mathbb{E}e^{s\mathrm{z}^2} &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-\frac{z^2}{2} + sz^2\right)\mathrm{d}z\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-\frac{1-2s}{2}z^2\right)\mathrm{d}z\\ &= (1-2s)^{-1/2}.\end{aligned}$$

consequently

$$\mathbb{E}e^{s(\mathrm{z}^2-1)} = e^{-s}(1-2s)^{-1/2},$$

and therefore

$$\log \mathbb{E}e^{s(\mathrm{z}^2-1)} = -s - \frac{1}{2}\log(1-2s).$$

It remains to prove the upper bound. Define $f(s) = -s - \frac{1}{2}\log(1-2s)$ and $g(s) = 2s^2 - f(s)$. clearly, $g(0) = 0$. The derivatives of these functions are given by,

$$f'(s) = -1 + \frac{1}{1-2s} = \frac{2s}{1-2s},$$
$$g'(s) = 4s - \frac{2s}{1-2s} = \frac{2s(1-4s)}{1-2s}.$$

For $0 \le s \le 1/4$, all terms in the numerator and denominator of $g'(s)$ are non-negative, which implies

$$g'(s) \ge 0.$$

Hence, $g(s)$ is increasing on this interval, and we obtain $g(s) \ge g(0) = 0$. On the other hand, for $-1/4 \le s \le 0$, we have

$$g'(s) \le 0.$$

Thus, $g(s)$ is decreasing as s increases on this interval. Equivalently, $g(s) \ge g(0) = 0$. Combing the two cases yields

$$g(s) \ge g(0) = 0, \quad |s| \le 1/4,$$

which is equivalent to

$$f(s) \le 2s^2.$$

This completes the proof. □

Lemma F.6. *Let $p_1, \cdots, p_m \ge 0$ satisfy $\sum_{j=1}^m = 1$, and let $r_1, \cdots, r_m > 0$ satisfy $\sum_{j=1}^m p_j r_j = 1$. Suppose that, for some $\kappa \ge 1$, $\frac{r_j}{r_\ell} \le \kappa$, $\forall j, \ell \in [m]$. Then*

$$\sum_{j=1}^m p_j |r_j - 1| \le 2\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

In particular, if $\kappa = e^\beta$ for some $\beta \ge 0$, then

$$\sum_{j=1}^m p_j |r_j - 1| \le 2\tanh\left(\frac{\beta}{4}\right) \le \beta/2.$$

Proof of Lemma F.6. Define

$$a := \min_{j\in[m]} r_j, \quad b := \max_{j\in[m]} r_j.$$

Since $\sum_j p_j r_j = 1$, we have $a \le 1 \le b$. If $a = b$, then $r_j = 1$ for all j, and the conclusion is immediate.

Otherwise, by convexity of $x \mapsto |x-1|$, for every $x \in [a, b]$,

$$|x-1| \le \frac{b-x}{b-a}(1-a) + \frac{x-a}{b-a}(b-1).$$

Taking $x = r_j$, multiplying by p_j, and summing over j, we obtain

$$\sum_{j-1}^m p_j |r_j - 1| \le \frac{2(1-a)(b-1)}{b-1}.$$

Let $t := b/a \le \kappa$. Since $b = ta$,

$$\begin{aligned}
\frac{2(1-a)(b-1)}{b-a} &= \frac{2}{t-1}(t+1-ta-1/a) \\
&\le \frac{2}{t-1}(t+1-2\sqrt{t}) \\
&= 2\frac{\sqrt{t}-1}{\sqrt{t}+1} \\
&\le 2\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},
\end{aligned}$$

where the first inequality follows from

$$ta + 1/a \geq 2\sqrt{t}.$$

Finally, when $\kappa = e^{\beta}$,

$$2\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 2\frac{e^{\beta/2}-1}{e^{\beta/2}+1} = 2\tanh(\beta/4) \leq \beta/2.$$

□

G Technical Details in Section 4

G.1 Proof of Theorem 4.1

Proof of Theorem 4.1. We first establish an identifiability property of the effective matrices. Suppose that

$$\mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}_1, \boldsymbol{C}_1) = \mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}_2, \boldsymbol{C}_2), \qquad \forall\, \boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}},$$

where $\boldsymbol{C}_1 \neq \boldsymbol{0}$. We prove that

$$\boldsymbol{B}_1 = \boldsymbol{B}_2, \qquad \boldsymbol{C}_1 = \boldsymbol{C}_2.$$

For any $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}}$, set

$$\boldsymbol{X} = \boldsymbol{1}_L \boldsymbol{x}^\top.$$

For either mask, every value row equals $\boldsymbol{x}^\top \boldsymbol{C}_j$, and each attention row sums to one. Hence

$$\mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}_j, \boldsymbol{C}_j) = \boldsymbol{1}_L \boldsymbol{x}^\top \boldsymbol{C}_j.$$

Equality of the two attention functions gives

$$\boldsymbol{x}^\top \boldsymbol{C}_1 = \boldsymbol{x}^\top \boldsymbol{C}_2, \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{hid}}},$$

so that

$$\boldsymbol{C}_1 = \boldsymbol{C}_2 =: \boldsymbol{C}.$$

Because $\boldsymbol{C} \neq \boldsymbol{0}$, there exists $\boldsymbol{a} \in \mathbb{R}^{d_{\mathrm{hid}}}$ such that

$$\boldsymbol{c} := \boldsymbol{C}\boldsymbol{a} \neq \boldsymbol{0}.$$

Let $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{hid}}}$ be arbitrary, and let $\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{hid}}}$ satisfy

$$\boldsymbol{w}^\top \boldsymbol{c} \neq 0.$$

Construct an input whose rows are

$$\boldsymbol{x}_1 = \boldsymbol{v} + \boldsymbol{w}, \qquad \boldsymbol{x}_2 = \cdots = \boldsymbol{x}_L = \boldsymbol{v}.$$

Consider the last query position $i = L$. Under the effective score matrix $\boldsymbol{B}_j$, its score for the first token is

$$\boldsymbol{v}^\top \boldsymbol{B}_j (\boldsymbol{v} + \boldsymbol{w}),$$

whereas its score for each of the remaining $L - 1$ tokens is

$$\boldsymbol{v}^\top \boldsymbol{B}_j \boldsymbol{v}.$$

Thus, its attention weight on the first token is

$$\alpha_j = \frac{\exp\big(\boldsymbol{v}^\top \boldsymbol{B}_j (\boldsymbol{v} + \boldsymbol{w})\big)}{\exp\big(\boldsymbol{v}^\top \boldsymbol{B}_j (\boldsymbol{v} + \boldsymbol{w})\big) + (L-1)\exp\big(\boldsymbol{v}^\top \boldsymbol{B}_j \boldsymbol{v}\big)} = \frac{\exp\big(\boldsymbol{v}^\top \boldsymbol{B}_j \boldsymbol{w}\big)}{\exp\big(\boldsymbol{v}^\top \boldsymbol{B}_j \boldsymbol{w}\big) + L - 1}.$$

Consequently,

$$\mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}_j, \boldsymbol{C})_{i,:} = \boldsymbol{v}^\top \boldsymbol{C} + \alpha_j \boldsymbol{w}^\top \boldsymbol{C}.$$

Multiplying on the right by $\boldsymbol{a}$ yields

$$\mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}_j, \boldsymbol{C})_{i,:} \boldsymbol{a} = \boldsymbol{v}^\top \boldsymbol{c} + \alpha_j \boldsymbol{w}^\top \boldsymbol{c}.$$

Since the two attention functions are identical and $\boldsymbol{w}^\top \boldsymbol{c} \neq 0$, we obtain

$$\alpha_1 = \alpha_2.$$

The function

$$g(t) := \frac{e^t}{e^t + L - 1}$$

is strictly increasing. Therefore,

$$\boldsymbol{v}^\top \boldsymbol{B}_1 \boldsymbol{w} = \boldsymbol{v}^\top \boldsymbol{B}_2 \boldsymbol{w}.$$

Equivalently,

$$\boldsymbol{v}^\top (\boldsymbol{B}_1 - \boldsymbol{B}_2) \boldsymbol{w} = 0.$$

Because $\boldsymbol{v}$ is arbitrary, it follows that

$$(\boldsymbol{B}_1 - \boldsymbol{B}_2) \boldsymbol{w} = \mathbf{0}$$

for every $\boldsymbol{w}$ satisfying $\boldsymbol{w}^\top \boldsymbol{c} \neq 0$.

The set

$$\left\{ \boldsymbol{w} \in \mathbb{R}^{d_{\text{hid}}} : \boldsymbol{w}^\top \boldsymbol{c} \neq 0 \right\}$$

is dense in $\mathbb{R}^{d_{\text{hid}}}$. Since the map $\boldsymbol{w} \mapsto (\boldsymbol{B}_1 - \boldsymbol{B}_2)\boldsymbol{w}$ is continuous, it must vanish on all of $\mathbb{R}^{d_{\text{hid}}}$. Hence

$$\boldsymbol{B}_1 = \boldsymbol{B}_2.$$

We have thus proved the identifiability implication

$$\text{Attn}(\boldsymbol{X}; \boldsymbol{B}_1, \boldsymbol{C}_1) = \text{Attn}(\boldsymbol{X}; \boldsymbol{B}_2, \boldsymbol{C}_2) \text{ for all } \boldsymbol{X} \quad \Longrightarrow \quad (\boldsymbol{B}_1, \boldsymbol{C}_1) = (\boldsymbol{B}_2, \boldsymbol{C}_2),$$

provided that $\boldsymbol{C}_1 \neq \mathbf{0}$.

We now prove the lower bound. Suppose that a K=V head of dimension s exactly implements the teacher attention function. By the identifiability result,

$$\widetilde{\boldsymbol{B}} = \boldsymbol{B}, \qquad \widetilde{\boldsymbol{C}} = \boldsymbol{C}.$$

Therefore,

$$\boldsymbol{B}^\top = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_Q^\top, \qquad \boldsymbol{C} = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_O,$$

and hence

$$[\boldsymbol{B}^\top, \boldsymbol{C}] = \widetilde{\boldsymbol{W}} [\widetilde{\boldsymbol{W}}_Q^\top, \widetilde{\boldsymbol{W}}_O].$$

It follows that

$$\text{rank}([\boldsymbol{B}^\top, \boldsymbol{C}]) \leq \text{rank}(\widetilde{\boldsymbol{W}}) \leq s.$$

Since this holds for every exact K=V realization,

$$d^* \geq \text{rank}([\boldsymbol{B}^\top, \boldsymbol{C}]).$$

For the matching upper bound, let

$$r := \text{rank}([\boldsymbol{B}^\top, \boldsymbol{C}]).$$

Since $\boldsymbol{C} \neq \mathbf{0}$, we have $r \geq 1$. Take a rank factorization

$$[\boldsymbol{B}^\top, \boldsymbol{C}] = \widetilde{\boldsymbol{W}} \boldsymbol{H},$$

where

$$\widetilde{\boldsymbol{W}} \in \mathbb{R}^{d_{\text{hid}} \times r}, \qquad \boldsymbol{H} \in \mathbb{R}^{r \times 2d_{\text{hid}}}.$$

Partition $\boldsymbol{H}$ according to the two block columns:

$$\boldsymbol{H} = [\boldsymbol{H}_K, \boldsymbol{H}_V], \qquad \boldsymbol{H}_K, \boldsymbol{H}_V \in \mathbb{R}^{r \times d_{\text{hid}}}.$$

Define

$$\widetilde{\boldsymbol{W}}_Q := \boldsymbol{H}_K^\top, \qquad \widetilde{\boldsymbol{W}}_O := \boldsymbol{H}_V.$$

Then

$$\boldsymbol{B}^\top = \widetilde{\boldsymbol{W}} \boldsymbol{H}_K = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_Q^\top,$$

so

$$\boldsymbol{B} = \widetilde{\boldsymbol{W}}_Q \widetilde{\boldsymbol{W}}^\top,$$

and

$$\boldsymbol{C} = \widetilde{\boldsymbol{W}} \boldsymbol{H}_V = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_O.$$

Thus, a K=V head of dimension r exactly realizes the teacher attention function. Consequently,

$$d^* \le r.$$

Combining the upper and lower bounds gives

$$d^* = \operatorname{rank}([\boldsymbol{B}^\top, \boldsymbol{C}]).$$

Finally,

$$\operatorname{Col}([\boldsymbol{B}^\top, \boldsymbol{C}]) = \operatorname{Col}(\boldsymbol{B}^\top) + \operatorname{Col}(\boldsymbol{C}),$$

and therefore

$$d^* = \dim\left(\mathcal{S}_K + \mathcal{S}_V\right).$$

For an untied head of dimension s, the identities $\operatorname{rank}(\boldsymbol{B}) \le s$ and $\operatorname{rank}(\boldsymbol{C}) \le s$ are necessary, while separate rank factorizations of $\boldsymbol{B}$ and $\boldsymbol{C}$, padded to a common dimension, are sufficient. Since $\operatorname{rank}(\boldsymbol{B}) = r_K$ and $\operatorname{rank}(\boldsymbol{C}) = r_V$, its minimum dimension is therefore $\max\{r_K, r_V\}$.

If $\operatorname{rank}(\boldsymbol{W}_Q) = d$, then $\boldsymbol{W}_Q^\top$ has full row rank, which implies

$$\operatorname{Col}(\boldsymbol{W}_K \boldsymbol{W}_Q^\top) = \operatorname{Col}(\boldsymbol{W}_K).$$

Similarly, if $\operatorname{rank}(\boldsymbol{W}_O) = d$, then

$$\operatorname{Col}(\boldsymbol{W}_V \boldsymbol{W}_O) = \operatorname{Col}(\boldsymbol{W}_V).$$

Because

$$\boldsymbol{B}^\top = \boldsymbol{W}_K \boldsymbol{W}_Q^\top, \qquad \boldsymbol{C} = \boldsymbol{W}_V \boldsymbol{W}_O,$$

the final assertion follows. □

G.2 Approximate realization

Theorem 4.1 characterizes the head dimension required for exact realization, but its rank-based conclusion does not distinguish between directions of different magnitudes. For approximate realization on a bounded input domain, weak directions in the joint address–content matrix $\boldsymbol{M} = [\boldsymbol{B}^\top, \boldsymbol{C}]$ may be discarded with only a small change in the attention output. This suggests replacing the algebraic rank of $\boldsymbol{M}$ by its effective numerical rank. The following corollary makes this intuition quantitative: truncating the singular value decomposition of $\boldsymbol{M}$ produces a lower-dimensional K=V head whose uniform approximation error is controlled explicitly by the first discarded singular value.

Corollary G.1 (Approximate realization by joint low-rank compression)**.** *Consider the setting of Theorem 4.1, and let $\sigma_1(\boldsymbol{M}) \ge \cdots \ge \sigma_{d_{\mathrm{hid}}}(\boldsymbol{M}) \ge 0$ denote the singular values of $\boldsymbol{M}$, and set $\sigma_{d_{\mathrm{hid}}+1}(\boldsymbol{M}) := 0$. Then, for every $R > 0$ and every $1 \le s \le d_{\mathrm{hid}}$, there exists a $\boldsymbol{W}_K = \boldsymbol{W}_V$ student head of dimension s, with effective matrices $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{C}}$, such that*

$$\sup_{\substack{\boldsymbol{X} \in \mathbb{R}^{L \times d_{\mathrm{hid}}} \\ \|\boldsymbol{X}\|_F \le R}} \left\| \mathrm{Attn}(\boldsymbol{X}; \widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{C}}) - \mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}, \boldsymbol{C}) \right\|_F \le \left(\frac{R^3}{2} \|\boldsymbol{C}\|_2 + \sqrt{L} R \right) \sigma_{s+1}(\boldsymbol{M}). \quad \text{(G.1)}$$

In particular, for $\varepsilon > 0$, define the minimum head dimension required for uniform ε-approximation on the radius-R input ball by

$$d_{\varepsilon,R}^{\mathrm{app}} := \min \Bigg\{ s \in \mathbb{N} : s \ge 1, \text{ and there exist } \widetilde{\boldsymbol{W}}_Q, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}}_O \text{ such that } \sup_{\|\boldsymbol{X}\|_F \le R} \left\| \mathrm{Attn}(\boldsymbol{X}; \widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{C}}) - \mathrm{Attn}(\boldsymbol{X}; \boldsymbol{B}, \boldsymbol{C}) \right\|_F \le \varepsilon \Bigg\},$$

where $\widetilde{\boldsymbol{B}} = \widetilde{\boldsymbol{W}}_Q \widetilde{\boldsymbol{W}}^\top, \widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_O$. *Then*

$$d_{\varepsilon,R}^{\mathrm{app}} \le \min \left\{ 1 \le s \le d_{\mathrm{hid}} : \sigma_{s+1}(\boldsymbol{M}) \le \frac{\varepsilon}{\frac{R^3}{2}\|\boldsymbol{C}\|_2 + \sqrt{L}R} \right\}. \tag{G.2}$$

Consequently, if the singular values of the joint address–content matrix decay rapidly, a head dimension strictly smaller than $\mathrm{rank}(\boldsymbol{M})$ *may suffice for approximate realization. When* $s \ge \mathrm{rank}(\boldsymbol{M})$*, the right-hand side of equation G.1 is zero, recovering the exact realization in Theorem 4.1.*

Proof. Let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be a singular value decomposition of $\boldsymbol{M}$, and let $\boldsymbol{U}_s \in \mathbb{R}^{d_{\mathrm{hid}} \times s}$ contain the first s left singular vectors. Define the orthogonal projection

$$\boldsymbol{P}_s := \boldsymbol{U}_s \boldsymbol{U}_s^\top.$$

By the Eckart–Young–Mirsky theorem (Golub & Van Loan, 2013),

$$\|(\boldsymbol{I} - \boldsymbol{P}_s)\boldsymbol{M}\|_2 = \sigma_{s+1}(\boldsymbol{M}). \tag{G.3}$$

We construct a K=V student by setting

$$\widetilde{\boldsymbol{W}} := \boldsymbol{U}_s, \qquad \widetilde{\boldsymbol{W}}_Q := \boldsymbol{B}\boldsymbol{U}_s, \qquad \widetilde{\boldsymbol{W}}_O := \boldsymbol{U}_s^\top \boldsymbol{C}.$$

These matrices have dimensions

$$\widetilde{\boldsymbol{W}}_Q, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{d_{\mathrm{hid}} \times s}, \qquad \widetilde{\boldsymbol{W}}_O \in \mathbb{R}^{s \times d_{\mathrm{hid}}}.$$

Their effective matrices are

$$\widetilde{\boldsymbol{B}} = \widetilde{\boldsymbol{W}}_Q \widetilde{\boldsymbol{W}}^\top = \boldsymbol{B}\boldsymbol{U}_s\boldsymbol{U}_s^\top = \boldsymbol{B}\boldsymbol{P}_s,$$
$$\widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{W}}_O = \boldsymbol{U}_s\boldsymbol{U}_s^\top \boldsymbol{C} = \boldsymbol{P}_s\boldsymbol{C}.$$

Thus,

$$[\widetilde{\boldsymbol{B}}^\top, \widetilde{\boldsymbol{C}}] = [\boldsymbol{P}_s\boldsymbol{B}^\top, \boldsymbol{P}_s\boldsymbol{C}] = \boldsymbol{P}_s\boldsymbol{M}.$$

It follows from equation G.3 that

$$\|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_2 = \|(\boldsymbol{I} - \boldsymbol{P}_s)\boldsymbol{B}^\top\|_2 \le \sigma_{s+1}(\boldsymbol{M}), \tag{G.4}$$
$$\|\boldsymbol{C} - \widetilde{\boldsymbol{C}}\|_2 = \|(\boldsymbol{I} - \boldsymbol{P}_s)\boldsymbol{C}\|_2 \le \sigma_{s+1}(\boldsymbol{M}). \tag{G.5}$$

We next record the Lipschitz property of the row-wise Softmax. For $z \in \mathbb{R}^L$, write $p = \mathrm{softmax}(z)$. Its Jacobian is

$$\boldsymbol{J}(z) = \mathrm{diag}(p) - pp^\top.$$

The i-th diagonal entry is $p_i(1-p_i)$, while the sum of the absolute values of the off-diagonal entries in the i-th row is also $p_i(1-p_i)$. Hence, by the Gershgorin circle theorem (Horn & Johnson, 2012),

$$\|\boldsymbol{J}(z)\|_2 \le \max_{1 \le i \le L} 2p_i(1-p_i) \le \frac{1}{2}.$$

The integral form of the mean-value theorem therefore gives

$$\|\,\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_2 \le \frac{1}{2}\|z - z'\|_2.$$

Applying this inequality on each row's allowed coordinates yields, for any finite score matrices $\boldsymbol{S}, \widetilde{\boldsymbol{S}} \in \mathbb{R}^{L \times L}$,

$$\left\|\mathrm{Softmax}(\boldsymbol{S}) - \mathrm{Softmax}(\widetilde{\boldsymbol{S}})\right\|_F \le \frac{1}{2}\|\boldsymbol{S} - \widetilde{\boldsymbol{S}}\|_F. \tag{G.6}$$

Fix an input $\boldsymbol{X}$ satisfying $\|\boldsymbol{X}\|_F \le R$, and define

$$\boldsymbol{A} := \mathrm{Softmax}(\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^\top), \qquad \widetilde{\boldsymbol{A}} := \mathrm{Softmax}(\boldsymbol{X}\widetilde{\boldsymbol{B}}\boldsymbol{X}^\top).$$

Using equation G.6 and equation G.4, we obtain

$$\begin{aligned}\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_F &\le \frac{1}{2}\|\boldsymbol{X}(\boldsymbol{B}-\widetilde{\boldsymbol{B}})\boldsymbol{X}^\top\|_F \\ &\le \frac{1}{2}\|\boldsymbol{X}\|_F\|\boldsymbol{B}-\widetilde{\boldsymbol{B}}\|_2\|\boldsymbol{X}^\top\|_2 \\ &\le \frac{R^2}{2}\sigma_{s+1}(\boldsymbol{M}).\end{aligned}$$

Moreover, every row of $\widetilde{\boldsymbol{A}}$ is a probability vector. Therefore,

$$\|\widetilde{\boldsymbol{A}}\|_2 \le \|\widetilde{\boldsymbol{A}}\|_F \le \sqrt{L}.$$

The output difference can now be decomposed as

$$\begin{aligned}&\mathrm{Attn}(\boldsymbol{X};\boldsymbol{B},\boldsymbol{C}) - \mathrm{Attn}(\boldsymbol{X};\widetilde{\boldsymbol{B}},\widetilde{\boldsymbol{C}}) \\ &\quad = (\boldsymbol{A}-\widetilde{\boldsymbol{A}})\boldsymbol{X}\boldsymbol{C} + \widetilde{\boldsymbol{A}}\boldsymbol{X}(\boldsymbol{C}-\widetilde{\boldsymbol{C}}).\end{aligned}$$

Consequently,

$$\begin{aligned}&\Big\|\mathrm{Attn}(\boldsymbol{X};\boldsymbol{B},\boldsymbol{C}) - \mathrm{Attn}(\boldsymbol{X};\widetilde{\boldsymbol{B}},\widetilde{\boldsymbol{C}})\Big\|_F \\ &\le \|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_F\|\boldsymbol{X}\boldsymbol{C}\|_2 + \|\widetilde{\boldsymbol{A}}\|_2\|\boldsymbol{X}(\boldsymbol{C}-\widetilde{\boldsymbol{C}})\|_F \\ &\le \frac{R^2}{2}\sigma_{s+1}(\boldsymbol{M})\cdot R\|\boldsymbol{C}\|_2 + \sqrt{L}\,R\|\boldsymbol{C}-\widetilde{\boldsymbol{C}}\|_2 \\ &\le \left(\frac{R^3}{2}\|\boldsymbol{C}\|_2 + \sqrt{L}R\right)\sigma_{s+1}(\boldsymbol{M}),\end{aligned}$$

where the final step uses equation G.5. Taking the supremum over all $\boldsymbol{X}$ satisfying $\|\boldsymbol{X}\|_F \le R$ proves equation G.1.

If s satisfies the singular-value condition in equation G.2, the constructed student has uniform error at most ε, proving the claimed upper bound on $d^{\mathrm{app}}_{\varepsilon,R}$. Finally, when $s \ge \mathrm{rank}(\boldsymbol{M})$, we have $\sigma_{s+1}(\boldsymbol{M}) = 0$, so that

$$\widetilde{\boldsymbol{B}} = \boldsymbol{B}, \qquad \widetilde{\boldsymbol{C}} = \boldsymbol{C}.$$

The construction then gives an exact realization, consistent with Theorem 4.1.

□