SAC-Copula: Quality-Preserving Watermarking for Diffusion Language Models via Smooth Correlated Gumbel Fields

Anonymous ACL submission

Abstract

Watermarking diffusion language models (DLMs) requires mechanisms compatible with iterative parallel unmasking rather than autoregressive decoding. Existing sampling-based watermarking methods typically inject position-wise i.i.d. perturbations, which can be poorly aligned with DLM decoding dynamics and degrade generation quality. We propose SAC-Copula, a quality-preserving watermarking method for DLMs based on smooth, locally correlated Gumbel perturbation fields constructed via a Gaussian copula. We further develop a SAC-aware detector using covariance-aware filtering and native-sample calibration. Mechanism-level analysis shows that local correlation reduces perturbation roughness and better matches iterative refinement dynamics. Experiments on LLaDA demonstrate that SAC-Copula improves overall generation quality and stability relative to existing baselines while maintaining strong watermark detectability. Additional token-edit stress tests further characterize residual watermark evidence under controlled synchronization drift.

1 Introduction

As Large Language Models (LLMs) become widely deployed, it becomes increasingly critical to establish a safety line between LLMs and humans, where one important brick is to distinguish the model-generated content from the human-created one. Watermarking offers a practical route: a generator embeds a private statistical signal during content decoding, and a detector later tests whether a text carries that signal via a shared secret key.

Most existing language-model watermarking methods are developed for autoregressive LLMs, including green-list token biasing, sampling-based or distortion-free schemes, robust watermarking variants, and multi-bit identification mechanisms (Kirchenbauer et al., 2023; Kuditipudi et al., 2024; Zhao et al., 2023; Yoo et al., 2024). In these methods, watermark signals are typically synchronized with the left context and tokens are generated sequentially from left to right conditioned on the prefix and watermark signal. In contrast, Diffusion Language Models (DLMs), such as LLaDA (Nie et al., 2025), generate text through iterative masked denoising, where multiple positions are updated in parallel and previously generated tokens can be revised across steps. This non-sequential and dynamically coupled decoding process makes watermark design fundamentally different from the autoregressive setting. In particular, watermarking schemes based on i.i.d. injected signals are poorly matched to DLM decoding dynamics, which can disrupt local refinement consistency across denoising steps. This may lead to degraded generation quality, including unstable neighboring predictions, semantic inconsistency, degraded fluency, and refinement oscillation. Figure 1 shows an illustrative example that compares i.i.d. and correlated perturbations.

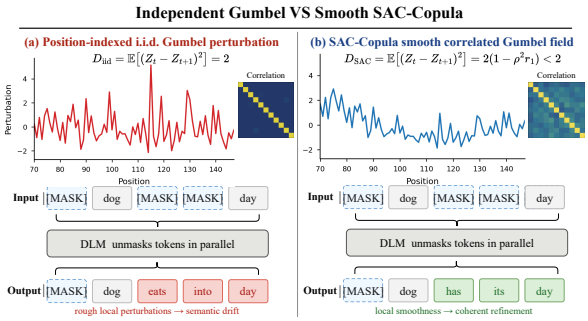

Figure 1: Motivation for smooth watermark fields in DLM decoding. Position-indexed i.i.d. Gumbel perturbations create independent high-frequency noises across neighboring positions, whereas SAC-Copula introduces controlled low-lag dependence while preserving one-coordinate Gumbel marginals. Token examples are illustrative rather than deterministic generation outcomes.

1.1 Related Works

Autoregressive and sampling-based watermarking. A large body of LLM watermarking work

embeds keyed statistical signals during autoregressive decoding and detects them under a false-positive constraint (Kirchenbauer et al., 2023; Li et al., 2025). Stronger watermark signals can improve detection power but may degrade fluency, semantic fidelity, or diversity, motivating selective, adaptive, and quality-aware objectives (Wang et al., 2025; Giboulot and Furon, 2024; He et al., 2025a). A related line uses Gumbel-max, inverse-transform, or distortion-free sampling ideas to embed keyed randomness under specific sampling and security assumptions (Aaronson, 2023; Fairoze et al., 2023; Christ et al., 2023). This paper follows the sampling-based watermarking paradigm, while explicitly adapting the design to DLM decoding dynamics.

DLM watermarking. Some recent works have started to explore designing watermarking methods adapted to DLMs. Gloaguen et al. (2026) propose the first watermark tailored for DLMs by applying the green-red list watermarking (Kirchenbauer et al., 2023) in expectation over partially observed contexts and promoting tokens that strengthen future watermark signals. To address the non-sequential decoding process, some works (Wu et al., 2025; Raban et al., 2026) further consider synchronizing watermark signals using both left and right contexts. Instead of directly perturbing token probabilities, Hong and No (2026) embed watermark by steering the decoding order to improve quality preservation in DLMs.

More closely related to our work, Bagchi et al. (2025) apply Gumbel-max watermarking (Aaronson, 2023) to diffusion decoding using sequence-position-indexed randomness. Although this approach is theoretically distribution-preserving, the injected i.i.d. Gumbel perturbations are not well aligned with DLM decoding dynamics, as discussed earlier, and can degrade generated text quality in practice. Existing Gumbel-based methods primarily focus on how randomness is indexed across positions. In contrast, we argue that the *joint structure* of the perturbation field is itself a critical design dimension for DLM watermarking. Under iterative refinement decoding, independent high-frequency perturbations across neighboring positions can conflict with local consistency. Motivated by this observation, we shift the design focus from injecting independent randomness to designing structured perturbation fields whose geometry and smoothness better align with diffusion decoding dynamics and detector evidence geometry.

The main insight of our paper is that locally smooth and correlated watermark perturbations are more compatible with DLM decoding than high-frequency i.i.d. perturbations, leading to improvement in semantic coherence and output stability. Based on this insight, we propose **SAC-Copula**, which constructs correlated Gumbel perturbation fields via a Gaussian copula (cf. Figure 1). Under suitable assumptions, SAC-Copula preserves Gumbel marginals while introducing controlled local dependence across neighboring token positions. This structured perturbation field better aligns with DLM refinement dynamics and improves generation quality, while maintaining effective watermark detection.

1.2 Main Contributions

Our contributions can be summarized as follows:

- We introduce **SAC-Copula**, a DLM watermarking method with **S**mooth and **A**uto-**C**orrelated Gumbel Copula perturbations through a Gaussian **Copula**, which preserves one-coordinate Gumbel marginals under mild assumptions. This is a DLM-aware watermarking perspective that treats the joint geometry of the perturbation field as a key design dimension for DLM decoding.
- We provide a mechanism-level analysis showing how local correlation reduces perturbation roughness and improves compatibility with DLM refinement. In particular, positive low-lag dependence decreases the latent local-difference energy from $D_{\text{i.i.d}} = 2$ to $D_{\text{SAC}} = 2(1 - \rho^2 r_1)$.
- We develop a SAC-aware detector that reads the induced local evidence geometry through covariance-aware filtering and native-sample calibration. To improve detection robustness, we further include an offset-scan rule to recover mismatched token–signal alignment.
- Experiments on LLaDA show that SAC-Copula substantially improves generation quality and stability relative to other baselines, while retaining strong clean detectability. We further report mild token-edit stress tests as diagnostic analyses of residual watermark evidence.

2 Watermark Injection and Detection

SAC-Copula replaces position-indexed i.i.d. Gumbel perturbations with a smooth, locally correlated Gumbel field for DLM decoding. The generation

side defines this field, while the detector provides a matched readout for the induced evidence geometry. Figure 2 summarizes the pipeline.

Baseline i.i.d. Gumbel Watermarking. Let $\mathcal{V}$ be the token vocabulary, $j \in \mathcal{V}$ a token, $t \in [T]$ a generated-token position, and n a DLM denoising step. The direct baseline applies position-indexed i.i.d. Gumbel perturbations (Bagchi et al., 2025; Aaronson, 2023), $G_{\text{i.i.d.}}(t, j) \sim \text{Gumbel}(0, 1)$, to masked-position logits:

$$\widetilde{\ell}_{n,t,j} = \ell_{n,t,j} + \alpha G_{\text{i.i.d.}}(t, j), \quad t \in [T], j \in \mathcal{V}.$$

Here $\ell_{n,t,j}$ is the clean logit and α is the perturbation scale. SAC-Copula reduces to this baseline when the correlated branch is removed.

2.1 SAC-Copula Perturbation Field

SAC-Copula constructs a locally correlated Gaussian latent field and maps each coordinate to a standard Gumbel variable through a Gaussian copula. It starts from two independent standard Gaussian tapes $Z_{\text{white},1}, Z_{\text{white},2}$, with $Z_{\text{white},r}(t, j) \sim \mathcal{N}(0, 1)$. Let W be an odd window, $h = (W - 1)/2$, and σ_K the kernel shape parameter. For $k \in \{-h, \ldots, h\}$, define

$$K_W[k] = \frac{\exp\big(-k^2/(2\sigma_K^2)\big)}{\sum_{m=-h}^{h} \exp\big(-m^2/(2\sigma_K^2)\big)}.$$

For full-window positions, the normalized spatial branch is

$$\widetilde{Z}_{\text{spatial}}(t, j) = \frac{\sum_{k=-h}^{h} K_W[k] Z_{\text{white},1}(t - k, j)}{\sqrt{\sum_{k=-h}^{h} K_W[k]^2}}.$$

The final latent field mixes the independent white noise and the normalized spatial stream:

$$Z_{\text{final}}(t, j) = \sqrt{1 - \rho^2}\, Z_{\text{white},2}(t, j) + \rho\, \widetilde{Z}_{\text{spatial}}(t, j),$$

where $\rho \in [0, 1]$ controls the local dependence strength. Under the full-window or boundary-normalized convention, this normalization keeps each valid coordinate of Z_{final} standard normal. Appendices A.1 and A.2 state the boundary convention and prove the resulting one-coordinate marginal exactness.

The correlated Gumbel noise tape is obtained by the Gaussian-copula transform

$$G_{\text{corr}}(t, j) = F_G^{-1}(\Phi(Z_{\text{final}}(t, j))) \sim \text{Gumbel}(0, 1),$$

where Φ is the standard Gaussian cumulative distribution function (CDF) and F_G is the standard Gumbel CDF. This coordinate-wise transform preserves one-coordinate Gumbel marginals under assumptions stated in Appendix A.1, while introducing the joint noise tape structure. Appendix A.3 derives the finite-window latent autocorrelation used by the detector design.

2.2 Injection into LLaDA Decoding

SAC-Copula noise is injected only into currently masked generated positions. At a masked generated position t, SAC-Copula perturbs logits before applying the decoding rule:

$$\widetilde{\ell}_{n,t,j} = \ell_{n,t,j} + \alpha G_{\text{corr}}(t, j), \quad t \in [T], j \in \mathcal{V},$$

where $\ell_{n,t,j}$ denotes the original clean logit and $G_{\text{corr}}(t, j)$ is the SAC-Copula perturbation. The perturbation scale α is an experimental configuration parameter. In the main configuration it matches the generation temperature, but this equality is not a theoretical requirement. Prompt tokens are not perturbed, and remasking confidence is computed from the unperturbed logits. The same token-position perturbation $G_{\text{corr}}(t, \cdot)$ is reused across denoising steps, indexed by generated token position and token but not by denoising step.

2.3 SAC-Aware Detector

For a clean fixed-alignment sequence $y = (y_1, \ldots, y_T)$, the detector extracts token-level watermark evidence $\mathbf{s}(y) = (s_1(y), \ldots, s_T(y))^\top$, where

$$s_t(y) = G_{\text{corr}}(t, y_t).$$

The legacy detector treats these values as exchangeable i.i.d. Gumbel evidence and applies an equal-weight z-score; Appendix B.1 gives the full old-score form. SAC-Copula intentionally induces low-lag evidence structure, so relying only on an i.i.d. equal-weight readout can discard useful local geometry. We therefore use the Full Filtered Ridge (FFR) detector, defined as a ridge-regularized matched linear statistic after SAC-aware local filtering and covariance calibration on native samples.

FFR first centers raw evidence using the Gumbel mean μ_G and a residual mean $\widehat{\boldsymbol{\mu}}_{x,0}$ estimated from held-out native (aka. unwatermarked) samples generated under the same decoding protocol:

$$\mathbf{x}(y) = \mathbf{s}(y) - \mu_G \mathbf{1} - \widehat{\boldsymbol{\mu}}_{x,0}.$$

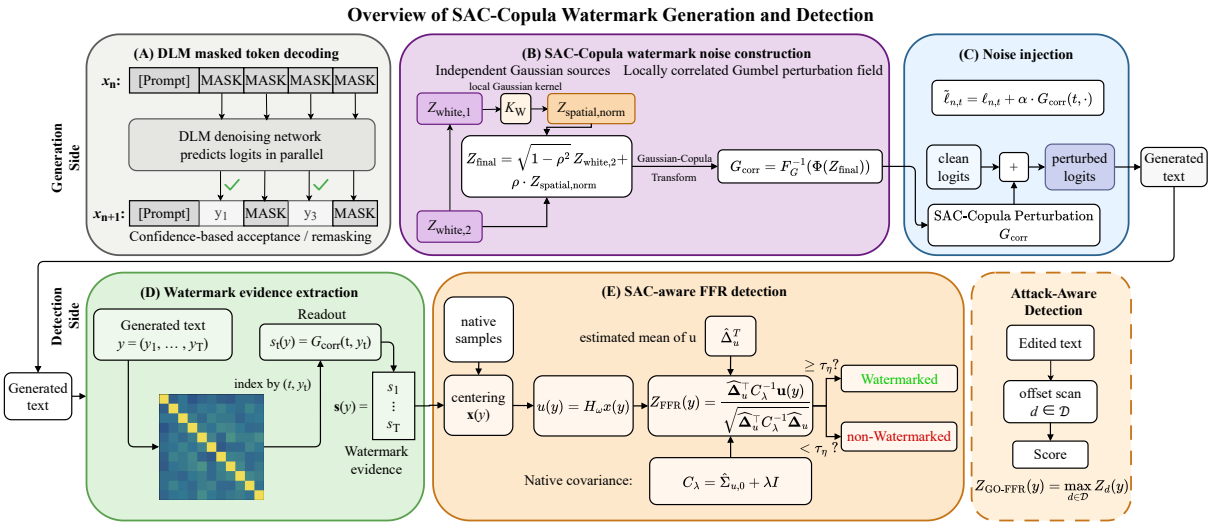

Figure 2: Overview of SAC-Copula. The generation side constructs a locally correlated Gumbel tape through variance-normalized convolution, dual-stream mixing, and a Gaussian-copula transform, then injects it at currently masked generated positions during DLM refinement. The detection side reads token-level evidence and applies SAC-aware filtering with thresholds calibrated on native, unwatermarked samples. GO-FFR is shown only as an auxiliary diagnostic for controlled insertion/deletion edits.

It then applies a SAC-aware local feature map

$$\mathbf{u}(y) = H_\omega \mathbf{x}(y),$$
$$\omega_\tau \propto \exp\left(-\frac{\tau^2}{4\sigma_K^2}\right) \mathbf{1}\{|\tau| \leq B\},$$

where H_ω a filter is motivated by the autocorrelation envelope of the SAC kernel (cf. Appendix B.2) and $B = \min(W-1, T-1)$. The lag template mirrors the Gaussian-envelope smoothing used by SAC-Copula, so H_ω aggregates evidence over the local low-lag structure introduced by the correlated perturbation field.

The filtered native-sample covariance and ridge covariance are

$$\widehat{\Sigma}_{u,0} = H_\omega \widehat{\Sigma}_{x,0} H_\omega^\top, \qquad C_\lambda = \widehat{\Sigma}_{u,0} + \lambda I.$$

Here $\widehat{\Sigma}_{x,0}$ is estimated from held-out native samples under the same detector configuration. Let $\widehat{\mathbf{\Delta}}_u$ be the filtered mean direction estimated from development or cross-fit data, not evaluation samples. The FFR score is

$$Z_{\text{FFR}}(y) = \frac{\widehat{\mathbf{\Delta}}_u^\top C_\lambda^{-1}\mathbf{u}(y)}{\sqrt{\widehat{\mathbf{\Delta}}_u^\top C_\lambda^{-1}\widehat{\mathbf{\Delta}}_u}}.$$

For a deployment-style binary decision, the detector flags a sequence when

$$Z_{\text{FFR}}(y) \geq \tau_\eta.$$

The threshold τ_η can be calibrated from held-out native samples under the same detector configuration. In our experiments, we report ROC-based AUC and TPR at target FPR levels using the corresponding Native-score distribution. FFR should be understood as a matched readout for SAC-induced evidence geometry, not as a full-distribution optimal detector or the central contribution. Appendix B–B.3 give the normalization, covariance rationale, filtering details, and relation to the old detector.

Furthermore, as one step towards robust detection, we introduce another detector, GO-FFR, extended from FFR by considering controlled token-editting attacks:

$$Z_{\text{GO-FFR}}(Y) = \max_{d \in \mathcal{D}} Z_d(Y),$$

where $\mathcal{D}$ is a scan window and $Z_d(Y)$ is a test statistic defined in Appendix B.5. It scans a small family of global offsets to partially recover mismatched token–signal alignment, reuses the FFR scoring direction, and calibrates its threshold on native samples under the same offset-scan rule. Appendix B.5 and Appendix B.6 give the full analyses.

3 Why Correlated Perturbations Help DLMs

This section provides a mechanism-level explanation for why smooth, locally correlated perturbations are better aligned with DLM refinement dynamics than high-frequency i.i.d. perturbations. We

analyze the geometry of the latent perturbation field and the induced detector evidence, and derive empirical predictions on generation quality and detectability that are validated in Section 4.

3.1 DLM Decoding Couples Neighboring Decisions

Autoregressive watermarking perturbs a left-to-right sampler whose prefix is fixed once a token is emitted. DLM decoding is different. LLaDA-style generation repeatedly predicts multiple masked positions in parallel, accepts or remasks tokens according to confidence, and revises a partially generated sequence over denoising steps (Nie et al., 2025; Ghazvininejad et al., 2019). More recent masked-diffusion and non-Markovian discrete-diffusion work likewise emphasizes iterative denoising and refinement-sensitive generation dynamics (Zhang et al., 2025b; Sahoo et al., 2024). As a result, neighboring token positions are not isolated one-step decisions: local choices can be compared, retained, or revised together during refinement.

This difference makes the geometry of the perturbation field relevant. Beyond its marginal strength at each position, a watermark also determines how perturbations vary across adjacent positions. An i.i.d. Gumbel field can vary sharply across neighboring masked positions that are refined together, creating a plausible mismatch with the local consistency preference of iterative masked refinement. A smoother locally correlated field makes nearby perturbation preferences more coherent, which motivates the quality-stability hypothesis tested in Section 4. Figure 1 illustrates this contrast, while the following subsections formalize it using a latent local-difference proxy.

3.2 Independent Latent Fields Create Local Roughness

We first define a latent local-difference proxy before the coordinate-wise Gumbel transform. Fix a vocabulary item j, and consider the Gaussian latent field $Z(t, j)$ that is later mapped monotonically into Gumbel noise. We measure lag-one local variation by

$$\mathcal{D} = \mathbb{E}\Big[(Z(t+1, j) - Z(t, j))^2\Big],$$

A larger $\mathcal{D}$ indicates sharper adjacent changes in the latent perturbation field.

For an i.i.d. latent Gaussian field, $\mathrm{Var}(Z(t, j)) = 1$ and $\mathrm{Cov}(Z(t, j), Z(t+1, j)) = 0$. Therefore, the difference proxy for i.i.d. latent Gaussian field is

$$\mathcal{D}_{\text{i.i.d}} = 2.$$

3.3 SAC-Copula Reduces Local Perturbation Variation

SAC-Copula changes the joint structure of the perturbation field rather than the one-coordinate marginal. Let $r_\tau(W, \sigma_K)$ denote the finite-window autocorrelation of the normalized spatial branch defined in Appendix A.3. This quantity is large at small lags when the smoothing kernel strongly overlaps with its shifted copy, and decays as the lag exceeds the kernel support. For the final latent field, the nonzero-lag covariance is

$$\mathrm{Cov}(Z_{\text{final}}(t, j), Z_{\text{final}}(t+\tau, j)) = \rho^2 r_\tau(W, \sigma_K).$$

Let $r_1 = r_{\tau=1}(W, \sigma_K)$, and write $\Delta_t = Z_{\text{final}}(t+1, j) - Z_{\text{final}}(t, j)$, the lag-one local difference energy becomes

$$\mathcal{D}_{\text{SAC}} = \mathbb{E}[\Delta_t^2] = 2(1 - \rho^2 r_1).$$

When $r_1 > 0$ and $\rho > 0$, SAC-Copula has lower latent local variation than the i.i.d. field:

$$\mathcal{D}_{\text{SAC}} < \mathcal{D}_{\text{i.i.d}}.$$

This lower local variation gives a plausible quality mechanism for DLM decoding. LLaDA-style models refine neighboring masked positions jointly, so when clean logits are locally uncertain, abrupt independent perturbation changes can push nearby positions toward inconsistent random preferences. SAC-Copula does not reduce the single-coordinate probability of large Gumbel values: after the monotone Gaussian-copula-to-Gumbel transform, each coordinate still has the standard Gumbel marginal under the assumptions in Appendix A.2. Instead, it reduces abrupt local changes across positions by inducing positive low-lag dependence. This motivates the empirical hypothesis that smoother perturbation fields may improve quality stability under iterative masked refinement.

Because the inverse-Gumbel transform is nonlinear, this latent local-difference proxy characterizes only the dependence structure before the copula transform and should not be interpreted as invariance of the full generated-text distribution.

Figure 3 visualizes these perturbation-field diagnostics. The autocorrelation view shows that

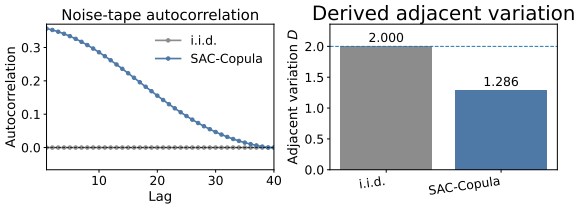

Figure 3: Perturbation-field smoothness diagnostics. SAC-Copula induces positive low-lag autocorrelation and lowers the latent adjacent-variation proxy from $\mathcal{D}_{\text{iid}} = 2$ to $\mathcal{D}_{\text{SAC}} \approx 1.286$. These are latent-field diagnostics, not text-quality measurements.

SAC-Copula introduces a positive low-lag dependence, while the adjacent-variation view shows the corresponding reduction in the latent roughness proxy.

Together, these results suggest that locally correlated perturbations better support DLM decoding quality while preserving watermark detectability. Section 4 evaluates these predictions empirically.

4 Experiments

This section evaluates whether SAC-Copula yields a practical DLM watermarking profile. We organize the experiments around three questions: 1) whether SAC-Copula improves clean generation quality and stability, 2) whether clean low-FPR detectability is preserved, and 3) what residual evidence remains under controlled token edits. **Code:** `https://anonymous.4open.science/status/review-artifact-90EF`

4.1 Experimental Setup

We evaluate generation from LLaDA-8B-Instruct (Nie et al., 2025) on ELI5 prompts (Fan et al., 2019). The comparison includes native decoding, the direct i.i.d. Gumbel baseline, and three LLaDA-adapted external references: KGW-style (Kirchenbauer et al., 2023), Unigram (Zhao et al., 2023), and PatternMark-style (Wu et al., 2025). We use green-list fraction $\gamma = 0.25$ and perturbation strength $\delta \in \{1, 2, 3\}$, treating these baselines as deterministic LLaDA-adapted references rather than exhaustive reproductions of all original variants.

Unless otherwise stated, SAC-Copula adopts the selected configuration $W = 39, \sigma_K = 15.0, \rho = 0.6$, where W, σ_K, and ρ control the correlation range, kernel shape, and correlated-branch strength, respectively. Figure 10 in Appendix C.5 visualizes the parameter sensitivity sweep used for operating-point selection, and Table 11 reports the corresponding raw-metric summary. We use this configuration as a balanced clean quality–detectability operating point. Full generation, watermark, detector, calibration, seed, and evaluator details are reported in Appendix C.2.

Quality metrics include mean perplexity (PPL) for average fluency, PPL tail collapse as the fraction of generations with PPL > 100, semantic preservation (SBERT), MAUVE-style distributional fidelity, token-distribution drift (Token TV), repetition (Rep3), and diversity (Distinct 3). (Pillutla et al., 2021). Clean detectability is measured by AUC, TPR@1%FPR and TPR@5%FPR, with detector scores calibrated against held-out native samples under the same detector configuration (Li et al., 2025; He et al., 2025b).

4.2 Text Quality and Stability

Text quality is the primary empirical target. We ask whether replacing position-wise i.i.d. Gumbel perturbations with a locally correlated Gumbel field improves clean DLM generation stability while preserving clean watermark detectability. Table 1 provides a compact numerical summary: Relative to the direct i.i.d. Gumbel baseline, SAC-Copula improves clean low-FPR detectability while substantially reducing tail instability, as reflected by mean PPL and collapse indicators.

We interpret quality through a multi-metric profile rather than a single scalar score. Figure 4 visualizes this profile with direction-normalized axes, where larger values indicate better or more native-like behavior. SAC-Copula shows a balanced clean-quality profile, with gains over i.i.d. Gumbel concentrated in tail-stability and repetition-sensitive dimensions. The radar chart is a visualization summary only; raw metrics and normalization details are reported in Table 1 and Appendix Tables 6–9.

Table 1 shows that SAC-Copula improves the clean quality–detectability trade-off relative to the direct i.i.d. Gumbel baseline. It increases TPR@FPR5 from 0.855 to 0.985, while reducing mean PPL from 636.1 to 12.2 and collapse from 3.5% to 1.0%. The main gain is therefore stronger clean detection together with substantially better tail stability. Compared with stronger static-bias baselines, SAC-Copula avoids much of the semantic, distributional, repetition, diversity, and token-drift cost. The Unigram sweep is used only as a contrastive appendix diagnostic for static-bias tuning. Some Unigram strengths obtain favorable PPL

Method	Strength	AUC ↑	TPR@1%FPR ↑	TPR@5%FPR ↑	Mean PPL ↓	Collapse ↓	SBERT ↑	MAUVE ↑	Token TV ↓	Rep3 ↓	Distinct3 ↑	Quality ↑
Native	–	–	–	–	16.840	0.005	1.000	1.000	0.000	0.406	0.594	0.960
i.i.d. Gumbel	–	0.966	0.815	0.855	636.069	0.035	0.784	0.991	0.163	0.415	0.585	0.661
KGW	$\delta = 1$	0.670	0.060	0.120	5.32×10^4	0.030	0.794	**0.999**	**0.146**	0.414	0.586	0.694
KGW	$\delta = 2$	0.820	0.200	0.315	6.98×10^4	0.045	0.776	0.998	0.173	0.450	0.550	0.553
KGW	$\delta = 3$	0.936	0.515	0.660	8.30×10^4	0.050	0.762	0.982	0.228	0.510	0.489	0.379
Unigram	$\delta = 1$	0.896	0.385	0.555	7.714	**0.000**	**0.803**	0.981	0.233	**0.395**	**0.605**	**0.798**
Unigram	$\delta = 2$	0.988	0.850	0.930	**7.630**	**0.000**	0.755	0.841	0.360	0.455	0.545	0.545
Unigram	$\delta = 3$	**0.997**	**0.985**	**0.995**	9.464	0.005	0.690	0.601	0.439	0.560	0.440	0.150
PatternMark	$\delta = 1$	0.818	0.510	0.625	3883.179	0.010	0.781	0.983	0.153	0.426	0.574	0.721
PatternMark	$\delta = 2$	0.926	0.760	0.800	3898.094	0.015	0.779	0.996	0.161	0.422	0.578	0.714
PatternMark	$\delta = 3$	0.938	0.760	0.805	3901.117	0.020	0.768	0.982	0.198	0.480	0.520	0.555
SAC-Copula (Ours)	–	0.990	0.980	0.985	12.164	0.010	0.798	**0.999**	0.163	0.414	0.586	0.758

Table 1: Clean detection and text-quality trade-off. Detection is reported by AUC and low-FPR TPR; quality is reported by raw fluency, collapse, semantic, distributional, token-drift, repetition, and diversity metrics. The normalized Quality score is descriptive only: it averages direction-normalized SBERT, native-token F1, Token TV, Rep3, Distinct3, and PPL-collapse components, with lower-is-better metrics inverted. **Best** and second-best results among watermarked methods are marked in each column.

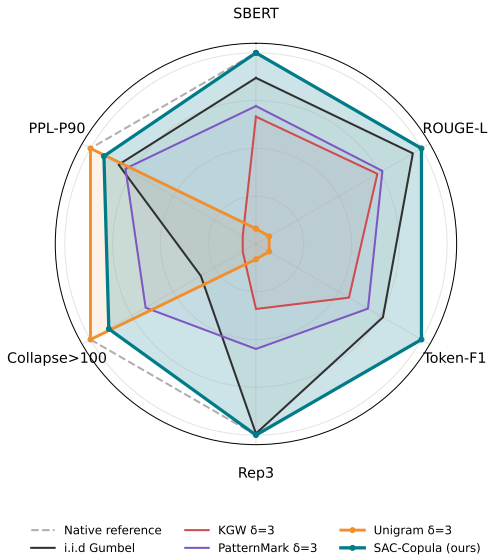

Figure 4: Overview of generation quality profile. The radar chart compares SAC-Copula, i.i.d. Gumbel, and representative external-baseline settings on direction-normalized quality axes. Larger values indicate better or more native-like behavior.

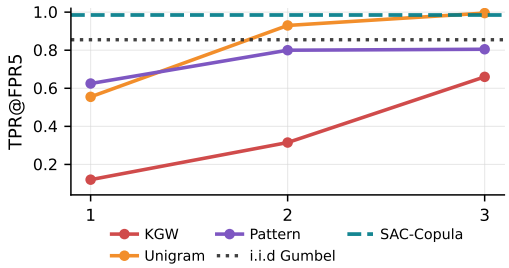

Figure 5: Clean TPR@5%FPR across watermark strengths. External baselines vary $\delta \in \{1, 2, 3\}$; i.i.d. Gumbel and SAC-Copula are fixed references. Full detection metrics are shown in Appendix C.1, Figure 8.

or collapse values, but stronger settings trade these gains against semantic, distributional, repetition, diversity, and token-drift dimensions. Appendix C.3, Figure 9, reports the full diagnostic figure.

4.3 Clean Detection

The second question is whether smoothing the perturbation field weakens the watermark signal under clean, fixed-alignment decoding. Figure 5 summarizes clean low-FPR detection using TPR@5%FPR, while Table 1 reports the full results. SAC-Copula remains strongly detectable at low false-positive rates, despite replacing position-wise independent Gumbel perturbations with a locally correlated field.

This result should be interpreted together with the clean quality results in Table 1. SAC-Copula is not merely a stronger detector setting: it preserves a strong clean signal while avoiding much of the quality cost seen in stronger perturbation baselines.

Ablation on Detector Readout. To separate generation-side structure from detector-side readout, we also evaluate SAC-Copula with the legacy equal-weight z-score detector. Under this old readout, SAC-Copula improves AUC from 0.966 to 0.987 and TPR@FPR5 from 0.855 to 0.950 relative to i.i.d. Gumbel. The FFR readout further increases TPR@5%FPR to 0.985 and TPR@1%FPR to 0.980 by matching the local evidence geometry induced by SAC-Copula. This ablation supports the interpretation that the clean-detection gain is not solely a detector-side artifact. Appendix C.6, Table 12, reports the compact detector comparison.

4.4 Residual Evidence under Controlled Token Edits

We evaluate deletion, insertion, and substitution as controlled token-level stress tests. These edits are not used to claim adversarial robustness; they diagnose residual evidence under token corruption

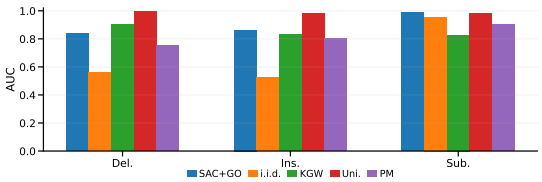

Figure 6: Compact AUC comparison at a fixed 20% token-edit rate. The panels report deletion, insertion, and substitution for SAC-Copula with GO-FFR, the direct i.i.d. Gumbel baseline, and representative external clean-text baselines. AUC is used as a threshold-agnostic residual-evidence diagnostic; the full detector/readout comparison is reported in Appendix C.7, Figure 11. This is a controlled edit diagnostic, not a broad robustness claim.

and synchronization drift. Figure 6 reports a compact AUC comparison at a fixed 20% edit rate, including SAC-Copula with GO-FFR, the direct i.i.d. Gumbel baseline, and representative external clean-text baselines. AUC is used as a threshold-agnostic residual-evidence diagnostic; stricter low-FPR TPR sweeps and the full detector/readout comparison are reported in Appendix C.7, Figure 11.

Deletion and insertion induce synchronization drift because the observed token index no longer matches the original tape position after accumulated edits. The full detector/readout comparison in Appendix C.7, Figure 11, shows that GO-FFR improves over fixed-alignment FFR and the old readout in these settings. However, the remaining gap and the low-FPR TPR curves in Appendix C.7 show that cumulative local drift is still unresolved. Substitution is less destructive because it corrupts token identities without shifting subsequent positions. External baselines provide contextual clean-text references in this diagnostic setting, not evidence of superiority over dedicated robust watermarking schemes.

Limitation Analysis. Deletion and insertion expose the main limitation: residual local evidence can remain, but the detector must still align observed tokens to the private tape. GO-FFR mitigates coarse global shifts but does not solve cumulative local drift. Because low-FPR TPR under token edits remains far below clean detection, we treat these results as stress-test diagnostics rather than evidence of edit robustness. Appendix C.7 reports the full attack-rate curves, and Appendix C.8 gives representative high-rate rows.

Semantic rewriting and document-level paraphrasing are stronger failures because they can replace, reorder, or regenerate the evidence-bearing sequence. Since SAC-Copula reads token-level evidence from a private Gumbel tape, these transformations break the fixed tape-readout assumption more directly than local token corruption. We therefore do not claim paraphrase robustness; these experiments define the method boundary rather than supporting the clean quality or clean detection claims (Krishna et al., 2023; Hou et al., 2024). Appendix C.8 reports the corresponding rows.

5 Discussion and Limitations

The main takeaway is that watermark smoothness matters in DLMs. SAC-Copula studies a DLM-specific design axis: the joint geometry of the Gumbel perturbation field. It preserves one-coordinate Gumbel marginals while introducing local dependence, but this marginal compatibility should not be read as invariance of the complete generated-text distribution.

The detector has the same boundary. FFR is a SAC-aware, covariance-matched readout of extracted evidence under a Gaussian mean-shift rationale, not an optimal detector for the full text distribution. GO-FFR is only a coarse-offset diagnostic for mild insertion/deletion settings; it is not a full edit-path alignment method, and cumulative local drift remains unresolved.

Several practical limitations remain. Detection depends on native-sample calibration and development data for the empirical direction, so deployment requires a reproducible calibration protocol and consistent access to the private tape construction. The current study is also limited to LLaDA-style masked refinement and one selected SAC configuration; broader DLM families, parameter sensitivity, and runtime cost require separate evaluation before deployment (Zhao et al., 2025).

SAC-Copula is therefore not a complete solution to adversarial edit robustness. Mild token-level edits are useful stress tests because they reveal residual evidence under limited desynchronization, but paraphrastic rewriting, public detectability, strong adversarial edits, and stealing attacks remain separate problems (Krishna et al., 2023; Hou et al., 2024; Zhang et al., 2025a; Jovanović et al., 2024). The contribution is a focused study of quality-preserving watermarking for diffusion language models through smooth, locally correlated Gumbel perturbation fields.

References

Scott Aaronson. 2023. Watermarking of large language models. https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17. Accessed: 2023-08.

Avi Bagchi, Akhil Bhimaraju, Moulik Choraria, Daniel Alabi, and Lav R. Varshney. 2025. Watermarking discrete diffusion language models. *Preprint*, arXiv:2511.02083.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *Preprint*, arXiv:2306.09194.

Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. 2023. Publicly-Detectable Watermarking for Language Models. Cryptology ePrint Archive, Paper 2023/1661.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *Preprint*, arXiv:1904.09324.

Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *Preprint*, arXiv:2403.04808.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. 2026. Watermarking diffusion language models. *Preprint*, arXiv:2509.24368.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2025a. Theoretically grounded framework for llm watermarking: A distribution-adaptive approach. *Preprint*, arXiv:2410.02890.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2025b. Universally optimal watermarking schemes for LLMs: from theory to practice.

Pyo Min Hong and Albert No. 2026. dgmark: Decoding-guided watermarking for diffusion language models. *Preprint*, arXiv:2601.22985.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *Preprint*, arXiv:2310.03991.

Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Watermark stealing in large language models. *Preprint*, arXiv:2402.19361.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Preprint*, arXiv:2303.13408.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Preprint*, arXiv:2307.15593.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. 2025. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1).

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *Preprint*, arXiv:2502.09992.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Preprint*, arXiv:2102.01454.

Ofek Raban, Ethan Fetaya, and Gal Chechik. 2026. Lr-dwm: Efficient watermarking for diffusion language models. *Preprint*, arXiv:2601.12376.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Preprint*, arXiv:2406.07524.

Chenrui Wang, Junyi Shu, Billy Chiu, Yu Li, Saleh Alharbi, Min Zhang, and Jing Li. 2025. Learning to watermark: A selective watermarking framework for large language models via multi-objective optimization. *Preprint*, arXiv:2510.15976.

Linyu Wu, Linhao Zhong, Wenjie Qu, Yuexin Li, Yue Liu, Shengfang Zhai, Chunhua Shen, and Jiaheng Zhang. 2025. Dmark: Order-agnostic watermarking for diffusion large language models. *Preprint*, arXiv:2510.02902.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing beyond identification: Multi-bit watermark for large language models. *Preprint*, arXiv:2308.00221.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2025a. Watermarks in the sand: Impossibility of strong watermarking for generative models. *Preprint*, arXiv:2311.04378.

Yangtian Zhang, Sizhuang He, Daniel Levine, Lawrence Zhao, David Zhang, Syed A Rizvi, Shiyang Zhang, Emanuele Zappala, Rex Ying, and David van Dijk. 2025b. Non-markovian discrete diffusion with causal language models. *Preprint*, arXiv:2502.09767.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *Preprint*, arXiv:2306.17439.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. 2025. Sok: Watermarking for ai-generated content. *Preprint*, arXiv:2411.18479.

A Watermark Properties

This appendix gives proof details behind the SAC-Copula field in Section 2.1. The claims are about the latent watermark tape. They are not claims about invariance of the complete output-text law.

A.1 Boundary Convention

The current convolution convention uses zero padding and divides by the global full-window energy

$$S_2 = \sum_{k=-h}^{h} K_W[k]^2.$$

Therefore exact unit variance is guaranteed for full-window interior positions. Boundary positions can have smaller variance because fewer nonzero taps contribute.

An all-position idealization would define

$$\mathcal{K}_t = \{k \in [-h, h] : 1 \le t - k \le L_{\max}\},$$
$$S_2(t) = \sum_{k \in \mathcal{K}_t} K_W[k]^2,$$

and divide the boundary convolution by $\sqrt{S_2(t)}$. In this paper, strict marginal claims are made under the full-window or boundary-normalized convention. Boundary deviations under zero padding are treated as implementation checks rather than as part of the exact marginal claim.

A.2 Marginal Gaussian and Gumbel Exactness

At a full-window position,

$$Z_{\text{spatial}}(t, j) = \sum_{k=-h}^{h} K_W[k] Z_{\text{white},1}(t - k, j).$$

This is a linear combination of independent Gaussian variables, hence Gaussian. Since the source variables have unit variance,

$$\text{Var}(Z_{\text{spatial}}(t, j)) = \sum_{k=-h}^{h} K_W[k]^2 = S_2.$$

Thus

$$\widetilde{Z}_{\text{spatial}}(t, j) = \frac{Z_{\text{spatial}}(t, j)}{\sqrt{S_2}} \sim \mathcal{N}(0, 1).$$

The same proof applies to the boundary-normalized ideal form after replacing S_2 by $S_2(t)$.

The white stream and spatial stream are independent. Therefore

$$Z_{\text{final}}(t, j) = \sqrt{1 - \rho^2} Z_{\text{white},2}(t, j) + \rho \widetilde{Z}_{\text{spatial}}(t, j)$$

is Gaussian with zero mean and variance $(1 - \rho^2) + \rho^2 = 1$. Hence each valid coordinate of Z_{final} is $\mathcal{N}(0, 1)$.

By the probability integral transform,

$$U(t, j) = \Phi(Z_{\text{final}}(t, j)) \sim \text{Uniform}(0, 1).$$

Applying the inverse standard Gumbel CDF gives

$$G_{\text{corr}}(t, j) = F_G^{-1}(U(t, j)) \sim \text{Gumbel}(0, 1).$$

For the standard Gumbel distribution, $F_G^{-1}(u) = -\log(-\log u)$. In numerical evaluation, U is clamped to avoid floating-point extremes, so exact unbounded Gumbel statements refer to the ideal transform.

A.3 Finite-Window Latent Autocorrelation

Let

$$b_k = \frac{K_W[k]}{\sqrt{S_2}}, \qquad b_k = 0 \quad \text{for } k \notin [-h, h].$$

Then

$$\widetilde{Z}_{\text{spatial}}(t, j) = \sum_k b_k Z_{\text{white},1}(t - k, j).$$

Expanding the covariance and using independence of the white source gives the finite-window autocorrelation

$$r_\tau(W, \sigma_K) = \sum_k b_k b_{k+\tau}.$$

Thus $r_0 = 1$, $r_{-\tau} = r_\tau$, and $r_\tau = 0$ for $|\tau| > W - 1$. For the final latent field and $\tau \neq 0$,

$$\begin{aligned}&\mathrm{Cov}(Z_{\text{final}}(t,j), Z_{\text{final}}(t+\tau,j))\\&\quad = \rho^2 r_\tau(W, \sigma_K),\end{aligned}$$

because the white stream contributes no cross-position covariance. This latent autocorrelation motivates the detector-side H_ω template, while the detector covariance itself is estimated from extracted Native evidence.

For the lag-one local-difference proxy used in Section 3.3, the unit marginal variance and the covariance above give

$$\begin{aligned}\mathcal{D}_{\text{SAC}} &= \mathbb{E}\Big[(Z_{\text{final}}(t+1,j) - Z_{\text{final}}(t,j))^2\Big]\\&= 2 - 2\rho^2 r_1(W, \sigma_K) = 2(1 - \rho^2 r_1).\end{aligned}$$

Thus the Section 3.3 expression is a direct consequence of the finite-window autocorrelation and the preserved one-coordinate latent variance.

A.4 Latent Tape KL Proxy

Let $P_0 = \mathcal{N}(0, I_T)$ be an i.i.d. latent Gaussian tape and $P_\rho = \mathcal{N}(0, \Sigma_T)$ the SAC latent Gaussian tape, where

$$\begin{aligned}\Sigma_T &= (1-\rho^2) I_T + \rho^2 R_T,\\[ab]_{ab} &= r_{a-b}(W, \sigma_K).\end{aligned}$$

Assume $\Sigma_T \succ 0$, so the log-determinant is well-defined. This holds for $\rho < 1$ when R_T is positive semidefinite; the endpoint $\rho = 1$ additionally requires nonsingularity of R_T.

For zero-mean Gaussians,

$$D_{\text{KL}}(P_\rho \parallel P_0) = \frac{1}{2}\left(\operatorname{tr}\Sigma_T - T - \log\det\Sigma_T\right).$$

Since $\operatorname{diag}(\Sigma_T) = 1$, $\operatorname{tr}\Sigma_T = T$, and

$$\mathcal{L}^{(T)}_{\text{tape}} = \frac{1}{T} D_{\text{KL}}(P_\rho \parallel P_0) = -\frac{1}{2T}\log\det\Sigma_T.$$

This is a latent process diagnostic, not semantic KL and not output-text KL.

This expansion is an explanatory approximation obtained from the log-determinant series around the identity covariance. Set $A = R_T - I_T$. Since $\Sigma_T = I_T + \rho^2 A$ and $\operatorname{tr}(A) = 0$,

$$\mathcal{L}^{(T)}_{\text{tape}} = \frac{\rho^4}{4T}\operatorname{tr}(A^2) + O(\rho^6).$$

For the Toeplitz finite-window case,

$$\operatorname{tr}(A^2) = \sum_{\tau \neq 0} (T - |\tau|) r_\tau(W, \sigma_K)^2.$$

The leading term is $O(\rho^4)$ because the diagonal marginal variances are preserved and the trace term cancels.

A.5 Implementation Checks

Implementation checks should include:

- marginal Gaussian and Gumbel QQ plots;
- empirical versus theoretical autocorrelation;
- low-frequency energy of the noise tape;
- effective sample size estimates;
- boundary variance curves.

For a diagnostic sequence a, with discrete Fourier transform $\widehat{a}(f)$, define the low-frequency energy fraction as

$$\mathrm{LFE}(a) = \frac{\sum_{f \in \mathcal{F}_{\text{low}}} |\widehat{a}(f)|^2}{\sum_{f \in \mathcal{F}} |\widehat{a}(f)|^2}.$$

These checks validate the noise-tape construction, not the distributional invariance of generated text.

B Detector Rationale

This appendix expands the detector rationale in Section 2.3. It clarifies why the detector is matched to the evidence geometry induced by the SAC-Copula perturbation field.

B.1 From i.i.d. Evidence to Covariance-Aware Calibration

The old detector aggregates raw evidence with an equal-weight i.i.d. Z-score:

$$Z_{\text{old}} = \frac{\mathbf{1}^\top(\mathbf{s} - \mu_G \mathbf{1})}{\sigma_G \sqrt{T}}.$$

It implicitly assumes

$$\Sigma_{x,0} \approx \sigma_G^2 I, \qquad \Delta_x \propto \mathbf{1}.$$

For non-i.i.d. evidence, however, the null variance of the equal-weight sum is

$$\mathrm{Var}_0(\mathbf{1}^\top \mathbf{x}) = \mathbf{1}^\top \Sigma_{x,0} \mathbf{1}.$$

Positive local dependence increases this variance relative to the i.i.d. value, reducing effective sample size. HAC or covariance-aware calibration corrects this null variance while keeping the evidence direction $\mathbf{1}$. If a calibration-only score is a positive rescaling,

$$Z_{\text{h0_hac}} = c_d Z_{\text{old}}, \qquad c_d > 0,$$

then within the same detector group and fixed calibration setting it does not change ROC ordering or AUC. This explains why calibration alone is insufficient when the useful watermark shift is not aligned with $\mathbf{1}$.

B.2 Why Filtering is Needed for SAC-Copula Evidence

SAC-Copula injects a locally correlated, low-lag perturbation field. The resulting evidence is not merely a uniform mean shift across positions. Full Filtered Ridge maps centered raw evidence into a SAC-aligned filtered feature:

$$\mathbf{u} = H_\omega \mathbf{x}.$$

The filter H_ω is motivated by the autocorrelation envelope of the SAC kernel, using weights proportional to $\exp(-\tau^2/(4\sigma_K^2))$ with boundary renormalization. Concretely, with

$$a_\tau = \exp\left(-\frac{\tau^2}{4\sigma_K^2}\right)\mathbf{1}\{|\tau| \le B\},$$

a row-normalized filtered readout can be written as

$$[H_\omega]_{t,s} = \frac{a_{s-t}}{\sum_{r:\,1\le t+r\le T} a_r}.$$

where $a_r = 0$ outside $[-B, B]$. This filter is a detector-side transform, not generation noise.

Native text and the decoding process may also introduce background correlations. Therefore the latent tape covariance should not be used directly as the detector's null covariance. The detector estimates the observed-evidence Native covariance and projects it into filtered space:

$$\widehat{\Sigma}_{u,0} = H_\omega \widehat{\Sigma}_{x,0} H_\omega^\top.$$

Thus FFR combines a SAC-motivated feature map with Native H_0 observed-evidence covariance whitening.

B.3 Ridge-Regularized Matched Direction

In the detector statistic, raw evidence is centered before filtering, so $\mathbf{u} = H_\omega \mathbf{x}$ is already H_0-centered up to finite-sample error. The surrogate $T_w(\mathbf{u}) = w^\top(\mathbf{u} - \boldsymbol{\mu}_{u,0})$ writes the same idea in abstract filtered-space notation.

In filtered feature space, use the Gaussian mean-shift surrogate

$$\begin{aligned} H_0 :&\quad \mathbf{u} \sim (\boldsymbol{\mu}_{u,0}, \Sigma_{u,0}),\\ H_1 :&\quad \mathbf{u} \sim (\boldsymbol{\mu}_{u,0} + \boldsymbol{\Delta}_u, \Sigma_{u,0}). \end{aligned}$$

For a linear score

$$T_w(\mathbf{u}) = w^\top(\mathbf{u} - \boldsymbol{\mu}_{u,0}),$$

the standardized separation is

$$\mathrm{SNR}(w) = \frac{w^\top \boldsymbol{\Delta}_u}{\sqrt{w^\top \Sigma_{u,0} w}}.$$

Let $b = \Sigma_{u,0}^{1/2} w$, so $w = \Sigma_{u,0}^{-1/2} b$, and

$$a = \Sigma_{u,0}^{-1/2} \boldsymbol{\Delta}_u.$$

Then

$$\begin{aligned} \mathrm{SNR}(w) &= \frac{b^\top a}{\|b\|_2}\\ &\le \|a\|_2. \end{aligned}$$

Equality holds when $b \propto a$, giving

$$w^\star \propto \Sigma_{u,0}^{-1} \boldsymbol{\Delta}_u.$$

The practical direction is the ridge-regularized plug-in estimate

$$w_\lambda = (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \widehat{\boldsymbol{\Delta}}_u.$$

Because the main detector centers evidence in raw space before filtering, the implemented scalar score can be written without an additional filtered-space mean subtraction:

$$Z_{\mathrm{FFR}}(y) = \frac{\widehat{\boldsymbol{\Delta}}_u^\top (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \mathbf{u}(y)}{\gamma_{\mathrm{FFR}}}.$$

For scores expressed directly in centered raw-evidence space, the corresponding linear direction is

$$\begin{aligned} \mathbf{q} &= \frac{H_\omega^\top w_\lambda}{\gamma_{\mathrm{FFR}}},\\ \gamma_{\mathrm{FFR}} &= \sqrt{\widehat{\boldsymbol{\Delta}}_u^\top (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \widehat{\boldsymbol{\Delta}}_u}. \end{aligned}$$

The ridge is finite-sample regularization for an empirical covariance inverse. The direction $\widehat{\boldsymbol{\Delta}}_u$ must be estimated from a development split or cross-fitting protocol without evaluation leakage. The optimality statement is limited to the linear detector class under the shared-covariance Gaussian surrogate.

B.4 Relation to the Original Detector

FFR reduces to the old detector under

$$H_\omega = I, \qquad \widehat{\Sigma}_{u,0} = \sigma_G^2 I,$$
$$\widehat{\boldsymbol{\Delta}}_u \propto \mathbf{1}, \qquad \lambda = 0,$$
$$\widehat{\boldsymbol{\mu}}_{x,0} = 0.$$

In this case the FFR direction is proportional to $\mathbf{1}$, and the score becomes the old equal-weight Gumbel evidence sum. FFR is therefore a covariance-aware, direction-aware, SAC-aware generalization of the original i.i.d. detector.

B.5 Global Offset Extension

FFR assumes fixed token-to-tape alignment $j = i$. Insertion and deletion can instead induce

$$j = i + \delta(i),$$

where $\delta(i)$ is an unknown drift. For an alignment A, the ideal alignment-aware centered evidence would be

$$x_i^A = s(y_{A(i)}; i) - \mu_G - \widehat{\mu}_{x,0,i},$$

with score

$$Z(A) = \mathbf{q}^\top \mathbf{x}^A.$$

Searching over full edit paths is outside the scope of GO-FFR. The extension restricts alignment to the global-offset family

$$\mathcal{A}_{\text{GO}} = \{A_d(i) = i + d : \ d \in \mathcal{D}\}.$$

The resulting score is

$$Z_{\text{GO-FFR}}(Y) = \max_{d \in \mathcal{D}} Z_d(Y),$$
$$Z_d(Y) = \mathbf{q}^\top \mathbf{x}^{(d)}(Y).$$

This is a GLRT-style scan over a small alignment family. The max over offsets inflates the Native H_0 right tail. The GO-specific threshold is the empirical $1 - \eta$ quantile of $Z_{\text{GO-FFR}}$ on Native H_0 samples under the same offset-scan rule. GO-FFR recovers only coarse global synchronization. It is not a full edit-path alignment algorithm, cannot solve cumulative local drift, and should not be presented as paraphrase robustness.

B.6 Scope and Limitations

FFR adapts detection to correlated SAC-Copula evidence. GO-FFR provides coarse synchronization recovery for mild insertion/deletion edits while reusing the same FFR scoring direction. Neither score is an optimal detector for the full text distribution, and neither addresses semantic paraphrase rewriting. GO-FFR should therefore be reported as an auxiliary edit analysis rather than as the paper's primary claim. All low-FPR metrics require detector-specific Native H_0 calibration, and $\widehat{\boldsymbol{\Delta}}_u$ must be estimated without evaluation leakage. Edit results should be framed as partial mild-edit behavior and failure analysis.

C Additional Experiment Evidence

This appendix records supporting clean-setting summaries and frozen attack-summary evidence used for the mild-edit and failure-analysis discussion in Section 4. The clean quality and clean detection claims in the main text are supported by the frozen clean multi-baseline sweep summaries, not by the attack-summary evidence. The attack sources are summary-only artifacts derived from token-level edit outputs and semantic/paraphrase attack summaries; no per-sample generated text, exact noise tapes, or full detector-score files are used here.

C.1 Clean Quality and Detection Details

This subsection records supporting material for the clean multi-baseline comparison in Section 4.2 and Section 4.3. The source tables are frozen clean-text sweep summaries used to construct the main trade-off table, quality-score components, radar normalization, and heatmap points. These tables explain the compact quality score and visualization normalization; they do not introduce additional experimental claims beyond the clean-setting trade-off reported in the main text. Table 3 defines metric directions, Tables 6–7 report the full raw quality metrics, Tables 8–9 document the radar normalization, and Figures 7–8 provide compact clean quality–detection and clean detection views.

C.2 Full Experimental Configuration

This appendix records implementation-level details for reproducibility. Fields not exposed by the frozen scripts or summary metadata are marked as unavailable rather than inferred.

Baseline adaptation notes. KGW is adapted as a context-greenlist bias baseline with the configured context width, PRF, and detector listed in Table 4 and Table 5. Unigram is used as a static-greenlist clean-text trade-off baseline with its unique-token

Evidence item	Appendix location	Source / scope note
Full experimental setup	Tables 4–5	implementation scripts, run metadata, and frozen summary files.
Metric definitions and orientation	Table 3	Defines raw metrics, direction, and visualization-only quality score.
Full clean quality metrics	Tables 6–7	frozen clean multi-baseline quality summaries.
Radar normalization	Tables 8–9	Display normalization only; not primary evidence.
Clean quality–detection view	Figure 7	Compact appendix view of clean quality and detection trade-offs.
Unigram strength trade-off	Figure 9	Contrastive static-bias sweep.
SAC parameter sensitivity	Table 11	Raw-metric support for the selected SAC operating point.
Detector comparison	Table 12	Clean old-readout / FFR / GO-FFR sanity check.
Token-edit sweeps	Figures 12–14	AUC, TPR@FPR1, and TPR@FPR5 across 0–50% edit rates.
Failure analysis	Tables 14–15	Boundary evidence for edit drift and semantic rewriting.
Detector derivation	Appendix B	Full detector definitions are not duplicated in this experiment appendix.

Table 2: Experiment appendix evidence index. Each row points to the compact appendix evidence used to support Section 4. The index is organizational only and does not introduce new experimental values.

Metric	Definition / source	Direction	Use in this paper
PPL	Perplexity under the configured evaluator model.	Lower is better	Fluency and tail-stability evidence.
PPL-P90	90th-percentile conditional PPL.	Lower is better	Radar tail-stability axis.
Collapse / tail collapse	Fraction with PPL > 100.	Lower is better	Severe fluency-tail diagnostic.
SBERT	Sentence-embedding similarity to the Native reference.	Higher is better	Semantic preservation.
BERTScore	Token-level contextual similarity.	Higher is better	Semantic preservation in full quality table.
ROUGE-L	Longest-common-subsequence lexical overlap.	Higher is better	Radar lexical overlap axis.
Token-F1	Native token-set overlap F1.	Higher is better	Radar token-overlap axis.
MAUVE	Distributional similarity against Native reference generations.	Higher is better	Distributional fidelity.
Token TV / Token JS	Unigram token-distribution divergence from Native.	Lower is better	Token-distribution drift.
Top-k overlap	Overlap with Native top-k token set.	Higher is better	Token-distribution fidelity.
Rep3	Repeated 3-gram rate.	Lower is better	Repetition diagnostic.
Distinct3	Distinct 3-gram ratio.	Higher is better	Lexical diversity.
Ent3	3-gram entropy.	Higher is better	Diversity / spread diagnostic.
AUC	Threshold-independent detector ranking metric.	Higher is better	Clean and attack diagnostic detection.
TPR@FPR1 / TPR@FPR5	True-positive rate at 1% / 5% false-positive rate.	Higher is better	Low-FPR clean detection and attack diagnostics.
Quality	Direction-normalized visualization summary.	Higher is better	Descriptive radar/heatmap display only; not primary evidence.

Table 3: Metric definitions and orientations used in Section 4. Raw metrics are the primary evidence. The normalized Quality score is a descriptive display summary constructed from direction-normalized components and should not be interpreted as an additional ground-truth quality label.

detector. PatternMark is the MarkovPattern ReferenceLite variant used in the clean evaluation, not a claim of a complete independent reimplementation of every PatternMark setting. Detector derivations for the old readout, covariance-aware calibration, filtering, ridge matched direction, and GO-FFR are provided in Appendix B.1, Appendix B.2, Appendix B.3, and Appendix B.5; this experiment appendix only records how those detectors are used in the experiments.

Quality–detection heatmap. Figure 7 summarizes the same clean-setting trends using a compact heatmap. The quality score averages direction-normalized components for semantic preservation, MAUVE-based distributional fidelity, token-distribution fidelity, low repetition, lexical diversity, and PPL-tail stability. This visualization is intended as a compact summary of the quality–detection balance rather than a replacement for the raw metric values in Table 6.

The external-baseline clean detection sweep over $\delta \in \{1, 2, 3\}$ is summarized in the main compact TPR@FPR5 figure and fully visualized below in Figure 8. Figure 7 and Table 12 further connect the clean delta sweep to the raw quality tables and detector-readout sanity checks without duplicating the full CSV rows.

Full clean detection sweep. Figure 8 reports the full clean detection visualization for AUC, TPR@FPR1, and TPR@FPR5 across the external-baseline strength sweep. This appendix figure complements the compact main-text TPR@FPR5 view and does not introduce new experimental values.

C.3 Unigram Strength Trade-off

Figure 9 keeps the Unigram strength sweep as a contrastive reference for static-bias tuning. It is included in the appendix so that the main text can focus on the SAC-Copula clean quality and clean detection profile. The figure shows that stronger Unigram settings can improve clean detectability, but the same sweep changes semantic, distributional, token-drift, repetition, and diversity dimensions. This supports the main-text interpretation of Unigram as a trade-off baseline, not as the central object of study.

C.4 Traceable Clean-Generation Example

Table 10 provides one traceable qualitative example from the frozen generated-text JSONL artifact. The

Category	Field	Setting
Generation	Model name	`LLaDA-8B-Instruct`
Generation	Prompt source	`ELI5.jsonl`
Generation	Prompt format	LLaDA chat template
Generation	Split sizes	Native $H_0 = 500$, Delta-dev $= 200$, eval $= 200$
Generation	Operating GPU	NVIDIA RTX PRO 6000
Generation	Generation length	256 tokens
Generation	Block length	32
Generation	Diffusion steps	$\max(1, \lfloor 0.5\times \text{ generation length} \rfloor) = 128$
Generation	Temperature	1.0
Generation	top-p / top-k	not available in the frozen configuration fields
Generation	Mask id	126336
Generation	Seeds and key	fixed random seeds and a fixed private watermark key; exact values are kept in the implementation configuration
Generation	Cached artifacts	generated texts and exact perturbation tapes are cached for reproducibility in the implementation artifacts
Watermark	Main SAC-Copula	$W = 39, \sigma_K = 15.0, \rho = 0.6$
Watermark	SAC sensitivity rows	$W = 35, \sigma_K = 11.5, \rho = 0.7$; $W = 35, \sigma_K = 13.5, \rho = 0.7$; $W = 39, \sigma_K = 15.0, \rho = 0.6$
Watermark	Perturbation scale	$\alpha = 1.0$, passed as the generation temperature in the generation call
Watermark	Tape length	256 generated positions in the clean evaluation
Watermark	Exact tapes	saved under `exact_tapes`; missing-tape reconstruction disabled
Baselines	Direct mechanism baseline	i.i.d. position-indexed Gumbel
Baselines	External baselines	KGW context-green bias, Unigram static-green bias, PatternMark MarkovPattern ReferenceLite
Baselines	External sweep	$\gamma = 0.25, \delta \in \{1, 2, 3\}$ for KGW, Unigram, and PatternMark
Baselines	KGW details	conv-kernel PRF, kernel $[-1]$, context width 1, sumhash seeding, no-bias fallback
Baselines	Unigram details	static greenlist bias; unique-token detector is the main detector
Baselines	PatternMark details	MarkovPattern ReferenceLite, $k = 2$, pattern length 4, 10000 null samples

Table 4: Generation, watermark, and baseline configuration for the clean multi-baseline evaluation. Values are taken from the frozen generation/detection script and collected run metadata.

Category	Field	Setting
Detection	Canonical detector mapping	i.i.d. Gumbel: old z-score; KGW: KGWGreenlistZDetectorV2; Unigram: UnigramUniqueGreenlistDetectorV2; PatternMark: PatternMarkMarkovPatternDetector
Detection	SAC detector rows	old z-score, Full Filtered Ridge, Global Offset Full Filtered Ridge
Detection	FFR ridge	$\lambda = 10^{-4}$
Detection	Cross-fitting	2 folds with crossfit seed 20260506 for empirical direction estimation
Detection	Native H_0 calibration	detector-specific calibration on held-out Native H_0 samples
Detection	Target FPRs	1% and 5%
Detection	GO offset setting	max offset 96 and offset penalty 0.0; clean GO-FFR scoring uses offset 0
Detection	Tape device	generation tape on runtime device; detection tape on CPU
Quality	PPL model	Llama-3.1-8B-Instruct evaluator checkpoint
Quality	PPL collapse threshold	PPL > 100
Quality	PPL trim quantile	0.95
Quality	SBERT model	all-MiniLM-L6-v2 sentence-embedding model
Quality	BERTScore model	`roberta-large`, English, no baseline rescaling
Quality	MAUVE setup	Native reference group, `gpt2-large` features, max text length 256, 10 buckets
Quality	Token drift	unigram JS, TV, and top-k overlap for $k \in \{50, 100, 500, 1000\}$
Quality	Repetition/diversity	Rep3, Distinct3, Ent3, with Self-BLEU 3/4 available in multidimensional quality metadata
Quality	Normalized quality score	descriptive direction-normalized summary for visualization, not a primary metric

Table 5: Detector, calibration, and evaluator configuration for the clean multi-baseline evaluation. Values are taken from the implementation scripts, collected run metadata, and existing summary tables.

example is included only to make the automatic quality metrics more concrete; it is not used as primary evidence and is not selected to prove a general qualitative claim.

C.5 SAC Parameter Sensitivity

This subsection is included to justify the SAC operating point used in the main experiments, rather than to claim that one configuration dominates every raw metric. The three SAC parameters have distinct roles: W controls the correlation window or range, σ_K controls the kernel decay and smoothness, and ρ controls the strength of the correlated branch. Table 11 compares several SAC configurations under the same clean-text evaluation slice so that the selected main-text setting is not treated as an arbitrary default.

Detection metrics come from the FFR clean-detection summary, and quality metrics come from the multidimensional quality and token-drift summaries in the shared clean-text evaluation slice indexed as $\delta = 1$ for the external-baseline strength sweep. This δ index is not a SAC parameter; the SAC parameters are W, σ_K, ρ. The normalized quality score is not included in these delta-sweep summary files, so it remains marked as unavailable rather than reconstructed, and the interpretation should rely on the raw listed metrics.

The table should be read as a compact trade-off summary. The $W = 35, \rho = 0.7$ variants can improve some central fluency or collapse indicators, especially for $W = 35, \sigma_K = 13.5, \rho = 0.7$, but they also shift diversity, token-drift, and low-FPR detection behavior. The selected $W = 39, \sigma_K = 15.0, \rho = 0.6$ setting is used because it retains strong low-FPR detection while maintaining favorable MAUVE, repetition/diversity, and token-drift behavior; it is a balanced operating point rather than a raw-metric dominance claim.

This appendix evidence has a limited scope. It is not a full hyperparameter search and does not prove the DLM refinement mechanism by itself.

Method	Strength	Mean PPL	Median PPL	Collapse	SBERT	BERTScore	MAUVE	Token TV
Native	–	16.840	6.381	0.005	1.000	1.000	1.000	0.000
i.i.d. Gumbel	–	636.069	8.446	0.035	0.784	0.868	0.991	0.163
KGW	$\delta = 1$	5.32×10^4	6.697	0.030	0.794	0.869	0.999	0.146
KGW	$\delta = 2$	6.98×10^4	8.020	0.045	0.776	0.867	0.998	0.173
KGW	$\delta = 3$	8.30×10^4	10.543	0.050	0.762	0.863	0.982	0.228
Unigram	$\delta = 1$	7.714	6.145	0.000	0.803	0.867	0.981	0.233
Unigram	$\delta = 2$	7.630	6.535	0.000	0.755	0.851	0.841	0.360
Unigram	$\delta = 3$	9.464	6.585	0.005	0.690	0.836	0.601	0.439
PatternMark	$\delta = 1$	3883.179	6.766	0.010	0.781	0.869	0.983	0.153
PatternMark	$\delta = 2$	3898.094	7.263	0.015	0.779	0.866	0.996	0.161
PatternMark	$\delta = 3$	3901.117	8.938	0.020	0.768	0.863	0.982	0.198
SAC-Copula	main	12.164	8.260	0.010	0.798	0.868	0.999	0.163

Table 6: Full automatic quality metrics for the clean multibase comparison, part 1. The table reports fluency and stability, semantic preservation, distributional fidelity, and token-distribution drift. Lower values are better for PPL, collapse, and token TV; higher values are better for semantic similarity and MAUVE.

Method	Strength	Token JS	Top50	Top100	Rep3	Distinct3	Ent3	Quality
Native	–	0.000	1.000	1.000	0.406	0.594	3.796	0.960
i.i.d. Gumbel	–	0.059	0.940	0.830	0.415	0.585	3.680	0.661
KGW	$\delta = 1$	0.057	0.920	0.900	0.414	0.586	3.798	0.694
KGW	$\delta = 2$	0.060	0.900	0.850	0.450	0.550	3.555	0.553
KGW	$\delta = 3$	0.068	0.860	0.830	0.510	0.489	3.179	0.379
Unigram	$\delta = 1$	0.074	0.860	0.760	0.395	0.605	3.967	0.798
Unigram	$\delta = 2$	0.120	0.660	0.600	0.455	0.545	3.819	0.545
Unigram	$\delta = 3$	0.164	0.600	0.430	0.560	0.440	3.408	0.150
PatternMark	$\delta = 1$	0.058	0.940	0.840	0.426	0.574	3.686	0.721
PatternMark	$\delta = 2$	0.061	0.900	0.800	0.422	0.578	3.809	0.714
PatternMark	$\delta = 3$	0.066	0.880	0.810	0.480	0.520	3.442	0.555
SAC-Copula	main	0.061	0.920	0.810	0.414	0.586	3.745	0.758

Table 7: Full automatic quality metrics for the clean multibase comparison, part 2. Lower values are better for token JS and repetition; higher values are better for vocabulary-overlap, distinctness, entropy, and the normalized quality score.

It supports the narrower claim that the reported SAC-Copula configuration is a deliberate quality–detectability operating point rather than an arbitrary default.

C.6 Detector Comparison for SAC Evidence

This subsection is included to separate generation-side structure from detector-side readout in the clean setting. The old detector is the i.i.d. equal-weight baseline readout, while FFR is the SAC-aware filtered statistic used to read the local low-lag evidence geometry induced by SAC-Copula. GO-FFR is included only for continuity with the mild insertion/deletion stress tests; it is not introduced here as a separate clean-detection advantage.

Table 12 reports the available clean detector comparison rows from the clean detection summary. The comparison should be read in two steps: if SAC evidence already improves over i.i.d. Gumbel under the old detector, that supports a generation-side signal difference; if FFR improves further on SAC evidence, that supports a matched readout for SAC structure. This appendix therefore serves as a sanity check for the clean-detection result rather than as a new main-text experiment or a detector-only claim.

The clean fixed-alignment setting also explains why GO-FFR should match or closely track FFR here: offset recovery is not needed when alignment is intact. GO-FFR is included to connect the clean setting to the controlled token-edit stress tests, not to claim an additional clean-detector gain. More broadly, this table is not a robustness proof, not a statement of global FFR optimality, and not evidence that synchronization is solved by GO-FFR.

Radar axis	Raw metric source	Original direction	Normalized score meaning
SBERT	SBERT	Higher is better	Embedding-level semantic closeness to Native text.
ROUGE-L	Mean ROUGE-L vs. Native	Higher is better	Lexical-sequence overlap with Native generations.
Token-F1	Native token-set F1	Higher is better	Native token-set overlap.
Rep3	Mean Rep3	Lower is better	Lower repeated 3-gram rate after inverted min–max normalization.
Collapse>100	CollapseRate_GT100	Lower is better	Fewer severe PPL-tail collapse cases after inverted min–max normalization.
PPL-P90	Conditional PPL P90	Lower is better	Lower 90th-percentile conditional perplexity after inverted min–max normalization.

Table 8: Metric orientation used in the v3.8 quality radar. All radar axes are direction-normalized so that larger values indicate better or more Native-like behavior. Lower-is-better metrics, including repetition, collapse rate, and PPL-P90, are inverted before normalization. The normalization is used only for visualization; raw quality metrics remain the primary evidence in Table 1, Table 6, and Table 7.

Method	SBERT	ROUGE-L	Token-F1	Rep3	Collapse>100	PPL-P90
Native	1.000	1.000	1.000	1.000	1.000	1.000
i.i.d. Gumbel	0.869	0.948	0.767	0.992	0.333	0.830
KGW $\delta = 3$	0.666	0.734	0.562	0.340	0.080	0.080
PatternMark $\delta = 3$	0.721	0.764	0.677	0.549	0.667	0.788
Unigram $\delta = 3$	0.080	0.080	0.080	0.080	1.000	1.000
SAC-Copula	1.000	1.000	1.000	1.000	0.889	0.919

Table 9: Display-normalized scores used in the v3.8 quality radar. A small display floor is applied to low-scoring axes to avoid degenerate polygons in the visualization. These scores are descriptive display values, not additional primary quality metrics.

C.7 Controlled Token-Level Edit Details

Table 13 and Figures 12–14 support Section 4.4. These results are intended as controlled token-edit diagnostics rather than a claim of adversarial robustness. Across selected deletion, insertion, and substitution settings, the SAC-aware detector retains partial residual evidence, and the Global Offset variant provides a coarse realignment benefit. The full sweep figures report AUC, TPR@FPR1, and TPR@FPR5 across 0–50% attack rates. Low-FPR TPR remains substantially harder than AUC under deletion and insertion, which supports the failure analysis rather than a broad edit-robustness claim. These conclusions come from the frozen token-edit summaries, relative-gain, and global-offset summaries. They should not be inferred from the multi-baseline clean-text results.

The 20% cross-method AUC comparison is shown in Figure 6, with the full detector/readout comparison in Figure 11. Table 13 focuses on the detector-readout rows used to interpret SAC OLD, FFR, and GO-FFR behavior. Substitution does not induce cumulative position drift, so the fixed-alignment rows are the relevant diagnostic; no separate GO-FFR substitution row is needed for the synchronization-drift interpretation.

C.8 Failure Analysis under Edit Drift

Table 14 and Table 15 support the failure-analysis discussion in Section 4. The high-rate edit results expose the limitation behind the token-edit diagnostics: deletion and insertion accumulate synchronization drift. Global Offset scanning can recover coarse shifts, but it does not solve cumulative local drift, leaving edit-path alignment as an open limitation of the current detector. The relevant frozen attack-summary artifacts are the relative-drop, token-attack, and global-offset summaries.

Semantic and paraphrase rewriting are treated as failure-analysis evidence, not as a successful robustness setting. In the frozen semantic/paraphrase summary, the relevant summaries cover paraphrase results, valid-only filtering, realized TER bins, relative gains, and detector-only comparisons. These summaries show that semantic rewriting degrades detection and that offset-style diagnostics are only partial attempts, not a solution to semantic rewriting.

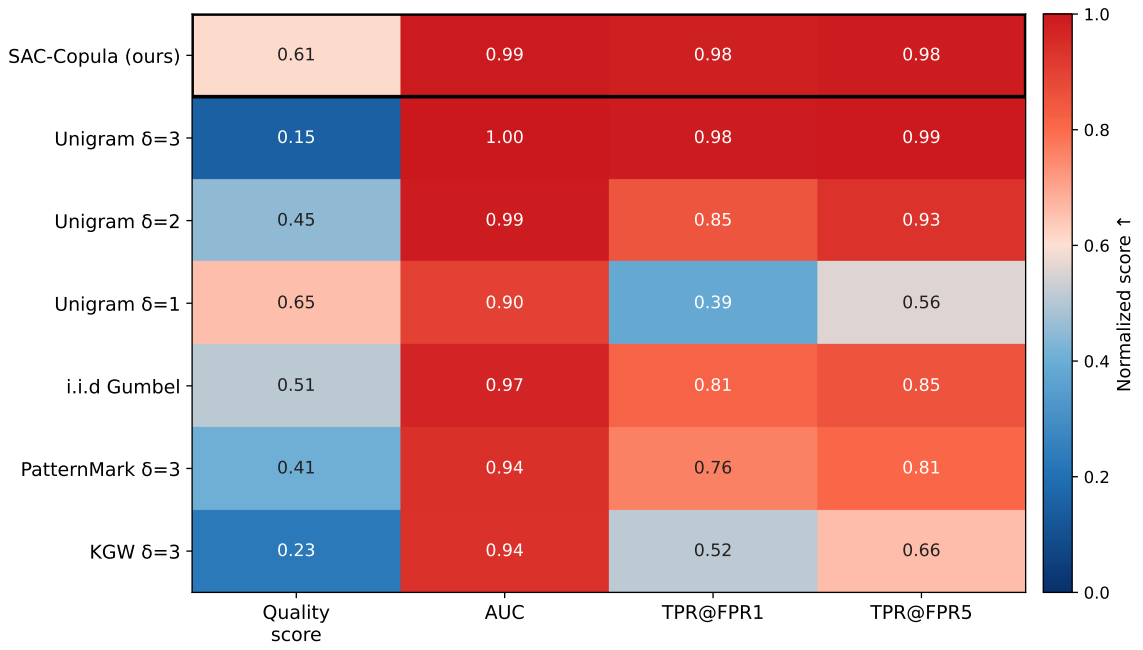

Figure 7: Quality–detection heatmap for the clean multibase comparison. Rows correspond to representative watermark settings, and columns show the normalized quality-preservation score, AUC, TPR@FPR1, and TPR@FPR5. The quality score averages semantic preservation, MAUVE, token-distribution fidelity, low repetition, lexical diversity, and tail stability after direction-aware normalization. SAC-Copula maintains a favorable balance between clean detectability and quality preservation compared with high-strength external baselines.

System	Output excerpt	
Native	`Eval|Native|701|256`	Dollar stores turn a profit primarily selling items at prices of only a dollar each. This allows them to sell a large quantity of items at a...
i.i.d. Gumbel	`Eval|i.i.d_Baseline|701|256`	Dollar stores turn a profit primarily by selling low-cost items at a lower price point. These items are usually bulk from larger retailers a...
SAC-Copula	`Eval|SAC_W39_S15.0_R0.6|701|256`	Dollar stores turn a profit through selling a wide variety of items at very low prices. They can offer a wide selection of items in from foo...
KGW	`Eval|KGW_ContextGreen_Bias|701|256`	Dollar stores turn a profit by by offering items at discounted prices, which allows them to sell more items per dollar spent. This is done b...
Unigram	`Eval|Unigram_StaticGreen_Bias|701|256`	Dollar stores turn a profit by selling a wide range of goods at low prices. They compete by offering a diverse selection of goods, including...
PatternMark ReferenceLite	`Eval|PatternMark_MarkovPattern_ReferenceLite|701|256`	Dollar stores make a profit through a combination of factors. 1. Lower pricing: Dollar stores sell items at very low prices, often as little...

Table 10: Traceable clean-generation example for the prompt "How do dollar stores turn a profit?" Excerpts are copied from the frozen clean-generation artifact and truncated for space. The table is illustrative only and does not replace the automatic quality metrics.

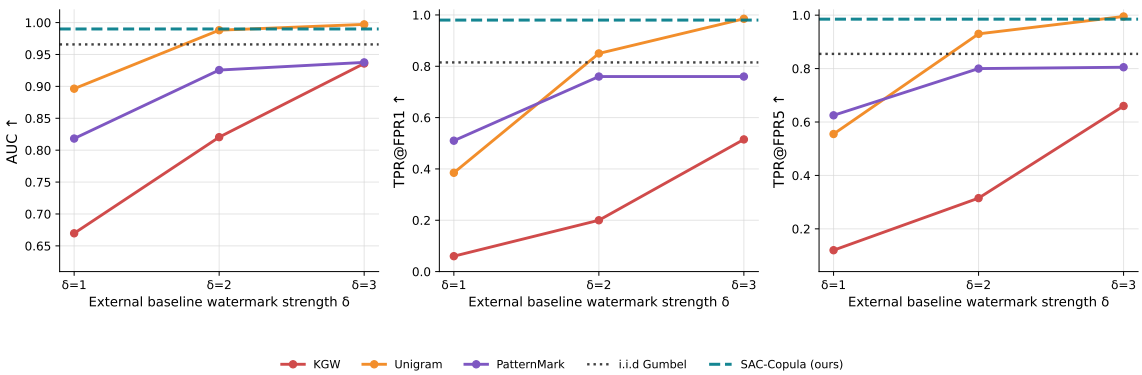

Figure 8: Full clean detection sweep across watermark strengths. The panels report AUC, TPR@FPR1, and TPR@FPR5 for KGW, Unigram, and PatternMark over $\delta \in \{1, 2, 3\}$, with i.i.d. Gumbel and SAC-Copula shown as reference operating points. This figure reports clean fixed-alignment detection only.

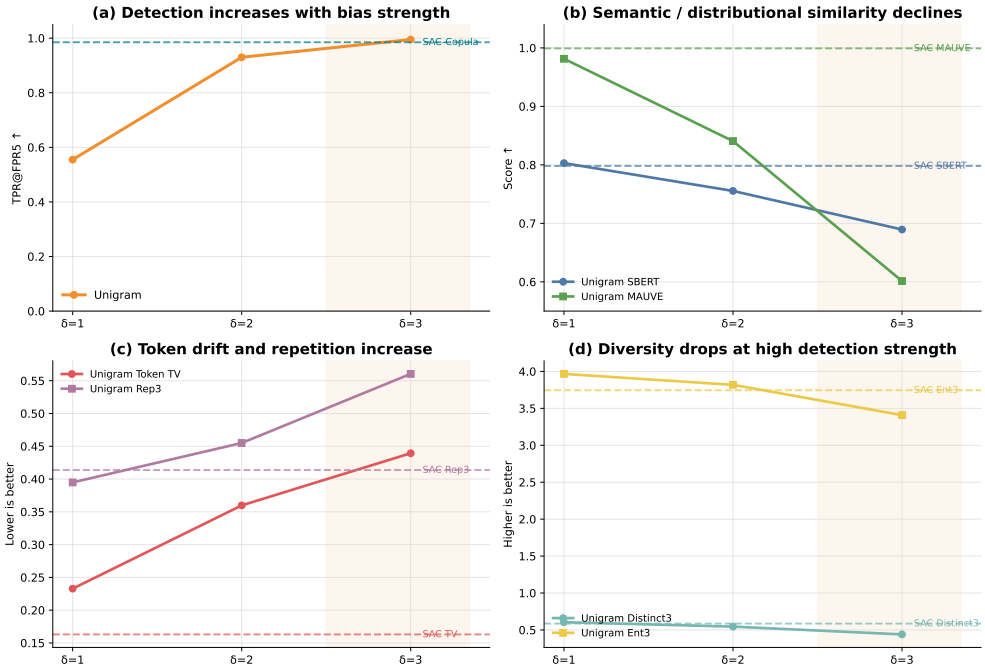

Figure 9: SAC-Copula avoids the high-strength static-bias quality trade-off. The Unigram strength sweep is included as a contrastive reference: Panel (a) shows clean TPR@FPR5 increasing with Unigram strength δ, while Panels (b)–(d) show the accompanying semantic, distributional, token-drift, repetition, and diversity costs. SAC-Copula is shown as the reference operating point.

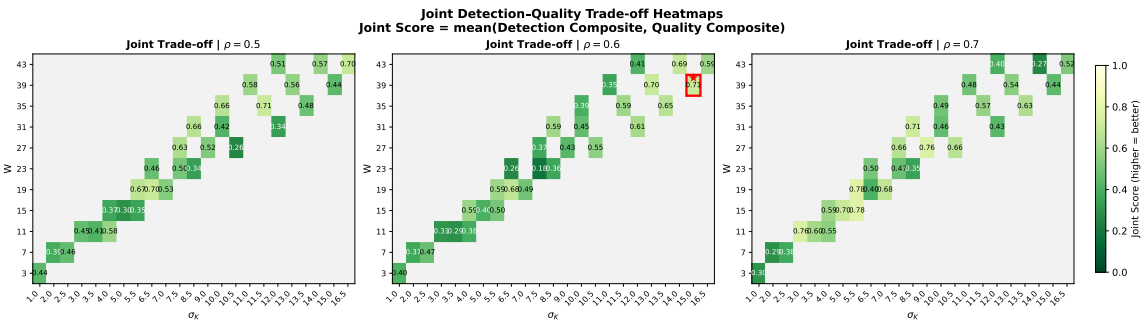

Figure 10: SAC parameter sensitivity for clean operating-point selection. The heatmap groups configurations by ρ, with W and σ_K on the axes and color indicating the joint quality–detectability score. The selected setting $W = 39, \sigma_K = 15.0, \rho = 0.6$ is used as the main SAC-Copula configuration. This visualization guides operating-point selection and is not a claim of global optimality or raw-metric dominance.

Configuration	W	σ_K	ρ	AUC	TPR@1%	TPR@5%	Mean PPL	Collapse	Quality	MAUVE	Rep3	Distinct3	Token TV
SAC W35 S11.5 R0.7	35	11.5	0.7	0.984	0.895	0.965	13.240	0.005	–	0.998	0.443	0.557	0.184
SAC W35 S13.5 R0.7	35	13.5	0.7	0.991	0.950	0.970	11.934	0.000	–	0.980	0.446	0.554	0.189
Selected SAC-Copula	39	15.0	0.6	0.990	0.980	0.985	12.163	0.010	–	0.999	0.414	0.586	0.163

Table 11: Compact appendix summary of SAC parameter sensitivity under clean DLM generation. The selected SAC-Copula row corresponds to the main-text configuration $W = 39, \sigma_K = 15.0, \rho = 0.6$. The normalized composite Quality score is unavailable in the delta-sweep summary CSVs used for this parameter table, so the Quality column is left blank and the table should be interpreted through the listed raw metrics. The $\delta = 1$ data slice is a shared clean-text evaluation slice, not a SAC parameter.

Method / group	Detector	AUC	TPR@1%	TPR@5%	Role
i.i.d. Gumbel	old z-score	0.966	0.815	0.855	Legacy baseline detector.
SAC-Copula	old z-score	0.987	0.915	0.950	Sanity check with the old readout.
SAC-Copula	FFR	0.990	0.980	0.985	Main SAC-aware clean readout.
SAC-Copula	GO-FFR	0.990	0.980	0.985	Offset-enabled variant; same clean score here.
SAC W35 S11.5 R0.7	FFR	0.984	0.895	0.965	Parameter-sweep SAC readout.
SAC W35 S13.5 R0.7	FFR	0.991	0.950	0.970	Parameter-sweep SAC readout.

Table 12: Compact appendix summary of available clean detector comparison rows for i.i.d. Gumbel and SAC-Copula evidence. FFR and GO-FFR read the local correlated evidence generated by SAC, while the old-detector rows provide sanity checks against the legacy i.i.d. readout.

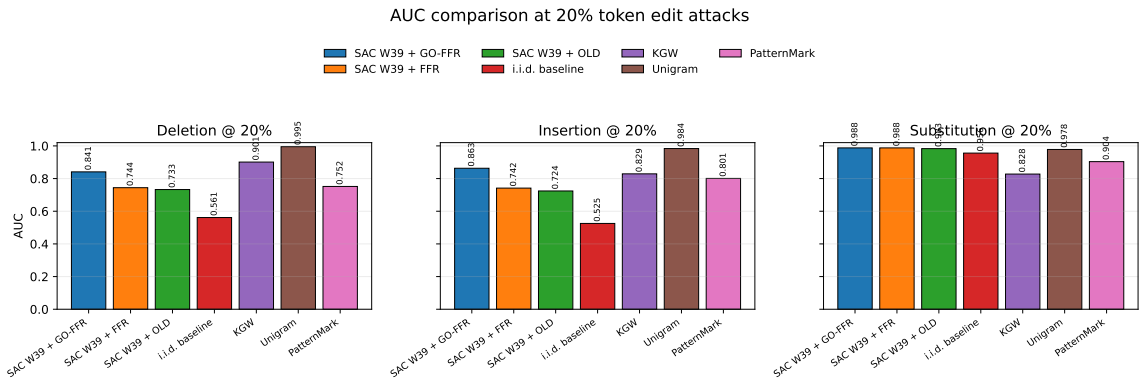

Figure 11: Full AUC comparison under 20% token-level edit attacks. The three panels report deletion, insertion, and substitution for SAC-Copula readouts, the direct i.i.d. Gumbel baseline, and external clean-text baselines. This full comparison complements the compact main-text AUC view and is reported as a residual-evidence diagnostic, not as a broad robustness claim.

Edit	Detector row	AUC	TPR@1%	TPR@5%	Best offset	Reading
Deletion 20%	i.i.d. old	0.561	0.025	0.110	0.000	Alignment-changing edit strongly weakens i.i.d. evidence.
Deletion 20%	SAC FFR	0.780	0.150	0.370	0.000	SAC-aware filtering retains partial residual evidence.
Deletion 20%	SAC GO-FFR	0.892	0.460	0.590	-27.295	Offset scanning partially recovers coarse shift.
Insertion 20%	i.i.d. old	0.525	0.010	0.045	0.000	Insertions disrupt fixed token-to-tape indexing.
Insertion 20%	SAC FFR	0.842	0.195	0.435	0.000	SAC-aware readout remains informative under controlled insertion.
Insertion 20%	SAC GO-FFR	0.884	0.505	0.670	18.230	Offset scanning gives a coarse synchronization gain.
Substitution 20%	i.i.d. old	0.956	0.755	0.825	0.000	Substitution is less damaging because sequence alignment is preserved.
Substitution 20%	SAC FFR	0.989	0.890	0.975	0.000	SAC evidence remains detectable in the indexed substitution setting.

Table 13: Selected low-rate token-level edit diagnostics from the frozen token-edit summary. These results are reported as partial residual-evidence diagnostics rather than broad robustness claims; the table is not evidence of full edit-path alignment or semantic rewrite robustness.

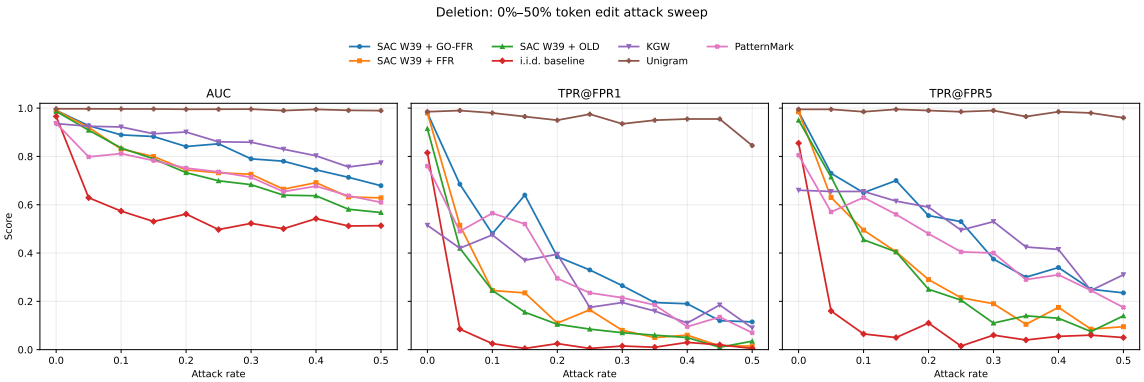

Figure 12: Deletion stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5 for SAC-Copula readouts, i.i.d. Gumbel, and external clean-text baselines. The curves are residual-evidence diagnostics: GO-FFR can partially recover coarse offset shifts, while low-FPR TPR remains limited under accumulated deletion drift.

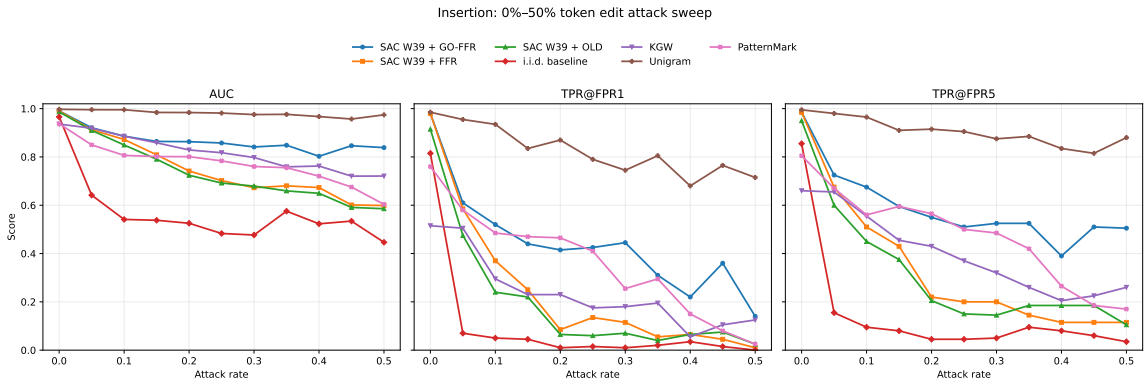

Figure 13: Insertion stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5. Insertion induces synchronization drift by shifting token-to-tape indexing; GO-FFR provides a coarse-offset diagnostic improvement but does not solve cumulative local drift.

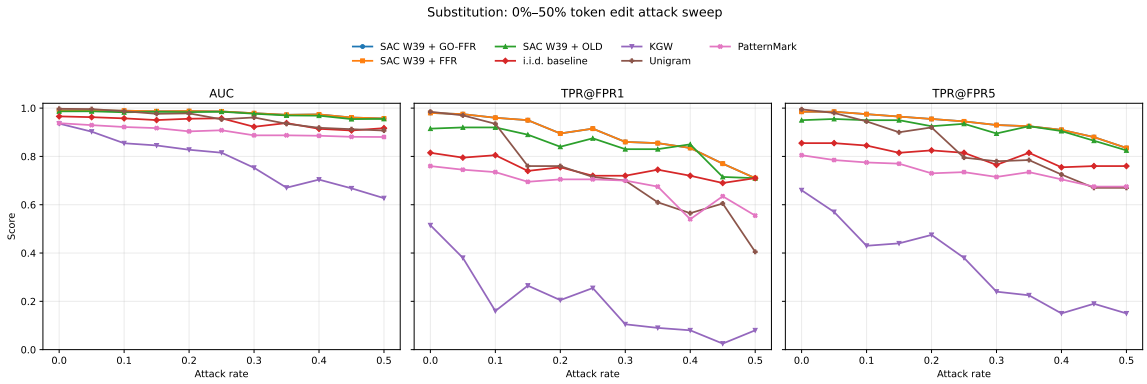

Figure 14: Substitution stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5. Substitution corrupts token identities without shifting all subsequent positions, so it is a controlled token-corruption diagnostic rather than a synchronization-drift setting.

Edit	Detector row	AUC	TPR@1%	TPR@5%	Best offset	Failure reading
Deletion 50%	i.i.d. old	0.513	0.005	0.050	0.000	Near-null detection after severe deletion drift.
Deletion 50%	SAC FFR	0.623	0.035	0.085	0.000	Fixed alignment is strongly degraded.
Deletion 50%	SAC GO-FFR	0.727	0.180	0.285	-51.945	Coarse offset helps but leaves large residual degradation.
Insertion 50%	i.i.d. old	0.447	0.000	0.035	0.000	High-rate insertion can invert or erase useful evidence.
Insertion 50%	SAC FFR	0.664	0.085	0.160	0.000	Fixed-alignment FFR remains limited under accumulated drift.
Insertion 50%	SAC GO-FFR	0.878	0.335	0.550	38.135	Offset scanning helps, but it is not full edit-path recovery.

Table 14: High-rate deletion and insertion diagnostics from the frozen token-edit summary. GO-FFR partially improves coarse synchronization, but cumulative local drift remains unresolved.

Attack	Detector row	AUC	TPR@1%	TPR@5%	Valid rate	Failure reading
DIPPER R20	i.i.d. row	0.580	0.075	0.105	0.545	Semantic rewrite substantially weakens i.i.d. evidence.
DIPPER R20	SAC row	0.745	0.130	0.260	0.495	SAC may retain partial signal, but detection is degraded.
DIPPER R30	i.i.d. row	0.526	0.005	0.070	0.330	Stronger rewriting approaches near-null detection.
DIPPER R30	SAC row	0.633	0.020	0.140	0.320	Residual signal remains limited under semantic rewriting.

Table 15: Semantic/paraphrase failure-analysis rows from the frozen semantic/paraphrase summary. These rows are included to define the method boundary and should not be interpreted as paraphrase robustness.