Objective-Induced Conditional Mismatch in Sequence Diffusion Models

Anonymous Authors[1]

Abstract

The choice of training objective shapes the geometry of the learned representation space by determining which conditional dependencies are consistently reinforced during optimisation. We study how causal (AR), symmetric diffusion (MDLM), and position-biased corruption objectives reshape the learned residual geometry and the associated mechanistic circuits, using a combination of loss-curvature inspection and circuit-level probes.

We formalise a *future marginalisation barrier*: non-causal denoising objectives optimise lower-entropy future-conditioned distributions, while prefix-conditioned inference requires marginalising over latent futures. We find that this mismatch is associated with more isotropic residual geometry, weaker directional OV circuits, and reduced context compression.

We introduce a position-biased corruption prior that masks later positions more frequently, encouraging suffix prediction from cleaner prefixes while preserving the tractable tokenwise diffusion-ELBO. This partially restores directional representation structure and improves context-conditioned prediction in controlled diffusion LM settings.

Our results suggest that objective structure is an important determinant of representation geometry and of prefix-conditioned inference behaviour in sequence diffusion models.

1. Introduction

Structured diffusion language models (DLMs) are trained under one conditional family and queried under another. This mismatch is especially sharp for sequence DLMs (Austin et al., 2021; Sahoo et al., 2024; von Rütte et al., 2025): training optimises local denoising conditionals over partially corrupted sequences, while many downstream uses require causal prefix-conditioned prediction. The unknown suffix is then a structured latent variable, and recovering the desired conditional requires marginalising over future completions.

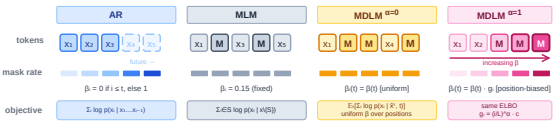


Figure 1. **Conditioning structures.** AR directly trains prefix-conditioned prediction. MLM and symmetric MDLM train bidirectional denoising conditionals. Position-biased MDLM ($\alpha>0$) masks later tokens more often, shifting training toward suffix-from-prefix denoising and partially restoring directional geometry (Section 5).

This paper studies that gap as both a conditional inference problem and a representation geometry problem. The two are linked: the corruption objective does not merely define a likelihood surrogate—it determines which conditional dependencies are consistently reinforced during optimisation, and thereby shapes the geometry of the learned residual representations. We ask: what uncertainty is introduced when future-conditioned denoising conditionals are used for prefix-conditioned inference? What geometric signature does this mismatch leave in the learned representations? And can a structured corruption prior reduce both?

Core thesis. *Non-causal denoising objectives bias optimisation toward more isotropic, rotation-tolerant representation geometry; causal objectives tend toward progressively anisotropic, directionally-specialised geometry. In our controlled settings, this geometry difference co-varies with conditional inference performance. We do not claim the geometric quantities are formally equivalent—only that they share a common objective-level origin.*

Figure 1 illustrates the conditioning structures across objective families.

Contributions.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. **Inference identity.** We formalise a future marginalisation entropy gap—a conditional mutual information identity—between future-conditioned denoising and prefix-conditioned prediction (Proposition 2.1).
2. **Representation geometry characterisation.** We show empirically that the barrier co-varies with a geometric signature: AR training favours directionally anisotropic representations, symmetric MDLM favours more isotropic/rotation-tolerant geometry, and position-biased MDLM partially restores anisotropy (Section 3).
3. **Empirical diagnostics.** One-step conditional-likelihood probes, denoising traces, and circuit decompositions across MDLM, GIDD, Dream, LLaDA, AR, and MLM (Section 4).
4. **Structured corruption prior.** Position-biased MDLM partially reduces the mismatch by shifting corruption toward later positions while preserving the tractable tokenwise diffusion-ELBO (Section 5).

Paper structure. Section 2 presents the conditional entropy gap. Section 3 characterises the objective-induced representation geometry. Section 4 reports learned-conditional diagnostics and circuit decompositions. Section 5 introduces the structured corruption prior and reports improvements. Appendices retain proofs, task details, geometry controls, and additional figures.

2. The Future Marginalisation Barrier

Causal vs. non-causal objectives. Autoregressive models parameterise $p_\theta(x_{1:L}) = \prod_{i=1}^{L} p_\theta(x_i \mid x_{<i})$, and at the population optimum satisfy $p_\theta(x_i \mid x_{<i}) = p^\star(x_i \mid x_{<i})$ for all i. DLMs learn denoising conditionals over corrupted sequences, which are non-causal. Using such a model for prefix-conditioned prediction requires marginalising over the unknown future:

$$p^\star(x_{k+1} \mid x_{\le k}) = \sum_{x_{k+2:L}} p^\star(x_{k+1} \mid x_{\le k}, x_{k+2:L}) \cdot p^\star(x_{k+2:L} \mid x_{\le k}). \quad (1)$$

This marginalisation is not free: each future contributes a lower-entropy conditional, and averaging over futures reinstates the uncertainty that conditioning on the future removed.

Proposition 2.1 (Future Marginalisation Barrier)**.** *Let $p^\star$ be a joint distribution over $X_{1:L}$, with $X_{k+2:L} \sim p^\star(\cdot \mid X_{\le k})$. Then*

$$H(X_{k+1} \mid X_{\le k}) - \mathbb{E}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})] = I(X_{k+1}; X_{k+2:L} \mid X_{\le k}) \ge 0,$$

with strict inequality iff X_{k+1} and the suffix carry conditional mutual information given the prefix.

We state this as the conditional-MI identity rather than an independent theorem (proof in Appendix D): the gap is the standard non-negativity of conditional mutual information, applied to the specific futures-as-latents view. The contribution is in naming the quantity and tying it to a concrete diagnostic. DLMs are trained on the lower-entropy future-conditioned distributions; recovering the causal marginal requires marginalising over latent futures that were never explicit training targets.

Context-compression diagnostic. We operationalise the barrier empirically via a one-step macro-ICL score: average NLL at position 500 minus position 50 in 512-token OpenWebText sequences. Negative values indicate that later tokens become easier to predict as context grows. AR is strongly negative; symmetric MDLM and GIDD remain positive, indicating that additional prefix context does not reduce predictive uncertainty under their learned conditionals.

3. Objective-Induced Representation Geometry

The marginalisation barrier is not only a conditional inference problem—it is an *optimisation landscape* problem. The corruption objective determines which conditional dependencies are consistently reinforced across all training steps, and this shapes the geometry of the learned residual representations in measurable ways.

Geometric signatures of objective family. We probe geometry along observables we can measure directly in our controlled settings (Figure 2):

- **Final-loss curvature**: top Hessian eigenvalue at convergence; sharper optima indicate a more directionally constrained, low-dimensional minimum.
- **Rotation invariance**: KL between outputs before and after a random orthogonal residual rotation; low KL means no privileged basis (isotropic), high KL means basis-privileged.
- **Embedding Hamming alignment**: in the TMS control, whether embedding distances track feature-set overlap.
- **OV circuit gap and context-compression slope**: downstream consequences (Section 4) that should co-vary with the above if the geometry story is right.

The geometry gradient across objectives. We find a consistent pattern across our Toy Models of Superposition (TMS) controls and matched small-model objective comparisons:

AR → directional, anisotropic manifold. Causal training reinforces forward temporal dependencies at every posi-

tion and layer. The gradient signal at position i always comes from prefix positions $j < i$; no future context is available. In our controlled settings, this tends to develop a *privileged temporal direction* in the residual space: the dominant eigenvector of the residual covariance encodes future-predictive information, OV circuits develop high rank-1 fractions aligned with the unembedding direction, and attention-routing specialises along causal offsets. Relative-position attention entropy is low in our small-model suite (AR achieves 2.94 nats on ICL recall vs. 3.97 for MDLM), indicating more concentrated, directional routing.

Symmetric MDLM (MDLM-F, α=0) → more isotropic, rotation-tolerant geometry. Bidirectional denoising reinforces symmetric conditional dependencies: each position can attend to all others. The gradient does not consistently privilege any temporal direction, and the residual space tends not to develop a dominant eigenvector aligned with the causal direction in our experiments. Feature directions are distributed across many dimensions (higher PR), and representations show greater rotation tolerance (downstream logits change less under random residual rotations). These are not representational failures: the TMS controls show that symmetric MDLM learns well-conditioned, high-decodability feature geometry—it tends to lack the *directionality* that causal objectives impose.

Position-biased MDLM → partial directional recovery. Shifting corruption toward later positions restores partial directional gradient signal: the suffix (where loss is computed) must now be predicted from a cleaner prefix (which provides one-directional evidence). This partially reinstates the temporal asymmetry of the gradient field. The recovery is metric-dependent: α=1 does *not* reintroduce basis-privilege (rotation invariance is essentially unchanged from α=0; Figure 2B), but it does improve Hamming alignment at high sparsity (Figure 2C), feature decodability (Appendix N), and the OV gap and context-compression slope (Section 4)—while preserving the tractable symmetric ELBO.

Why geometry matters for conditional inference. An isotropic residual space cannot easily develop the low-rank, directionally-aligned circuits that support induction-head-like retrieval. The OV-circuit evidence in Section 4 supports this: OV gap co-varies with objective directionality while QK routing does not—value projection geometry, not attention routing, is the bottleneck under non-causal objectives.

4. Empirical Diagnostics

One-step conditional-likelihood probe. We isolate what each objective amortises, independent of iterative sampling. For AR: teacher-forced next-token logit. For MDLM: keep prefix clean, mask only the target, score the masked position in a single forward pass. This asks what conditional the model has learned, not how hard the sampler works at inference. The full protocol, t-conditioning sweep, and mask-distractor probe (following Piskorz et al. 2025) are in Appendices B and L; the synthetic construction motivating the marginalisation gap is in Appendix E.

Table 1. **Circuit gaps on OverrideKV (haystack).** OV gap tracks objective directional geometry; QK gap does not.

Model	Peak QK gap	Peak OV gap
AR small	0.037	41.4
LLaDA-8B	0.028	33.2
Dream-7B	0.025	2.4
MDLM-F α=1 small	0.004	3.0
MDLM-F α=0 small	0.003	1.9

Figure 3 shows the macro context-compression scores. The pattern is consistent with the geometric prediction: more directional training yields more negative scores. Position-biased MDLM (α=1) recovers to the range of Dream-7B despite being $\sim$40$\times$ smaller, supporting the claim that the geometry fix is training-time rather than capacity-driven.

Geometric circuit diagnostics. Table 1 summarises peak QK recency gap and OV amplification gap from the OverrideKV circuit decomposition (full results in Appendix I). The OV gap—measuring whether the attended value is projected into the output with appropriate amplification—tracks the geometric anisotropy story: AR has the largest OV gap (41.4 on haystack), LLaDA-8B the next largest (33.2), while symmetric small MDLM has weak OV (1.9). The macro-ICL ranking mirrors the OV gap rather than the QK gap, suggesting the bottleneck is *value projection geometry* rather than attention routing. *Caveat.* The large-model comparison is between off-the-shelf instances with different training data and initialisations (Dream-7B inits from Qwen2.5-7B, whose OV weights are already pre-optimised for causal projection).

Residual geometry probes. In the matched small-model suite (AR, MLM, faithful MDLM-F with $\alpha \in \{0, 1\}$), the next-vs-previous residual probe shows higher asymmetry under AR (privileged temporal direction) and the layer-1 QK weight norms are larger; the attention-entropy comparison reported in Section 3 corroborates this. Removing learned positional embeddings restores AR length extrapolation but does not rescue MLM/MDLM (Appendix G): the representation geometry difference is objective-induced, not position-encoding-induced.

Uncertainty amplification under iterative denoising. The probes above evaluate the one-step learned conditional; iterative denoising—the native diffusion-LM infer-

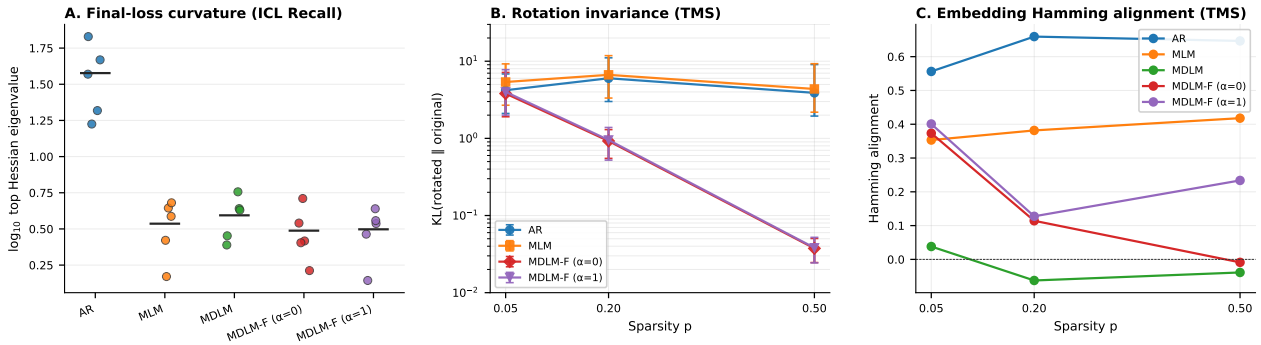


Figure 2. **Objective-induced representation geometry.** *A. Final-loss curvature (ICL Recall).* Top Hessian eigenvalue at end of training, per seed, with seed-mean bar. AR converges to a markedly sharper optimum (typically $\sim$10$\times$ larger top eigenvalue) than MLM and MDLM variants, consistent with a directionally-constrained, anisotropic minimum. *B. Rotation invariance (TMS).* KL divergence between the model's output distribution before and after a random orthogonal rotation of the residual stream, vs. feature sparsity p. Low KL means outputs do not depend on a privileged residual basis. AR and MLM remain basis-privileged across sparsities (KL $\sim$4–7 nats); symmetric and position-biased MDLM variants become rotation-tolerant for $p \geq 0.2$ (KL drops to $\sim 10^{-2}$–10^{0}). The empirical finding is non-trivial: bidirectional gradients could in principle pick out a privileged basis, but in our settings they do not. *C. Embedding Hamming alignment (TMS).* Hamming alignment of learned embeddings vs. feature sparsity p. AR maintains high alignment across sparsities, MLM is intermediate, symmetric MDLM collapses to $\approx$0, and MDLM-F α=1 partially recovers above α=0 at low sparsity.

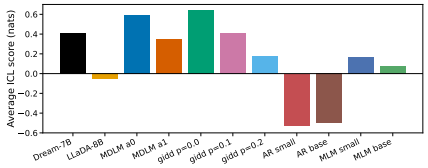


Figure 3. **Context-compression diagnostic.** Negative scores mean later tokens are predicted with lower uncertainty. Symmetric MDLM and GIDD remain positive; α=1 substantially recovers context compression. Co-varies with the anisotropy measures in Figure 2.

ence procedure—can amplify the barrier when conditioning on model-generated futures. For α=0, Binary FSM (K=16), clean-prefix decoding starts with gold probability 0.690 but degrades to 0.333 as generated suffix tokens accumulate; entropy rises from 0.66 to 3.14 nats. Free generation commits to a wrong hypothesis early (gold probability 0.00043 $\rightarrow$ 0.00023) while entropy falls. Full denoising-regime comparisons (free, teacher-forced, clean-prefix) and the high-cardinality amplification analysis are in Appendices H and M.

5. Position-Biased MDLM: Restoring Directional Geometry

Method. Standard MDLM applies position-independent masking probability $\beta(t)$. We make it position-dependent:

$$\beta_i(t) = \mathrm{clip}(\beta(t) \cdot g_i,\, 0,\, 1)\,, \quad g_i = \frac{(i/L)^\alpha}{\frac{1}{L}\sum_{j=1}^{L}(j/L)^\alpha}, \tag{2}$$

for $\alpha \geq 0$. At α=0 we recover vanilla MDLM; at α>0 later tokens are masked more often. This retains the same tractable tokenwise diffusion-ELBO form—no new architecture parameters, no inference-time changes. The intervention is an *objective-level geometry fix*: by shifting the gradient signal toward suffix positions that must be predicted from cleaner prefixes, we partially reinstate the directional gradient flow of causal training.

Geometry and context-compression recovery. As shown in Figures 3 and 2, α=1 substantially reduces the macro context-compression score and improves Hamming alignment at high sparsity (further TMS metrics including feature decodability and participation ratio in Appendix N). The improvement appears early in training and persists (Appendix J).

Few-shot generalisation. Figure 4 shows few-shot accuracy on Cipher (latent rule inference under a fresh random symbol permutation per episode). Position-biased MDLM (α=1) benefits more reliably from additional context than the symmetric baseline (α=0), consistent with its more directional representation geometry. The improvement is real but partial—and not uniform across tasks: at K=8 on one-step OverrideKV and Bigram Repeat, α=0 outperforms α=1 (Appendix M), so α=1 trades one-step task accuracy on some probes for better context compression and few-shot scaling on others. The causal AR objective remains the stronger inductive bias for directional ICL overall. The same pattern holds on Binary FSM; Key–Value and Override-KV results are in Appendix K.

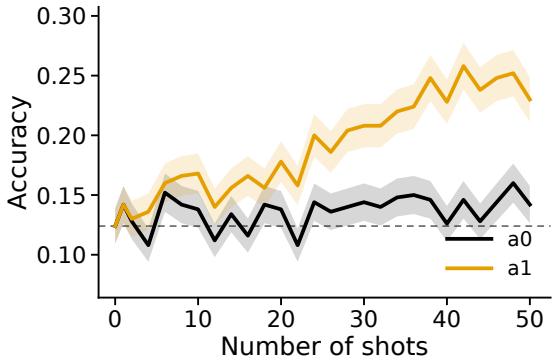


Figure 4. **Few-shot accuracy on Cipher** ($k \in \{0, \ldots, 50\}$ demonstrations, chance $1/8 = 0.125$, 500 episodes per point). α=1 benefits more reliably from additional demonstrations than α=0, consistent with partial restoration of directional representation geometry.

6. Related Work

Structured conditional inference. Our identity (Proposition 2.1) is closest in spirit to amortised-inference mismatch in latent-variable models: the model is trained on one conditional family and queried under another. Here the latent variable is the unobserved future suffix. The entropy gap quantifies how much uncertainty must be propagated when future-conditioned denoising conditionals are converted into prefix-conditioned predictions.

Representation geometry in neural networks. Representation anisotropy in language models has been documented empirically: Ethayarajh (2019) show that contextual embeddings concentrate most variance in few directions; Aghajanyan et al. (2021) measure the intrinsic dimensionality of fine-tuning objectives. Random-matrix theory characterises how feature covariance spectra evolve during training (Pennington & Wakhtel, 2018). Toy models of superposition (Elhage et al., 2022) show that objectives and capacity jointly determine feature coding; Papyan et al. (2020) study neural collapse in terminal representations. We extend these ideas to *temporal objectives*, showing that causal vs. non-causal training correlates with qualitatively different anisotropy regimes in controlled settings.

Discrete diffusion LMs. MDLM (Sahoo et al., 2024), GIDD (von Rütte et al., 2025), Dream (Ye et al., 2025), and LLaDA (Nie et al., 2025) are the primary models under study. Feng et al. (2025) study formal-language correctness in masked diffusion; Piskorz et al. (2025) study locality bias from suffix masks. Our contribution is to frame the conditional mismatch as a representation geometry question and to introduce a structured corruption prior that partially addresses it.

Evaluation methodology and adaptive inference. Diffusion models are often evaluated via iterative denoising chains. For conditional-inference questions this is necessary but not sufficient: a sampler may compensate for a poor learned conditional by spending more steps reconstructing context. We therefore separate one-step conditional likelihood (what the objective learned) from full denoising regimes (how much the inference chain can recover), following the same distinction as Piskorz et al. (2025).

7. Discussion

Conditional mismatch as uncertainty propagation. The converging evidence—entropy theorem, one-step diagnostics, full denoising traces, position-biased intervention, and geometry controls—points to a unified conclusion: symmetric denoising is a poor conditional match for prefix-conditioned language use. The training objective determines which conditionals are easy for the model to expose. If the objective does not directly support prefix-conditioned prediction, the sampler must spend test-time computation reconstructing or approximating the missing future marginalisation.

Diffusion models and directional structure. The failure is specific to directional/causal structure. TMS controls (Appendix N) show that diffusion objectives *can* learn coherent symmetric, rotation-invariant geometry comparable to AR—they are objective-level mismatched for directional tasks, not representationally limited. The geometry framing generalises: any objective that symmetrises the conditioning structure tends to produce representations less biased toward the relevant information direction. Position-biased corruption is one minimal training-time lever; learned schedules, response-only diffusion, prefix-clamping, and hybrid AR/diffusion objectives are natural extensions.

Limitations. Main MDLM experiments use ∼170M-parameter models and a small α sweep ($\{0, 1, 1.5\}$); large-model evidence for Dream/LLaDA is diagnostic rather than a controlled retraining study. TMS controls use sparse binary features; naturalistically structured features may differ.

References

Aghajanyan, A., Gupta, S., and Zettlemoyer, L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 7319–7328, 2021.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Oord, A. Structured denoising diffusion models in dis-

crete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 55–65, 2019.

Feng, G., Geng, Y., Guan, J., Wu, W., Wang, L., and He, D. Theoretical benefit and limitation of diffusion language model. *arXiv preprint arXiv:2502.09622*, 2025.

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 *Advances in Neural Information Processing Systems*, 2025.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Pennington, J. and Wakhtel, P. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1924–1932. PMLR, 2018.

Piskorz, J., Pinneri, C., Correia, A., Alfarra, M., Garrepalli, R., and Louizos, C. Masks can be distracting: On context comprehension in diffusion language models. *arXiv preprint arXiv:2511.21338*, 2025.

Sahoo, S. S., Arriola, M., Gokaslan, A., Marroquin, E. M., Rush, A. M., Schiff, Y., Chiu, J. T., and Kuleshov, V. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

von Rütte, D., Fluri, J., Ding, Y., Orvieto, A., Schölkopf, B., and Hofmann, T. Generalized interpolating discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025.

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.

A. Technical Appendices and Supplementary Material

The appendix is organised as follows. Appendix B justifies the one-step conditional-likelihood evaluation protocol. Appendix C gives architecture, training, and task-construction details. Appendix D contains full proofs. Appendix E describes the minimal synthetic example and Appendix F provides numerical verification. Appendix H extends the analysis to iterative denoising with model-generated futures. Appendix J reports eval-loss training curves. Appendix K reports additional few-shot tasks. Appendix G contains the matched small-model control results and mechanistic probes. Appendix L reports robustness checks. Appendix M reports full iterative denoising under three decoding regimes. Appendix I reports the OverrideKV QK/OV circuit decomposition. Appendix N reports TMS geometry controls.

B. Evaluation Probes and Scoring Protocols

We use several probes because the question has two distinct parts: what conditional distribution the trained model exposes, and what an iterative denoising sampler can recover at inference time. The primary results use *one-step conditional likelihood*; generation and circuit probes are supplementary diagnostics.

One-step conditional likelihood. For diffusion language models we use *one-step* conditional likelihood. Given a prefix-conditioned prediction problem, we keep the semantic prefix clean, replace only the target position with `[MASK]`, supply a scalar noise level t, and read the model's logits at the masked target in a single forward pass. This asks the same question as AR teacher-forced next-token scoring: with the demonstrations already available, how much probability does the model assign to the correct next answer? For AR we score the next-token logit under teacher forcing; for MLM and MDLM-style models we score the masked target position. This avoids any inference-time sampling procedure that could conflate *what is learned* with *how hard we work at inference*.

Macro-ICL score. The macro score is the average loss at a late position minus the average loss at an early position, using positions 500 and 50 in 512-token packed sequences. Negative values mean the model predicts later tokens better than earlier tokens, consistent with context compression. Positive values mean additional context does not reduce predictive uncertainty. This is a coarse two-point statistic; it is used because it is architecture-agnostic: AR, MLM, MDLM, GIDD, Dream, and LLaDA can all be scored with the same held-out tokens under their native one-step conditionals.

Few-shot one-step scoring. For few-shot tasks, each prompt contains K demonstrations and one query. The demonstrations are kept clean; the answer token for the query is masked for MLM/MDLM-style models and scored as the next token for AR. Accuracy is the fraction of episodes where the highest-scoring candidate label is the true answer. Each plotted point averages 500 independently generated episodes.

Noise-level and mask-suffix probes. The one-step MDLM probe requires supplying a scalar diffusion time t. We sweep $t \in \{0.01, 0.05, 0.1, 0.2\}$ while holding the visible mask pattern fixed, separating dependence on the explicit time input from dependence on the actual observed corruption pattern. We also append suffix `[MASK]` tokens after the query to test whether additional active denoising targets distract the model. These probes are reported in Appendix L: the key finding is that fixed-mask predictions are largely insensitive to t, and suffix masks have task-dependent rather than uniformly negative effects.

Circuit probes. The OverrideKV circuit analysis decomposes performance into QK routing and OV amplification. QK gaps ask whether attention routes to the recent/override binding rather than stale bindings. OV gaps ask whether the value/output path from the attended binding pushes the correct label logit. This distinction matters because a model can attend to the right location but fail to copy the right value into the output distribution.

Why not iterative denoising as the primary metric? One-step denoising evaluates a learned conditional directly with no feedback loop, no posterior sharpening over multiple steps, and no model-generated suffix. It therefore tests whether the objective has trained a prefix-conditioned predictor. Iterative denoising tests a different capability: whether an inference chain can reconstruct the useful prefix and avoid committing to wrong self-generated tokens. Under iterative denoising the marginalisation barrier can be *further amplified* (Appendix H).

C. Additional Experimental Details

Hyperparameters. All MDLM models use a 12-layer Transformer with hidden size 768, 12 attention heads, feedforward dimension 3072, conditional embedding dimension 128, dropout 0.0, max sequence length 512, and 169.7M total parameters (92.1M non-embedding). Sequences are tokenized using the GPT-2 tokenizer and packed to exactly 512 tokens with no padding.

Few-shot episode format. For each task we generate 500 evaluation episodes with $k \in \{0, 1, 2, 4, 6, 8, 10, 12, \ldots, 48, 50\}$. Each episode uses a fresh random seed and a pool of 60 demonstrations from which the first k are selected. The latent rule is resampled per episode; memorising a global mapping cannot solve the evaluation.

Key–Value retrieval. Keys are sampled as `wug`*NNNN* for *NNNN*$\in [0000, 9999]$. Each episode samples a fresh mapping from keys to one of 8 label tokens. The query repeats one of the demonstrated keys and asks for its label.

Cipher. Cipher uses a fresh random permutation per episode over the first 8 lowercase letters. The query asks for the output of a held-out or repeated input symbol under the same permutation.

Binary FSM. Binary FSM samples a two-state finite-state machine per episode with a random 2×2 transition table and random output flip parameter. Input strings are binary sequences of length 3–12; the query provides a new binary string and asks for its label.

OverrideKV and recency probes. OverrideKV separates retrieval from recency. In the *haystack* variant, the queried key is bound to a stale value three times early in the prompt and then to a recent override at a fixed intermediate position; later demonstrations use unrelated keys. In the *recency-last* variant, every demonstration uses the same key, with $K-1$ stale values followed by one recent value immediately before the query.

Matched toy-objective suite. The small AR/MLM/MDLM comparison uses the same transformer, tokenizer, data distribution, and context length for all objectives. Tasks: associative-recall ICL, induction (copy token following repeated prefix), directional prediction (left-to-right vs. right-to-left), and MQAR (content-addressed lookup).

Table 2. **Evaluation modes used in this paper.** One-step probes learned conditionals directly; generation probes whether iterative denoising reconstructs and preserves the useful prefix.

Mode	Prefix handling	Suffix handling	Purpose
One-step	Clean demonstrations	Target masked only	Learned conditional / primary metric
Free generation	Generated by model	Generated by model	Native denoising without help
Teacher-forced generation	Corrected as revealed	Generated by model	Partial prefix recovery
Clean-prefix generation	Clamped clean throughout	Generated by model	Best-case causal prefix access

D. Theoretical Proofs

Proposition D.1 (Autoregressive optimality)**.** *At the population optimum, an autoregressive model satisfies* $p_\theta(x_i \mid x_{<i}) = p^\star(x_i \mid x_{<i})$ *for all* $i \in \{1, \ldots, L\}$.

Proof. $\mathcal{L}_{\mathrm{AR}}(\theta) = \sum_{i=1}^{L} \mathbb{E}_{x_{1:i}\sim p^\star}[-\log p_\theta(x_i \mid x_{<i})]$ is a sum of independent cross-entropies, each minimized pointwise by matching the true conditional. □

D.1. Proof of Proposition 2.1 (Extended)

Proof. The result follows from the definition of conditional mutual information and the strict concavity of Shannon entropy:

$$\begin{aligned} I(X_{k+1}; X_{k+2:L} \mid X_{\le k}) &= H(X_{k+1} \mid X_{\le k}) \\ &\quad - H(X_{k+1} \mid X_{\le k}, X_{k+2:L}) \\ &= H(X_{k+1} \mid X_{\le k}) - \mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})]\,. \end{aligned}$$

If this conditional mutual information is positive, the stated strict inequality follows immediately. Equivalently, the marginal conditional $p^\star(x_{k+1} \mid x_{\le k})$ is a mixture over future-conditioned conditionals $p^\star(x_{k+1} \mid x_{\le k}, f)$ with weights $p^\star(f \mid x_{\le k})$; the entropy gap follows from the concavity of Shannon entropy, with strictness exactly when the future carries conditional information about X_{k+1}. □

Remark D.2 (Mutual Information). Equivalently, $H(X_{k+1} \mid X_{\le k}) - \mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})] = I(X_{k+1}; X_{k+2:L} \mid X_{\le k})$.

D.2. Proof of Proposition D.3

Proposition D.3. *In the synthetic example of Appendix E, with non-identical* $\{p_h\}$, $H(X_{k+1} \mid X_{\le k}) > \mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})]$.

Proof. The prefix gives a non-degenerate posterior $p(h \mid x_{\le k})$, so $p^\star(x_{k+1} \mid x_{\le k}) = \sum_h p(h \mid x_{\le k}) p_h(x_{k+1})$. For any realised suffix, Bayes gives $p(h \mid x_{\le k}, x_{k+2:L}) \propto p(x_{k+2:L} \mid h) p(h \mid x_{\le k})$, which generally sharpens the posterior because suffix likelihoods differ across h. The future-conditioned distribution $\sum_h p(h \mid x_{\le k}, x_{k+2:L}) p_h(x_{k+1})$ therefore varies with the suffix. Marginalising yields a convex mixture of distinct conditionals; by Proposition 2.1, the marginal entropy strictly exceeds the expected conditional entropy. □

E. A Minimal Synthetic Example

We construct a data distribution where AR training exhibits context compression but joint objectives provably fail, even with perfect joint modelling. Let $h \in \{1, \dots, H\}$ be a latent hypothesis sampled from $\pi(h)$. A sequence $x_{1:L}$ is generated by sampling a prefix $x_{1:k}$ from a hypothesis-dependent distribution $p_{\text{pre}}(x_{1:k} \mid h)$ providing informative but non-degenerate evidence about h, then drawing each suffix token x_i $(i > k)$ i.i.d. from a hypothesis-specific p_h. We assume $\{p_h\}$ are not all identical.

For an AR model: $p^\star(x_{k+1} \mid x_{\le k}) = \sum_h p(h \mid x_{\le k}) p_h(x_{k+1})$, and as the prefix accumulates evidence and $p(h \mid x_{\le k})$ concentrates, $H(X_{k+1} \mid X_{\le k})$ decreases. AR training thus directly optimises prefix-conditioned prediction and enables context compression as latent uncertainty is reduced.

For a joint objective, prefix-conditioned prediction requires marginalising over suffix realisations (Proposition D.3): joint objectives entangle hypothesis uncertainty with suffix variation, so even though the prefix provides informative evidence about h, averaging over futures that convey *different* information dilutes the prefix signal. This is an objective-level bias, not a capacity issue.

F. Simulation Verification of the Future Marginalisation Barrier

To validate Proposition 2.1 numerically we instantiate the construction in Appendix E with $h \in \{1, 2, 3\}$ over a vocabulary of 10 token values, with h=1 favouring low values, h=2 middle, and h=3 high. Figure 5 shows: panel (a) the three hypothesis distributions $\{p_h\}$; panel (b) the prefix-derived posterior $p(h \mid x_{\le k})$, concentrated on h=2 but uncertain; panel (c) the causal conditional, diffuse with $H = 2.254$ nats; panels (d)–(i) future-conditioned distributions sharpened by Bayes' rule with $H \approx 1.92$–2.13 nats.

The expected future-conditioned entropy is ≈ 2.077 nats, giving an entropy gap $\Delta = 0.177$ nats and confirming Proposition 2.1. This gap is the *cost of marginalisation*: the causal conditional must average over all possible futures, each providing different information about the next token. Diffusion training observes futures and learns the sharp conditionals (panels d–i); at inference with unknown futures the model must implicitly marginalise to recover the diffuse conditional (panel c)—precisely where the barrier bites.

G. Matched Small-Model Control: Full Toy ICL Suite

Models. We train matched small transformers under AR, MLM, a simplified toy MDLM, and faithful MDLM-F variants with $\alpha \in \{0, 1\}$. All models use the same sequence format, vocabulary, context length, optimiser settings, and training examples; only the objective and corruption process change. MDLM-F is the repo-faithful implementation using `gidd.diffusion_process.MaskedDiffusion` and `gidd.loss.MDLMLoss`.

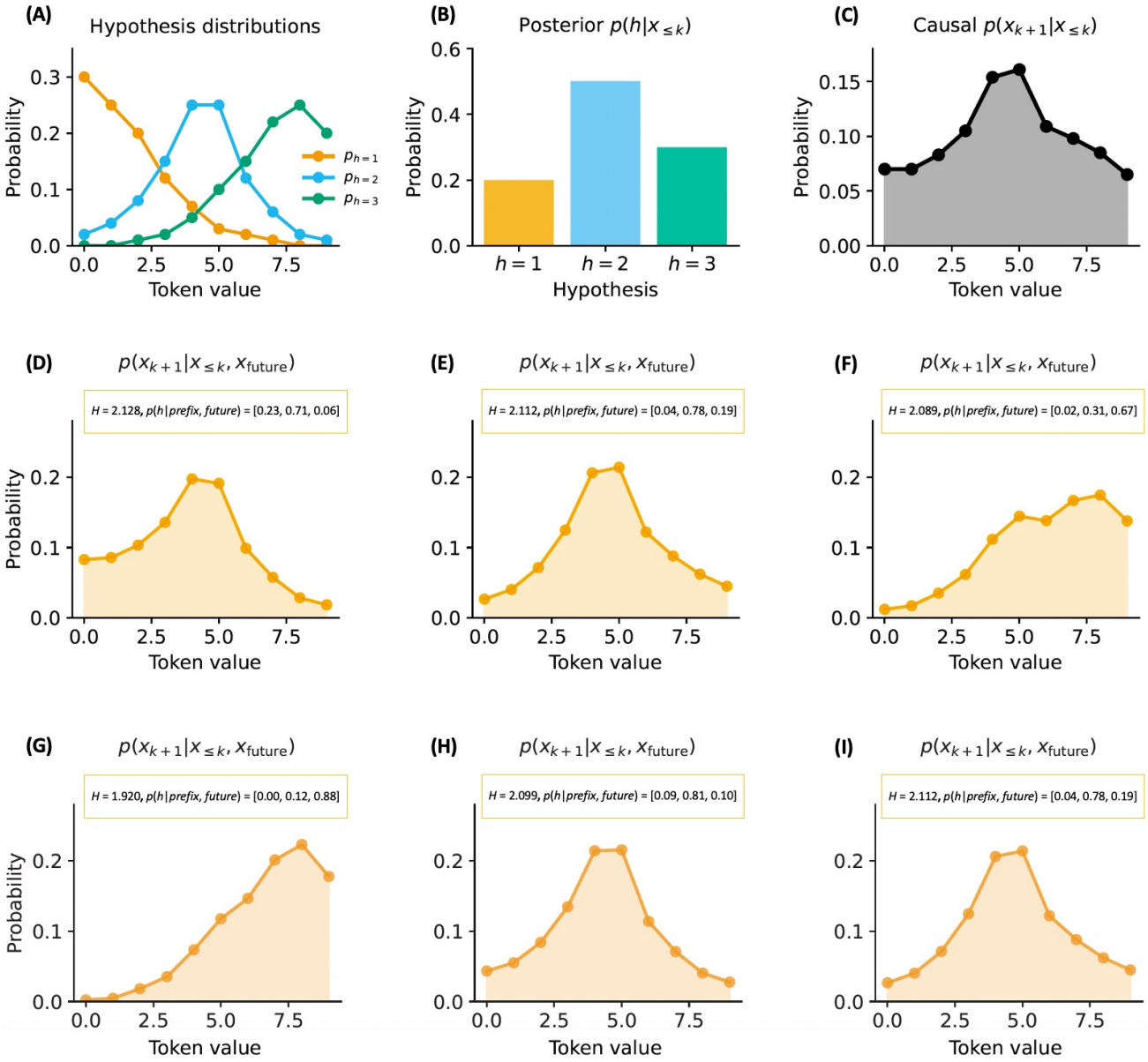


Figure 5. **Numerical verification of Proposition 2.1.** Panel (a): hypothesis distributions $\{p_h\}$. Panel (b): prefix posterior, concentrated on h=2. Panel (c): causal conditional, diffuse ($H = 2.254$ nats). Panels (d)–(i): future-conditioned distributions with $H \approx 1.92$–2.13 nats. Marginalisation gap $\Delta = 0.177$ nats.

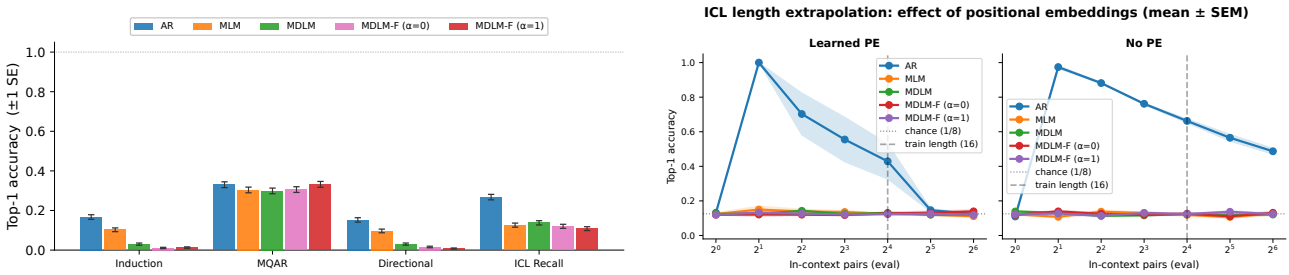
Table 3. **Objective variants in the matched toy suite.**

Variant	Loss / scoring objective	Time/noise handling	Context/corruption structure
AR	Causal next-token cross-entropy	None	Causal attention, prefix only
MLM	Masked-token cross-entropy	None	Bidirectional context, uniformly masked targets
MDLM	Simplified inline absorbing-mask ELBO	Sampled scalar schedule	Uniform position masking
MDLM-F α=0	Repo-faithful `MDLMLoss`	Faithful MDLM components	Uniform position masking
MDLM-F α=1	Repo-faithful `MDLMLoss`	Faithful MDLM components	Later tokens masked more often

Toy ICL tasks. *Associative-recall ICL* samples fresh key–value mappings in context and queries a held-out key. *Induction* tests copying the token following a repeated prefix. *Directional prediction* compares left-to-right vs. right-to-left token prediction. *MQAR* tests content-addressed retrieval and serves as a control where bidirectional objectives can succeed.

Result. Across all settings, only AR learns associative-recall ICL; MLM, MDLM, and MDLM-F variants remain at chance (Figure 6a). Removing learned positional embeddings rescues AR length extrapolation but does not improve MLM/MDLM. Because both simplified MDLM and faithful MDLM-F fail, the result is an objective-level failure, not an implementation artefact.

Mechanistic probes. We measure attention entropy, relative-position attention, the linear next-vs-previous residual probe, QK recency gaps, OV logit gaps, residual effective rank, Hessian curvature, and LMC barriers. The central pattern is that AR develops directional routing and residual asymmetry, while MLM/MDLM remain more symmetric and diffuse—these are the geometric signatures discussed in Section 3.



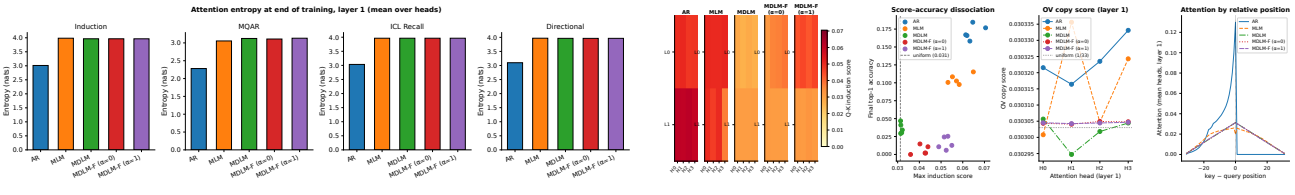
(a) Final accuracy across toy tasks.

(b) Associative-recall length extrapolation.

Figure 6. **Matched-objective behavioural controls.** AR outperforms bidirectional/diffusion objectives on induction, directional prediction, and associative-recall ICL, while MQAR remains largely objective-agnostic. Removing positional embeddings improves AR extrapolation but does not rescue MLM/MDLM.

Table 4. **Layer-1 relative-position attention entropy (nats).** Lower entropy means attention concentrates on fewer relative offsets.

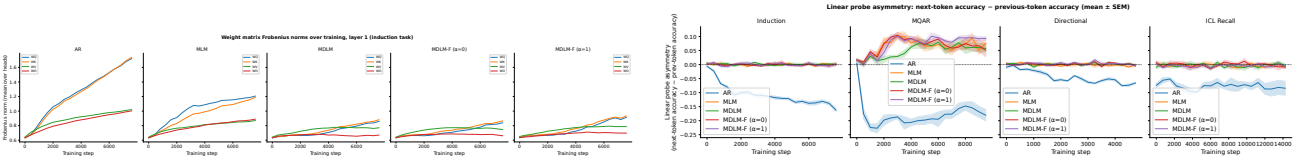
Task	AR	MLM	MDLM	MDLM-F α=0	MDLM-F α=1
Induction	3.00	4.00	3.97	3.97	3.96
MQAR	2.28	3.07	3.13	3.11	3.08
Directional	3.13	3.97	3.97	3.97	3.97
ICL recall	2.94	3.96	3.97	3.97	3.97



(a) Attention entropy.

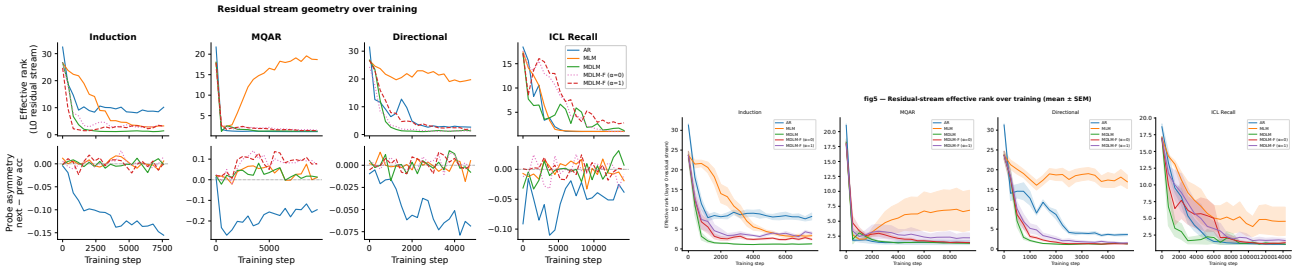
(b) Induction routing diagnostics.

Figure 7. **Attention-routing diagnostics.** AR develops lower-entropy, position-specialised attention; MDLM variants remain diffuse. Induction diagnostics show a partial QK induction signal and sharper AR relative-position routing.



(a) Attention weight norms. *(b)* Next-vs-previous linear probe.

Figure 8. **Training dynamics and residual directionality.** On induction, AR grows substantially larger layer-1 QK norms than MLM/MDLM. The residual probe shows that AR develops a nonzero temporal orientation, while bidirectional/diffusion objectives remain approximately symmetric.



(a) Residual rank and probe asymmetry. *(b)* Residual effective-rank spectrum.

Figure 9. **Residual geometry.** The cleaner signal is the probe asymmetry: AR develops a privileged temporal direction and MLM/MDLM remain close to symmetric except on MQAR.

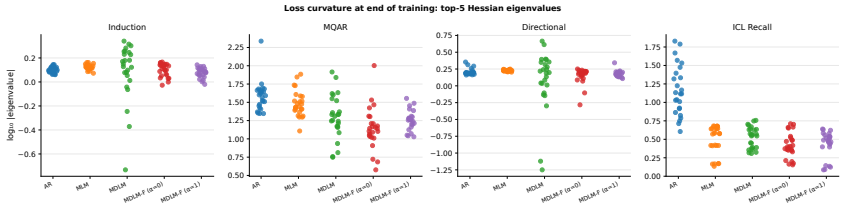


Figure 10. **Final-loss curvature.** AR is sharper on ICL Recall and MQAR; MDLM-F variants are generally flatter, consistent with a less constrained, higher-dimensional optimum.

H. Iterative Denoising Amplification

Appendix F illustrated the marginalisation barrier under prefix-only prediction. We now show how, for diffusion-style inference, *iterative denoising can further exacerbate this gap* when the model conditions on *model-generated* (rather than ground-truth) futures.

Setting. We reuse Appendix E: $h \in \{1, 2, 3\}$, prefix $x_{\leq k}$, suffix tokens drawn from p_h. The future-conditioned distributions $p^\star(x_{k+1} \mid x_{\leq k}, x_{k+2:L})$ vary with the realised future, so marginalising over futures increases uncertainty (Proposition 2.1).

Diffusion perspective. A discrete diffusion model defines $p_\theta(x^{t-1}_{1:L} \mid x^t_{1:L})$ and iteratively denoises $x^T_{1:L} \to \cdots \to x^0_{1:L}$. With prefix coordinates clamped to ground truth, the future coordinates are corrupted and progressively filled in by the model. At any intermediate step the model conditions on a sampled future $\hat{x}_{k+2:L}$ produced by its own dynamics. If $\hat{x}_{k+2:L}$ is systematically biased relative to $p^\star(\cdot \mid x_{\leq k})$, the induced conditional at $k{+}1$ can be *worse* than the marginalised baseline.

Reference quantities. Fix a prefix $x_{\leq k}$. (i) Ground-truth future conditioning gives the training-like quantity $\mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\leq k}, X_{k+2:L})] \approx 1.87$ nats in our toy instantiation. (ii) The marginalised causal conditional gives $H(X_{k+1} \mid X_{\leq k}) \approx 2.19$ nats, larger by Proposition 2.1. Iterative denoising replaces $X_{k+2:L} \sim p^\star(\cdot \mid x_{\leq k})$ with $\hat{X}_{k+2:L}$ from the model's own induced distribution.

Three error modes. (A) *Miscalibration:* the model's posterior assigns mass to a wrong h, producing sharp but wrong predictions—"confident incorrectness". (B) *Uninformative futures:* $\hat{x}_{k+2:L}$ carries little information about h, so the induced conditional is close to the marginalised baseline—"no benefit from iterative conditioning". (C) *Systematic bias:* denoising favours futures consistent with a particular h regardless of evidence, frequently conditioning on highly mismatched futures and yielding large cross-entropies. In our toy setup, always conditioning as if $h{=}1$ gives $\mathbb{E}_h[\mathrm{CE}(p_h, p_1)] \approx 2.84$ nats, well above the marginalised baseline. This is the *amplification effect*: model-generated futures can make prefix-conditioned prediction strictly worse than the intrinsic marginalisation barrier.

High-cardinality settings are more brittle. The 3-hypothesis case understates what can happen in few-shot settings, where many plausible latent "tasks" may have nearly disjoint label distributions. With $H{=}1000$, $V{=}10000$, true posterior 0.90 on $h^\star$, and model sampling $h^\star$ with probability 0.85: wrong hypotheses placing $\sim 1/V$ on the correct token give $\mathbb{E}[\mathrm{NLL}] \approx 0.85 \cdot 2.08 + 0.15 \cdot 9.21 \approx 3.15$ nats. Small calibration errors in the induced future distribution thus have outsized impact when the hypothesis space is large and wrong hypotheses are strongly incompatible with the correct label.

Bias parameterisation. For $p_\theta(h \mid x_{\leq k}) = (1 - \beta)p^\star(h \mid x_{\leq k}) + \beta[1, 0, 0]$, expected cross-entropy increases monotonically with β (Figure 12); values above the marginalised baseline correspond to amplification beyond Proposition 2.1.

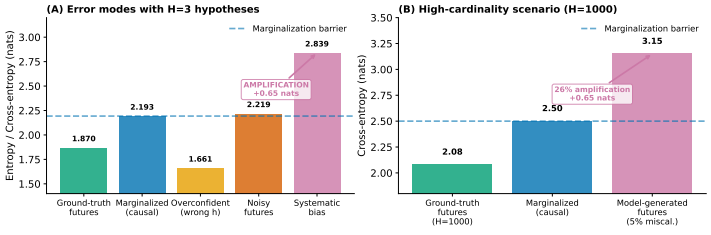


Figure 11. **Iterative denoising amplification.** Ground-truth future conditioning (green) is the training-like target; the marginalised causal conditional (blue dashed) is the Proposition 2.1 barrier. Scenario A (orange): miscalibration. Scenario B (red): uninformative futures. Scenario C (purple): systematic bias, exceeding the marginalisation baseline. The mechanism becomes more severe in high-cardinality settings.

Relation to AR. Under teacher-forced evaluation, AR conditions on the ground-truth prefix at every position, so there is no analogous dependence on model-generated futures. Diffusion's prefix-conditioned evaluation naturally involves unknown futures; iterative denoising replaces these with model-imputed futures, introducing additional distribution shift beyond the marginalisation barrier.

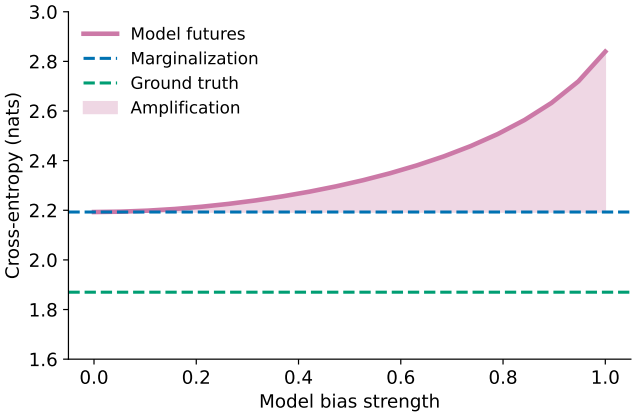


Figure 12. **Cross-entropy with increasing bias.** Expected cross-entropy increases monotonically with bias strength β; the dashed line is the marginalised baseline $H(X_{k+1} \mid X_{\leq k})$.

I. OverrideKV Circuit Decomposition

Tasks. Haystack: the queried key is bound to a stale value early and to a recent override at a fixed intermediate position; later demonstrations use unrelated keys. Recency-last: all demonstrations use the same key, $K-1$ stale values followed by one recent value immediately before the query.

Key finding. The macro-ICL ranking mirrors the OV gap (value projection strength) rather than the QK gap (attention routing). Dream-7B has QK routing comparable to LLaDA-8B but OV gap $\sim 15\times$ smaller, and correspondingly lower accuracy. AR has the strongest QK+OV combination. This supports the geometric story: OV gap measures how much the value circuit has specialised its projection direction toward the unembedding (high rank-1 OV anisotropy), while QK gap measures attention routing separately. Directional representation geometry determines the former; the latter can be learned without it.

Table 5. **Peak QK and OV recency gaps (full OverrideKV).**

Task	Model	Peak QK gap	Peak OV gap	Interpretation
Haystack	AR small	0.037	41.4	strong QK + OV
Haystack	Dream-7B	0.025	2.4	strong QK, weak OV
Haystack	LLaDA-8B	0.028	33.2	strong QK + OV
Haystack	MDLM α=1	0.004	3.0	weak QK, mod. OV
Haystack	MDLM α=0	0.003	1.9	weak QK, weak OV
Recency-last	AR small	0.076	88.7	strongest circuit
Recency-last	Dream-7B	0.056	3.3	strong QK, weak OV
Recency-last	LLaDA-8B	0.051	49.9	strong QK + OV
Recency-last	MDLM α=1	0.010	4.8	weak QK, best MDLM OV
Recency-last	MDLM α=0	0.008	3.7	weak QK, mod. OV

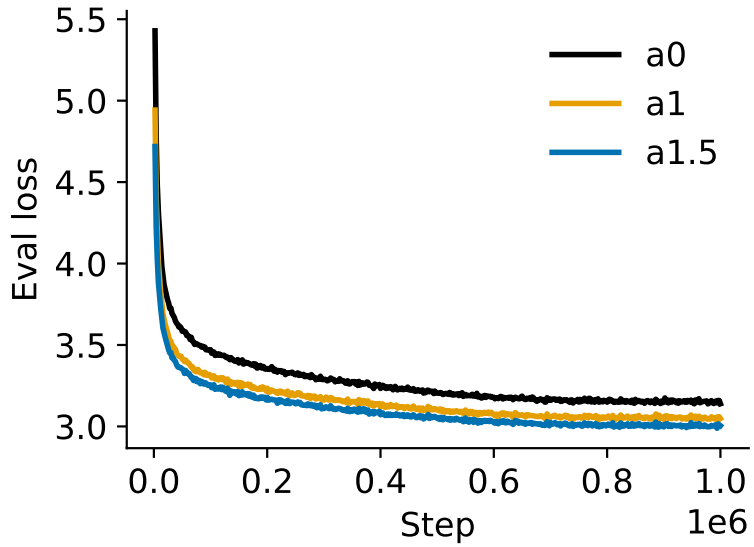
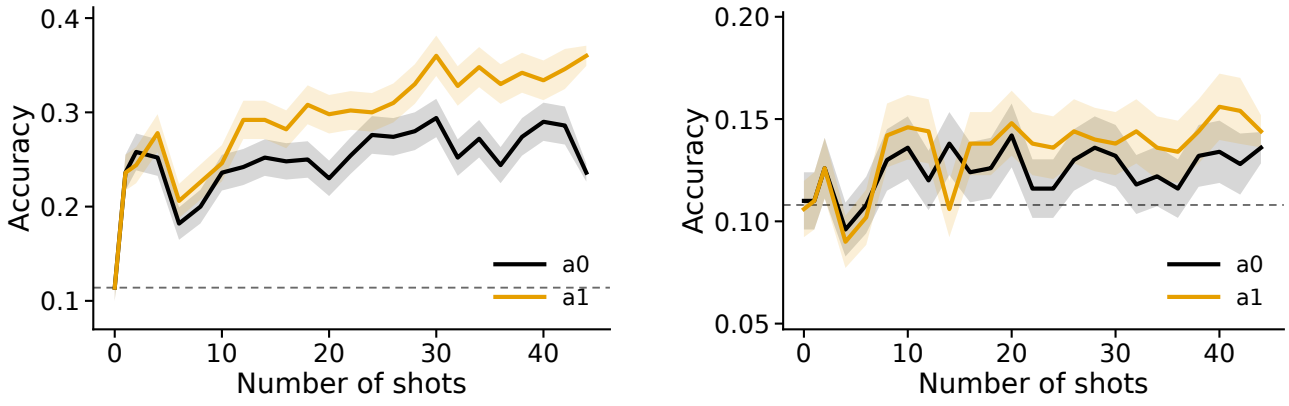


Figure 13. **Eval loss during training.** Higher α consistently yields lower loss, with improvement appearing early and persisting to convergence.



(a) Key–value retrieval.

(b) Override KV.

Figure 14. **Additional few-shot tasks.** α=1 benefits more reliably from additional context than α=0 across KV retrieval and Override KV.

J. Training Curves

K. Additional Few-Shot Tasks

L. Robustness: t-Conditioning and Mask Distractors

t-conditioning insensitivity. The primary MDLM evaluation uses one-step conditional denoising: demonstrations and query prefix are clean, the answer token is replaced by `[MASK]`, and logits at that mask are scored. Sweeping the supplied noise level $t \in \{0.01, 0.05, 0.1, 0.2\}$ while keeping this mask pattern fixed leaves accuracy essentially unchanged for both α=0 and α=1. This makes the one-step probe directly comparable to MLM masked-token scoring and AR next-token scoring: the model's answer is controlled by which positions are masked, not by the scalar noise label supplied to the network.

Mask-distractor probe. To test whether extra masks distract MDLM at inference time, we append suffix `[MASK]` tokens after the query and compare with clean fill tokens. Contrary to a simple "masks always hurt" hypothesis, the effect is task-dependent: suffix masks do not produce universal collapse and can sometimes help by rerouting attention, complementing the mask-distractor failure mode documented by Piskorz et al. (2025). In the faithful large-scale MDLM evaluations, recency/override tasks are neutral to slightly positive under suffix masks, while harder tasks such as Binary FSM are close to flat. The distinction is important: appended masks are active denoising targets, not padding, but their effect depends on whether they dilute the useful prefix signal or create uninformative positions that reroute attention toward demonstrations.

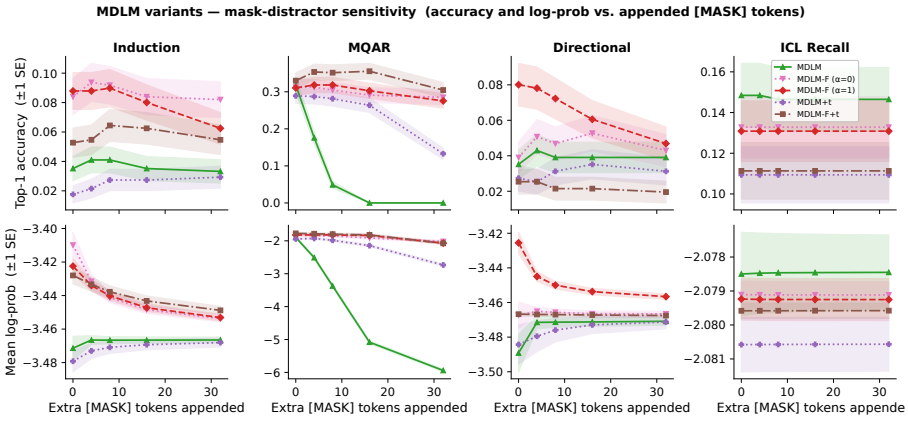


Figure 15. t**-conditioning insensitivity and mask-distractor probe.** Sweeping the supplied noise level t at fixed mask pattern leaves predictions essentially unchanged. Appending suffix `[MASK]` tokens has task-dependent effects, sometimes improving accuracy by rerouting attention.

M. Generation: Iterative Denoising Under Three Regimes

We complement one-step evaluation with full iterative denoising. One-step scoring asks whether the trained model can use a clean prefix in a single conditional prediction. Full generation asks whether a denoising chain can make that prefix available and then avoid corrupting the query prediction with its own generated suffix. We run 64 denoising steps and compare uniform versus biased unmasking schedules crossed with three prefix-handling regimes. *Free generation:* the model generates its own prefix and answer tokens. *Teacher forcing:* prefix tokens are corrected only after the chain reveals them. *Clean-prefix decoding:* demonstrations and query prefix are clamped clean throughout the chain.

We report OverrideKV and a Bigram Repeat control. The hierarchy is stable: free generation usually fails on the synthetic ICL tasks, teacher forcing partially recovers, and clean-prefix decoding performs best. In the clean-prefix regime, the model starts with the correct demonstrations visible and assigns high probability to the gold answer, but confidence decreases as model-generated suffix tokens accumulate. In free generation, the prefix starts masked and early self-generated prefix tokens can commit the chain to a wrong hypothesis; after that, entropy may fall while gold probability remains near zero.

These generation results should not be read as a monotone win for α=1. At K=8, α=0 has higher one-step accuracy on

both OverrideKV and Bigram Repeat, while $\alpha{=}1$ slightly surpasses $\alpha{=}0$ only under clean-prefix generation on OverrideKV and remains worse on Bigram Repeat. This is a useful failure mode: one-step scoring measures whether the model can use an already clean prefix, whereas generation additionally tests whether iterative denoising preserves that prefix and avoids suffix contamination.

Table 6. **One-step versus full generation at $K{=}8$ for small MDLMs.** Generation columns report final denoising accuracy. Chance is $1/8 = 0.125$.

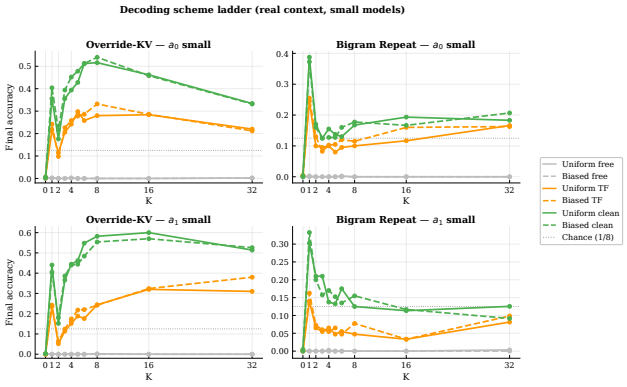
Task	Model	One-step	Free gen	TF gen	Clean-prefix gen
OverrideKV	$\alpha{=}0$	0.570	0.000	0.332	0.540
OverrideKV	$\alpha{=}1$	0.242	0.002	0.244	0.582
Bigram repeat	$\alpha{=}0$	0.373	0.000	0.115	0.178
Bigram repeat	$\alpha{=}1$	0.208	0.000	0.078	0.155

Table 7. **Binary FSM denoising traces for $\alpha{=}0$, $K{=}16$.** Clean-prefix conditioning gives high initial confidence but confidence decays as generated suffix tokens accumulate.

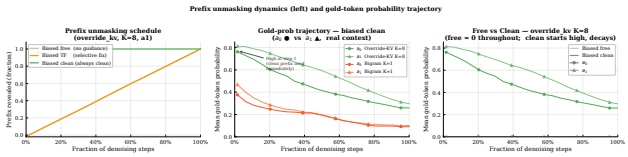
Regime	Final acc.	Gold prob. start	Gold prob. end	Entropy start/end	Prefix reveal
Uniform free	0.000	0.00043	0.00023	7.10 / 2.32	$0 \to 1$
Uniform TF	0.356	0.00043	0.201	7.10 / 3.80	$0 \to 1$
Uniform clean-prefix	0.504	0.690	0.333	0.66 / 3.14	$1 \to 1$

Table 8. **Average binary-FSM recovery for $\alpha{=}0$.**

Setting	One-step	Free gen	TF gen	Clean-prefix gen
Avg. all K incl. 0	0.700	0.000	0.222	0.371
$K{=}16$	0.742	0.000	0.356	0.504
Avg. $K \in \{8, 16, 32\}$	0.742	0.000	0.332	0.505



(a) Decode regimes vs. accuracy.



(b) Unmasking dynamics.

Figure 16. **Iterative denoising regimes.** Free generation fails; teacher forcing partially recovers; clean-prefix decoding performs best but degrades as generated suffix tokens accumulate.

N. TMS Geometry Controls

The TMS controls use sparse binary feature structure to test whether non-causal objectives can learn coherent global representations at all. Each example is generated from a sparse set of latent binary features. The model must represent which features are present and map them to output statistics under the same training distribution across objectives. Sparsity controls the degree of superposition pressure: at low sparsity, features can be represented with little interference; at high sparsity, the model must pack many possible features into limited residual dimensions.

TMS model construction. We compare AR, MLM, symmetric MDLM, and position-biased MDLM variants on the same sparse-feature data. The architecture and data distribution are fixed; only the objective and masking/corruption process differ. AR receives a causal prediction objective, MLM predicts uniformly masked positions, symmetric MDLM denoises uniformly corrupted positions over a diffusion-time schedule, and position-biased MDLM shifts corruption toward later coordinates while keeping the same tokenwise diffusion-loss form. These runs are not intended as language models; they are controlled representation-learning systems where geometry can be measured directly.

TMS probes. We measure rotation sensitivity, feature decodability, feature-direction inner products, Hamming alignment of embeddings, embedding effective rank, participation ratio, and feature dimensionality. Rotation sensitivity asks whether downstream logits depend on a privileged residual basis. Feature decodability asks whether individual latent features can be recovered linearly. Feature-direction inner products measure whether feature directions are aligned, orthogonal, or antipodal. Hamming alignment asks whether embedding distances track feature-set overlap. Effective rank and participation ratio measure how many dimensions are actively used. These are not ICL tests; they ask whether diffusion objectives can learn structured symmetric geometry. The result is that they can: bidirectional/diffusion objectives produce rotation-invariant, well-conditioned feature geometries comparable to AR. This rules out a generic "representation learning" explanation for the macro-ICL failure: the failure is specific to causal/directional ICL rather than an inability to learn structure.

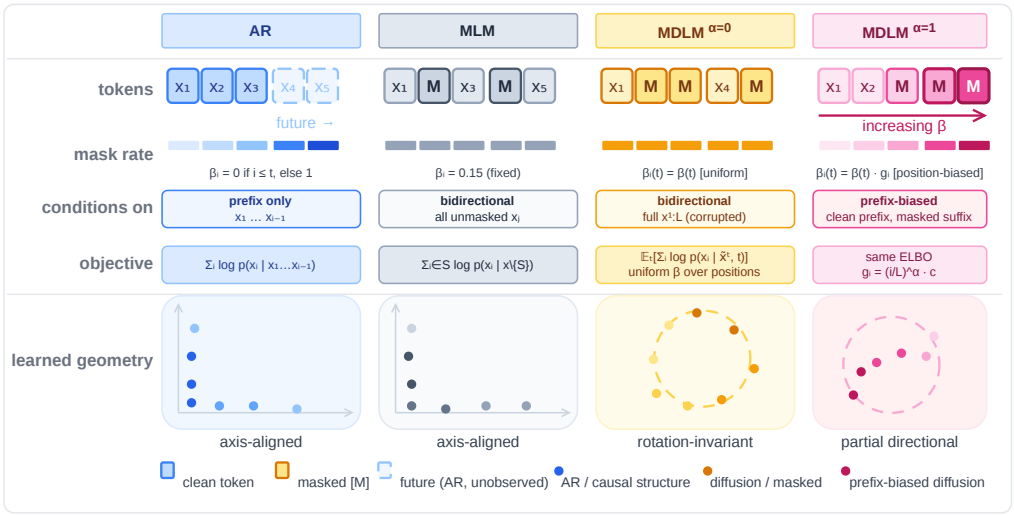
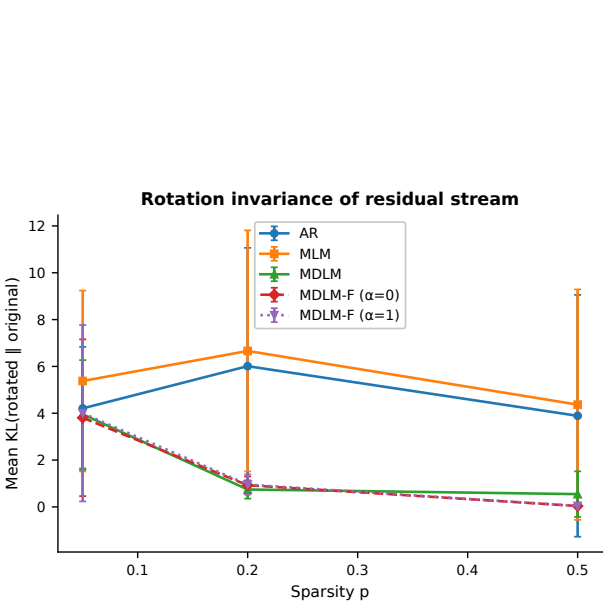
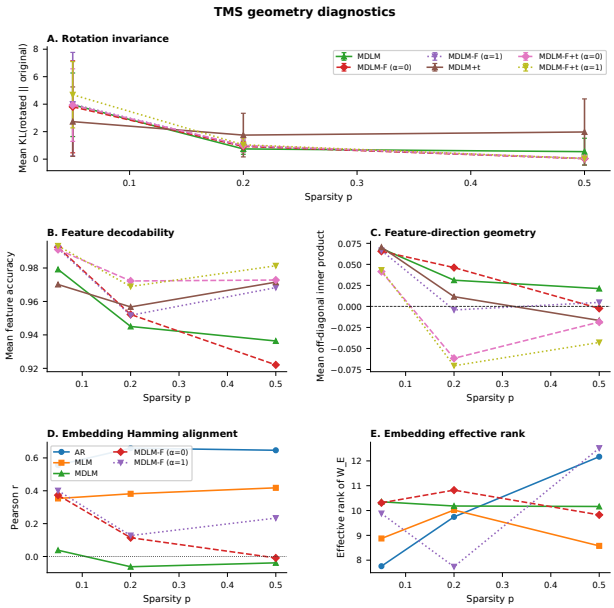


Figure 17. **TMS objective schematic.** The toy models of superposition compare AR, MLM, symmetric MDLM, and position-biased MDLM under the same sparse-feature data distribution. The diagram summarises which tokens are clean or masked, which conditionals the objective trains, and the kind of geometry each objective tends to produce.



(a) Rotation sensitivity.



(b) MDLM-family geometry.

Figure 18. **TMS geometry for MDLM variants.** At high sparsity, diffusion objectives can become much less basis-privileged than AR/MLM under random residual rotations. The grid shows that MDLM variants learn structured feature geometry, including feature decodability, feature-direction organisation, Hamming alignment, and embedding effective rank.

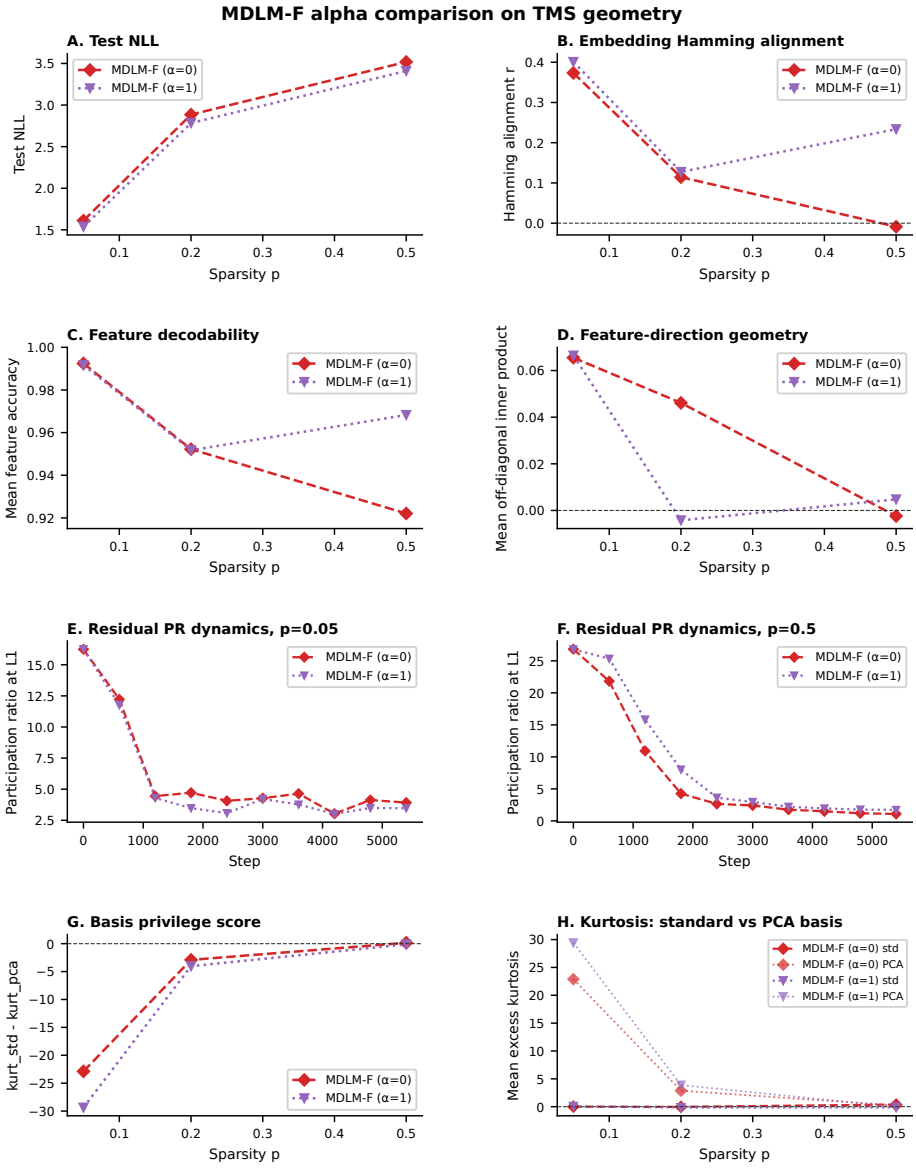


Figure 19. **Position-biased versus uniform MDLM-F geometry.** Position-biased masking (α=1) preserves richer high-sparsity geometry than uniform masking (α=0): better Hamming alignment, higher feature decodability, broader residual participation, and weakly more antipodal feature directions.