A Theoretical Analysis of Why Masked Diffusion Models Mitigate the Reversal Curse

Moongyu Jeon [*1] **Sangwoo Shin** [*1] **Bumjun Kim** [1] **Kyelim Lee** [1] **Albert No** [†1]

Abstract

Autoregressive language models (ARMs) suffer from the reversal curse: after learning "A is B," they often fail on the reverse query "B is A." Masked diffusion language models (MDMs) exhibit this failure in a much weaker form, but the underlying reason has remained unclear. A common explanation attributes this mitigation to their any-order masked training objective. However, observing "[**M**] is B" during training teaches recovery of A from B in one positional configuration, and does not by itself explain why the learned evidence should transfer to the reverse prompt "B is [**M**]." We provide a theoretical analysis showing that this transfer arises from a parameter-level coupling between forward and reverse positional conditionals: shared Transformer parameters store token-pair evidence, while relative positional encodings route attention through queries and keys without changing the value-side evidence being retrieved. In a one-layer MDM, we prove that forward masked training strengthens evidence that is reusable in reverse queries, induces correlated forward–reverse attention routes, and yields a positively aligned shared-storage gradient component that decreases the reverse loss to first order. Controlled one-layer experiments and large-scale LLaDA/Dream experiments verify these signatures and show that they translate into improved reverse prediction.

1. Introduction

Autoregressive language models (ARMs) are the dominant paradigm for large-scale language modeling (Vaswani et al., 2017; Radford et al., 2018; 2019; Brown et al., 2020). Despite their empirical success, they exhibit the *reversal curse*: after learning a relational fact in one order, a model may fail to answer the corresponding order-reversed query (Berglund et al., 2024). For example, a model trained on "The capital of France is Paris" may answer "What is the capital of France?" but fail to answer "Which country has Paris as its capital?" This failure exposes a gap between storing a fact in the learned weights and making it accessible under order reversal: when the fact is explicitly provided in context, strong models can often infer the reverse relation, but the same relation need not be retrievable from the weights alone (Berglund et al., 2024).

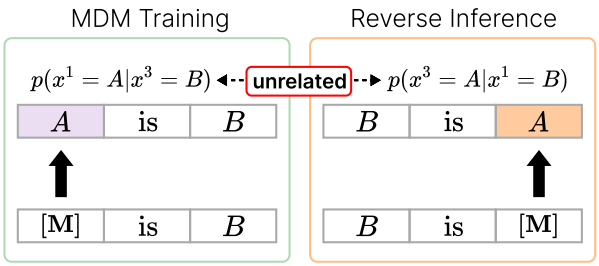


Figure 1. Masked training teaches one conditional, but the objective alone does not explain its link to the reverse conditional.

The failure is natural under left-to-right training. An ARM trained on "A is B" reinforces predicting B from a context containing A, whereas the reverse query asks for A from a context containing B, corresponding to a different conditional. Zhu et al. (2024) formalize this asymmetry through training dynamics, showing that training on one direction does not strengthen the reverse direction.

Masked diffusion language models (MDMs) use a random masking objective that reconstructs arbitrary masked subsets, supporting flexible any-order generation rather than fixed left-to-right generation (Hoogeboom et al., 2021; Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024; Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025; Nie et al., 2025b). Recent work reports that MDMs exhibit a substantially weaker reversal curse than comparable ARMs (He et al., 2026), often attributed to this any-order training objective (Kitouni et al., 2024; Nie et al., 2025a;b).

However, we argue that this objective-only explanation is incomplete, as illustrated in Fig. 1. As a representative

[*]Equal contribution [1]Department of Artificial Intelligence, Yonsei University, Seoul, Korea. Correspondence to: Albert No <albertno@yonsei.ac.kr>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

forward example, suppose "A is B" is observed during training, with A to the left of B. Random masking can create the forward masked example "[**M**] is B," where the masked position i has target A, the evidence token B appears at position j, and $i < j$, providing supervision for the forward conditional $p_\theta(x^i = A|x^j = B)$. The reverse query "$B$ is [**M**]" again asks for A from B, but in a different positional configuration, schematically corresponding to the reverse conditional $p_\theta(x^{i'} = A|x^{j'} = B)$, where $j' < i'$. Thus, random masking supervises recovering A from B in the forward conditional, but does not explain why the learned evidence should carry over to the reverse conditional.

Our central claim is that MDMs link these positional conditionals through *position-invariant storage of input-target evidence* and *attention-based routing*. With relative positional encodings such as RoPE (Su et al., 2024), position affects query-key attention but not the value vector of an input token. Thus, forward masked training can store B-to-A evidence independently of where B appears, while attention determines whether the reverse query routes the mask to B strongly enough to retrieve it.

We formalize this mechanism in a simplified one-layer MDM, where the contribution of input token B to the A-logit decomposes as $\alpha_{i,j}Y_{A,B}$. Here $Y_{A,B}$ is the shared storage term for B-to-A evidence, and $\alpha_{i,j}$ is the attention weight routing that evidence to the mask. We prove that forward masked training increases $Y_{A,B}$, relative positional encodings preserve positively correlated attention routes between forward and reverse queries, and the forward and reverse losses have positively aligned gradients so that a forward update decreases the reverse loss to first order.

Controlled one-layer experiments show that the ingredients identified by our theory are present in trained models: forward–reverse attention correlation, gradient alignment, and reverse-probability improvement under forward-only training. Positional encoding interventions further validate the analysis by changing the storage-routing components and producing the predicted changes in reverse transfer. Large-scale LLaDA (Nie et al., 2025b) and Dream (Ye et al., 2025) experiments show the same qualitative signatures in practical multi-layer diffusion LLMs.

2. Background and Related Work

2.1. Autoregressive Models and the Reversal Curse

Autoregressive models (ARMs) (Radford et al., 2018; 2019) generate a sequence $\mathbf{x} = x^1x^2\ldots x^L$ from left to right, factorizing $p_\theta(\mathbf{x}) = \prod_{i=1}^{L} p_\theta(x^i|\mathbf{x}^{<i})$. They are trained with a cross-entropy loss for next-token prediction, where each position is optimized to predict x^i from its left context $\mathbf{x}^{<i}$. In Transformer ARMs, this objective is implemented with causal self-attention, so the prediction at position i can use only $\mathbf{x}^{<i}$. Thus, the training signal is intrinsically directional.

This directional signal helps explain the *reversal curse* (Berglund et al., 2024). If the forward direction "A is B" is observed, then training directly reinforces predicting B from a context containing A. The reverse direction instead asks the model to predict A from a context containing B, a distinct conditional that is not directly reinforced by the forward example. Here and throughout, "A is B" is only a schematic convention for paired relations whose observed and queried orders differ, not a restriction to the literal verb "is." (See App. E.3 for a more detailed formulation.)

Several recent works seek to mitigate the reversal curse through reverse training, dataset design, relation-aware augmentation, and broader analyses of order-sensitive generalization (Golovneva et al., 2024; Lin et al., 2024; Lu et al., 2024; Wang & Sun, 2026). These efforts highlight the reversal curse as an important failure mode, but leave its mechanism largely unexplained. Closest to this goal, Zhu et al. (2024) analyze simplified one-layer ARMs through training dynamics and use effective-weight asymmetry to explain why forward training does not improve the reverse direction.

2.2. Masked Diffusion Models and Reversal Curse Mitigation

Masked diffusion models (MDMs) generate text by starting from masked sequences and iteratively recovering masked positions (Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025). During training, a clean sequence $\mathbf{x} = x^1x^2\ldots x^L$ is partially masked into $\mathbf{x}_t$ by independently replacing tokens with [**M**] with probability $t \in (0, 1]$. The model is trained with a masked-token cross-entropy objective for any-order generation: at each masked position i, it predicts x^i from $\mathbf{x}_t$, i.e., $p_\theta(x^i|\mathbf{x}_t)$. Unlike ARMs, MDMs use full attention, so a masked position can attend to unmasked tokens on either side.

A useful point of comparison is BERT-style masked language modeling (Devlin et al., 2019). Current text MDMs can be viewed operationally as generative masked language models: they use full-attention masked reconstruction during training and iteratively replace mask tokens during generation, without requiring explicit time embeddings in the Transformer backbone (Sahoo et al., 2024; Ou et al., 2025; Nie et al., 2025b; Ye et al., 2025). (See App. E.1 for details.)

Recent work reports that MDMs exhibit a weaker reversal curse than ARMs. Kitouni et al. (2024) show that uniform-rate masked language models, which are closely related to MDMs, mitigate reversal failures in controlled settings. Nie et al. (2025a) and Nie et al. (2025b) further report reversal advantages in masked diffusion LLMs. These results moti-

vate a natural explanation based on the any-order objective: random masking can provide supervision in multiple directions, including directions unavailable to left-to-right ARM training.

However, this explanation is incomplete as illustrated in Fig. 1. Concretely, a forward statement "A is B" can yield a masked training example that provides supervision for the schematic conditional $p_\theta(x^i = A | x^j = B)$. The reverse query asks for the same target A from the same evidence token B, but in a different positional configuration, corresponding to $p_\theta(x^{i'} = A | x^{j'} = B)$. It is therefore not immediate why improving the first conditional should improve the second unless the model parameters couple the two. The remainder of the paper shows that MDMs provide exactly such a parameter-level coupling through position-invariant storage and attention routing.

3. One-Layer Setup: The Storage-Routing Decomposition

This section sets up the one-layer MDM used in our analysis and isolates the central storage-routing decomposition. It separates *what evidence is stored* in the parameters from *how that evidence is routed* by attention: the former will be represented by a position-invariant storage term $Y_{A,B}$, and the latter by a position-dependent attention weight $\alpha_{i,j}$. This decomposition will be the basis for the theoretical results in Sec. 4. See App. A.1 for explicit dimensions and discussion.

The one-layer setting is intentionally simple and follows a standard strategy in theoretical work on Transformer training dynamics (Tian et al., 2023; Li et al., 2023; Zhu et al., 2024; Huang et al., 2025b). It is not meant to fully describe deep MDMs, but to isolate a mechanism whose signatures can be tested in trained one-layer MDMs and large-scale diffusion LLMs in Secs. 5 and 6.

One-layer MDM with relative positional encoding. Consider a one-layer MDM applied to a partially masked sequence $\mathbf{x} = x^1 x^2 \dots x^L$. Let i be a masked position with $x^i = [\mathbf{M}]$, and let A be the target token to be recovered at this position. Let j be a position containing the unmasked input token $x^j = B$. For a token x, let $\mathbf{e}_x$ denote its one-hot vector and $\mathbf{h}_x = W_E \mathbf{e}_x$ its token embedding. The query, key, and value vectors are $\mathbf{q}_x = W_Q \mathbf{h}_x$, $\mathbf{k}_x = W_K \mathbf{h}_x$, $\mathbf{v}_x = W_V \mathbf{h}_x$.

We consider relative positional encodings such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2022), where position affects attention scores but not value vectors. To cover these encodings, write the attention score and attention weight from masked position i to position k as

$$S_{i,k} = S(\mathbf{q}_{x^i}, \mathbf{k}_{x^k}, k - i),$$

$$\alpha_{i,k} = \frac{\exp(S_{i,k}/\sqrt{D})}{\sum_{k'=1}^{L} \exp(S_{i,k'}/\sqrt{D})}.$$

Thus $\alpha_{i,k} > 0$ and $\sum_{k=1}^{L} \alpha_{i,k} = 1$. The logit vector at the masked position and the model probability are

$$\mathbf{z}_i = W_U \sum_{k=1}^{L} \alpha_{i,k} \mathbf{v}_{x^k},$$

$$\mathbf{p}_i = p_\theta(x^i = \cdot \,|\, \mathbf{x}) = \mathrm{softmax}(\mathbf{z}_i),$$

where W_U is the unembedding matrix mapping hidden states to vocabulary logits.

Effective reparameterization. Following prior analyses of Transformer training dynamics (Tian et al., 2023; Zhu et al., 2024), we analyze the effective weight

$$Y = W_U W_V W_E.$$

This reparameterization provides a useful coordinate system for isolating the storage component. In particular, the masked-position logit vector and its A-coordinate become

$$\mathbf{z}_i = \sum_{k=1}^{L} \alpha_{i,k} Y_{:,x^k}, \quad (\mathbf{z}_i)_A = \sum_{k=1}^{L} \alpha_{i,k} Y_{A,x^k}.$$

Thus, if the unmasked input token B appears at position j, i.e., $x^j = B$, its contribution to the A-logit at the masked position i is

$$\alpha_{i,j} Y_{A,B}.$$

This is the central decomposition. To interpret $Y_{A,B}$ in the original parameterization, let $\mathbf{u}_A^\top$ be the row of W_U corresponding to target token A. Then $Y_{A,B} = \mathbf{u}_A^\top \mathbf{v}_B$, so $Y_{A,B}$ measures how strongly the value vector contributed by input token B matches the unembedding direction for target token A. We interpret this scalar as the stored B-to-A evidence. Under relative positional encodings, $\mathbf{v}_B$ contains no positional component, and $\mathbf{u}_A$ is also shared across positions; hence the same $Y_{A,B}$ is used regardless of where B appears.

In contrast, $\alpha_{i,j} > 0$ is the attention weight that the masked position i assigns to the token at position j. It specifies how strongly the model attends to B when filling the mask at position i. Unlike $Y_{A,B}$, this routing weight depends on the positional configuration through the query-key attention score, including the relative offset between $[\mathbf{M}]$ and B. Thus, $Y_{A,B}$ stores the position-invariant input-B-to-target-A evidence, while $\alpha_{i,j}$ is the position-dependent attention weight that determines how much of this evidence is used at the masked position.

Positional encoding determines whether this decomposition yields reusable evidence. For the relative encodings we

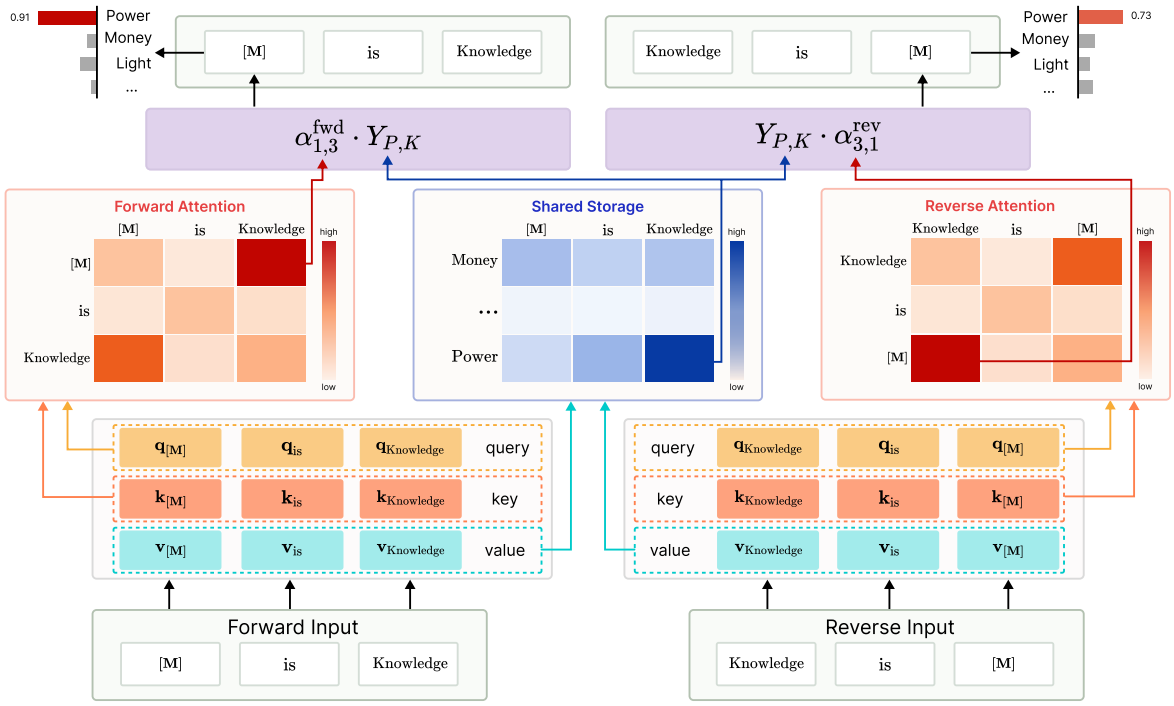


Figure 2. **Mechanism overview of the storage-routing decomposition.** From a forward statement "Power is Knowledge," random masking can produce the forward masked example "[**M**] is Knowledge," where Knowledge contributes to the Power logit through $\alpha_{1,3}^{\text{fwd}} Y_{P,K}$. The reverse query "Knowledge is [**M**]" uses the same stored evidence term through $\alpha_{3,1}^{\text{rev}} Y_{P,K}$. Thus $Y_{P,K}$ is shared across the two configurations, while the attention weights $\alpha_{1,3}^{\text{fwd}}$ and $\alpha_{3,1}^{\text{rev}}$ route this evidence with different strengths. Under relative positional encodings, the attention scores producing these weights remain positively correlated, enabling reverse reuse of the stored Knowledge-to-Power evidence $Y_{P,K}$.

study, moving B can change $\alpha_{i,j}$, but not the value vector $\mathbf{v}_B$: position changes how much B's evidence is routed, not what evidence is stored. By contrast, under standard additive absolute embeddings $W_E\mathbf{e}_B + \mathbf{p}_j$, the value vector $W_V(W_E\mathbf{e}_B + \mathbf{p}_j)$ depends on j, breaking a single position-invariant B-to-A evidence term shared across configurations. We focus on relative encodings because they are standard in modern LLMs, especially through RoPE (Su et al., 2024); absolute encodings are used for comparison in Sec. 5. (See App. E.2 for additional discussion.)

4. Theoretical Results: Storage, Routing, and Gradient Alignment

Sec. 3 decomposed the contribution of input token B to target logit A as $\alpha_{i,j} Y_{A,B}$. We use this decomposition to analyze storage, routing, and prediction transfer. Sec. 4.1 shows that learning target A from token B increases the shared evidence term $Y_{A,B}$, independent of where B appears. Sec. 4.2 shows that relative positional encodings preserve positively correlated attention scores to B across forward and reverse configurations, making the reverse route non-arbitrary. Sec. 4.3 then connects these ingredients to training dynamics: on shared storage parameters, forward and reverse losses have positively aligned gradients, so a forward update decreases the reverse loss to first order. Apps. A, B, and C provide details, proofs, and additional analyses.

4.1. Storage: Forward Training Reinforces a Reverse-Usable Relation

In this subsection, we ask how a forward masked example changes the shared evidence term $Y_{A,B}$. Consider a masked training instance from a forward fact, such as "[**M**] is B," where the masked position is i, the target token to be recovered is A, and the unmasked input token at position j is $x^j = B$. The corresponding reverse query may place the same evidence token B at a new position j' and ask the model to recover the same target A at a new masked position i', as in "B is [**M**]." Although these prompts have different, order-reversed positional configurations, both can rely on the same storage $Y_{A,B}$. We therefore ask how the forward masked update modifies this shared entry.

For the cross-entropy loss of recovering A at masked position i from the partially unmasked input $\mathbf{x}$,

$$\mathcal{L}_i = -\log p_i(A) = -\log p_\theta(x^i = A|\mathbf{x}),$$

we have the following direct gradient calculation.

Proposition 4.1 (Forward Masked Training Strengthens the Shared B-to-A Relation)**.** *Consider the one-layer MDM above and treat Y as the effective storage parameter. If $x^i = [\mathbf{M}]$ has target token A and $x^j = B$ is an unmasked input token with attention weight $\alpha_{i,j} > 0$, then*

$$\frac{\partial \mathcal{L}_i}{\partial Y_{A,B}} \leq -\alpha_{i,j}(1 - p_i(A)) < 0,$$
$$\frac{\partial \mathcal{L}_i}{\partial Y_{C,B}} \geq \alpha_{i,j} p_i(C) > 0 \quad (C \neq A).$$

Thus, one gradient step increases $Y_{A,B}$ and decreases $Y_{C,B}$ for all $C \neq A$.

The proposition gives the storage-side mechanism. Training the model to recover A in the presence of B increases the position-invariant entry $Y_{A,B}$, which contributes to predicting A whenever B appears again as an input token. The attention weight $\alpha_{i,j}$ controls the magnitude of this update, but since $\alpha_{i,j} > 0$, it does not change its sign. Thus, the forward masked example stores B-to-A evidence in a way that is not tied to the particular positions.

This pinpoints why the any-order training objective alone is incomplete as an explanation. Random masking supplies a training instance in which A is recovered from B, but the resulting evidence becomes reverse-usable only because it is stored in the shared, position-invariant entry $Y_{A,B}$. The remaining question is whether the reverse prompt can still attend to B and use this stored evidence.

4.2. Routing: Attention Correlation under Relative Positional Encodings

We now study the routing term $\alpha_{i,j}$. Even if forward training increases the shared evidence term $Y_{A,B}$, a reverse configuration can use this evidence only if the masked position still attends to the same input token B. Since attention weights depend on all tokens in the sequence, we analyze the more local quantity that directly controls B's contribution: the attention score from the masked position to B. Our goal is to show that this score remains positively correlated across forward and reverse positional configurations.

The sequence may have arbitrary length L, and the masked position may attend to all positions. We therefore isolate the attention score assigned to a shared input token B appearing in paired forward and reverse prompts. When B is the relevant evidence token, this score controls how strongly the masked position can use the stored B-to-A evidence.

RoPE. We first analyze RoPE (Su et al., 2024), the dominant relative positional encoding in modern LLMs and the encoding used by the large MDMs (Nie et al., 2025b; Ye et al., 2025) studied in this paper. RoPE injects relative position into attention scores through block-wise rotations in two-dimensional frequency planes. Consider a forward prompt where B is Δ_1 positions to the right of $[\mathbf{M}]$, and a reverse prompt where B is Δ_2 positions to the left of $[\mathbf{M}]$. We study the attention score assigned to B, which controls how strongly the masked position can use B's stored evidence. Writing $\mathbf{q}_{[\mathbf{M}]}$ for the masked-token query and $\mathbf{k}_B$ for the key of token B, these scores are

$$S_{\text{fwd}} = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1)\mathbf{k}_B, \quad S_{\text{rev}} = \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B,$$

where $R(\Delta)$ is the RoPE rotation matrix. Reversal changes the relative rotation, so the two scores are not identical; we ask whether they remain positively correlated.

Assumption 4.2 (RoPE-Compatible Block Covariance)**.** *Let D be even, with $D/2$ two-dimensional RoPE planes. Assume $\mathbf{q}_{[\mathbf{M}]} \sim \mathcal{N}(0, \sigma^2 I_D)$. Within each plane, write the two-dimensional blocks as $\mathbf{q}_{[\mathbf{M}]}^{(s)}, \mathbf{k}_B^{(s)} \in \mathbb{R}^2$. Conditioned on $\mathbf{q}_{[\mathbf{M}]}$, the key blocks are independent across planes and have the same total conditional variance. Moreover, within each plane, the key block has at least as much conditional variance along the query direction as along the orthogonal direction.*

This assumption mirrors the geometry of RoPE: rotations act independently within two-dimensional frequency planes. It rules out an adversarial case where, within each plane, the key of B varies mainly in directions orthogonal to the masked-token query. (See App. A.3 for the formal assumption.) Under this non-adversarial covariance condition, reversal changes the RoPE rotation but preserves positive correlation between the forward and reverse scores.

Theorem 4.3 (RoPE Attention-Score Correlation)**.** *If Assumption 4.2 holds and $\Delta_1 = \Delta_2 = \Delta$, then*

$$\mathbb{E}_{\mathbf{q}_{[\mathbf{M}]}}\left[\operatorname{Corr}\left(S_{\text{fwd}}, S_{\text{rev}} \mid \mathbf{q}_{[\mathbf{M}]}\right)\right] \geq \frac{1}{D}\operatorname{Tr}(R(2\Delta)).$$

The right-hand side is positive for standard RoPE frequencies in the range $\Delta \leq 50$ and $D = 64, 128, 256$: its minimum over this range is 0.337; see App. B.1. Thus, Thm. 4.3 supports positive expected forward–reverse attention score correlation in the relevant distance range.

Remark 4.4. **The equal-distance assumption is not essential.** The condition $\Delta_1 = \Delta_2$ is used only to obtain the closed form above. For $\Delta_1 \neq \Delta_2$, the same analysis gives a numerically computable bound; for all $\Delta_1, \Delta_2 \leq 30$, the worst-case bound over admissible within-plane covariance strengths remains positive, with minima 0.363 for $D = 128$ and 0.395 for $D = 256$, respectively; see App. B.3.

RoPE does not make the forward and reverse attention routes identical, but it does preserve positive correlation between them. Thus, when the mask strongly attends to B in the forward configuration, the reverse configuration tends to

retain a nontrivial attention route to B. This explains how the stored B-to-A evidence can remain accessible under order reversal.

ALiBi. To show that the routing analysis is not specific to RoPE, we also consider ALiBi (Press et al., 2022). With ALiBi, position enters as an additive distance bias, so the attention scores assigned to B take the form

$$S_{\text{fwd}} = \mathbf{q}_{[\mathbf{M}]}^\top \mathbf{k}_B + b\Delta_1, \quad S_{\text{rev}} = \mathbf{q}_{[\mathbf{M}]}^\top \mathbf{k}_B - b\Delta_2.$$

Thus, the two scores share the same content term $\mathbf{q}_{[\mathbf{M}]}^\top \mathbf{k}_B$ and differ only by constants.

Proposition 4.5 (ALiBi Attention-Score Correlation)**.** *For ALiBi, suppose the shared content score $\mathbf{q}_{[\mathbf{M}]}^\top \mathbf{k}_B$ has nonzero variance. Then, for any joint distribution of $(\mathbf{q}_{[\mathbf{M}]}, \mathbf{k}_B)$,*

$$\text{Corr}\left(S_{\text{fwd}}, S_{\text{rev}}\right) = 1.$$

The result is immediate but useful: ALiBi preserves the forward and reverse attention scores perfectly at the correlation level because both scores contain the same varying content term. Together with Thm. 4.3, this shows that relative positional encodings can preserve a reverse attention route, with the strength of the correlation depending on how position enters the attention score.

4.3. Dynamics: Forward-to-Reverse Gradient Alignment

The previous subsections developed the storage-routing picture. We now ask whether it leads to an actual per-step training effect: does a gradient step on a forward masked loss increase the corresponding reverse probability? Following the training-dynamics perspective of Zhu et al. (2024), we answer this question to first order through gradient alignment.

Let $\mathbf{x}_{\text{fwd}}$ and $\mathbf{x}_{\text{rev}}$ denote paired masked prompts derived from the same underlying fact. For the schematic fact "A is B," these correspond to prompts such as "[**M**] is B" and "B is [**M**]," respectively. Note that the prompts may contain arbitrary additional context tokens. Both prompts ask the model to recover the same target token A at their masked positions, and both contain an informative unmasked token B. Let $\mathbf{p}_{\text{fwd}}$ and $\mathbf{p}_{\text{rev}}$ denote the prediction distributions at the masked positions, and define

$$\mathcal{L}_{\text{fwd}} = -\log p_{\text{fwd}}(A), \quad \mathcal{L}_{\text{rev}} = -\log p_{\text{rev}}(A).$$

For a forward update $\theta^+ = \theta - \eta \nabla_\theta \mathcal{L}_{\text{fwd}}$, a first-order expansion yields

$$\mathcal{L}_{\text{rev}}(\theta^+) - \mathcal{L}_{\text{rev}}(\theta) = -\eta \left\langle \nabla_\theta \mathcal{L}_{\text{rev}}, \nabla_\theta \mathcal{L}_{\text{fwd}} \right\rangle + O(\eta^2).$$

Thus, if the gradient inner product $\langle \nabla_\theta \mathcal{L}_{\text{rev}}, \nabla_\theta \mathcal{L}_{\text{fwd}} \rangle$ is positive, then the forward step decreases the reverse loss, equivalently increasing the reverse probability of target A, to first order.

We focus on the shared storage column $Y_{:,B}$ associated with the common input token B. Let α_B^{fwd} and α_B^{rev} denote the total attention mass assigned to all occurrences of B in the forward and reverse prompts $\mathbf{x}_{\text{fwd}}$ and $\mathbf{x}_{\text{rev}}$, respectively.

Theorem 4.6 (Shared-Column Gradient Alignment)**.** *Consider the one-layer MDM setup of Sec. 3. Suppose the unmasked token B appears in both the forward and reverse prompts, and both losses use the same target token A. Then the gradients of the forward and reverse losses with respect to the shared storage column $Y_{:,B}$ satisfy*

$$\begin{aligned} &\left\langle \nabla_{Y_{:,B}} \mathcal{L}_{\text{fwd}}, \nabla_{Y_{:,B}} \mathcal{L}_{\text{rev}} \right\rangle \\ &= \alpha_B^{\text{fwd}} \alpha_B^{\text{rev}} \left\langle \mathbf{p}_{\text{fwd}} - \mathbf{e}_A, \mathbf{p}_{\text{rev}} - \mathbf{e}_A \right\rangle > 0. \end{aligned}$$

Consequently, the $Y_{:,B}$ component of a forward gradient step decreases the reverse loss to first order.

The theorem gives a training-dynamics view of the shared Y-column update. A forward update that strengthens the B-to-A evidence in $Y_{:,B}$ also moves the reverse query toward higher probability on A. A column $Y_{:,C}$ receives a direct update only when token C appears in the input prompt, so tokens absent from the prompt have no direct Y-column contribution.

This theorem isolates the guaranteed Y-column component of forward-to-reverse transfer. A complete full-gradient statement must also account for query-key terms. We therefore check full-gradient behavior empirically in Secs. 5 and 6, where the measured full-parameter gradient cosine remains positive, showing that the remaining terms do not cancel the positive Y-column contribution in trained models. We complement this check theoretically: App. C.1 analyzes query-key routing within the effective reparameterization and explains why it does not systematically cancel the Y-column effect, App. C.2 analyzes the same tendency in the original parameter space, and App. C.3 proves, in a two-token setting, a probability-level convergence result in the spirit of Zhu et al. (2024).

5. One-Layer Validation

We test the storage-routing account in a one-layer MDM and use positional encoding interventions to validate the predictions specific to each positional encoding. See App. D.1 for details.

Setup. We instantiate a one-layer MDM using RADD (Ou et al., 2025), with RoPE as the default positional encoding. We construct length L sequences containing one lowercase–uppercase pair and fill the remaining $L-2$ positions with a placeholder `0`. The lowercase token is the target, and the

Table 1. In the one-layer setting, the MDM retains substantial reverse accuracy, whereas the ARM collapses.

	MDM		ARM	
L	Fwd.	Rev.	Fwd.	Rev.
10	99.31	**43.10**	99.83	0.00
20	97.36	**55.70**	99.80	0.00
30	96.91	**33.89**	99.93	0.00
40	97.27	**38.37**	99.93	0.00

uppercase token is the unmasked evidence token. Training uses only the forward order, where the lowercase token precedes its paired uppercase token; all reversed-order sequences are excluded. For example, with $L = 3$ and pair $(\texttt{d}, \texttt{D})$, the training set contains `dD0`, `d0D`, and `0dD`, but never `Dd0`, `D0d`, or `0Dd`. The placeholders introduce positional variation without adding semantic cues.

At each checkpoint, we evaluate paired forward and reverse masked queries from the same lowercase–uppercase pair. Both recover the lowercase target from the unmasked uppercase evidence token, but differ in order and position. We track attention-score correlation, full-parameter gradient cosine, and reverse-target probability.

Results. For paired forward and reverse queries, we first compute the attention score from the masked position to the uppercase evidence token. Fig. 3 (left) shows that these scores remain strongly positively correlated across positional configurations, consistent with Thm. 4.3. Thus, even when the mask and evidence token appear in reversed relative positions, the model preserves a correlated attention route from the mask to the evidence token.

We next compute the forward–reverse gradient cosine over all trainable parameters in the trained model, including query, key, and MLP. Fig. 3 (middle) shows consistently positive cosine similarity across sequence lengths. Thus, the shared-column alignment identified in Sec. 4.3 persists in the original parameterization, even after including query-key, MLP, and other parameters.

Finally, we ask whether these signals translate into improved reverse prediction. Fig. 3 (right) shows that, under forward-only training, the correct target probability also increases on reverse queries. Thus, the predicted attention correlation and gradient alignment are reflected in the model's output distribution. A one-layer GPT-2-style ARM baseline (Radford et al., 2019) learns the forward direction but fails to transfer, as shown in Tab. 1.

Positional encoding intervention. We next change only the positional encoding to test both sides of the theory. Relative encodings preserve position-invariant storage, while learned absolute embeddings (Gehring et al., 2017; Devlin et al., 2019) make value vectors position-dependent and weaken the single shared B-to-A evidence term from Sec. 4.1. Among relative encodings, RoPE and ALiBi differ in routing: RoPE yields partial but positive attention-score correlation, whereas ALiBi preserves the shared content score up to additive distance biases as analyzed in Sec. 4.2. Tab. 2 matches this pattern: learned absolute embeddings fit the forward direction but transfer poorly, RoPE transfers nontrivially but imperfectly, and ALiBi transfers almost perfectly.

Table 2. Relative encodings preserve position-invariant storage and therefore transfer better than learned absolute embeddings, while ALiBi gives the strongest transfer by preserving content-based attention routing.

L	Dir.	Absolute	RoPE	ALiBi
10	Fwd.	98.3	99.3	99.7
	Rev.	11.5	43.1	99.8
20	Fwd.	97.8	96.8	99.5
	Rev.	27.8	55.7	99.5

Together, the one-layer experiments provide a controlled test of storage-routing. Rather than checking the effective reparameterization alone, we measure the relevant signatures in the trained model's original parameters, including query, key, and MLP components. The results show that reverse transfer is accompanied by the predicted routing correlation and gradient alignment. Positional encoding interventions further change reverse transfer in the predicted direction. Thus, the experiments support the mechanism itself, not merely that MDMs outperform ARMs on reverse queries.

6. Large-Scale Validation

We now test whether the same signatures persist in diffusion LLMs, using LLaDA (Nie et al., 2025b). See App. D.2 for experimental details and additional Dream (Ye et al., 2025) results.

Setup. We perform forward-only fine-tuning on the Parent–Child dataset (Berglund et al., 2024). Tab. 7 reports the full reversal-evaluation task and accuracy, and we use the same task here to probe the analyzed mechanism. For each relation, we construct paired forward and reverse masked-prediction queries. In a P2C (Parent-to-Child) case, the forward query is "[**M**]'s child is Tom Holland", where the masked entity is the parent, and the paired reverse query is "Tom Holland's parent is [**M**]". Both queries ask for the same target entity from the same unmasked evidence entity, but place the target and evidence at different positions.

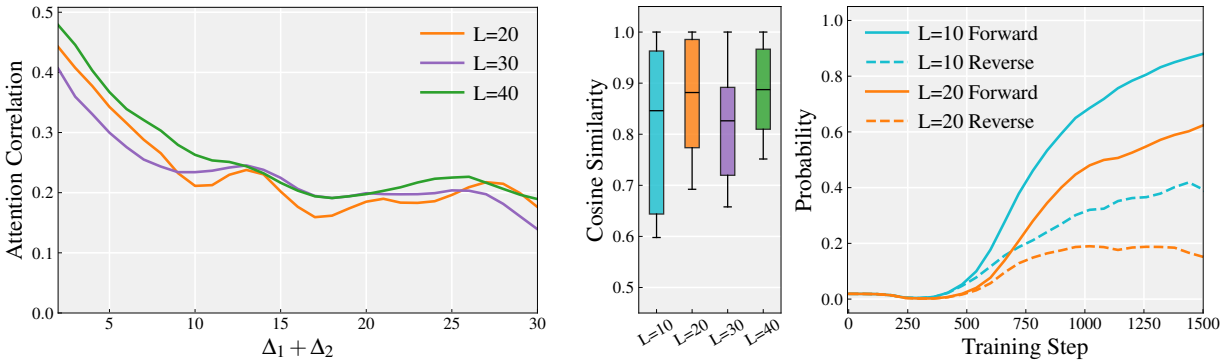


Figure 3. **Left:** Attention-score correlation in a one-layer RADD across $\Delta_1 + \Delta_2$. The correlations remain positive. **Middle:** Forward–reverse gradient cosine during training. It stays above 0.6, indicating aligned optimization directions. **Right:** Target-token probability under forward and reverse queries. The curves show that the analyzed effects translate into probability-level transfer.

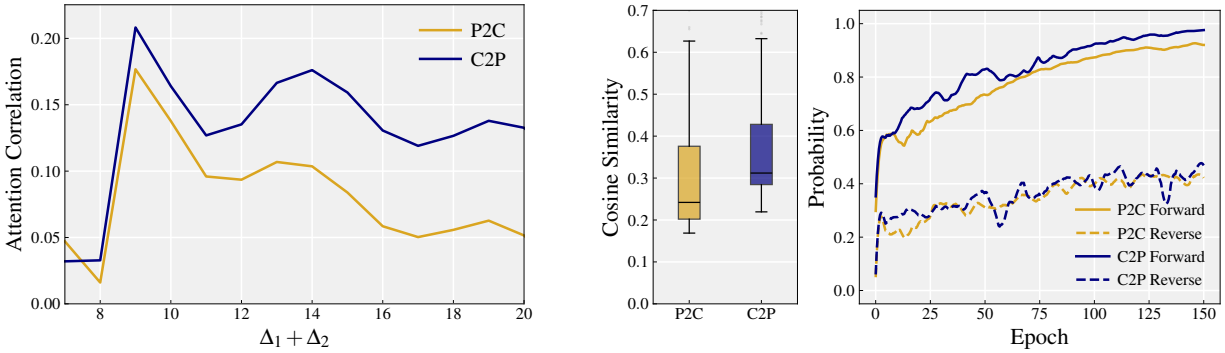


Figure 4. **Left:** LLaDA attention correlation. Forward and reverse attention scores assigned to the evidence entity remain positively correlated across relative-distance placements. **Middle:** LLaDA gradient alignment. Forward–reverse gradient cosine similarity remains positive. **Right:** LLaDA probability transfer. The correct reverse-entity probability increases during forward-only fine-tuning.

Results. Fig. 4 (left) shows that forward and reverse attention scores assigned to the evidence entity remain positively correlated in LLaDA across relative-distance placements. Thus, even in multi-layer diffusion LLMs, reverse prompts retain a nontrivial attention route to the unmasked evidence entity.

Fig. 4 (middle) further shows that the forward/reverse full-parameter gradient cosine remains positive. This indicates that the shared-storage alignment identified in the one-layer theory is not washed out by query-key parameters, MLP blocks, or multi-layer interactions.

Finally, Fig. 4 (right) shows that the correct reverse-entity probability steadily increases during forward-only fine-tuning. Reverse probabilities nevertheless remain below forward probabilities, as expected: attention routing is correlated but not identical across forward and reverse prompts, and the gradients are positively but not perfectly aligned. This gap is therefore consistent with the one-layer results in Sec. 5. Nevertheless, the observed probability trend shows that attention-score correlation and gradient alignment improve reverse prediction.

Together, the large-scale experiments show that the ingredients identified by the theory persist in practical multi-layer diffusion LLMs: the predicted attention route and gradient alignment are observed, and they translate into improved reverse probability under forward-only training.

7. Conclusion

We studied why MDMs mitigate the reversal curse, where ARMs typically collapse. Our analysis shows that the key issue is not only whether the objective exposes useful masked examples, but whether the resulting parameter updates create evidence that can be reused across positional configurations. Our one-layer theory identifies the mechanism enabling this reuse: position-invariant storage of input-target evidence, positively correlated attention routing under relative positional encodings, and transfer through forward-to-reverse gradient alignment. Controlled one-layer experiments and large-scale diffusion LLM experiments verify these signatures, showing that the same storage-routing account persists beyond the simplified setting.

Impact Statement

This paper studies the reversal curse in masked diffusion language models and provides a theoretical explanation for why masked training can improve reverse-direction generalization. Our results are intended to clarify model behavior rather than introduce a new deployment capability. In practice, better understanding of reverse retrieval may help develop language models that are more reliable in knowledge-intensive settings such as question answering, editing, and scientific assistance. At the same time, any improvement in model recall and generalization can also amplify downstream misuse, including more persuasive generation or automated content production at larger scale. As with other language-model research, the societal impact will depend on how the resulting methods are trained, evaluated, and deployed, and careful attention should be paid to safety, privacy, bias, and human oversight.

References

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *ICLR*, 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. In *NeurIPS*, 2022.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Dinkelbach, W. On nonlinear fractional programming. *Management Science*, 1967.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. In *ICML*, 2017.

Golovneva, O., Allen-Zhu, Z., Weston, J. E., and Sukhbaatar, S. Reverse training to nurse the reversal curse. In *COLM*, 2024.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021.

He, S., Wei, K., Zeng, X., Chen, X., Yang, X., Li, Z., Zhong, J., and Tian, Y. Differ: Diffusion entity-relation modeling for reversal curse in diffusion large language models. *arXiv preprint arXiv:2601.07347*, 2026.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. Argmax flows and multinomial diffusion: Learning categorical distributions. In *NeurIPS*, 2021.

Huang, R., Liang, Y., and Yang, J. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *ICML*, 2025a.

Huang, Y., Zhu, H., Guo, T., Jiao, J., Sojoudi, S., Jordan, M. I., Russell, S., and Mei, S. Generalization or hallucination? understanding out-of-context reasoning in transformers. In *NeurIPS*, 2025b.

Kitouni, O., Nolte, N., Bouchacourt, D., Williams, A., Rabbat, M., and Ibrahim, M. The factorization curse: Which tokens you predict underlie the reversal curse and more. In *NeurIPS*, 2024.

Li, Y., Li, Y., and Risteski, A. How do transformers learn topic structure: Towards a mechanistic understanding. In *ICML*, 2023.

Li, Z., Cao, Y., Gao, C., He, Y., Liu, H., Klusowski, J. M., Fan, J., and Wang, M. One-layer transformer provably learns one-nearest neighbor in context. *NeurIPS*, 2024.

Lin, Z., Fu, Z., Liu, K., Xie, L., Lin, B., Wang, W., Cai, D., Wu, Y., and Ye, J. Delving into the reversal curse: How far can large language models generalize? In *NeurIPS*, 2024.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In *ICML*, 2024.

Lu, Z., Jin, L., Li, P., Tian, Y., Zhang, L., Wang, S., Xu, G., Tian, C., and Cai, X. Rethinking the reversal curse of llms: a prescription from human knowledge reversal. In *EMNLP*, 2024.

Nie, S., Zhu, F., Du, C., Pang, T., Liu, Q., Zeng, G., Lin, M., and Li, C. Scaling up masked diffusion models on text. In *ICLR*, 2025a.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ZHOU, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. In *NeurIPS*, 2025b.

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *ICLR*, 2025.

Press, O., Smith, N., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. In *NeurIPS*, 2024.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and generalized masked diffusion for discrete data. In *NeurIPS*, 2024.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *NeurIPS*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, B. and Sun, H. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. In *ICLR*, 2026.

Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., and Kong, L. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Zhu, H., Huang, B., Zhang, S., Jordan, M., Jiao, J., Tian, Y., and Russell, S. J. Towards a theoretical understanding of the 'reversal curse' via training dynamics. In *NeurIPS*, 2024.

A. Proofs for the Main Theoretical Results

A.1. One-Layer MDM Setup and Effective Reparameterization

We restate the one-layer setup from Sec. 3, making the dimensions explicit. Let $\mathcal{V}$ be the vocabulary of clean tokens, and let $\mathbf{x} = x^1 x^2 \dots x^L$ be a partially masked sequence of length L, where each token satisfies $x^k \in \mathcal{V} \cup \{[\mathbf{M}]\}$. Let i be a masked position with $x^i = [\mathbf{M}]$, and let $A \in \mathcal{V}$ be the target token to be recovered at this position.

For any token $x \in \mathcal{V} \cup \{[\mathbf{M}]\}$, let $\mathbf{e}_x \in \mathbb{R}^{|\mathcal{V}|+1}$ be its one-hot vector, including the mask token. Let

$$W_E \in \mathbb{R}^{D \times (|\mathcal{V}|+1)}$$

be the token embedding matrix, and define the token embedding

$$\mathbf{h}_x = W_E \mathbf{e}_x \in \mathbb{R}^D.$$

Given query, key, and value matrices

$$W_Q, W_K, W_V \in \mathbb{R}^{D \times D},$$

we define

$$\mathbf{q}_x = W_Q \mathbf{h}_x, \quad \mathbf{k}_x = W_K \mathbf{h}_x, \quad \mathbf{v}_x = W_V \mathbf{h}_x.$$

For relative positional encodings such as RoPE or ALiBi, position affects the query-key attention score but not the value vector. We write the attention score from masked position i to position k as

$$S_{i,k} = S(\mathbf{q}_{x^i}, \mathbf{k}_{x^k}, k - i).$$

The corresponding attention weight is

$$\alpha_{i,k} = \frac{\exp(S_{i,k}/\sqrt{D})}{\sum_{k'=1}^{L} \exp(S_{i,k'}/\sqrt{D})}.$$

For finite scores, $\alpha_{i,k} > 0$ and $\sum_{k=1}^{L} \alpha_{i,k} = 1$.

Let

$$W_U \in \mathbb{R}^{|\mathcal{V}| \times D}$$

be the unembedding matrix mapping hidden states to logits over clean vocabulary tokens. The logit vector and prediction distribution at the masked position are

$$\mathbf{z}_i = W_U \sum_{k=1}^{L} \alpha_{i,k} \mathbf{v}_{x^k}, \quad \mathbf{p}_i = p_\theta(x^i = \cdot | \mathbf{x}) = \mathrm{softmax}(\mathbf{z}_i).$$

Thus $\mathbf{z}_i, \mathbf{p}_i \in \mathbb{R}^{|\mathcal{V}|}$.

Effective reparameterization. Following prior analyses of Transformer training dynamics (Tian et al., 2023; Zhu et al., 2024), we introduce the effective weights

$$Y = W_U W_V W_E \in \mathbb{R}^{|\mathcal{V}| \times (|\mathcal{V}|+1)}, \quad Q = W_Q W_E \in \mathbb{R}^{D \times (|\mathcal{V}|+1)}, \quad K = W_K W_E \in \mathbb{R}^{D \times (|\mathcal{V}|+1)}.$$

Then

$$\mathbf{q}_x = Q \mathbf{e}_x, \quad \mathbf{k}_x = K \mathbf{e}_x,$$

and the masked-position logits can be written as

$$\mathbf{z}_i = \sum_{k=1}^{L} \alpha_{i,k} Y \mathbf{e}_{x^k}, \quad (\mathbf{z}_i)_A = \sum_{k=1}^{L} \alpha_{i,k} Y_{A,x^k}.$$

Although the main text uses only Y, we also introduce Q and K here to make the query-key gradient analysis in App. C.1 explicit. Therefore, if $x^j = B$, the contribution of the unmasked input token B to the A-logit at masked position i is

$$\alpha_{i,j} Y_{A,B}.$$

This is the storage-routing decomposition from Sec. 3.

Role of the effective reparameterization. The effective reparameterization is an analytic coordinate system rather than the original training parameterization. Following Tian et al. (2023); Zhu et al. (2024), we collect the input-to-value-to-logit path into $Y = W_U W_V W_E$, so that the stored contribution of input token B to target logit A appears directly as $Y_{A,B}$. We also collect the input-to-query and input-to-key maps into $Q = W_Q W_E$ and $K = W_K W_E$, which determine attention routing. Thus, the effective coordinates make the storage-routing decomposition explicit: Y represents shared storage, while (Q, K) determine how attention routes the stored evidence.

We do not claim that gradient descent in these effective parameters exactly preserves the training dynamics of the factorized original parameters. Rather, the effective coordinates expose the mechanism in its cleanest form. A parallel analysis in the original parameterization gives a similar qualitative picture, but the resulting full-parameter alignment is less sign-deterministic and is naturally stated as a tendency rather than an exact identity; see App. C.2.

A.2. Proof of Proposition 4.1

Recall that

$$\mathbf{z}_i = \sum_{k=1}^{L} \alpha_{i,k} Y \mathbf{e}_{x^k},$$

where $Y \in \mathbb{R}^{|\mathcal{V}| \times (|\mathcal{V}|+1)}$ is the effective storage parameter, and $\mathbf{p}_i = \mathrm{softmax}(\mathbf{z}_i)$. For the loss

$$\mathcal{L}_i = -\log p_i(A) = -(\mathbf{z}_i)_A + \log \sum_{C \in \mathcal{V}} \exp((\mathbf{z}_i)_C),$$

we have

$$\frac{\partial \mathcal{L}_i}{\partial (\mathbf{z}_i)_C} = p_i(C) - \delta_{C,A}.$$

Moreover,

$$(\mathbf{z}_i)_C = \sum_{k=1}^{L} \alpha_{i,k} Y_{C,x^k}, \quad \frac{\partial (\mathbf{z}_i)_C}{\partial Y_{C,B}} = \sum_{k: x^k = B} \alpha_{i,k}.$$

Let

$$\beta_{i,B} := \sum_{k: x^k = B} \alpha_{i,k}.$$

If $x^j = B$, then $\beta_{i,B} \geq \alpha_{i,j}$. This is the step that turns the exact derivative into the inequalities stated in the proposition.

Combining the two derivatives gives

$$\frac{\partial \mathcal{L}_i}{\partial Y_{C,B}} = (p_i(C) - \delta_{C,A})\, \beta_{i,B}.$$

Hence

$$\frac{\partial \mathcal{L}_i}{\partial Y_{A,B}} = -\beta_{i,B}(1 - p_i(A)) \leq -\alpha_{i,j}(1 - p_i(A)) < 0,$$

and for $C \neq A$,

$$\frac{\partial \mathcal{L}_i}{\partial Y_{C,B}} = \beta_{i,B} p_i(C) \geq \alpha_{i,j} p_i(C) > 0.$$

Therefore, a gradient step increases $Y_{A,B}$ and decreases $Y_{C,B}$ for every $C \neq A$. □

A.3. Proof of Theorem 4.3

Formal covariance assumption. We use the following formal version of Assumption 4.2. Let D be even, and decompose the RoPE representation into $D/2$ two-dimensional planes:

$$\mathbf{q}_{[\mathbf{M}]} = \left((\mathbf{q}_{[\mathbf{M}]}^{(1)})^\top, \ldots, (\mathbf{q}_{[\mathbf{M}]}^{(D/2)})^\top \right)^\top, \quad \mathbf{k}_B = \left((\mathbf{k}_B^{(1)})^\top, \ldots, (\mathbf{k}_B^{(D/2)})^\top \right)^\top.$$

Assume $\mathbf{q}_{[\mathbf{M}]} \sim \mathcal{N}(0, \sigma^2 I_D)$. Conditioned on $\mathbf{q}_{[\mathbf{M}]}$, the blocks $\mathbf{k}_B^{(s)}$ are independent across RoPE planes, and

$$\mathrm{Cov}(\mathbf{k}_B | \mathbf{q}_{[\mathbf{M}]}) = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_{D/2}).$$

For each plane s, assume that there exist $\tau^2 > 0$ and $\rho_s = \rho_s(\mathbf{q}_{[\mathbf{M}]}) \in [0, 1]$ such that

$$\Sigma_s = \tau^2 \left(\frac{1-\rho_s}{2} I_2 + \rho_s \frac{\mathbf{q}_{[\mathbf{M}]}^{(s)} (\mathbf{q}_{[\mathbf{M}]}^{(s)})^\top}{\|\mathbf{q}_{[\mathbf{M}]}^{(s)}\|^2} \right),$$

with the convention that the last term is defined arbitrarily on the measure-zero event $\mathbf{q}_{[\mathbf{M}]}^{(s)} = \mathbf{0}$.

This assumption matches the two-dimensional plane structure of RoPE. The block-diagonal condition removes cross-plane covariance terms, and the equal-trace condition $\mathrm{Tr}(\Sigma_s) = \tau^2$ prevents the bound from being dominated by a few frequency planes. Within each plane, ρ_s allows the covariance to range from isotropic ($\rho_s = 0$) to query-aligned ($\rho_s = 1$). Thus the assumption does not require identical covariance blocks or a fixed alignment strength across planes; it only excludes the adversarial case where the key varies more in the direction orthogonal to the masked-token query than along the query direction.

Proof. Let $\mathbf{q} = \mathbf{q}_{[\mathbf{M}]}$ and $\mathbf{k} = \mathbf{k}_B$. For RoPE (Su et al., 2024), define

$$R_s(\Delta) = \begin{pmatrix} \cos(\theta_s \Delta) & -\sin(\theta_s \Delta) \\ \sin(\theta_s \Delta) & \cos(\theta_s \Delta) \end{pmatrix}, \quad R(\Delta) = \mathrm{diag}\left(R_1(\Delta), \ldots, R_{D/2}(\Delta)\right),$$

where the standard RoPE frequencies are

$$\theta_s = \theta_{\text{base}}^{-2(s-1)/D}, \quad \theta_{\text{base}} = 10000.$$

In the equal-distance case $\Delta_1 = \Delta_2 = \Delta$, the attention scores assigned to evidence token B are

$$S_{\text{fwd}} = \sum_{s=1}^{D/2} (\mathbf{q}^{(s)})^\top R_s(\Delta) \mathbf{k}^{(s)}, \quad S_{\text{rev}} = \sum_{s=1}^{D/2} (\mathbf{q}^{(s)})^\top R_s(-\Delta) \mathbf{k}^{(s)}.$$

Condition on $\mathbf{q}$. By the block independence assumption, cross-plane covariance terms vanish. Therefore,

$$\mathrm{Cov}(S_{\text{fwd}}, S_{\text{rev}} | \mathbf{q}) = \sum_{s=1}^{D/2} (\mathbf{q}^{(s)})^\top R_s(\Delta) \Sigma_s R_s(-\Delta)^\top \mathbf{q}^{(s)}.$$

Since $R_s(-\Delta)^\top = R_s(\Delta)$, the covariance becomes

$$\mathrm{Cov}(S_{\text{fwd}}, S_{\text{rev}} | \mathbf{q}) = \frac{\tau^2}{2} \sum_{s=1}^{D/2} \|\mathbf{q}^{(s)}\|^2 \left[\cos(2\theta_s \Delta) + \rho_s\right].$$

The calculation uses

$$(\mathbf{q}^{(s)})^\top R_s(2\Delta) \mathbf{q}^{(s)} = \|\mathbf{q}^{(s)}\|^2 \cos(2\theta_s \Delta), \quad \frac{\left((\mathbf{q}^{(s)})^\top R_s(\Delta) \mathbf{q}^{(s)}\right)^2}{\|\mathbf{q}^{(s)}\|^2} = \|\mathbf{q}^{(s)}\|^2 \cos^2(\theta_s \Delta).$$

Similarly,

$$\mathrm{Var}(S_{\text{fwd}} | \mathbf{q}) = \sum_{s=1}^{D/2} (\mathbf{q}^{(s)})^\top R_s(\Delta) \Sigma_s R_s(\Delta)^\top \mathbf{q}^{(s)},$$

so

$$\mathrm{Var}(S_{\text{fwd}} | \mathbf{q}) = \frac{\tau^2}{2} \sum_{s=1}^{D/2} \|\mathbf{q}^{(s)}\|^2 \left[1 + \rho_s \cos(2\theta_s \Delta)\right].$$

The same expression holds for $\mathrm{Var}(S_{\text{rev}} | \mathbf{q})$, since cosine is even.

Define

$$w_s = \frac{\|\mathbf{q}^{(s)}\|^2}{\|\mathbf{q}\|^2}, \quad \gamma_s = \cos(2\theta_s\Delta).$$

Then

$$\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) = \frac{\sum_{s=1}^{D/2} w_s(\gamma_s + \rho_s)}{1 + \sum_{s=1}^{D/2} w_s\rho_s\gamma_s}.$$

The denominator is positive whenever the conditional variances are nonzero.

We lower bound this expression. Since

$$\left|\sum_{s=1}^{D/2} w_s\gamma_s\right| \le 1, \quad \left|\sum_{s=1}^{D/2} w_s\rho_s\gamma_s\right| \le \sum_{s=1}^{D/2} w_s\rho_s,$$

we have

$$\left(\sum_{s=1}^{D/2} w_s\gamma_s\right)\left(\sum_{s=1}^{D/2} w_s\rho_s\gamma_s\right) \le \sum_{s=1}^{D/2} w_s\rho_s.$$

Hence

$$\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) \ge \sum_{s=1}^{D/2} w_s\gamma_s = \sum_{s=1}^{D/2} w_s\cos(2\theta_s\Delta).$$

Finally, since $\mathbf{q} \sim \mathcal{N}(0, \sigma^2 I_D)$, the quantities $\|\mathbf{q}^{(s)}\|^2$ are i.i.d. across RoPE planes. Thus the weights w_s are exchangeable and satisfy

$$\mathbb{E}_{\mathbf{q}}[w_s] = \frac{2}{D}.$$

Taking expectations gives

$$\mathbb{E}_{\mathbf{q}}\left[\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q})\right] \ge \frac{2}{D}\sum_{s=1}^{D/2}\cos(2\theta_s\Delta).$$

Since each two-dimensional block satisfies

$$\mathrm{Tr}(R_s(2\Delta)) = 2\cos(2\theta_s\Delta),$$

the right-hand side equals

$$\frac{1}{D}\,\mathrm{Tr}(R(2\Delta)).$$

This proves the theorem. □

A.4. Proof of Proposition 4.5

For ALiBi, the informative-token scores can be written as

$$S_{\mathrm{fwd}} = \mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B + b\Delta_1, \quad S_{\mathrm{rev}} = \mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B - b\Delta_2,$$

where b, Δ_1, Δ_2 are fixed. Thus the two scores differ from the same random content score $\mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B$ only by additive constants. Hence

$$\mathrm{Var}(S_{\mathrm{fwd}}) = \mathrm{Var}(S_{\mathrm{rev}}) = \mathrm{Var}(\mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B), \quad \mathrm{Cov}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}) = \mathrm{Var}(\mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B).$$

When $\mathrm{Var}(\mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B) > 0$, these identities immediately give

$$\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}) = 1.$$

□

Table 3. Minimum of $\mathrm{Tr}(R(2\Delta))/D$ over integer $0 \le \Delta \le 50$.

D	64	128	256
Minimum bound	0.337	0.384	0.410

A.5. Proof of Theorem 4.6

Let $\mathbf{x}_{\mathrm{fwd}}$ and $\mathbf{x}_{\mathrm{rev}}$ be paired masked prompts with the same target token A and the same unmasked evidence token B. Let $\mathbf{p}_{\mathrm{fwd}}$ and $\mathbf{p}_{\mathrm{rev}}$ be the prediction distributions at the corresponding masked positions, and define

$$\mathcal{L}_{\mathrm{fwd}} = -\log p_{\mathrm{fwd}}(A), \quad \mathcal{L}_{\mathrm{rev}} = -\log p_{\mathrm{rev}}(A).$$

For any prompt with prediction distribution $\mathbf{p}$, target A, and total attention mass α_B assigned to all occurrences of token B, the B-column gradient is

$$\nabla_{Y_{:,B}} \mathcal{L} = \alpha_B(\mathbf{p} - \mathbf{e}_A).$$

Therefore,

$$\nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{fwd}} = \alpha_B^{\mathrm{fwd}}(\mathbf{p}_{\mathrm{fwd}} - \mathbf{e}_A), \quad \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{rev}} = \alpha_B^{\mathrm{rev}}(\mathbf{p}_{\mathrm{rev}} - \mathbf{e}_A).$$

Taking the inner product gives

$$\left\langle \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{fwd}}, \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{rev}} \right\rangle = \alpha_B^{\mathrm{fwd}} \alpha_B^{\mathrm{rev}} \left\langle \mathbf{p}_{\mathrm{fwd}} - \mathbf{e}_A, \mathbf{p}_{\mathrm{rev}} - \mathbf{e}_A \right\rangle.$$

It remains to check that the last inner product is positive. Expanding over the vocabulary,

$$\langle \mathbf{p}_{\mathrm{fwd}} - \mathbf{e}_A, \mathbf{p}_{\mathrm{rev}} - \mathbf{e}_A \rangle = (1 - p_{\mathrm{fwd}}(A))(1 - p_{\mathrm{rev}}(A)) + \sum_{C \neq A} p_{\mathrm{fwd}}(C) p_{\mathrm{rev}}(C) > 0.$$

Since B appears in both prompts and the attention scores are finite, $\alpha_B^{\mathrm{fwd}}, \alpha_B^{\mathrm{rev}} > 0$. Hence

$$\left\langle \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{fwd}}, \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{rev}} \right\rangle > 0.$$

□

Consequently, a forward gradient step restricted to the shared column $Y_{:,B}$ decreases the reverse loss to first order:

$$\mathcal{L}_{\mathrm{rev}}\left(Y_{:,B} - \eta \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{fwd}}\right) - \mathcal{L}_{\mathrm{rev}}(Y_{:,B}) = -\eta \left\langle \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{rev}}, \nabla_{Y_{:,B}} \mathcal{L}_{\mathrm{fwd}} \right\rangle + O(\eta^2) < 0$$

for sufficiently small $\eta > 0$.

B. Details on Attention Correlation Bounds

B.1. Numerical Positivity of the RoPE Trace Bound

Thm. 4.3 lower bounds the expected forward–reverse attention score correlation by

$$\frac{1}{D} \mathrm{Tr}(R(2\Delta)).$$

Thus, to show that the theorem gives a nontrivial positive correlation in the distance range used in our experiments, it remains to check that this trace term is positive for standard RoPE frequencies.

For RoPE with base frequency $\theta_{\mathrm{base}} = 10000$, the normalized trace is

$$\frac{1}{D} \mathrm{Tr}(R(2\Delta)) = \frac{2}{D} \sum_{s=1}^{D/2} \cos\left(2\Delta \cdot 10000^{-2(s-1)/D}\right).$$

This quantity is exact and can be evaluated directly for any finite D. Tab. 3 reports the minimum over integer $0 \le \Delta \le 50$ for $D = 64, 128, 256$. The bound is positive throughout this range for all tested dimensions.

The large-D behavior is also consistent with this numerical pattern. The normalized trace is a Riemann sum:

$$\frac{2}{D}\sum_{s=1}^{D/2}\cos\left(2\Delta\cdot 10000^{-2(s-1)/D}\right)=\int_0^1\cos\left(\frac{2\Delta}{10000^x}\right)dx+O\left(\frac{1}{D}\right).$$

Using the cosine integral

$$\mathrm{Ci}(x)=-\int_x^\infty\frac{\cos t}{t}\,dt,$$

this limiting integral can be written as

$$\int_0^1\cos\left(\frac{2\Delta}{10000^x}\right)dx=\frac{\mathrm{Ci}(2\Delta)-\mathrm{Ci}(2\Delta/10000)}{\log 10000}.$$

Numerically, this limiting approximation is positive over the same range; at $\Delta=50$, it is approximately 0.437. Consequently, Thm. 4.3 gives a positive lower bound in the distance regime considered here.

B.2. A High-Probability Bound for RoPE Correlation

Thm. 4.3 gives a positive lower bound on the expected forward–reverse attention-score correlation in the distance range of interest. Here we show that this positivity is not only an average-case statement: the conditional correlation is also bounded away from zero with high probability.

Using the notation from the proof of Thm. 4.3, the conditional correlation satisfies

$$\mathrm{Corr}(S_{\mathrm{fwd}},S_{\mathrm{rev}}|\mathbf{q})\geq\sum_{s=1}^{m}w_s\cos(2\Delta\theta_s),\quad m=\frac{D}{2},$$

where

$$w_s=\frac{\|\mathbf{q}^{(s)}\|^2}{\|\mathbf{q}\|^2}.$$

Since $\mathbf{q}\sim\mathcal{N}(0,\sigma^2 I_D)$, the weights $(w_1,\ldots,w_m)$ follow $\mathrm{Dirichlet}(1,\ldots,1)$. Thus

$$\mathbb{E}[w_s]=\frac{1}{m},\quad \mathbb{E}[w_s w_t]=\frac{1+\mathbf{1}_{s=t}}{m(m+1)}.$$

Define

$$E_\Delta=\frac{1}{m}\sum_{s=1}^{m}\cos(2\Delta\theta_s),$$

and

$$V_\Delta=\frac{1}{m+1}\left[\frac{1}{m}\sum_{s=1}^{m}\cos^2(2\Delta\theta_s)-\left(\frac{1}{m}\sum_{s=1}^{m}\cos(2\Delta\theta_s)\right)^2\right].$$

Then

$$\mathbb{E}\left[\sum_{s=1}^{m}w_s\cos(2\Delta\theta_s)\right]=E_\Delta,$$

and

$$\mathrm{Var}\left[\sum_{s=1}^{m}w_s\cos(2\Delta\theta_s)\right]=V_\Delta.$$

For any threshold $0<c<E_\Delta$, Cantelli's inequality gives

$$\Pr\left[\sum_{s=1}^{m}w_s\cos(2\Delta\theta_s)\geq c\right]\geq\frac{(E_\Delta-c)^2}{V_\Delta+(E_\Delta-c)^2}.$$

Table 4. Cantelli lower bounds for $\Pr[\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) \geq c]$ under standard RoPE frequencies.

	$D/2 = 64$			$D/2 = 128$		
Δ	$c = 0.10$	$c = 0.20$	$c = 0.30$	$c = 0.10$	$c = 0.20$	$c = 0.30$
10	0.978	0.966	0.941	0.989	0.983	0.970
20	0.959	0.930	0.859	0.982	0.969	0.938
30	0.947	0.905	0.791	0.977	0.958	0.906
40	0.942	0.893	0.753	0.974	0.952	0.885
50	0.956	0.922	0.828	0.973	0.949	0.874

Since the conditional correlation is lower bounded by this weighted sum,

$$\Pr\left[\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) \geq c\right] \geq \frac{(E_\Delta - c)^2}{V_\Delta + (E_\Delta - c)^2}.$$

Tab. 4 shows that the lower bound remains high for the distances used in our experiments. For example, when $\Delta = 20$, $c = 0.3$, and $m = 128$, we obtain

$$\Pr\left[\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) \geq 0.3\right] \geq 0.938.$$

B.3. Numerical Check for Asymmetric RoPE Distances

This subsection provides the numerical check referenced in Remark 4.4. Thm. 4.3 gives a closed-form lower bound in the equal-distance case $\Delta_1 = \Delta_2$. We now check that the same positive-correlation behavior persists when the forward and reverse relative distances are asymmetric, i.e., $\Delta_1 \neq \Delta_2$.

Let $m = D/2$ be the number of RoPE planes. For each plane s, define

$$a_s = \cos((\Delta_1 + \Delta_2)\theta_s), \quad b_s = \cos((\Delta_1 - \Delta_2)\theta_s),$$

and

$$c_s = \cos(2\Delta_1\theta_s), \quad d_s = \cos(2\Delta_2\theta_s).$$

Using the same conditional covariance structure as in Assumption 4.2, the same two-dimensional calculation as in App. A.3 gives

$$\mathrm{Corr}(S_{\mathrm{fwd}}, S_{\mathrm{rev}}|\mathbf{q}) = \frac{\sum_{s=1}^m w_s(a_s + \rho_s b_s)}{\sqrt{(1 + \sum_{s=1}^m w_s\rho_s c_s)\,(1 + \sum_{s=1}^m w_s\rho_s d_s)}},$$

where

$$w_s = \frac{\|\mathbf{q}^{(s)}\|^2}{\|\mathbf{q}\|^2}, \quad \rho_s \in [0, 1].$$

Here ρ_s is the within-plane query-alignment strength from Assumption 4.2.

To obtain a conservative numerical check, we apply the arithmetic-geometric mean inequality to the denominator. Since

$$c_s + d_s = 2a_s b_s,$$

we define the surrogate

$$\underline{\kappa}(\mathbf{w}, \boldsymbol{\rho}) = \frac{2\sum_{s=1}^m w_s(a_s + \rho_s b_s)}{2 + \sum_{s=1}^m w_s\rho_s c_s + \sum_{s=1}^m w_s\rho_s d_s} = \frac{\sum_{s=1}^m w_s(a_s + \rho_s b_s)}{\sum_{s=1}^m w_s(1 + \rho_s a_s b_s)}.$$

Whenever this surrogate is positive, it is a lower bound on the exact conditional correlation, because the exact denominator is no larger than the arithmetic-geometric mean denominator.

For a candidate threshold $\ell \in (0, 1)$, the condition

$$\underline{\kappa}(\mathbf{w}, \boldsymbol{\rho}) \geq \ell$$

Table 5. Minimum Monte Carlo estimate of the asymmetric surrogate bound over the grid $0 \le \Delta_1, \Delta_2 \le 30$.

D	64	128	256
Minimum bound	0.275	0.363	0.395

is equivalent to

$$\sum_{s=1}^{m} w_s \left[a_s + \rho_s b_s - \ell(1 + \rho_s a_s b_s)\right] \ge 0.$$

For fixed ℓ, the inner term is linear in ρ_s:

$$a_s - \ell + \rho_s b_s (1 - \ell a_s).$$

Since $1 - \ell a_s > 0$ for $\ell \in (0, 1)$ and $a_s \in [-1, 1]$, the worst-case choice of ρ_s is

$$\rho_s^\star = \begin{cases} 1, & b_s < 0, \\ 0, & b_s \ge 0. \end{cases}$$

Thus, for each sampled $\mathbf{w}$, the certified threshold ℓ can be checked after minimizing over the admissible ρ_s's.

We estimate the expected worst-case surrogate lower bound

$$\mathbb{E}_{\mathbf{w}} \left[\inf_{\boldsymbol{\rho} \in [0,1]^m} \underline{\kappa}(\mathbf{w}, \boldsymbol{\rho}) \right], \quad \mathbf{w} \sim \mathrm{Dirichlet}(1, \ldots, 1),$$

over the grid $0 \le \Delta_1, \Delta_2 \le 30$. For each grid point, we solve the inner infimum over $\boldsymbol{\rho}$ using a Dinkelbach fractional programming iteration (Dinkelbach, 1967). Tab. 5 reports the minimum Monte Carlo estimate over this grid for $D = 64, 128, 256$, using $50{,}000$ samples of $\mathbf{w}$ at each grid point. The resulting estimates remain positive, indicating that the positive correlation behavior is not specific to the equal-distance case.

C. Additional Theoretical Analyses on Training Dynamics

This section collects additional analyses that complement the main training-dynamics results. App. C.1 analyzes gradient alignment for the query-key terms under the effective reparameterization, with the goal of understanding whether the routing-side components cancel the positive Y-column effect. App. C.2 then analyzes the same gradient-alignment question in the original parameter space, where the decomposition is less sign-determined but exhibits the same positive-transfer tendency. Finally, App. C.3 proves a probability-level convergence result analogous in goal to Zhu et al. (2024) in a single evidence token setting.

C.1. Full-Gradient Alignment Analysis

Thm. 4.6 proves a sign-determined alignment for the shared storage column $Y_{:,B}$: a forward masked update through the common evidence token B decreases the reverse loss to first order. To understand why the full-gradient cosine measured in Secs. 5 and 6 remains positive, we revisit the storage-side Y gradient and then analyze the attention-routing parameters

$$Q = W_Q W_E, \quad K = W_K W_E.$$

This is the first point where the effective query-key coordinates Q and K are used in the gradient analysis.

We use cosine similarity because it separates direction from scale and yields more interpretable decompositions. The Y-gradient cosine factors into a probability-error cosine and an attention-mass cosine. For the query-key terms, the corresponding cosine decomposes into a loss-sensitivity sign and a geometric score-gradient cosine. The analysis does not prove pointwise positivity for every full-gradient component; instead, it explains why the attention-routing terms do not systematically cancel the positive storage-side transfer.

Storage-side cosine decomposition. For a prompt with target token A, prediction distribution $\mathbf{p}$, and attention weights α_k, define the token-type attention vector

$$\mathbf{a} = \sum_{k=1}^{L} \alpha_k \mathbf{e}_{x^k} \in \mathbb{R}^{|\mathcal{V}|+1}.$$

The x-coordinate of $\mathbf{a}$ is the total attention mass assigned to all occurrences of token x. Since

$$\mathbf{z} = \sum_{k=1}^{L} \alpha_k Y \mathbf{e}_{x^k} = Y\mathbf{a},$$

the effective storage gradient is the outer product

$$\nabla_Y \mathcal{L} = (\mathbf{p} - \mathbf{e}_A)\mathbf{a}^\top.$$

Therefore,

$$\langle \nabla_Y \mathcal{L}_{\text{fwd}}, \nabla_Y \mathcal{L}_{\text{rev}} \rangle_F = \langle \mathbf{p}_{\text{fwd}} - \mathbf{e}_A, \mathbf{p}_{\text{rev}} - \mathbf{e}_A \rangle \langle \mathbf{a}_{\text{fwd}}, \mathbf{a}_{\text{rev}} \rangle .$$

Since the Frobenius norm of an outer product factorizes,

$$\|\nabla_Y \mathcal{L}\|_F = \|\mathbf{p} - \mathbf{e}_A\| \, \|\mathbf{a}\|,$$

we obtain the exact cosine decomposition

$$\cos_F (\nabla_Y \mathcal{L}_{\text{fwd}}, \nabla_Y \mathcal{L}_{\text{rev}}) = \cos (\mathbf{p}_{\text{fwd}} - \mathbf{e}_A, \mathbf{p}_{\text{rev}} - \mathbf{e}_A) \cos (\mathbf{a}_{\text{fwd}}, \mathbf{a}_{\text{rev}}) .$$

The probability-error cosine is positive because

$$\langle \mathbf{p}_{\text{fwd}} - \mathbf{e}_A, \mathbf{p}_{\text{rev}} - \mathbf{e}_A \rangle = (1 - p_{\text{fwd}}(A))(1 - p_{\text{rev}}(A)) + \sum_{C \neq A} p_{\text{fwd}}(C) p_{\text{rev}}(C) > 0.$$

The attention-mass cosine is nonnegative because $\mathbf{a}_{\text{fwd}}$ and $\mathbf{a}_{\text{rev}}$ have nonnegative entries, and it is strictly positive whenever the two prompts share at least one token receiving positive attention in both prompts. In particular, the shared evidence token B makes this cosine positive. Thus the full Y-gradient cosine is positive, strengthening the B-column inner-product result of Thm. 4.6.

Query-key cosine decomposition. We next isolate the component of the Q and K gradients that flows through the attention score assigned to the evidence token B. For a single masked position, write

$$\mathbf{z} = \sum_{\ell=1}^{L} \alpha_\ell Y \mathbf{e}_{x^\ell}, \quad \mathbf{p} = \text{softmax}(\mathbf{z}), \quad \mathcal{L} = -\log p(A).$$

Define the token-wise storage error

$$\mu_\ell = (\mathbf{p} - \mathbf{e}_A)^\top Y \mathbf{e}_{x^\ell}.$$

Since $\partial \mathcal{L} / \partial \alpha_\ell = \mu_\ell$ and $\alpha_\ell = \text{softmax}(S_\ell / \sqrt{D})$, the loss sensitivity of score S_k is

$$\frac{\partial \mathcal{L}}{\partial S_k} = \frac{\alpha_k}{\sqrt{D}} \left(\mu_k - \sum_{\ell=1}^{L} \alpha_\ell \mu_\ell \right).$$

To isolate the clean transfer term, we restrict the query-key calculation to the backpropagation path through the attention score assigned to the evidence token B. Thus the derivation below keeps the B-associated score-gradient contribution and leaves the other token contributions to the final interpretation paragraph.

For the evidence token B, define

$$\xi_B := \frac{\partial \mathcal{L}}{\partial S_B} = \frac{\alpha_B}{\sqrt{D}} \left(\mu_B - \sum_{\ell=1}^{L} \alpha_\ell \mu_\ell \right).$$

The sign of ξ_B is determined by

$$\mu_B - \sum_{\ell=1}^{L} \alpha_\ell \mu_\ell.$$

This contrast measures whether attending more to B would reduce the loss compared with the current attention mixture. For relative positional encodings, the value carried by token B does not depend on its position, so the forward and reverse prompts retrieve the same B-side evidence even though they route attention through different attention scores. Starting from a mean-zero initialization, forward training on recovering A from B makes this B-side evidence increasingly helpful for predicting A, so the contrast moves downward and becomes negative. Hence, once B is useful in both prompts, the forward and reverse score sensitivities have the same sign:

$$\mathrm{sgn}(\xi_B^{\mathrm{fwd}}\xi_B^{\mathrm{rev}}) = 1.$$

For a score $S_B = S(\mathbf{q}_{[\mathbf{M}]}, \mathbf{k}_B, \Delta)$, the B-route gradient components are

$$\nabla_Q \mathcal{L}^{(B)} = \xi_B(\nabla_{\mathbf{q}_{[\mathbf{M}]}} S_B)\mathbf{e}_{[\mathbf{M}]}^\top, \quad \nabla_K \mathcal{L}^{(B)} = \xi_B(\nabla_{\mathbf{k}_B} S_B)\mathbf{e}_B^\top.$$

Hence, whenever the relevant norms are nonzero,

$$\cos_F\left(\nabla_Q \mathcal{L}_{\mathrm{fwd}}^{(B)}, \nabla_Q \mathcal{L}_{\mathrm{rev}}^{(B)}\right) = \mathrm{sgn}\left(\xi_B^{\mathrm{fwd}}\xi_B^{\mathrm{rev}}\right)\cos\left(\nabla_{\mathbf{q}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{q}} S_B^{\mathrm{rev}}\right),$$

and

$$\cos_F\left(\nabla_K \mathcal{L}_{\mathrm{fwd}}^{(B)}, \nabla_K \mathcal{L}_{\mathrm{rev}}^{(B)}\right) = \mathrm{sgn}\left(\xi_B^{\mathrm{fwd}}\xi_B^{\mathrm{rev}}\right)\cos\left(\nabla_{\mathbf{k}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{k}} S_B^{\mathrm{rev}}\right).$$

Thus, once the sensitivity signs agree, the B-route query-key cosine is controlled by the geometry of the attention-score gradients.

RoPE. For RoPE,

$$S_B^{\mathrm{fwd}} = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1)\mathbf{k}_B, \quad S_B^{\mathrm{rev}} = \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B.$$

Therefore,

$$\nabla_{\mathbf{q}} S_B^{\mathrm{fwd}} = R(\Delta_1)\mathbf{k}_B, \quad \nabla_{\mathbf{q}} S_B^{\mathrm{rev}} = R(-\Delta_2)\mathbf{k}_B,$$

and

$$\nabla_{\mathbf{k}} S_B^{\mathrm{fwd}} = R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]}, \quad \nabla_{\mathbf{k}} S_B^{\mathrm{rev}} = R(\Delta_2)\mathbf{q}_{[\mathbf{M}]}.$$

Thus the geometric cosines are

$$\cos\left(\nabla_{\mathbf{q}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{q}} S_B^{\mathrm{rev}}\right) = \frac{\mathbf{k}_B^\top R(\Delta_1+\Delta_2)\mathbf{k}_B}{\|\mathbf{k}_B\|^2},$$

and

$$\cos\left(\nabla_{\mathbf{k}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{k}} S_B^{\mathrm{rev}}\right) = \frac{\mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1+\Delta_2)\mathbf{q}_{[\mathbf{M}]}}{\|\mathbf{q}_{[\mathbf{M}]}\|^2}.$$

These quantities are not pointwise positive for every vector, because RoPE rotates different frequency planes by different angles. However, under isotropic or RoPE-compatible covariance conditions, their expectations are controlled by normalized trace terms. For example, if $\mathbf{k}_B \sim \mathcal{N}(0, \sigma^2 I_D)$, then

$$\mathbb{E}\left[\frac{\mathbf{k}_B^\top R(\Delta_1+\Delta_2)\mathbf{k}_B}{\|\mathbf{k}_B\|^2}\right] = \frac{1}{D}\,\mathrm{Tr}(R(\Delta_1+\Delta_2)).$$

Thus RoPE gives partial, trace-controlled query-key alignment rather than exact alignment. It may weaken the B-route query-key cosine relative to the storage-side Y-cosine, but it does not introduce a systematic negative alignment in the distance regime where the corresponding trace terms remain positive.

ALiBi. For ALiBi,

$$S_B = \mathbf{q}_{[\mathbf{M}]}^\top \mathbf{k}_B + b\Delta,$$

The bias $b\Delta$ affects the attention score and hence the routing weight, but it vanishes when differentiating with respect to $\mathbf{q}_{[\mathbf{M}]}$ or $\mathbf{k}_B$:

$$\nabla_{\mathbf{q}} S_B = \mathbf{k}_B, \quad \nabla_{\mathbf{k}} S_B = \mathbf{q}_{[\mathbf{M}]}.$$

Consequently,

$$\cos\left(\nabla_{\mathbf{q}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{q}} S_B^{\mathrm{rev}}\right) = 1, \quad \cos\left(\nabla_{\mathbf{k}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{k}} S_B^{\mathrm{rev}}\right) = 1.$$

Thus, when the sensitivity signs agree, ALiBi gives exact B-route query-key cosine alignment. This sharply contrasts with RoPE: ALiBi changes the scalar attention scores by an additive distance bias, but leaves the query-key gradient directions identical across forward and reverse configurations.

Full-gradient interpretation. The Y-gradient cosine gives the clean storage-side signal, while the B-route Q, K analysis shows that the attention-routing terms do not inherently oppose it. The remaining Q, K terms involve other context tokens and cross-token interactions through the attention softmax, so their signs are not determined by the B-route calculation. These terms depend on token-specific embeddings and positional score gradients that are not consistently tied to the forward–reverse evidence pair. Under a non-adversarial high-dimensional geometry, such residual directions are expected to be weakly correlated on average, rather than systematically anti-aligned with the B-route signal.

Thus, this analysis should be read as a robustness check on Thm. 4.6: the guaranteed storage-side alignment is the main positive transfer signal, and the B-route query-key terms do not structurally cancel it. Under ALiBi they align exactly; under RoPE they retain partial, trace-controlled alignment. This is consistent with the positive full-parameter gradient cosine observed in Secs. 5 and 6.

C.2. Gradient Alignment in the Original Parameter Space

The main text analyzes the transfer mechanism through the shared storage parameter Y, and the preceding appendix section uses the effective query-key parameters Q and K to analyze routing. We now relate this effective-parameter picture to the original parameterization, where W_E, W_Q, W_K, W_V, and W_U are updated jointly. The goal is to show that the same positive transfer tendency observed in the effective coordinates also appears naturally when the gradient is decomposed over the original matrices.

Setup. To isolate the dynamics, we consider a single evidence token setting with arbitrary relative distances. Even though there is only one informative token, we keep the relative distances explicit so that the effect of positional encoding remains visible in the attention score. The forward prompt contains a masked target position and the evidence token B, with B lying Δ_1 positions to the right of the mask. The reverse prompt contains the same evidence token B, with B lying Δ_2 positions to the left of the mask. Both prompts ask the model to recover the same target token A.

Let

$$\mathbf{h}_{[\mathbf{M}]} = W_E \mathbf{e}_{[\mathbf{M}]}, \quad \mathbf{h}_B = W_E \mathbf{e}_B,$$

and

$$\mathbf{q}_{[\mathbf{M}]} = W_Q \mathbf{h}_{[\mathbf{M}]}, \quad \mathbf{k}_B = W_K \mathbf{h}_B, \quad \mathbf{v}_B = W_V \mathbf{h}_B.$$

Let

$$S_B = S(\mathbf{q}_{[\mathbf{M}]}, \mathbf{k}_B, \Delta)$$

denote the attention score assigned to the evidence token B. In this simplified two-route attention setting, we write the attention weight on B as

$$\alpha_B = \sigma(S_B/\sqrt{D}),$$

where σ is the sigmoid function. This notation can be understood as absorbing the competing mask-route score into the score difference defining S_B.

We focus on the contribution of the evidence-token value to the output logits. Equivalently, we approximate the masked-position logit vector by the part routed through B:

$$\mathbf{z} = W_U(\alpha_B \mathbf{v}_B), \quad \mathbf{p} = \mathrm{softmax}(\mathbf{z}), \quad \mathbf{u} = W_U^\top(\mathbf{p} - \mathbf{e}_A).$$

Proposition C.1 (Original-Parameter Gradient Decomposition)**.** *Under the single-evidence-token setup above, define*

$$I_W = \langle \nabla_W \mathcal{L}_{\mathrm{fwd}}, \nabla_W \mathcal{L}_{\mathrm{rev}} \rangle_F$$

for each original parameter matrix $W \in \{W_U, W_V, W_Q, W_K, W_E\}$. Then the unembedding, value, query, and key projection contributions satisfy

$$I_{W_U} = \alpha_B^{\mathrm{fwd}} \alpha_B^{\mathrm{rev}} \langle \mathbf{p}_{\mathrm{fwd}} - \mathbf{e}_A, \mathbf{p}_{\mathrm{rev}} - \mathbf{e}_A \rangle \|\mathbf{v}_B\|^2,$$

$$I_{W_V} = \alpha_B^{\mathrm{fwd}} \alpha_B^{\mathrm{rev}} \langle \mathbf{u}_{\mathrm{fwd}}, \mathbf{u}_{\mathrm{rev}} \rangle \|\mathbf{h}_B\|^2,$$

$$I_{W_Q} = \frac{\alpha_B^{\mathrm{fwd}}(1-\alpha_B^{\mathrm{fwd}})\alpha_B^{\mathrm{rev}}(1-\alpha_B^{\mathrm{rev}})}{D} (\mathbf{u}_{\mathrm{fwd}}^\top \mathbf{v}_B)(\mathbf{u}_{\mathrm{rev}}^\top \mathbf{v}_B) \|\mathbf{h}_{[\mathbf{M}]}\|^2 \left\langle \nabla_{\mathbf{q}} S_B^{\mathrm{fwd}}, \nabla_{\mathbf{q}} S_B^{\mathrm{rev}} \right\rangle,$$

and

$$I_{W_K} = \frac{\alpha_B^{\text{fwd}}(1-\alpha_B^{\text{fwd}})\alpha_B^{\text{rev}}(1-\alpha_B^{\text{rev}})}{D}(\mathbf{u}_{\text{fwd}}^\top \mathbf{v}_B)(\mathbf{u}_{\text{rev}}^\top \mathbf{v}_B)\|\mathbf{h}_B\|^2 \left\langle \nabla_{\mathbf{k}} S_B^{\text{fwd}}, \nabla_{\mathbf{k}} S_B^{\text{rev}} \right\rangle.$$

For the embedding matrix, the inner product decomposes as

$$I_{W_E} = I_{W_E}^{(Q)} + I_{W_E}^{(K)} + I_{W_E}^{(V)} + I_{W_E}^{(KV)},$$

where

$$I_{W_E}^{(Q)} = \frac{\alpha_B^{\text{fwd}}(1-\alpha_B^{\text{fwd}})\alpha_B^{\text{rev}}(1-\alpha_B^{\text{rev}})}{D}(\mathbf{u}_{\text{fwd}}^\top \mathbf{v}_B)(\mathbf{u}_{\text{rev}}^\top \mathbf{v}_B) \left\langle W_Q^\top \nabla_{\mathbf{q}} S_B^{\text{fwd}}, W_Q^\top \nabla_{\mathbf{q}} S_B^{\text{rev}} \right\rangle,$$

$$I_{W_E}^{(K)} = \frac{\alpha_B^{\text{fwd}}(1-\alpha_B^{\text{fwd}})\alpha_B^{\text{rev}}(1-\alpha_B^{\text{rev}})}{D}(\mathbf{u}_{\text{fwd}}^\top \mathbf{v}_B)(\mathbf{u}_{\text{rev}}^\top \mathbf{v}_B) \left\langle W_K^\top \nabla_{\mathbf{k}} S_B^{\text{fwd}}, W_K^\top \nabla_{\mathbf{k}} S_B^{\text{rev}} \right\rangle,$$

$$I_{W_E}^{(V)} = \alpha_B^{\text{fwd}}\alpha_B^{\text{rev}} \left\langle W_V^\top \mathbf{u}_{\text{fwd}}, W_V^\top \mathbf{u}_{\text{rev}} \right\rangle,$$

and $I_{W_E}^{(KV)}$ collects the cross-terms between the key and value paths inside the B-embedding column.

Proof of Proposition C.1. For $\mathcal{L} = -\log p(A)$,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}} = \mathbf{p} - \mathbf{e}_A.$$

Since $\mathbf{z} = W_U(\alpha_B \mathbf{v}_B)$, the unembedding and value projection gradients are

$$\nabla_{W_U}\mathcal{L} = \alpha_B(\mathbf{p} - \mathbf{e}_A)\mathbf{v}_B^\top,$$

and

$$\nabla_{W_V}\mathcal{L} = \alpha_B W_U^\top(\mathbf{p} - \mathbf{e}_A)\mathbf{h}_B^\top = \alpha_B \mathbf{u}\mathbf{h}_B^\top.$$

Taking Frobenius inner products of the corresponding forward and reverse gradients gives the stated formulas for I_{W_U} and I_{W_V}.

For the query and key projections,

$$\frac{\partial \alpha_B}{\partial S_B} = \frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}.$$

Also,

$$\frac{\partial \mathcal{L}}{\partial \alpha_B} = (\mathbf{p} - \mathbf{e}_A)^\top W_U \mathbf{v}_B = \mathbf{u}^\top \mathbf{v}_B.$$

Hence

$$\frac{\partial \mathcal{L}}{\partial S_B} = \frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}\mathbf{u}^\top \mathbf{v}_B.$$

By the chain rule,

$$\nabla_{W_Q}\mathcal{L} = \left[\frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}\mathbf{u}^\top \mathbf{v}_B\right](\nabla_{\mathbf{q}_{[\mathbf{M}]}} S_B)\mathbf{h}_{[\mathbf{M}]}^\top,$$

and

$$\nabla_{W_K}\mathcal{L} = \left[\frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}\mathbf{u}^\top \mathbf{v}_B\right](\nabla_{\mathbf{k}_B} S_B)\mathbf{h}_B^\top.$$

Their Frobenius inner products yield the stated formulas for I_{W_Q} and I_{W_K}.

It remains to track the embedding matrix. The masked-token embedding receives the query-path signal

$$\nabla_{\mathbf{h}_{[\mathbf{M}]}}\mathcal{L} = \left[\frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}\mathbf{u}^\top \mathbf{v}_B\right] W_Q^\top \nabla_{\mathbf{q}_{[\mathbf{M}]}} S_B.$$

The evidence-token embedding receives both key-path and value-path signals:

$$\nabla_{\mathbf{h}_B}\mathcal{L} = \left[\frac{\alpha_B(1-\alpha_B)}{\sqrt{D}}\mathbf{u}^\top \mathbf{v}_B\right] W_K^\top \nabla_{\mathbf{k}_B} S_B + \alpha_B W_V^\top \mathbf{u}.$$

Thus

$$\nabla_{W_E}\mathcal{L} = (\nabla_{\mathbf{h}_{[\mathbf{M}]}}\mathcal{L})\mathbf{e}_{[\mathbf{M}]}^\top + (\nabla_{\mathbf{h}_B}\mathcal{L})\mathbf{e}_B^\top.$$

The masked-token column and the B-token column are orthogonal because $\mathbf{e}_{[\mathbf{M}]}^\top\mathbf{e}_B = 0$, so the query-path contribution separates from the B-column contribution. Inside the B-column, the key and value paths contribute the K-path and V-path terms above, along with their cross-terms $I_{W_E}^{(KV)}$. This gives the stated decomposition. □

Storage-side interpretation. The decomposition above shows how the effective storage-side alignment appears inside the original parameterization. The W_U term is sign-determined:

$$\langle \mathbf{p}_{\text{fwd}} - \mathbf{e}_A, \mathbf{p}_{\text{rev}} - \mathbf{e}_A\rangle = (1 - p_{\text{fwd}}(A))(1 - p_{\text{rev}}(A)) + \sum_{C\neq A} p_{\text{fwd}}(C)p_{\text{rev}}(C) > 0.$$

Thus $I_{W_U} > 0$ whenever B receives positive attention in both prompts.

The W_V and $W_E^{(V)}$ terms depend on the backpropagated error signals

$$\langle \mathbf{u}_{\text{fwd}}, \mathbf{u}_{\text{rev}}\rangle = \left\langle W_U^\top(\mathbf{p}_{\text{fwd}} - \mathbf{e}_A), W_U^\top(\mathbf{p}_{\text{rev}} - \mathbf{e}_A)\right\rangle.$$

Under non-adversarial geometry, such as an approximately isotropic unembedding map, this inner product preserves the positive probability-error alignment. These terms are the original-parameter counterparts of the positive storage-side alignment captured by Y.

Query-key routing interpretation. The W_Q, W_K, and query/key paths of W_E share the scalar coefficient

$$(\mathbf{u}_{\text{fwd}}^\top\mathbf{v}_B)(\mathbf{u}_{\text{rev}}^\top\mathbf{v}_B).$$

The scalar $\mathbf{u}^\top\mathbf{v}_B$ measures whether increasing attention to B lowers the loss. When B is useful for predicting A, this scalar has the same sign in the forward and reverse prompts, so the product above is positive. The remaining factor is geometric and depends on the positional encoding.

For RoPE,

$$S_B^{\text{fwd}} = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1)\mathbf{k}_B, \quad S_B^{\text{rev}} = \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B.$$

Therefore,

$$\left\langle \nabla_{\mathbf{q}} S_B^{\text{fwd}}, \nabla_{\mathbf{q}} S_B^{\text{rev}}\right\rangle = \mathbf{k}_B^\top R(\Delta_1 + \Delta_2)\mathbf{k}_B,$$

and

$$\left\langle \nabla_{\mathbf{k}} S_B^{\text{fwd}}, \nabla_{\mathbf{k}} S_B^{\text{rev}}\right\rangle = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1 + \Delta_2)\mathbf{q}_{[\mathbf{M}]}.$$

These quadratic forms are not pointwise positive for every vector, but their expectations are controlled by the same RoPE trace terms analyzed in Apps. B.1 and B.3. Thus RoPE may weaken routing alignment, but it does not create a systematic negative alignment in the distance regimes considered here.

For ALiBi,

$$S_B = \mathbf{q}_{[\mathbf{M}]}^\top\mathbf{k}_B + b_{\text{pos}}(\Delta),$$

so the additive positional bias vanishes under differentiation:

$$\nabla_{\mathbf{q}} S_B = \mathbf{k}_B, \quad \nabla_{\mathbf{k}} S_B = \mathbf{q}_{[\mathbf{M}]}.$$

The geometric factors are therefore $\|\mathbf{k}_B\|^2$ and $\|\mathbf{q}_{[\mathbf{M}]}\|^2$, both positive. This explains why ALiBi preserves query-key routing alignment more strongly than RoPE.

Summary. Overall, the original-parameter decomposition supports the same conclusion as the effective-parameter analysis. The storage-side terms provide the main positive transfer signal, while the query-key routing terms do not structurally cancel it. The remaining embedding cross-terms $I_{W_E}^{(KV)}$ are not sign-determined, but they arise from interactions between distinct key and value paths and are best interpreted as residual terms rather than as a systematic negative force.

C.3. Probability-level Transfer in a Single Evidence Token Setting

The gradient-alignment results above show that a forward masked update moves the shared storage column $Y_{:,B}$ in a direction that decreases the reverse loss to first order. Here we give a minimal probability-level training-dynamics counterpart to the analysis of Zhu et al. (2024), adapted to the MDM setting. The result is intentionally idealized: it isolates a single-evidence-token mechanism and should not be read as a claim that reverse probabilities generally saturate in practical multi-token or multi-layer MDMs. The distinction is that, in an MDM, storing the B-to-A evidence is not by itself enough. Forward training can make the shared column $Y_{:,B}$ strongly favor A, but the reverse prompt benefits from this storage only if its masked position actually attends to B. This routing issue is absent or effectively trivial in the autoregressive setting analyzed by Zhu et al. (2024), where the relevant attention route is fixed. Our result therefore makes the two steps explicit: forward training separates $Y_{A,B}$ from $Y_{C,B}$ for $C \neq A$, and reverse transfer follows when the reverse route to B remains active.

Setup. We work with the effective parameters Y, Q, K. Both the forward and reverse prompts contain one masked position and one informative evidence token B. Although this is a single evidence token setting, we keep explicit relative distances in order to track the effect of positional encoding on the attention score. In the forward prompt, B lies Δ_1 positions to the right of the mask. In the reverse prompt, B lies Δ_2 positions to the left of the mask. Both prompts ask the model to recover the same target token A.

We isolate the evidence-token contribution to the masked-position logits:

$$\mathbf{z}_{\mathrm{fwd}}(t) = \alpha_B^{\mathrm{fwd}}(t) Y_{:,B}(t), \quad \mathbf{z}_{\mathrm{rev}}(t) = \alpha_B^{\mathrm{rev}}(t) Y_{:,B}(t).$$

Here

$$\alpha_B(t) = \sigma(S_B(t)/\sqrt{D}),$$

where σ is the sigmoid function. Equivalently, this is the single evidence token reduction of the two-route attention between the mask token and the evidence token B. The prediction distributions are

$$\mathbf{p}_{\mathrm{fwd}}(t) = \mathrm{softmax}(\mathbf{z}_{\mathrm{fwd}}(t)), \quad \mathbf{p}_{\mathrm{rev}}(t) = \mathrm{softmax}(\mathbf{z}_{\mathrm{rev}}(t)).$$

We use gradient flow on the forward loss

$$\mathcal{L}_{\mathrm{fwd}}(t) = -\log p_{\mathrm{fwd}}(A; t)$$

with respect to the shared storage column $Y_{:,B}$ and the routing variables specified below. For a clean statement, we initialize the shared column uniformly and write this as

$$Y_{:,B}(0) = \mathbf{0}.$$

RoPE setup. For RoPE, the query and key play symmetric roles in the bilinear attention score. To isolate the routing geometry in the simplest form, we fix the masked-token query $\mathbf{q}_{[\mathbf{M}]}$ and update only the evidence-token key $\mathbf{k}_B(t)$ by gradient flow on the forward loss. The forward and reverse scores are

$$S_{\mathrm{fwd}}(t) = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1) \mathbf{k}_B(t), \quad S_{\mathrm{rev}}(t) = \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2) \mathbf{k}_B(t).$$

We assume the RoPE geometric alignment condition

$$\mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1 + \Delta_2) \mathbf{q}_{[\mathbf{M}]} > 0.$$

This is the same trace-controlled geometry that appears in the RoPE attention-correlation analysis. For example, if $\mathbf{q}_{[\mathbf{M}]} \sim \mathcal{N}(0, \sigma^2 I_D)$, then

$$\mathbb{E}\left[\mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1 + \Delta_2) \mathbf{q}_{[\mathbf{M}]}\right] = \sigma^2 \,\mathrm{Tr}(R(\Delta_1 + \Delta_2)).$$

Thus, the corresponding quadratic form has positive expectation whenever the RoPE trace is positive. This is the same trace quantity with argument $r = \Delta_1 + \Delta_2$; App. B.1 verifies positivity over the relevant range.

Proposition C.2 (RoPE Probability Transfer with a Fixed Query)**.** *Assume the RoPE setup above. Then*

$$\lim_{t\to\infty} p_{\mathrm{fwd}}(A; t) = 1, \quad \lim_{t\to\infty} p_{\mathrm{rev}}(A; t) = 1.$$

ALiBi setup. For ALiBi, write the forward and reverse scores as

$$S_{\text{fwd}}(t) = \mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t) + b\Delta_1, \quad S_{\text{rev}}(t) = \mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t) - b\Delta_2.$$

Here b is the fixed ALiBi bias slope, and Δ_1, Δ_2 are the forward and reverse distances defined above. We allow the routing variables $\mathbf{q}_{[\mathbf{M}]}(t)$ and $\mathbf{k}_B(t)$ to be either fixed or updated by gradient flow on the forward loss.

Proposition C.3 (ALiBi Probability Transfer)**.** *Assume the ALiBi setup above. Then*

$$\lim_{t\to\infty} p_{\text{fwd}}(A;t) = 1, \quad \lim_{t\to\infty} p_{\text{rev}}(A;t) = 1.$$

Proof of Propositions C.2 and C.3. We first prove the shared storage part, which is common to both propositions. For the forward prompt,

$$(\mathbf{z}_{\text{fwd}}(t))_C = \alpha_B^{\text{fwd}}(t) Y_{C,B}(t),$$

and therefore

$$p_{\text{fwd}}(C;t) = \frac{\exp(\alpha_B^{\text{fwd}}(t) Y_{C,B}(t))}{\sum_{C'\in\mathcal{V}} \exp(\alpha_B^{\text{fwd}}(t) Y_{C',B}(t))}.$$

The gradient of the forward loss with respect to $Y_{C,B}$ is

$$\nabla_{Y_{C,B}} \mathcal{L}_{\text{fwd}} = \alpha_B^{\text{fwd}}(t) \left(p_{\text{fwd}}(C;t) - \delta_{C,A}\right).$$

Thus, under gradient flow with rate $\eta > 0$,

$$\frac{d}{dt} Y_{A,B}(t) = \eta \alpha_B^{\text{fwd}}(t) \left(1 - p_{\text{fwd}}(A;t)\right),$$

and, for $C \neq A$,

$$\frac{d}{dt} Y_{C,B}(t) = -\eta \alpha_B^{\text{fwd}}(t) p_{\text{fwd}}(C;t).$$

Hence

$$\frac{d}{dt}\left(Y_{A,B}(t) - Y_{C,B}(t)\right) = \eta \alpha_B^{\text{fwd}}(t) \left(1 - p_{\text{fwd}}(A;t) + p_{\text{fwd}}(C;t)\right) > 0.$$

Since $Y_{:,B}(0) = \mathbf{0}$, we have

$$Y_{A,B}(t) - Y_{C,B}(t) \geq 0 \quad (C \neq A)$$

for all t.

We next show that the forward route to B does not collapse. First,

$$\frac{\partial \mathcal{L}_{\text{fwd}}}{\partial \alpha_B^{\text{fwd}}} = (\mathbf{p}_{\text{fwd}}(t) - \mathbf{e}_A)^\top Y_{:,B}(t).$$

Using

$$1 - p_{\text{fwd}}(A;t) = \sum_{C\neq A} p_{\text{fwd}}(C;t),$$

we obtain

$$\begin{aligned}
(\mathbf{p}_{\text{fwd}}(t) - \mathbf{e}_A)^\top Y_{:,B}(t) &= \sum_{C\neq A} p_{\text{fwd}}(C;t) Y_{C,B}(t) - \left(1 - p_{\text{fwd}}(A;t)\right) Y_{A,B}(t) \\
&= \sum_{C\neq A} p_{\text{fwd}}(C;t) \left(Y_{C,B}(t) - Y_{A,B}(t)\right) \\
&= -\sum_{C\neq A} p_{\text{fwd}}(C;t) \left(Y_{A,B}(t) - Y_{C,B}(t)\right) \leq 0.
\end{aligned}$$

Therefore,

$$\frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} = \frac{\alpha_B^{\text{fwd}}(t)\left(1 - \alpha_B^{\text{fwd}}(t)\right)}{\sqrt{D}} \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial \alpha_B^{\text{fwd}}} \leq 0.$$

For RoPE, gradient flow on $\mathbf{k}_B(t)$ gives

$$\frac{d}{dt}\mathbf{k}_B(t) = -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]}.$$

Hence

$$\begin{aligned}\frac{d}{dt}S_{\text{fwd}}(t) &= \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1)\frac{d}{dt}\mathbf{k}_B(t) \\ &= -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1)R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]} \\ &= -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} \left\|\mathbf{q}_{[\mathbf{M}]}\right\|^2 \geq 0.\end{aligned}$$

Thus

$$S_{\text{fwd}}(t) \geq S_{\text{fwd}}(0),$$

and

$$\alpha_B^{\text{fwd}}(t) = \sigma(S_{\text{fwd}}(t)/\sqrt{D}) \geq \sigma(S_{\text{fwd}}(0)/\sqrt{D}) > 0.$$

For ALiBi, if the routing variables are fixed, then

$$\mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t) = \mathbf{q}_{[\mathbf{M}]}(0)^\top \mathbf{k}_B(0).$$

If they are updated by gradient flow on the forward loss, then

$$\frac{d}{dt}\mathbf{q}_{[\mathbf{M}]}(t) = -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} \mathbf{k}_B(t),$$

and

$$\frac{d}{dt}\mathbf{k}_B(t) = -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} \mathbf{q}_{[\mathbf{M}]}(t).$$

Therefore,

$$\begin{aligned}\frac{d}{dt}\left(\mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t)\right) &= \left(\frac{d}{dt}\mathbf{q}_{[\mathbf{M}]}(t)\right)^\top \mathbf{k}_B(t) + \mathbf{q}_{[\mathbf{M}]}(t)^\top \left(\frac{d}{dt}\mathbf{k}_B(t)\right) \\ &= -\eta \frac{\partial \mathcal{L}_{\text{fwd}}}{\partial S_{\text{fwd}}} \left(\|\mathbf{k}_B(t)\|^2 + \left\|\mathbf{q}_{[\mathbf{M}]}(t)\right\|^2\right) \\ &\geq 0.\end{aligned}$$

Thus, in either ALiBi case, $S_{\text{fwd}}(t)$ is bounded below, and hence

$$\alpha_B^{\text{fwd}}(t) = \sigma(S_{\text{fwd}}(t)/\sqrt{D})$$

is bounded away from zero.

We now prove that all storage margins diverge. Fix $C \neq A$. Suppose for contradiction that

$$Y_{A,B}(t) - Y_{C,B}(t)$$

is bounded above. Since it is nondecreasing, there exists $M < \infty$ such that

$$0 \leq Y_{A,B}(t) - Y_{C,B}(t) \leq M$$

for all t. Also, for every $C' \neq A$,

$$Y_{A,B}(t) - Y_{C',B}(t) \geq 0.$$

Therefore,

$$\begin{aligned}p_{\text{fwd}}(C;t) &= \frac{\exp\left(\alpha_B^{\text{fwd}}(t) Y_{C,B}(t)\right)}{\exp\left(\alpha_B^{\text{fwd}}(t) Y_{A,B}(t)\right) + \sum_{C' \neq A} \exp\left(\alpha_B^{\text{fwd}}(t) Y_{C',B}(t)\right)} \\ &= \frac{\exp\left(-\alpha_B^{\text{fwd}}(t)\left(Y_{A,B}(t) - Y_{C,B}(t)\right)\right)}{1 + \sum_{C' \neq A} \exp\left(-\alpha_B^{\text{fwd}}(t)\left(Y_{A,B}(t) - Y_{C',B}(t)\right)\right)}.\end{aligned}$$

Since $0 < \alpha_B^{\mathrm{fwd}}(t) \le 1$,

$$\exp\left(-\alpha_B^{\mathrm{fwd}}(t)\left(Y_{A,B}(t) - Y_{C,B}(t)\right)\right) \ge e^{-M}.$$

Moreover,

$$\exp\left(-\alpha_B^{\mathrm{fwd}}(t)\left(Y_{A,B}(t) - Y_{C',B}(t)\right)\right) \le 1 \quad (C' \ne A).$$

Hence

$$p_{\mathrm{fwd}}(C;t) \ge \frac{e^{-M}}{|\mathcal{V}|} > 0.$$

Using the positive lower bound on $\alpha_B^{\mathrm{fwd}}(t)$, we get

$$\begin{aligned}\frac{d}{dt}\left(Y_{A,B}(t) - Y_{C,B}(t)\right) &= \eta\alpha_B^{\mathrm{fwd}}(t)\left(1 - p_{\mathrm{fwd}}(A;t) + p_{\mathrm{fwd}}(C;t)\right)\\ &\ge \eta\alpha_B^{\mathrm{fwd}}(t)p_{\mathrm{fwd}}(C;t)\\ &\ge \eta\left(\inf_{s\ge 0}\alpha_B^{\mathrm{fwd}}(s)\right)\frac{e^{-M}}{|\mathcal{V}|} > 0.\end{aligned}$$

This contradicts the boundedness of

$$Y_{A,B}(t) - Y_{C,B}(t).$$

Therefore,

$$Y_{A,B}(t) - Y_{C,B}(t) \to +\infty \quad (C \ne A).$$

It follows that

$$\begin{aligned}p_{\mathrm{fwd}}(A;t) &= \frac{\exp\left(\alpha_B^{\mathrm{fwd}}(t)Y_{A,B}(t)\right)}{\sum_{C'\in\mathcal{V}}\exp\left(\alpha_B^{\mathrm{fwd}}(t)Y_{C',B}(t)\right)}\\ &= \frac{1}{1 + \sum_{C\ne A}\exp\left(-\alpha_B^{\mathrm{fwd}}(t)\left(Y_{A,B}(t) - Y_{C,B}(t)\right)\right)}.\end{aligned}$$

Since $\alpha_B^{\mathrm{fwd}}(t)$ is bounded away from zero and

$$Y_{A,B}(t) - Y_{C,B}(t) \to +\infty \quad (C \ne A),$$

each exponential term vanishes. Thus

$$p_{\mathrm{fwd}}(A;t) \to 1.$$

It remains to check the reverse route to B. For RoPE,

$$\frac{d}{dt}\mathbf{k}_B(t) = -\eta\frac{\partial\mathcal{L}_{\mathrm{fwd}}}{\partial S_{\mathrm{fwd}}}R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]},$$

with

$$-\eta\frac{\partial\mathcal{L}_{\mathrm{fwd}}}{\partial S_{\mathrm{fwd}}} \ge 0.$$

Therefore, for some nondecreasing $\lambda(t) \ge 0$,

$$\mathbf{k}_B(t) = \mathbf{k}_B(0) + \lambda(t)R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]}.$$

Then

$$\begin{aligned}S_{\mathrm{rev}}(t) &= \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B(t)\\ &= \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B(0) + \lambda(t)\mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)R(-\Delta_1)\mathbf{q}_{[\mathbf{M}]}\\ &= \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B(0) + \lambda(t)\mathbf{q}_{[\mathbf{M}]}^\top R(-(\Delta_1+\Delta_2))\mathbf{q}_{[\mathbf{M}]}.\end{aligned}$$

For any RoPE rotation $R(r)$,

$$\mathbf{q}_{[\mathbf{M}]}^\top R(-r)\mathbf{q}_{[\mathbf{M}]} = \left(\mathbf{q}_{[\mathbf{M}]}^\top R(-r)\mathbf{q}_{[\mathbf{M}]}\right)^\top = \mathbf{q}_{[\mathbf{M}]}^\top R(-r)^\top\mathbf{q}_{[\mathbf{M}]} = \mathbf{q}_{[\mathbf{M}]}^\top R(r)\mathbf{q}_{[\mathbf{M}]}.$$

Thus

$$\mathbf{q}_{[\mathbf{M}]}^\top R(-(\Delta_1+\Delta_2))\mathbf{q}_{[\mathbf{M}]} = \mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1+\Delta_2)\mathbf{q}_{[\mathbf{M}]}.$$

By the RoPE geometric alignment condition,

$$\mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1+\Delta_2)\mathbf{q}_{[\mathbf{M}]} > 0.$$

Since $\lambda(t) \geq 0$, this gives

$$S_{\mathrm{rev}}(t) \geq \mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B(0).$$

Therefore

$$\alpha_B^{\mathrm{rev}}(t) = \sigma(S_{\mathrm{rev}}(t)/\sqrt{D}) \geq \sigma\left(\mathbf{q}_{[\mathbf{M}]}^\top R(-\Delta_2)\mathbf{k}_B(0)/\sqrt{D}\right) > 0.$$

For ALiBi,

$$S_{\mathrm{fwd}}(t) = \mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t) + b\Delta_1,$$

and

$$S_{\mathrm{rev}}(t) = \mathbf{q}_{[\mathbf{M}]}(t)^\top \mathbf{k}_B(t) - b\Delta_2.$$

Hence

$$S_{\mathrm{rev}}(t) = S_{\mathrm{fwd}}(t) - b(\Delta_1+\Delta_2).$$

Since $S_{\mathrm{fwd}}(t)$ is bounded below and $b(\Delta_1+\Delta_2)$ is finite, $S_{\mathrm{rev}}(t)$ is bounded below. Thus

$$\alpha_B^{\mathrm{rev}}(t) = \sigma(S_{\mathrm{rev}}(t)/\sqrt{D})$$

is bounded away from zero.

Finally,

$$\begin{aligned} p_{\mathrm{rev}}(A;t) &= \frac{\exp\left(\alpha_B^{\mathrm{rev}}(t) Y_{A,B}(t)\right)}{\sum_{C'\in\mathcal{V}} \exp\left(\alpha_B^{\mathrm{rev}}(t) Y_{C',B}(t)\right)} \\ &= \frac{1}{1+\sum_{C\neq A} \exp\left(-\alpha_B^{\mathrm{rev}}(t)\left(Y_{A,B}(t) - Y_{C,B}(t)\right)\right)}. \end{aligned}$$

Since $\alpha_B^{\mathrm{rev}}(t)$ is bounded away from zero and

$$Y_{A,B}(t) - Y_{C,B}(t) \to +\infty \quad (C \neq A),$$

each exponential term vanishes. Therefore,

$$p_{\mathrm{rev}}(A;t) \to 1.$$

This proves both propositions. □

Interpretation. These propositions isolate probability-level transfer in a single evidence token setting. Forward training is not guaranteed to raise the forward and reverse probabilities at the same rate. However, the result is still useful: once training makes $Y_{:,B}$ increasingly favor A, the reverse probability also converges as long as the reverse attention to B stays active. Thus the result identifies the minimal storage and attention conditions under which forward learning becomes reusable in the reverse prompt.

ALiBi and RoPE differ in how naturally they keep this reverse attention active. For ALiBi, the forward and reverse scores differ only by a finite additive bias, so active forward attention implies active reverse attention. For RoPE, this depends on the geometric factor $\mathbf{q}_{[\mathbf{M}]}^\top R(\Delta_1+\Delta_2)\mathbf{q}_{[\mathbf{M}]}$, which is positive in the trace-positive regimes studied in App. B.1 but not for every possible query vector. Thus the RoPE statement is weaker: it captures the favorable routing geometry that typically supports transfer, rather than an unconditional guarantee.

Role of additional context tokens. The clean convergence result relies on the single evidence token setting. If additional context tokens are present, they can also receive attention, contribute to the logits, and interact with the routing dynamics. These terms may dilute the direct transfer path through $Y_{:,B}$, so the reverse probability need not saturate as cleanly as in the minimal proof. The value of the proposition is therefore mechanistic: it identifies a concrete probability-level route from forward learning to reverse prediction, while showing that the strength of this route depends on positional attention geometry.

Table 6. Hyperparameters for the one-layer experiments.

Hyperparameter	Value
Batch size	256
Learning rate	3×10^{-4}
Training steps	3,000
Gradient clipping	1.0
Weight decay	0.0
Dropout	0.02
Warmup steps	1,000
Hidden dimension	256
Attention heads	1
EMA decay for RADD	0.9999

D. Experimental Details and Additional Results

Compute resources. All experiments were conducted on a single NVIDIA L40S GPU with 48GB of GPU memory. The one-layer RADD experiments, GPT-2-style causal Transformer baselines, and positional-encoding interventions each used a single L40S GPU and took less than one hour per run. All large-scale LoRA fine-tuning experiments, including both MDMs and ARM baselines, also used a single L40S 48GB GPU per run. Each MDM fine-tuning run took approximately 7–8 hours, with Dream runs taking about 7.5 hours, while ARM fine-tuning runs were of comparable or shorter duration. Across all reported experiments, including repeated random seeds, the total compute was approximately 500 GPU-hours. Additional pilot and debugging runs used the same hardware and did not require substantially more compute than the reported experiments.

D.1. Details for One-Layer Validation

Data construction. For each sequence length L, we construct synthetic sequences containing one lowercase–uppercase pair and $L - 2$ placeholder tokens. The lowercase token is the target token, and the uppercase token is the evidence token. Training includes only forward-order sequences, where the lowercase target precedes the uppercase evidence token. All reverse-order sequences are excluded from training. At evaluation time, we query both forward and reverse configurations from the same pair. This construction isolates whether evidence learned from the forward order can be reused when the input and target roles are positionally reversed.

Models and hyperparameters. We train a one-layer RADD model as the MDM and a one-layer GPT-2-style causal Transformer as the ARM baseline. Unless otherwise stated, both models use hidden dimension 256, one attention head, batch size 256, learning rate 3×10^{-4}, dropout 0.02, gradient clipping 1.0, and $3{,}000$ training steps. The RADD model additionally uses EMA with decay 0.9999.

Measurements. We measure three quantities corresponding to the theory. First, we compute the correlation between forward and reverse attention scores assigned to the shared evidence token. Second, we compute the full-parameter gradient cosine similarity between paired forward and reverse losses. Third, we track the probability assigned to the correct target token under forward and reverse queries. For the positional encoding intervention, we keep the data, model size, and training protocol fixed while changing only the positional encoding among learned absolute embeddings, RoPE, and ALiBi.

Table 7. Large-scale reversal evaluation on Parent–Child, Person–Description, and T-REx. Results are exact-match accuracy (%) averaged across three random seeds. Fwd. denotes the trained forward direction, and Rev. denotes the held-out reverse direction. MDMs retain substantially stronger reverse accuracy, whereas ARMs, including the reasoning-oriented Qwen-3.5 model, largely fail under order reversal.

	MDM				ARM					
	LLaDA 8B		Dream 7B		LLaMA-3.1 8B		Qwen-2.5 7B		Qwen-3.5 4B	
Train Dataset	Fwd.	Rev.	Fwd.	Rev.	Fwd.	Rev.	Fwd.	Rev.	Fwd.	Rev.
Parent → Child (P2C)	76.7	48.3	62.0	50.0	89.9	15.9	89.9	0.5	99.0	3.0
Child → Parent (C2P)	87.7	43.7	83.0	54.0	95.9	6.9	89.0	1.4	100.0	1.0
Person → Description (P2D)	72.7	99.5	54.3	99.8	72.7	3.5	70.7	2.2	81.3	0.0
Description → Person (D2P)	99.7	41.3	100.0	27.2	83.0	1.8	80.0	1.5	100.0	0.0
T-REx	92.3	81.5	83.0	35.0	87.3	2.8	89.8	2.3	100.0	0.0

D.2. Details for Large-Scale Validation

Datasets and evaluation protocol. We evaluate on three relation-reversal benchmarks: Parent–Child, Person–Description, and T-REx. Parent–Child contains child–parent pairs annotated with parent type, Person–Description contains entity–description pairs, and T-REx contains entity–relation–description triples. For each dataset, models are fine-tuned only on the forward direction and evaluated on both the trained forward direction and the held-out reverse direction. For example, a Parent-to-Child training instance uses a parent entity as input and a child entity as target, while the reverse evaluation asks for the parent entity from the child entity. We report exact-match accuracy after minimal normalization, including lowercasing and whitespace stripping. For Parent–Child, either parent is accepted as correct when both father and mother are valid.

Models. We compare two masked diffusion language models and three autoregressive language models. For MDMs, we use LLaDA-8B and Dream-7B. For ARMs, we use LLaMA-3.1-8B, Qwen-2.5-7B, and Qwen-3.5-4B. The exact model identifiers are:

LLaDA-8B: `GSAI-ML/LLaDA-8B-Instruct`,
Dream-7B: `Dream-org/Dream-v0-Instruct-7B`,
LLaMA-3.1-8B: `meta-llama/Meta-Llama-3.1-8B-Instruct`,
Qwen-2.5-7B: `Qwen/Qwen2.5-7B-Instruct`,
Qwen-3.5-4B: `Qwen/Qwen3.5-4B`.

Dream-7B is added to test whether the MDM reversal advantage persists beyond LLaDA. Qwen-3.5-4B is added to test whether a reasoning-oriented ARM overcomes the same reversal failure.

Fine-tuning details. All models are fine-tuned with LoRA adapters. Unless otherwise stated, we use rank $r = 32$, LoRA scaling $\alpha = 64$, AdamW optimizer, weight decay 0.1, batch size 8, and train for 150 epochs. Each experiment is repeated with three random seeds, $\{1, 42, 1234\}$. Evaluation uses greedy decoding with temperature $T = 0$ and maximum generation length 32. Learning rates are selected per model and dataset:

LLaDA-8B: 5×10^{-5} for Parent–Child and Person–Description, 2×10^{-4} for T-REx,
Dream-7B: 5×10^{-5} for Parent–Child and Person–Description, 2×10^{-5} for T-REx,
LLaMA-3.1-8B: 5×10^{-5} for Parent–Child and Person–Description, 9×10^{-5} for T-REx,
Qwen-2.5-7B: 1×10^{-4} for Parent–Child and Person–Description, 5×10^{-5} for T-REx,
Qwen-3.5-4B: 5×10^{-5} for Parent–Child and Person–Description, 5×10^{-5} for T-REx.

Forward and reverse accuracy. Tab. 7 reports large-scale exact-match accuracy averaged over three random seeds. Across all datasets, ARMs achieve high forward accuracy but almost collapse in the reverse direction. This pattern holds not only for LLaMA-3.1 and Qwen-2.5, but also for Qwen-3.5, suggesting that reasoning-oriented autoregressive models do not

automatically escape the directional limitation induced by left-to-right training. In contrast, LLaDA retains substantially stronger reverse accuracy across all datasets. Dream also shows clear reverse-transfer behavior compared with ARMs, although its performance is less uniform than LLaDA. One plausible reason is that Dream is initialized from an autoregressive model, so part of the left-to-right inductive bias may remain after conversion and fine-tuning. Overall, the results support the main empirical claim: under the same forward-only training protocol, MDMs show substantially stronger reverse inference than ARMs.

Mechanistic probes in large models. For LLaDA and Dream, we additionally measure the three signatures predicted by the theory. First, we compute the correlation between forward and reverse attention scores assigned to the shared evidence entity. Second, we compute the cosine similarity between forward and reverse gradients over trainable parameters. Third, we track the probability assigned to the correct reverse entity during forward-only fine-tuning. These probes test whether the storage-routing signatures observed in the one-layer model persist in practical multi-layer MDMs.

In LLaDA, the main text shows that forward and reverse attention scores remain positively correlated, forward and reverse gradients remain positively aligned, and the reverse-entity probability increases during forward-only fine-tuning. Fig. 6 reports the corresponding measurements for Dream. Dream shows the same qualitative pattern: forward/reverse attention scores are positively correlated, gradient alignment remains positive, and the reverse probability increases during training. The signatures are somewhat weaker and less uniform than in LLaDA, which is consistent with Dream's autoregressive initialization: although Dream is trained and used as a masked diffusion model, its initialization may retain a residual left-to-right bias. Nevertheless, the Dream results support the same conclusion as the LLaDA experiments: the storage-routing signatures are not artifacts of the one-layer model and persist in large masked diffusion LLMs.

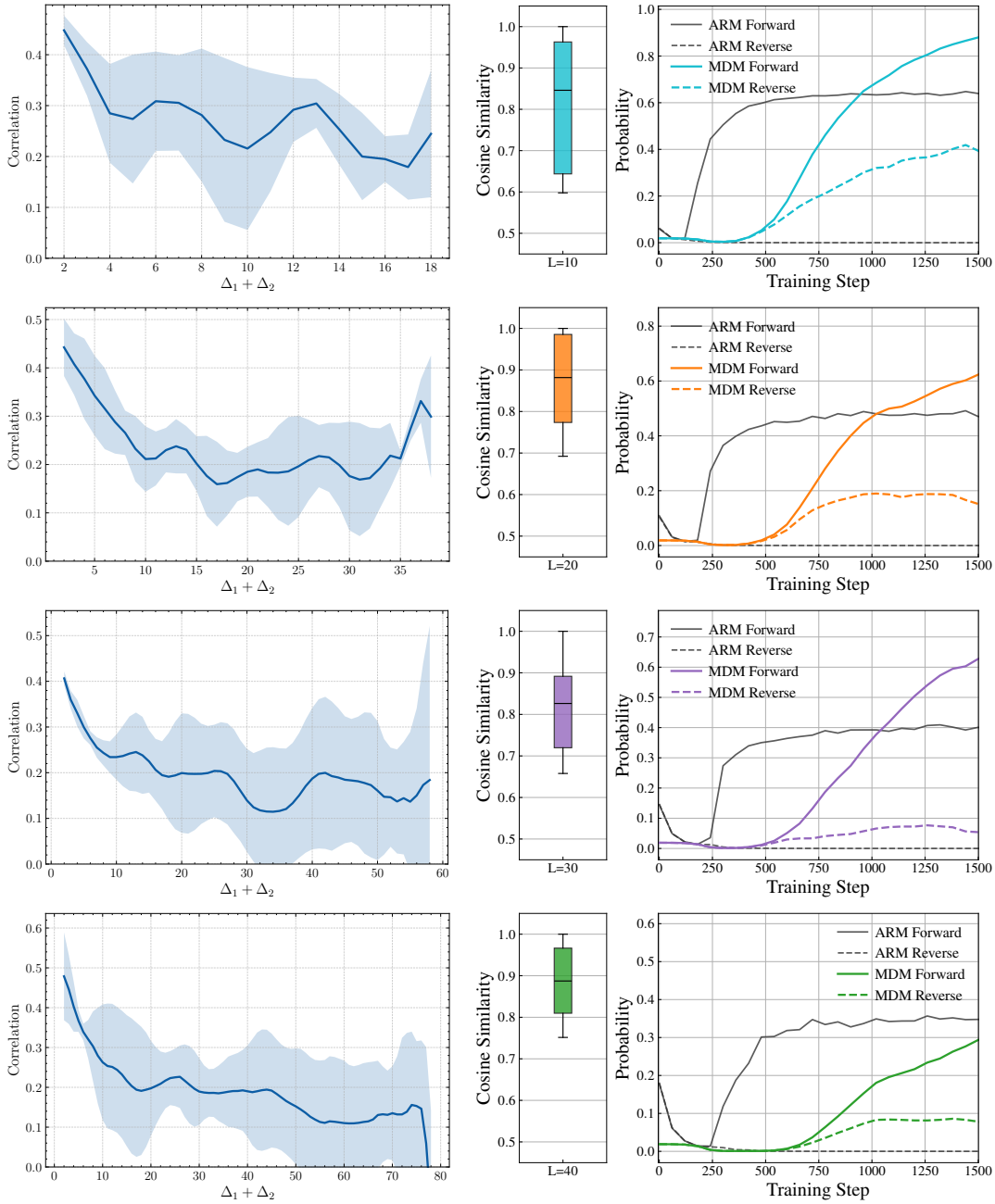
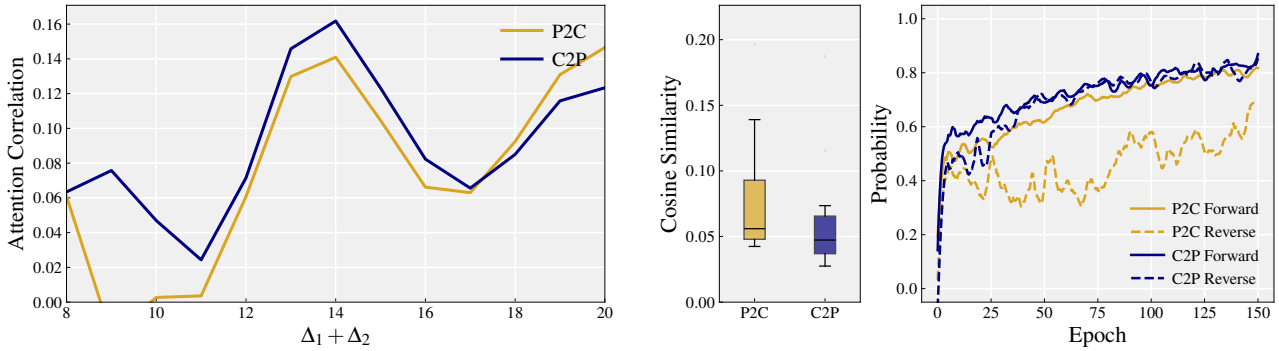


Figure 5. Additional one-layer validation across sequence lengths $L = 10, 20, 30, 40$. **Left column:** correlation of forward and reverse attention scores as a function of total relative distance $\Delta_1 + \Delta_2$. **Right column:** gradient alignment and probability transfer during training. Across sequence lengths, forward and reverse attention scores remain positively correlated, forward and reverse gradients remain aligned, and the reverse target probability increases during forward-only training.



(a) Dream attention correlation.

(b) Dream gradient and probability transfer.

Figure 6. Large-scale mechanistic signatures in Dream. Forward and reverse attention scores remain positively correlated across relative-distance placements. Forward and reverse full-parameter gradients remain positively aligned. The reverse target probability rises during forward-only fine-tuning. The signatures are weaker than in LLaDA, which may reflect residual left-to-right bias inherited from autoregressive initialization, but they remain consistent with the storage-routing account.

E. Discussion and Limitations

E.1. The MDM Framing of Our Analysis

We clarify why the paper is framed around masked diffusion models, even though our analysis identifies a more specific set of ingredients behind reversal curse mitigation. Our claim is not that the word "diffusion" itself explains the effect. Rather, the relevant mechanism is the interaction of three ingredients: *any-order masked supervision, full attention, and position-invariant value vectors.*

This framing is natural because modern masked diffusion language models instantiate these ingredients in a scalable generative setting. Early discrete diffusion models for text were formulated through a forward noising process and a learned reverse denoising process (Austin et al., 2021; Lou et al., 2024). Recent masked diffusion language models, however, are implemented in a simpler time-free form (Sahoo et al., 2024; Ou et al., 2025; Nie et al., 2025b; Ye et al., 2025). They take partially or fully masked sequences as input, predict distributions over the original tokens at masked positions, and are trained with a weighted masked-token cross-entropy loss. In this operational sense, they are generative masked language models: they retain the masked-token prediction structure of BERT-style models (Devlin et al., 2019), but use it generatively by iteratively filling masked positions. While BERT-style MLMs may share this mechanism, they are not standalone generative LLMs; modern MDMs are the practical setting where masked reconstruction, full attention, and position-invariant value vectors are jointly realized.

Thus, the MDM framing is not an appeal to diffusion terminology. It identifies the practical model class in which the relevant ingredients appear together, while our analysis isolates which of those ingredients drive reversal transfer.

E.2. Clarification on Absolute and Relative Positional Encodings

We clarify that the absolute/relative distinction is formally about how position is represented, not about whether positional information enters the value vector. Absolute encodings use position-specific signals, whereas relative encodings represent pairwise offsets or distances. For our storage-routing mechanism, however, the relevant issue is whether the value-output path remains position-free. Widely used absolute encodings (Vaswani et al., 2017; Gehring et al., 2017) are added before the query, key, and value projections, making the value vector position-dependent. Modern relative encodings (Dai et al., 2019; Raffel et al., 2020; He et al., 2021) such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2022) instead inject position through the query-key attention score or attention bias, leaving the value vector position-free. Thus, although "relative" does not mean "position-invariant value vectors" by definition, the two distinctions largely coincide for the modern positional encodings.

E.3. The Reversal Curse Formulation

The reversal curse tests whether a model that has learned a relation in one order can use the same learned evidence under an order-reversed query. Following Berglund et al. (2024), we use "A is B" and "B is A" only as schematic notation, not as a restriction to the literal word "is" or to symmetric equality relations. More generally, training may present a fact of the form $A\,R\,B$, while evaluation asks for the corresponding inverse query $B\,R^{-1}\,A$. Thus, the issue is not whether the surface relation is symmetric, but whether evidence acquired from the forward statement is stored in a form usable by the reverse query.

E.4. Parametric Memory versus In-Context Evidence

The reversal curse should be distinguished from in-context learning. If the relevant premise or inverse examples are supplied at test time, a model may answer the reverse query by using information newly provided in the prompt. This is useful, but it is a different problem from the one studied here. We ask whether a relation learned during training, without being restated in the prompt, becomes accessible through the model's parameters under an order-reversed query. This parametric-memory setting is important because a model cannot always rely on having the relevant fact or inverse examples in context. The same distinction applies to chain-of-thought prompting and reasoning-oriented inference: such methods can help manipulate evidence that is already available, but they do not by themselves guarantee that forward-learned evidence is retrievable from parameters under a reverse query.

E.5. Scope of the Claims

What the theory guarantees. We clarify the level of guarantee provided by our theory. Secs. 4.1 and 4.2 identify the storage-routing mechanism by which forward and reverse configurations share evidence and maintain a correlated route to that evidence. These results explain why reverse transfer can occur, but they do not predict the final reverse accuracy. Sec. 4.3 gives a local first-order alignment result: along the shared storage column, a forward gradient step decreases the corresponding reverse loss to first order. This identifies the sign of the transfer effect in the isolated component, but does not characterize the full training trajectory, the exact magnitude of the effect, or the final probability assigned to the reverse answer.

In particular, our theory should not be read as claiming that the reverse probability must approach one. The routing analysis in Sec. 4.2 shows that forward and reverse attention routes are positively correlated, not identical. Thus, the reverse query can retrieve some of the same stored evidence, but generally through a different and weaker route. From this perspective, imperfect reversal is not a failure of the theory; it is the natural outcome of shared storage being accessed through non-identical routing. This is also consistent with our positional encoding intervention: different positional encodings preserve the same value-side storage mechanism but induce different routing strengths, leading to different degrees of reverse transfer (Tab. 2). The perfect-transfer result in App. C.3 should therefore be interpreted as an idealized single-informative-token case, where no competing context tokens interfere with retrieval.

What reversal transfer does not imply. A related clarification is that stronger reversal transfer should not be interpreted as stronger abstract logical inversion. Our analysis concerns how a model learns to predict a target token from visible evidence tokens, and whether the resulting evidence is stored and routed in a form that remains usable under an order-reversed query. In this sense, MDMs have an advantage over ARMs because masked reconstruction with full attention can create input-target supervision without the left-to-right directional restriction imposed by autoregressive training. This is different from claiming that MDMs learn logical inverse relations better in general. Establishing such a claim would require a separate analysis of reasoning behavior beyond the storage-routing mechanism studied here.

The same caution applies to positional encodings. Our ALiBi results show that stronger preservation of the content score can yield stronger reverse transfer in our reversal task. They do not imply that ALiBi is universally preferable to RoPE or other positional encodings. A positional encoding that produces high reverse transfer may also induce different tradeoffs on tasks requiring fine positional discrimination, long-context extrapolation, or order-sensitive reasoning. Thus, the positional encoding intervention should be read as evidence that routing geometry affects parametric reversibility, not as a general ranking of positional encodings.

E.6. The One-Layer Theory versus Deep Models

Our theory follows the standard practice of isolating Transformer training dynamics in a one-layer setting (Tian et al., 2023; Zhu et al., 2024; Huang et al., 2025b; Li et al., 2023; 2024; Huang et al., 2025a). The goal is not to model every component of a deep MDM, but to expose the minimal storage-routing mechanism. Thus, the analysis should be read as a mechanistic explanation rather than a complete multilayer proof. Deep MDMs contain residual paths, MLP blocks, layer composition, and head specialization, whose net effect may modulate, amplify, or dilute the isolated signal. We therefore use large-scale LLaDA and Dream experiments to test whether the predicted signatures, including attention correlation, gradient alignment, and coupled probability transfer, persist in practical models (Sec. 6). A precise theory of how these mechanisms compose across layers is left for future work.