# DIFFSCORE: Text Evaluation Beyond Autoregressive Likelihood

**Wen Lai**[1,2*] **Yingli Shen**[3*] **Dingnan Jin**[1] **Qing Cui**[1] **Jun Zhou**[1†]
**Maosong Sun**[2] **Alexander Fraser**[4]

[1]Ant Group [2]Tsinghua University
[3]Beijing International Study University [4]Technical University of Munich
laiwen.lai@antgroup.com

## Abstract

Autoregressive language models are widely used for text evaluation, however, their left-to-right factorization introduces positional bias. Scoring early tokens under limited leftward context conflates architectural asymmetry with intrinsic text quality. To address this, we propose *masked reconstruction*, an alternative paradigm that evaluates every token using full bidirectional context, instantiated as DIFFSCORE via Masked Large Diffusion Language Models. By measuring text recoverability across continuous masking rates, DIFFSCORE eliminates positional bias and naturally establishes an evaluation hierarchy from local fluency to global coherence. Furthermore, it unlocks diagnostic tools unavailable to autoregressive approaches, including *multi-timestep quality profiles* and *bidirectional pointwise mutual information (PMI) decomposition* to disentangle source-conditioned faithfulness from source-independent fluency. Comprehensive experiments across ten benchmarks demonstrate that DIFFSCORE consistently outperforms autoregressive baselines in zero-shot and fine-tuned settings. Our code is available at https://github.com/wenlai-lavine/DiffScore.

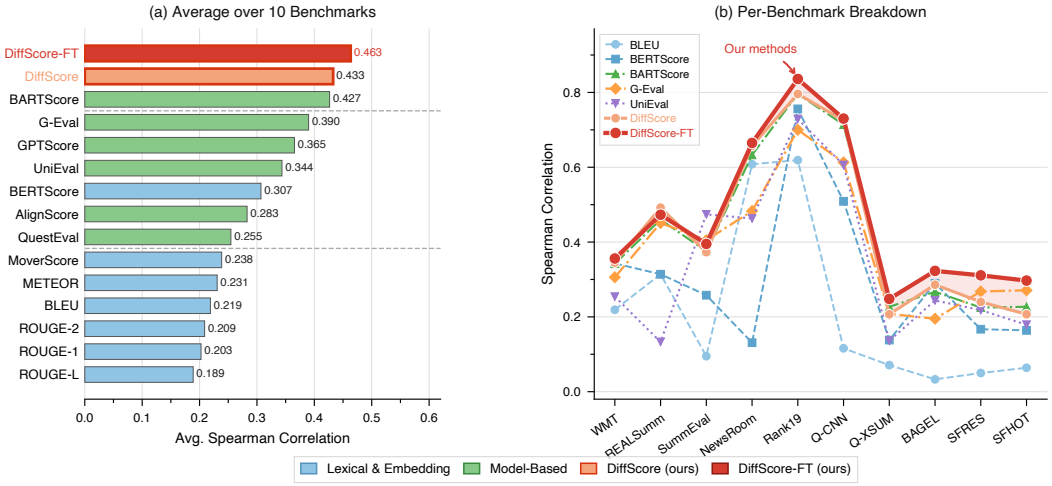



Figure 1: DIFFSCORE consistently outperforms all baselines across 10 diverse evaluation benchmarks.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

# 1 Introduction

Evaluating natural language generation (NLG) remains challenging due to high lexical variation among semantically equivalent outputs [1, 2]. Evaluation paradigms have evolved from $n$-gram overlap (e.g., BLEU [3], ROUGE [4]) and embedding matching (e.g., BERTScore [5], MoverScore [6]) to generative scoring (e.g., BARTScore [7], GPTScore [8]) and LLM-as-a-Judge methods (e.g., G-Eval [9]). While autoregressive (AR) scoring via conditional log-likelihood is principled, its left-to-right factorization conditions tokens solely on preceding context. This introduces *positional bias*, evaluating early tokens under information-impoverished contexts with lower signal-to-noise ratios (§6.4). Furthermore, the "Reversal Curse" [10], where AR models fail to infer "B is A" from "A is B," exposes an intrinsic directional asymmetry, suggesting that AR scores reflect architectural artifacts alongside text quality.

To overcome these limitations, we formulate text evaluation as *masked reconstruction*, measuring the ability to recover candidate text under varying context corruption. Intuitively, high-quality text conforms to linguistic regularities and is reliably reconstructible across arbitrary masking patterns. Masked Large Diffusion Language Models (MDLLMs) [11–13] provide an ideal backbone for this paradigm with three structural advantages. First, **bidirectionality** eliminates positional bias by conditioning predictions on bilateral context. Second, **multi-granularity signals** across continuous masking rates naturally form an evaluation hierarchy from local fluency to global coherence. Third, **native objective alignment** directly matches pretraining to the evaluation task.

We instantiate this paradigm as DIFFSCORE, a masked reconstruction NLG evaluation framework with two variants. **DIFFSCORE-Zero** performs zero-shot quality estimation via the MDLLM evidence lower bound (ELBO). **DIFFSCORE-FT** is fine-tuned on task corpora without human ratings, isolating the paradigm's intrinsic gains. To enhance interpretability, we introduce two diagnostic tools. First, *multi-timestep quality profiles* trace evaluation scores across masking rates. Second, *bidirectional pointwise mutual information (PMI) decomposition* disentangles source-conditioned faithfulness from source-independent recoverability (a fluency proxy). Unlike in AR frameworks, both terms are structurally symmetric under bidirectional scoring.

Extensive experiments across ten benchmarks in summarization, machine translation, and data-to-text generation validate our approach. As Figure 1 shows, DIFFSCORE-FT outperforms BARTScore-CNN in most settings, while DIFFSCORE-Zero surpasses zero-shot GPTScore, highlighting MDLLMs as potent out-of-the-box evaluators. Controlled comparisons matching model backbone (AR-LLaDA), parameter scale (0.6B), and latency budget (enriched BARTScore) confirm these gains stem directly from bidirectional masked reconstruction (§5.2).

Our contributions are summarized as follows. **(1)** We formulate NLG evaluation as masked reconstruction and introduce DIFFSCORE, an MDLLM-based framework mitigating the positional biases of AR scoring. **(2)** We propose two intrinsic diagnostic tools, multi-timestep quality profiles and bidirectional PMI decomposition, enabling fine-grained quality assessment beyond AR capabilities. **(3)** Built on open-weight MDLLMs, DIFFSCORE offers a competitive, reproducible alternative to proprietary LLM evaluators (e.g., G-Eval [9]) without commercial APIs or brittle prompt engineering.

# 2 Preliminaries

We first establish our notation and review the two foundations of our framework: autoregressive text evaluation (§2.1) and masked large diffusion language models (§2.2).

**Notation.** Let $\mathcal{V}$ denote a finite vocabulary, $\texttt{[M]} \notin \mathcal{V}$ be a mask token, and $\bar{\mathcal{V}} \triangleq \mathcal{V} \cup \{\texttt{[M]}\}$. A sequence of length $L$ is denoted by $\mathbf{x} = (x^1, \ldots, x^L) \in \mathcal{V}^L$. In text evaluation, we assess the quality of a *candidate* $\mathbf{c} = (c^1, \ldots, c^{L_c})$ given a *source* $\mathbf{s} = (s^1, \ldots, s^{L_s})$.

## 2.1 Autoregressive Text Evaluation

An autoregressive (AR) language model factorizes the sequence log-probability from left to right:

$$\log p_{\mathrm{AR}}(\mathbf{x}) = \sum_{n=1}^{L} \log p_{\mathrm{AR}}(x^n \mid x^{<n}). \tag{1}$$

BARTScore [7] adapts this for text evaluation by scoring a candidate $\mathbf{c}$ conditioned on a source $\mathbf{s}$:

$$\text{BARTSCORE}(\mathbf{c} \mid \mathbf{s}) = \sum_{n=1}^{L_c} \log p_{\text{AR}}(c^n \mid c^{<n}, \mathbf{s}), \tag{2}$$

where $p_{\text{AR}}$ is parameterized by a task-finetuned BART-large model [14]. Common variants include the *marginal* score $\text{BARTSCORE}(\mathbf{c})$ for fluency, the *reverse* score $\text{BARTSCORE}(\mathbf{s} \mid \mathbf{c})$ for coverage, and their *bidirectional* average for holistic quality. Despite its effectiveness, AR factorization introduces two structural limitations [10]. The first is *positional bias*, where unidirectional conditioning evaluates early tokens under information-impoverished contexts. The second is *directional asymmetry*, where the fixed generation order conflates architectural artifacts with intrinsic text quality.

### 2.2 Masked Large Diffusion Language Models

Masked Large Diffusion Language Models (MDLLMs) [11–13] define a discrete diffusion process over token sequences with `[M]` as an absorbing state.

**Forward (masking) process.** At a continuous timestep $t \in [0, 1]$, the forward process independently masks each token of a clean sequence $\mathbf{x}_0 \in \mathcal{V}^L$ with probability $t$:

$$q(x_t^i \mid x_0^i) = (1 - t)\mathbb{1}[x_t^i = x_0^i] \; + \; t\mathbb{1}[x_t^i = \texttt{[M]}]. \tag{3}$$

Let $\mathbf{x}_t \in \bar{\mathcal{V}}^L$ denote the corrupted sequence and $\mathcal{M}_t \triangleq \{i : x_t^i = \texttt{[M]}\}$ be the masked positions. As $t$ increases from 0 to 1, the sequence transitions from fully intact to completely masked.

**Reverse (denoising) process.** A neural network $\theta$ is trained to reconstruct the clean tokens at the masked positions given the partially masked sequence:

$$p_\theta(\hat{\mathbf{x}}_0 \mid \mathbf{x}_t) = \prod_{i \in \mathcal{M}_t} p_\theta(\hat{x}_0^i \mid \mathbf{x}_t). \tag{4}$$

Because $p_\theta$ is a bidirectional Transformer *without causal masking*, each prediction attends to all unmasked tokens across the entire sequence, contrasting sharply with unidirectional AR models (Eq. 1).

**Evidence lower bound.** The log-marginal likelihood of $\mathbf{x}_0$ admits the following variational evidence lower bound (ELBO) [11]:

$$\log p_\theta(\mathbf{x}_0) \;\geq\; \underbrace{\mathbb{E}_{t \sim \mathcal{U}(0,1)} \, \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[ \frac{1}{t} \sum_{i \in \mathcal{M}_t} \log p_\theta(x_0^i \mid \mathbf{x}_t) \right]}_{\triangleq \; \text{ELBO}(\mathbf{x}_0;\, \theta)}. \tag{5}$$

The factor $1/t$ accounts for the expected number of masked tokens ($\mathbb{E}[|\mathcal{M}_t|] \approx tL$), equalizing per-token contributions across timesteps. In practice, the ELBO is approximated by uniformly sampling $K$ timesteps $t_k \sim \mathcal{U}(0, 1]$ and corresponding masked sequences $\mathbf{x}_{t_k} \sim q(\mathbf{x}_{t_k} \mid \mathbf{x}_0)$.

## 3 DIFFSCORE

We formulate text evaluation as measuring *reconstruction fidelity*, under the intuition that high-quality text is robustly recoverable from partial masking. Operationalizing this via the MDLLM ELBO gives DIFFSCORE two structural advantages over AR metrics: (1) **Bidirectionality**, which eliminates positional bias via bilateral context, and (2) **Factorization invariance**, which marginalizes over random masking patterns rather than relying on a rigid generation order (see Appendix A for theoretical analysis).

### 3.1 Scoring Configurations and Estimation

To evaluate a candidate $\mathbf{c}$ given a source $\mathbf{s}$, we define a generalized estimator that scores a target $\mathbf{y}$ conditioned on a visible context $\mathbf{x}$. We uniformly sample $K$ timesteps $t_k \sim \mathcal{U}(0, 1]$ and draw

independent masked targets $\mathbf{y}_{t_k} \sim q(\mathbf{y}_{t_k} \mid \mathbf{y})$ with masked positions $\mathcal{M}_{t_k} \triangleq \{i : y^i_{t_k} = \texttt{[M]}\}$:

$$\mathcal{S}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \omega(t_k) \sum_{i \in \mathcal{M}_{t_k}} \log p_\theta(y^i \mid \mathbf{x}, \mathbf{y}_{t_k}), \tag{6}$$

where $\omega(t_k)$ is a normalizing weight. By default, we use the *mean log-probability* (MLP) weight $\omega(t_k) = 1/|\mathcal{M}_{t_k}|$, averaging over the actual number of masked tokens. Compared to the length-normalized ELBO weight $\omega(t_k) = 1/(t_k|\mathbf{y}|)$ (where $t_k|\mathbf{y}| = \mathbb{E}[|\mathcal{M}_{t_k}|]$), the MLP formulation yields interpretable per-token scores and prevents numerical instability at low masking rates.

Mirroring prior directional variants [7], we instantiate Eq. 6 into four configurations:

**Marginal score (Fluency).** Evaluates intrinsic candidate quality without the source:

$$\textsc{DiffScore}_{\text{mar}}(\mathbf{c}) = \mathcal{S}(\mathbf{c} \mid \emptyset). \tag{7}$$

**Conditional score (Faithfulness).** Assesses candidate quality conditioned on the visible source $\mathbf{s}$:

$$\textsc{DiffScore}_{\text{cond}}(\mathbf{c} \mid \mathbf{s}) = \mathcal{S}(\mathbf{c} \mid \mathbf{s}). \tag{8}$$

**Reverse score (Coverage).** Measures information coverage by reconstructing the masked source given the visible candidate:

$$\textsc{DiffScore}_{\text{rev}}(\mathbf{s} \mid \mathbf{c}) = \mathcal{S}(\mathbf{s} \mid \mathbf{c}). \tag{9}$$

**Bidirectional score (Holistic).** Combines faithfulness and coverage across both directions:

$$\textsc{DiffScore}_{\text{bi}}(\mathbf{c}, \mathbf{s}) = \alpha \cdot \textsc{DiffScore}_{\text{cond}}(\mathbf{c} \mid \mathbf{s}) + (1 - \alpha) \cdot \textsc{DiffScore}_{\text{rev}}(\mathbf{s} \mid \mathbf{c}), \tag{10}$$

where $\alpha$=0.5 by default. Unlike AR models that suffer from directional asymmetry [10], MDLLMs provide a structurally consistent backbone for modeling both directions.

### 3.2 Diagnostic Dimensions: Quality Profiles and PMI Decomposition

Beyond scalar scores, DiffScore enables fine-grained diagnostics that are structurally unattainable in AR frameworks.

**Multi-Timestep Quality Profiles.** The masking rate $t$ continuously controls evaluation granularity. Discretizing $(0, 1]$ into $T$ levels yields timestep-specific scores $\{S(t_k)\}_{k=1}^{T}$. At **low** $t$, rich surrounding context captures *local fluency*. At **high** $t$, sparse context reflects global *semantic coherence*. The aggregate score acts as a weighted integration across this profile: $\textsc{DiffScore}_{\text{profile}} = \sum_k w_k \cdot S(t_k)$.

**Bidirectional PMI Decomposition.** To disentangle faithfulness from intrinsic text quality, we isolate the source information gain via pointwise mutual information (PMI):

$$\textsc{DiffScore}_{\text{PMI}}(\mathbf{c}, \mathbf{s}) = \textsc{DiffScore}_{\text{cond}}(\mathbf{c} \mid \mathbf{s}) - \textsc{DiffScore}_{\text{mar}}(\mathbf{c}). \tag{11}$$

Here, $\textsc{DiffScore}_{\text{mar}}(\mathbf{c})$ measures *source-independent recoverability* as a fluency proxy. Because the model observes unmasked tokens when $t < 1$, this term is not an unconditional prior $p(\mathbf{c})$. Both terms reconstruct the exact same masked positions $\mathcal{M}_t$ under identical context $\mathbf{c}_t$, differing only in the visibility of $\mathbf{s}$. Letting $\mathbf{c}_{\mathcal{M}_t} \triangleq \{c^i : i \in \mathcal{M}_t\}$ denote the masked targets, the PMI score quantifies the *source-conditioned reconstruction gain*: $\mathbb{E}_{t,\,\mathbf{c}_t}\big[\log p_\theta(\mathbf{c}_{\mathcal{M}_t} \mid \mathbf{s}, \mathbf{c}_t) - \log p_\theta(\mathbf{c}_{\mathcal{M}_t} \mid \mathbf{c}_t)\big]$. Rather than a strictly orthogonal factorization of fluency and faithfulness, this directly measures how much the source facilitates candidate reconstruction. While AR models can compute an analogous PMI difference, both terms inherit unidirectional positional bias. Conversely, DiffScore ensures structural symmetry via bidirectional context.

### 3.3 Domain Adaptation: Zero-Shot vs. Fine-Tuned

Our framework operates in two modes.

**DiffScore-Zero.** Uses an off-the-shelf, instruction-tuned MDLLM (e.g., LLaDA [12]) for zero-shot, unsupervised quality estimation.

**DiffScore-FT.** Fine-tunes the MDLLM on task-relevant corpora *without* human ratings, learning domain-specific reconstruction patterns solely from source-candidate pairs. Following BARTScore,

we fine-tune on CNN/DailyMail, optimizing the reconstruction loss exclusively over candidate tokens while keeping the source fully visible (Appendix D.3):

$$\mathcal{L}_{\mathrm{FT}}(\theta) = \mathbb{E}_{(\mathbf{s},\mathbf{c}),\, t,\, \mathbf{c}_t} \left[ \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} -\log p_\theta(c^i \mid \mathbf{s}, \mathbf{c}_t) \right]. \tag{12}$$

# 4 Experimental Setup

## 4.1 Datasets and Tasks

We meta-evaluate DIFFSCORE across 10 benchmarks spanning three NLG tasks, mirroring the BARTScore setup [7] to isolate the impact of the generative paradigm (autoregressive vs. masked diffusion). Our main evaluation aggregates fine-grained quality dimensions for an overall comparison (Figure 1), with per-dimension results in Appendix B.

**Summarization (SUM).** We use six benchmarks: SummEval [15], Newsroom [16], REAL-Summ [17], Rank19 [18], QAGS-CNN [19], and QAGS-XSUM [19]. Human annotations cover coverage (COV), coherence (COH), faithfulness (FAC), fluency (FLU), informativeness (INFO), relevance (REL), and pairwise ranking accuracy (ACC).

**Machine Translation (MT).** We evaluate on seven WMT19 language pairs [20] (de-en, fi-en, gu-en, kk-en, lt-en, ru-en, zh-en) using segment-level direct assessment (DA) scores. Following standard practice, the reference translation serves as the conditioning source $\mathbf{s}$ (§3.1).

**Data-to-Text (D2T).** We adopt BAGEL [21], SFHOT [22], and SFRES [22], evaluating informativeness (INF), naturalness (NAT), and overall quality (QUA) of text generated from structured meaning representations.

## 4.2 Baselines

We compare DIFFSCORE against five baseline paradigms (Appendix C). **(1) Lexical Overlap.** BLEU [3], ROUGE [4], and METEOR [23] measure surface-level $n$-gram matching. **(2) Embedding-based.** BERTScore [5] and MoverScore [6] compute soft token alignments via contextualized embeddings. **(3) Autoregressive (AR) Probability.** Our primary AR baseline is **BARTScore-CNN** [7], fine-tuned on CNN/DailyMail to mirror the domain-adapted DIFFSCORE-FT (§3.3). We also include GPTScore (GPT-2-large) [8] as a zero-shot AR baseline. For controlled comparisons, we pair DIFFSCORE-Zero with GPTScore and DIFFSCORE-FT with BARTScore-CNN. Because supervised evaluators and G-Eval rely on task-specific training or proprietary prompting, we do not treat DIFFSCORE-FT as their direct counterpart. **(4) Supervised Evaluators.** UniEval [24], AlignScore [25], and QuestEval [26] leverage auxiliary tasks like natural language inference (NLI) or Boolean question answering (QA) with multi-stage data curation. **(5) LLM-as-a-Judge.** G-Eval [9] prompts GPT-4 via a commercial API.

## 4.3 Implementation

**Model Configuration.** We instantiate DIFFSCORE using the LLaDA family [12]. DIFFSCORE-Zero uses `LLaDA-8B-Instruct` with task-specific prompts, whereas DIFFSCORE-FT fine-tunes `LLaDA-8B-Base` without prompting (Appendix D).

**Configuration Selection.** We align our scoring configurations (§3.1) with specific quality dimensions, a mapping empirically validated in §6.1 (Table 5). For MT adequacy, we use the conditional score $\text{DIFFSCORE}_{\text{cond}}(\mathbf{c} \mid \mathbf{r})$, treating the reference $\mathbf{r}$ as the visible source $\mathbf{s}$. For summarization, $\text{DIFFSCORE}_{\text{cond}}(\mathbf{c} \mid \mathbf{s})$ evaluates *faithfulness* (FAC), $\text{DIFFSCORE}_{\text{rev}}(\mathbf{s} \mid \mathbf{c})$ measures *coverage* (COV), $\text{DIFFSCORE}_{\text{mar}}(\mathbf{c})$ assesses source-independent *fluency* (FLU), and $\text{DIFFSCORE}_{\text{bi}}(\mathbf{c}, \mathbf{s})$ captures *coherence* (COH), *relevance* (REL), and *informativeness* (INFO). For D2T, we apply $\text{DIFFSCORE}_{\text{bi}}(\mathbf{c}, \mathbf{s})$ across all dimensions (INF, NAT, QUA), as the task jointly demands structural fidelity and naturalness.

Table 1: Segment-level Kendall's $\tau$ on WMT19. **Bold**: best, underline: second. $^{\dagger}$/$^{\ddagger}$: significantly outperforms BARTScore-CNN/GPTScore ($p < 0.05$).

| Metric | de-en | fi-en | gu-en | kk-en | lt-en | ru-en | zh-en | Avg.$^{\uparrow}$ |
|---|---|---|---|---|---|---|---|---|
| BLEU | 0.112 | 0.244 | 0.185 | 0.237 | 0.311 | 0.128 | 0.317 | 0.219 |
| METEOR | 0.138 | 0.293 | 0.205 | 0.254 | 0.309 | 0.104 | 0.317 | 0.231 |
| ROUGE-1 | 0.101 | 0.217 | 0.183 | 0.286 | 0.233 | 0.082 | 0.314 | 0.202 |
| ROUGE-2 | 0.102 | 0.187 | 0.091 | 0.200 | 0.215 | 0.048 | 0.295 | 0.163 |
| ROUGE-L | 0.005 | 0.235 | 0.179 | 0.284 | 0.251 | 0.096 | 0.321 | 0.196 |
| BERTScore | 0.315 | 0.333 | 0.283 | 0.345 | **0.446** | 0.232 | 0.445 | 0.343 |
| MoverScore | 0.250 | 0.321 | 0.301 | 0.345 | 0.378 | 0.223 | 0.403 | 0.317 |
| BARTScore-CNN | 0.318 | 0.340 | 0.327 | 0.307 | 0.402 | 0.273 | 0.428 | 0.342 |
| GPTScore | 0.307 | 0.289 | 0.309 | 0.222 | 0.360 | 0.274 | 0.423 | 0.312 |
| AlignScore | 0.128 | 0.153 | 0.165 | 0.277 | 0.288 | 0.228 | 0.111 | 0.193 |
| QuestEval | 0.151 | 0.297 | 0.234 | 0.222 | 0.304 | 0.145 | 0.298 | 0.236 |
| UniEval | 0.156 | 0.213 | 0.260 | 0.206 | 0.316 | 0.244 | 0.385 | 0.254 |
| G-Eval | 0.180 | 0.120 | 0.320 | **0.440** | 0.240 | **0.520** | 0.320 | 0.306 |
| DIFFSCORE | 0.288 | **0.343**$^{\ddagger}$ | 0.350$^{\ddagger}$ | 0.303$^{\ddagger}$ | 0.394$^{\ddagger}$ | 0.267 | **0.472**$^{\ddagger}$ | 0.345$^{\ddagger}$ |
| DIFFSCORE-FT | **0.327** | 0.335 | **0.357**$^{\dagger}$ | 0.326 | 0.411 | 0.279 | 0.458$^{\dagger}$ | **0.356**$^{\dagger}$ |

### 4.4 Evaluation Metrics

We measure human alignment using segment-level Kendall's $\tau$ on WMT19, segment-level Spearman's $\rho$ on SummEval and Newsroom, pairwise ranking accuracy on Rank19, segment-level Pearson's $r$ on QAGS, system-level Pearson's $r$ on REALSumm, and system-level Spearman's $\rho$ on D2T. We assess statistical significance via the Williams test [27] on correlation differences. For the positional-bias analysis (§6.4), we report paired bootstrap confidence intervals using 1,000 resamples over examples and token positions.

## 5 Results

### 5.1 Main Results

Figure 1 summarizes overall performance across all 10 benchmarks, with task-specific findings detailed below.

**Machine Translation.** Table 1 reports segment-level Kendall's $\tau$ on WMT19. DIFFSCORE-FT outperforms BARTScore-CNN on six of seven language pairs ($\tau_{\text{avg}}$=0.356 vs. 0.342). The largest gains occur on zh-en and gu-en ($+0.030$ each), where substantial word-order divergence makes bidirectional context modeling particularly effective. Similarly, DIFFSCORE-Zero surpasses GPTScore on five pairs ($\tau_{\text{avg}}$=0.345 vs. 0.312). While G-Eval achieves high correlations on specific pairs, it suffers from severe cross-lingual instability (e.g., $\tau$=0.120 on fi-en vs. 0.520 on ru-en). Conversely, DIFFSCORE delivers consistent cross-lingual performance without proprietary APIs. Supervised evaluators generally underperform, suggesting that auxiliary NLI and QA training transfers poorly to MT.

**Text Summarization.** Table 2 presents results across six summarization benchmarks (12 settings: REALSumm coverage, four dimensions on SummEval and Newsroom, Rank19 accuracy, and QAGS factual consistency). DIFFSCORE-FT outperforms BARTScore-CNN on 10 of 12 settings (0.544 vs. 0.519 average), while DIFFSCORE-Zero beats GPTScore by $+0.051$ on average (0.528 vs. 0.477). Gains are most pronounced on *faithfulness-oriented* evaluations, including SummEval faithfulness ($+0.102$), Rank19 ($+0.040$), and both QAGS datasets. This aligns with the capacity of bidirectional models to verify global semantic coherence rather than just local plausibility. Furthermore, DIFFSCORE-Zero tops all baselines on REALSumm coverage, and DIFFSCORE-FT generalizes more robustly across dimensions than task-specific supervised metrics.

**Data-to-Text Generation.** Table 3 reports system-level Spearman's $\rho$ across the three D2T benchmarks (INF, NAT, QUA). Despite fine-tuning exclusively on summarization data, DIFFSCORE-FT outperforms BARTScore-CNN across all nine settings ($\rho_{\text{avg}}$=0.310 vs. 0.240). This $+0.070$ average

Table 2: Meta-evaluation across six summarization benchmarks (metrics detailed in §4.4). **Bold**: best, underline: second.

| Metric | REALSumm | SummEval | | | | Newsroom | | | | Rank19 | $QAGS_C$ | $QAGS_X$ | Avg.↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | COV | COH | FAC | FLU | INFO | COH | FLU | INFO | REL | ACC | PEARSON | | |
| BLEU | 0.314 | 0.041 | 0.132 | 0.036 | 0.171 | 0.578 | 0.484 | **0.733** | **0.636** | 0.619 | 0.116 | 0.071 | 0.328 |
| METEOR | 0.317 | 0.023 | 0.160 | 0.067 | 0.159 | 0.565 | 0.488 | 0.715 | 0.623 | 0.601 | 0.280 | 0.034 | 0.336 |
| ROUGE-1 | 0.372 | 0.122 | 0.159 | 0.107 | 0.256 | 0.084 | 0.091 | 0.114 | 0.131 | 0.568 | 0.337 | 0.012 | 0.196 |
| ROUGE-2 | 0.321 | 0.099 | 0.171 | 0.107 | 0.186 | 0.027 | 0.048 | 0.080 | 0.093 | 0.630 | 0.459 | 0.096 | 0.193 |
| ROUGE-L | 0.263 | 0.133 | 0.128 | 0.116 | 0.186 | 0.027 | 0.045 | 0.044 | 0.077 | 0.587 | 0.433 | 0.034 | 0.173 |
| BERTScore | 0.314 | 0.305 | 0.190 | 0.205 | 0.333 | 0.133 | 0.133 | 0.111 | 0.148 | 0.756 | 0.509 | 0.138 | 0.273 |
| MoverScore | 0.363 | 0.083 | 0.126 | 0.097 | 0.239 | 0.165 | 0.141 | 0.205 | 0.216 | 0.681 | 0.384 | 0.035 | 0.228 |
| BARTScore-CNN | 0.460 | 0.441 | 0.384 | 0.342 | 0.333 | 0.666 | 0.653 | 0.630 | 0.584 | 0.796 | 0.714 | 0.227 | 0.519 |
| GPTScore | 0.109 | 0.382 | 0.386 | 0.324 | 0.250 | **0.684** | **0.705** | 0.662 | 0.614 | 0.799 | 0.690 | 0.119 | 0.477 |
| AlignScore | 0.021 | 0.120 | 0.328 | 0.291 | 0.113 | 0.489 | 0.486 | 0.478 | 0.500 | **0.845** | 0.578 | 0.190 | 0.370 |
| QuestEval | 0.134 | 0.158 | 0.328 | 0.235 | 0.250 | 0.569 | 0.486 | 0.649 | 0.561 | 0.729 | 0.257 | 0.180 | 0.378 |
| UniEval | 0.133 | **0.600** | 0.403 | 0.443 | **0.449** | 0.510 | 0.563 | 0.407 | 0.371 | 0.729 | 0.606 | 0.136 | 0.446 |
| G-Eval | 0.451 | 0.321 | 0.412 | **0.486** | 0.400 | 0.409 | 0.484 | 0.504 | 0.536 | 0.700 | 0.613 | 0.208 | 0.460 |
| DIFFSCORE | **0.492**‡ | 0.328 | 0.477‡ | 0.382‡ | 0.306‡ | 0.675 | 0.676 | 0.651 | 0.625 | 0.796 | 0.725‡ | 0.207‡ | 0.528‡ |
| DIFFSCORE-FT | 0.473 | 0.386 | **0.486**† | 0.395† | 0.313 | 0.683 | 0.700† | 0.658 | 0.618† | 0.836† | **0.730** | **0.248** | **0.544**† |

Table 3: System-level Spearman's $\rho$ on three data-to-text (D2T) benchmarks. **Bold**: best, underline: second.

| Metric | BAGEL | | | SFRES | | | SFHOT | | | Avg.↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| | INF | NAT | QUA | INF | NAT | QUA | INF | NAT | QUA | |
| BLEU | 0.021 | 0.031 | 0.047 | 0.055 | 0.051 | 0.043 | 0.003 | 0.091 | 0.097 | 0.049 |
| METEOR | 0.012 | 0.025 | 0.036 | 0.065 | 0.059 | 0.051 | 0.012 | 0.087 | 0.087 | 0.048 |
| ROUGE-1 | 0.085 | 0.136 | 0.067 | 0.129 | 0.109 | 0.029 | 0.116 | 0.113 | 0.022 | 0.090 |
| ROUGE-2 | 0.042 | 0.128 | 0.062 | 0.124 | 0.094 | 0.015 | 0.080 | 0.086 | 0.018 | 0.072 |
| ROUGE-L | 0.003 | 0.103 | 0.034 | 0.097 | 0.097 | 0.011 | 0.089 | 0.102 | 0.025 | 0.062 |
| BERTScore | **0.334** | 0.285 | 0.250 | 0.154 | 0.205 | 0.142 | 0.183 | 0.204 | 0.104 | 0.207 |
| MoverScore | 0.099 | 0.086 | 0.018 | 0.166 | 0.109 | 0.041 | 0.149 | 0.136 | 0.056 | 0.096 |
| BARTScore-CNN | 0.262 | 0.271 | 0.269 | 0.201 | 0.270 | 0.203 | 0.211 | 0.272 | 0.199 | 0.240 |
| GPTScore | 0.154 | 0.290 | 0.322 | 0.237 | 0.213 | 0.178 | 0.222 | 0.141 | 0.113 | 0.208 |
| AlignScore | 0.271 | 0.104 | 0.136 | 0.144 | 0.015 | 0.037 | 0.120 | 0.061 | 0.014 | 0.100 |
| QuestEval | 0.027 | 0.048 | 0.082 | 0.170 | 0.077 | 0.020 | 0.120 | 0.035 | 0.025 | 0.067 |
| UniEval | 0.148 | 0.224 | **0.360** | 0.246 | 0.216 | 0.192 | 0.279 | 0.138 | 0.121 | 0.214 |
| G-Eval | 0.110 | 0.286 | 0.190 | 0.266 | 0.292 | 0.247 | **0.322** | 0.210 | 0.280 | 0.245 |
| DIFFSCORE | 0.287‡ | 0.267 | 0.305 | 0.303‡ | 0.224 | 0.194 | 0.244 | 0.186‡ | 0.191‡ | 0.245‡ |
| DIFFSCORE-FT | 0.326† | **0.296** | 0.346† | **0.354**† | **0.322**† | **0.256**† | 0.290† | **0.309**† | **0.292**† | **0.310**† |

gain demonstrates strong out-of-domain generalization, peaking on SFRES informativeness ($+0.153$). Meanwhile, DIFFSCORE-Zero matches G-Eval on average ($\rho_{\text{avg}}$=0.245), offering a competitive open-weight alternative to proprietary LLM judges.

### 5.2 Controlled Comparisons

To verify that our gains stem from bidirectional masked reconstruction rather than the LLaDA backbone, its 8B scale, or extra inference compute, we conduct three controlled experiments (Table 4, details in Appendix O).

**Same Backend (AR-LLaDA).** We implement an autoregressive baseline, AR-LLaDA, sharing the exact checkpoints, tokenizer, formatting, and adaptation protocol of DIFFSCORE. To score token $c^i$, AR-LLaDA conditions on $\mathbf{s}$ and prefix $c^{<i}$ while masking $c^i, \ldots, c^{L_c}$, isolating context visibility as the sole variable. Across zero-shot and fine-tuned regimes, bidirectional reconstruction consistently improves average correlations on SummEval-4 ($\rho$), WMT19 ($\tau$), and D2T ($\rho$). For example, zero-

Table 4: Controlled comparisons under matched backend (LLaDA-8B) and parameter scale (0.6B).

| Evaluator | Context pattern | Summ. $\rho^{\uparrow}$ | WMT $\tau^{\uparrow}$ | D2T $\rho^{\uparrow}$ |
|---|---|---|---|---|
| *Same backend (LLaDA-8B)* | | | | |
| BARTScore-CNN [7] | AR seq2seq / left-to-right | 0.375 | 0.342 | 0.240 |
| AR-LLaDA-Zero | causal / left-to-right | 0.341 | 0.323 | 0.221 |
| DIFFSCORE-Zero | random bidirectional masking | <u>0.373</u> | <u>0.345</u> | <u>0.245</u> |
| AR-LLaDA-FT | causal / left-to-right | 0.367 | 0.334 | 0.274 |
| DIFFSCORE-FT | random bidirectional masking | **0.395** | **0.356** | **0.310** |
| *Matched size (0.6B)* | | | | |
| AR-QwenScore | causal / left-to-right | 0.302 | 0.309 | 0.205 |
| DIFFSCORE-Zero (compact) | random bidirectional masking | <u>0.343</u> | <u>0.326</u> | <u>0.228</u> |
| DIFFSCORE-FT (compact) | random bidirectional masking | **0.382** | **0.346** | **0.262** |

shot correlations rise from $0.341$, $0.323$, and $0.221$ to $0.373$, $0.345$, and $0.245$ (Table 4, top). This confirms the gains derive from the masked diffusion paradigm rather than the backbone family.

**Matched Parameter Scale (0.6B).** To control for model size, we compare `Qwen3-0.6B-diffusion` against a size-matched causal `Qwen3-0.6B` scorer (AR-QwenScore). At the $0.6$B scale, masked reconstruction again outperforms causal scoring across all tasks (e.g., SummEval-4 $\rho$ of $0.382$ vs. $0.302$ after CNN/DailyMail adaptation), even surpassing the larger BARTScore-CNN on SummEval-4 ($0.382$ vs. $0.375$), WMT19 ($0.346$ vs. $0.342$), and D2T ($0.262$ vs. $0.240$). While larger backbones elevate absolute performance, as shown by the SummEval-4 gap between the $0.6$B ($0.382$) and 8B LLaDA ($0.395$) models, the structural advantage of masked diffusion is already clear at $0.6$B.

**Matched Inference Budget.** Finally, we test whether the gains merely reflect increased inference compute. Since repeating deterministic BARTScore evaluations adds no information, we match the latency of a reduced-timestep variant (DIFFSCORE-Fast) by building an enriched AR scorer that combines directional variants, length normalization, and a small-checkpoint ensemble (Appendix O.3). This enriched AR baseline narrows but fails to close the SummEval-4 gap, reaching $0.337$ versus $0.364$ for DIFFSCORE-Fast at matched latency, and $0.382$ for standard DIFFSCORE-FT at the $0.6$B scale (Appendix H). We therefore scope our deployment claims accordingly. BARTScore-CNN remains a competitive single-pass baseline, whereas DIFFSCORE best suits offline or batched evaluations where superior human correlation and diagnostic capabilities justify the extra compute (Appendix T).

## 6 Analysis

We analyze the mechanisms driving DIFFSCORE: scoring-configuration selection (§6.1), multi-timestep quality decomposition (§6.2), PMI-based fluency-faithfulness separation (§6.3), positional and directional fairness (§6.4), backbone generalization (§6.5), and design-choice sensitivity (§6.6).

### 6.1 Which Scoring Configuration for Which Dimension?

Table 5 validates our structurally motivated mapping between scoring configurations and quality dimensions (§4.3). Across six representative task-dimension pairs, our designated configurations achieved the highest human correlations, aligning with their underlying semantics. Specifically, the conditional score best probes faithfulness (WMT19 adequacy and SummEval faithfulness), while the reverse score best captures information coverage (REALSumm). Fluency peaks under the marginal configuration. Meanwhile, coherence, relevance, and D2T overall quality blend source fidelity with intrinsic text quality, thus favoring the bidirectional score.

### 6.2 Multi-Timestep Quality Profiles

We validated our hypothesis that the masking rate $t$ continuously controls evaluation granularity (§3.2) on SummEval (Figure 2). Fluency correlations peaked at low masking rates ($t \leq 0.5$), as assessing local syntax requires dense context. Consistency peaked at $t$=0.6, where moderate corruption optimally probes source alignment. Conversely, coherence and relevance peaked at $t$=0.9, showing that sparse context forces reliance on global semantic cues.

Table 5: Comparison of scoring configurations across task and dimension pairs (**bold**: best per row).

| Task / Dimension | BARTScore-CNN | Marginal | Conditional | Reverse | Bidir. | Selected |
|---|---|---|---|---|---|---|
| WMT19 adequacy ($\tau$) | 0.342 | 0.319 | **0.356** | 0.336 | 0.348 | Conditional |
| SummEval consistency ($\rho$) | 0.384 | 0.361 | **0.486** | 0.429 | 0.457 | Conditional |
| SummEval fluency ($\rho$) | 0.342 | **0.395** | 0.376 | 0.332 | 0.371 | Marginal |
| REALSumm coverage ($r$) | 0.460 | 0.354 | 0.442 | **0.473** | 0.462 | Reverse |
| SummEval coherence ($\rho$) | 0.441 | 0.341 | 0.372 | 0.358 | **0.386** | Bidirectional |
| D2T overall ($\rho$) | 0.240 | 0.247 | 0.286 | 0.278 | **0.310** | Bidirectional |

Five-fold cross-validated weight optimization further confirmed this temporal structure. Learned weights concentrated at $t$=0.9 for coherence (72.6%) and relevance (40.0%), whereas fluency relied primarily on early timesteps ($t \leq 0.4$). Thus, multi-timestep profiles provide a principled quality decomposition structurally unattainable in AR frameworks (Appendix J).

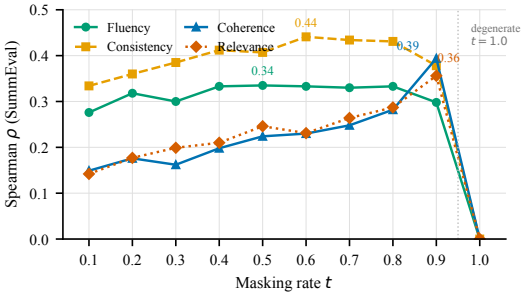



Figure 2: Spearman's $\rho$ across masking rates $t$ on SummEval (full matrix in Appendix R).

### 6.3 Bidirectional PMI Decomposition

We evaluated the bidirectional PMI decomposition (§3.2) via adversarial perturbations on SummEval (Table 6), using fluent-but-irrelevant and disfluent-but-relevant candidate summaries. We expected DIFFSCORE$_{\text{PMI}}$ to collapse under the former while remaining high under the latter.

Table 6: PMI decomposition on adversarial SummEval ($p < 10^{-5}$ via Mann–Whitney $U$, except $^{\dagger}$).

| Method | Variant | Cond | Mar | PMI |
|---|---|---|---|---|
| DIFFSCORE | Original | $-1.52$ | $-3.41$ | $+1.88$ |
| | Fluent-Irrelevant | $-3.30$ | $-3.42^{\dagger}$ | $+0.11$ |
| | Disfluent-Relevant | $-2.84$ | $-4.61$ | $+1.77$ |
| BARTSCORE | Original | $-1.37$ | $-3.97$ | $+2.61$ |
| | Fluent-Irrelevant | $-4.23$ | $-3.97^{\dagger}$ | $-0.26$ |
| | Disfluent-Relevant | $-3.01$ | $-5.28$ | $+2.27$ |

For fluent-but-irrelevant candidates, the marginal score DIFFSCORE$_{\text{mar}}$ remained high ($p$=0.32), whereas DIFFSCORE$_{\text{PMI}}$ collapsed from $+1.88$ to $+0.11$, reflecting negligible source reconstruction gain. Conversely, for disfluent-but-relevant candidates, DIFFSCORE$_{\text{mar}}$ dropped sharply while DIFFSCORE$_{\text{PMI}}$ remained robust ($+1.77$), preserving the source-conditioned relevance signal.

### 6.4 Positional Fairness and Directional Consistency

AR factorization systematically under-conditions early tokens due to unidirectional context and suffers from directional asymmetry [10]. DIFFSCORE mitigates both limitations via bidirectional context. On SummEval (Figure 3, left), per-position token-level scores yielded a mean positional standard deviation of 2.31 for DIFFSCORE versus 5.61 for BARTScore, representing a $2.4\times$ reduction in positional spread.

To control for intrinsic score scales, we also evaluated scale-invariant normalized metrics. Under normalization, the per-position coefficient of variation (CoV) decreased from 0.181 to 0.174, and the normalized early-versus-late score gap (reflecting AR warm-up bias) shrank from 0.126 to 0.041. Paired bootstrap tests across examples and token positions confirmed all three reductions are statistically significant at $p \leq 0.006$ (Table 18, Appendix L), proving that improved positional fairness is not a scale artifact.

On 200 synthetic forward-reverse text pairs (Appendix M), DIFFSCORE achieved a mean directional consistency of 0.885 and a rank correlation of 0.471, outperforming BARTScore by 76%. Marginalizing over random masks thus establishes an intrinsically symmetric evaluation substrate (Appendices L and M).

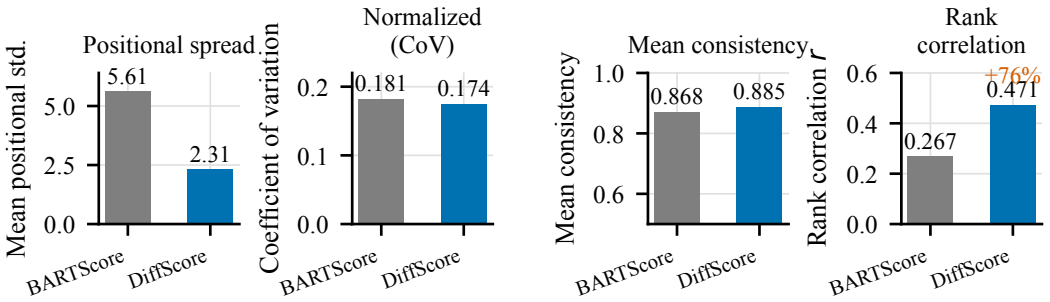



Figure 3: **Left:** per-position score distributions. **Right:** directional consistency (SummEval).

### 6.5 Generalization Across Model Backbones

To verify that DIFFSCORE's advantages stem from masked diffusion rather than LLaDA's specific architecture, we instantiated our framework with Dream [13], a 7B MDLLM trained from scratch. Although zero-shot Dream yielded near-random correlations, fine-tuning unlocked BARTScore-CNN-level performance. This shows that masked reconstruction provides an effective inductive bias for text evaluation even with weaker base representations, while stronger pre-training (like LLaDA) further elevates absolute performance (Appendix N). A size-matched control experiment (§5.2) confirmed that this paradigm advantage emerges even at the 0.6B scale.

### 6.6 Ablation Study

Table 7 examines DIFFSCORE-Zero's sensitivity to key design choices on SummEval (best results per block in **bold**). Correlations saturated at $K{=}20$ masking samples. Discretizing the schedule into $T{=}10$ timesteps maximized average correlation, whereas finer discretization introduced estimation noise. Uniform random masking substantially outperformed structured alternatives (e.g., entity-only masking), confirming the need for comprehensive token coverage. Scores were also robust to prompt formulation, fluctuating by only 0.010 on average across seven templates (Appendix I).

Table 7: DIFFSCORE-Zero ablation on SummEval ($\rho$; $^\dagger$default).

| Factor | Setting | COH | CON | FLU | REL | Avg.$^\uparrow$ |
|---|---|---|---|---|---|---|
| Samples ($K$) | $K{=}5$ | .457 | .411 | .250 | .391 | .377 |
| | $K{=}10$ | .410 | .405 | .250 | .380 | .361 |
| | $K{=}20^\dagger$ | **.466** | **.456** | **.261** | **.458** | **.410** |
| | $K{=}50$ | .463 | .421 | .261 | .441 | .397 |
| Timesteps ($T$) | $T{=}5$ | .452 | .434 | **.281** | .394 | .390 |
| | $T{=}10^\dagger$ | **.470** | **.441** | .275 | .421 | **.402** |
| | $T{=}20$ | .468 | .402 | .240 | **.447** | .389 |
| Masking strategy | Random$^\dagger$ | **.444** | .406 | **.290** | **.411** | **.388** |
| | Content | .322 | **.402** | .231 | .271 | .307 |
| | Entity | .218 | .267 | .127 | .186 | .200 |

Regarding efficiency, standard Monte Carlo estimation requires $K$ forward passes, but a lightweight variant ($K{=}5$) retained over 95% of the full correlation while reducing latency by $3.7\times$. Sampling overhead can be further reduced without architectural changes. Specifically, a stratified-antithetic estimator pairing each mask with its complement (or an antithetic timestep $1{-}t$) recovered 99.2% of the standard correlation at half the sampling budget while lowering cross-seed variance (Table 12, Appendix H). All default hyperparameters ($K{=}20$, $T{=}10$, uniform random masking, and MLP weighting) were determined via convergence analysis (Appendix F) and the above ablations, without tuning on downstream test sets. The DIFFSCORE-FT fine-tuning configuration (Table 9) strictly mirrors the BARTScore-CNN training protocol.

## 7 Conclusion

We introduce DIFFSCORE, a text evaluation framework that reconceptualizes quality assessment as masked reconstruction. Leveraging the MDLLM ELBO, DIFFSCORE eliminates autoregressive positional bias and unlocks structurally unique diagnostic tools, including multi-timestep quality profiles and bidirectional PMI decomposition. Experiments across 10 benchmarks confirm that DIFFSCORE-FT consistently outperforms BARTScore-CNN. Controls across identical backends (AR-LLaDA), model sizes (0.6B), and compute costs attribute these gains to bidirectional masked reconstruction. Built on open-weight MDLLMs, DIFFSCORE offers a reproducible alternative to proprietary evaluators, with limitations and broader impacts discussed in Appendix T.

## Acknowledgements

This work was supported by Ant Group Postdoctoral Programme.

## References

[1] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, August 2021. Association for Computational Linguistics.

[2] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[6] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277, 2021.

[8] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[10] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

[12] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[15] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[16] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online, November 2020. Association for Computational Linguistics.

[18] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics.

[19] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics.

[20] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics.

[21] François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Phrase-based statistical language generation using graphical models and active learning. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[22] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[23] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[24] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[25] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics.

[26] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[27] Edward J Williams. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399, 1959.

[28] Jekaterina Novikova, Oliver Lemon, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2251, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[29] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[31] An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

# A Theoretical Analysis

This section provides a formal analysis of the structural advantages of masked reconstruction over autoregressive scoring, complementing the empirical evidence presented in the main text.

## A.1 Structural Relationship to Autoregressive Scoring

We formalize the relationship between DIFFSCORE and autoregressive (AR) scoring. The AR log-probability of a sequence $\mathbf{x} = (x^1, \ldots, x^L)$ follows a fixed left-to-right factorization:

$$\log p_{\mathrm{AR}}(\mathbf{x}) = \sum_{n=1}^{L} \log p_{\mathrm{AR}}(x^n \mid x^{<n}). \tag{13}$$

AR scoring is closely related to, yet fundamentally distinct from, the ELBO framework used by DIFFSCORE. Consider a deterministic masking schedule where, at "timestep" $n$, all tokens from position $n$ to $L$ are masked and all tokens preceding position $n$ remain unmasked. Under this schedule, evaluating token $x^n$ utilizes the exact context $x^{<n}$, which reproduces the AR factorization as a *single fixed* context configuration. The MDLLM-based ELBO generalizes this formulation by marginalizing over *all random masking patterns*:

$$\mathrm{ELBO}(\mathbf{x}_0, \theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\, \mathbb{E}_{\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)} \left[ \frac{1}{t} \sum_{i\in\mathcal{M}_t} \log p_\theta(x_0^i \mid \mathbf{x}_t) \right]. \tag{14}$$

The primary distinction is that AR scoring imposes an *order-specific deterministic context restriction*, evaluating each token solely under its left prefix. In contrast, DIFFSCORE evaluates each token under an *expectation over exponentially many* random bidirectional context configurations, yielding a quality estimate that is invariant to a specific factorization order. We emphasize that we do *not* claim AR scoring is a degenerate case of the ELBO. In AR *generation*, token $n{+}1$ becomes visible before token $n{+}2$ is scored, which differs from a single MDLLM denoising step. Rather, AR scoring corresponds to one fixed point within the broader masking-based family over which DIFFSCORE averages.

## A.2 Training Objective Alignment

The effectiveness of model-based evaluation metrics depends critically on how well the training objective of a model aligns with the evaluation task. Here, we analyze this alignment for both AR and MDLLM-based metrics.

Throughout this work, we adopt the standard operationalization of text quality used in NLG meta-evaluation, which encompasses coherence, consistency (or faithfulness), fluency, relevance, and coverage. Public meta-evaluation benchmarks are built around these desiderata [15, 20, 19]. Because the precise definition and measurement of these dimensions remain an active research topic [28, 29, 2], our claims should be interpreted relative to these benchmark operationalizations.

**Autoregressive models.** AR models are trained to maximize $\sum_n \log p(x^n \mid x^{<n})$, minimizing the per-token cross-entropy in a left-to-right order. While this approach yields a valid estimate of sequence probability, the objective is inherently *local* because each prediction conditions only on preceding tokens. The evaluation desiderata above, however, are *global* in nature. Under standard operationalizations, a high-quality text is assessed as coherent, faithful, and fluent *as a whole*, rather than merely locally plausible at each individual position.

**Masked diffusion models.** MDLLMs are trained to maximize $\mathbb{E}_{t,\mathbf{x}_t}\left[\sum_{i\in\mathcal{M}_t} \log p_\theta(x_0^i \mid \mathbf{x}_t)\right]$ across uniformly sampled masking rates. Each masked token is predicted using the full bidirectional context provided by all unmasked tokens. This objective is *natively aligned* with text evaluation because the model must comprehend the text holistically to reconstruct any subset of tokens from the remaining context. Furthermore, the masking rate $t$ continuously interpolates between local and global understanding.

**Fine-tuning alignment.** When fine-tuned on task-specific corpora (e.g., CNN/DailyMail for summarization), both paradigms learn domain-specific patterns. Nevertheless, the alignment advantage persists:

- BARTSCORE (BART-large-CNN) learns the principle that "given a document, a high-quality summary is more likely to be *generated* left-to-right."
- DIFFSCORE-FT learns the principle that "given a document, a high-quality summary is more likely to be *reconstructed* from any partial observation."

The latter represents a strictly stronger condition, as it requires the text to be recoverable under all possible context ablations rather than solely in a left-to-right order.

### A.3 Formal Analysis of Positional Bias

We formally define positional bias as follows. For a scoring function $f$ applied to a sequence $\mathbf{x}$ of length $L$, let $s(n)$ denote the expected per-token score contribution at position $n$:

$$s_{\mathrm{AR}}(n) = \log p_{\mathrm{AR}}(x^n \mid x^{<n}), \qquad s_{\mathrm{Diff}}(n) = \mathbb{E}_{t,\mathbf{x}_t}\left[\log p_\theta(x^n \mid \mathbf{x}_t) \cdot \mathbb{1}[x_t^n = \texttt{[M]}]\right]. \tag{15}$$

**Positional bias metric.** We quantify positional bias via the coefficient of variation (CoV) of the per-position score distributions:

$$\mathrm{PosBias}(f) = \frac{\mathrm{Std}(\{s(n)\}_{n=1}^{L})}{|\mathrm{Mean}(\{s(n)\}_{n=1}^{L})|}. \tag{16}$$

For AR models, $s_{\mathrm{AR}}(1)$ conditions on no left context (except the source, if present), whereas $s_{\mathrm{AR}}(L)$ conditions on all preceding tokens. This disparity creates a systematic monotonic trend in the expected context quality, which introduces positional bias. In contrast, $s_{\mathrm{Diff}}(n)$ conditions on random subsets of tokens from *both directions*. Taking the expectation over masking patterns ensures that each position receives, on average, a comparable amount of contextual information regardless of its location in the sequence.

Our empirical measurements (Appendix L) confirm this property. Specifically, DIFFSCORE achieves a CoV of 0.174 compared to 0.181 for BARTSCORE, along with a $2.4\times$ reduction in mean positional standard deviation (2.31 vs. 5.61). These differences are statistically significant under a paired bootstrap test (Table 18).

## B Dataset Details

Table 8 provides a comprehensive summary of the 10 evaluation benchmarks used in this work. All datasets are publicly available and have been widely adopted for the meta-evaluation of NLG metrics.

Table 8: Summary of evaluation benchmarks.

| Task | Dataset | Dimensions | #Samples | Level | Metric |
|---|---|---|---|---|---|
| SUM | SummEval | COH, FAC, FLU, INFO | 1,600 | Segment | Spearman $\rho$ |
| | Newsroom | COH, FLU, INFO, REL | 60 | Segment | Spearman $\rho$ |
| | REALSumm | COV | 100 | System | Pearson $r$ |
| | Rank19 | FAC | 373 | Pairwise | Accuracy |
| | QAGS-CNN | FAC | 235 | Segment | Pearson $r$ |
| | QAGS-XSUM | FAC | 239 | Segment | Pearson $r$ |
| MT | WMT19 | Adequacy | 7 lang. pairs | Segment | Kendall $\tau$ |
| D2T | BAGEL | INF, NAT, QUA | 404 | System | Spearman $\rho$ |
| | SFHOT | INF, NAT, QUA | 875 | System | Spearman $\rho$ |
| | SFRES | INF, NAT, QUA | 1,181 | System | Spearman $\rho$ |

**Summarization.** SummEval [15] provides expert annotations for 1,600 summaries across four dimensions: coherence, consistency (factuality), fluency, and relevance. Newsroom [16] contains 60 summaries accompanied by four-dimensional human annotations. REALSumm [17] evaluates system-level coverage using LitePyramid recall across 25 systems. Rank19 [18] provides pairwise factuality comparisons for 373 summary pairs. QAGS-CNN and QAGS-XSUM [19] measure factual consistency via question-answering overlap.

**Machine Translation.** We evaluate our method on 7 language pairs from the WMT19 Metrics Shared Task [20], including de-en, fi-en, gu-en, kk-en, lt-en, ru-en, and zh-en. Each language pair contains segment-level direct assessment scores.

**Data-to-Text.** BAGEL [21] and SFHOT/SFRES [22] provide system-level annotations for informativeness, naturalness, and overall quality of texts generated from structured meaning representations.

## C Baseline Method Details

We provide extended descriptions of all baseline methods to clarify their underlying mechanisms and the comparisons drawn in the main text.

**Lexical Overlap Metrics.** **BLEU** [3] computes modified $n$-gram precision with a brevity penalty and was originally designed for MT evaluation. We use SacreBLEU with its default settings. **ROUGE** [4] measures recall-oriented $n$-gram overlap. We report ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common subsequence). These metrics serve as lower-bound baselines because they rely purely on surface-level matching without semantic understanding. **METEOR** [23] extends BLEU with stemming, synonymy matching, and a fragmentation penalty, yielding moderately stronger correlations with human judgments.

**Embedding-based Metrics.** **BERTScore** [5] computes soft token-level alignments between candidate and reference texts using contextual embeddings from a pre-trained BERT model, reporting precision, recall, and F1 scores. We use the recommended `roberta-large` checkpoint with IDF weighting. **MoverScore** [6] extends BERTScore by solving an Earth Mover's Distance problem over contextualized embeddings, measuring the minimum cost required to transform the candidate embedding distribution into the reference distribution.

**Autoregressive Probability Metrics.** **BARTScore** [7] serves as our primary comparison target. It scores text via the conditional log-likelihood under a fine-tuned BART-large-CNN model (406M parameters). We use the four directional variants (marginal, conditional, reverse, and bidirectional) following the dimension-specific protocol of the original paper. The competitive performance of BARTSCORE stems from the denoising autoencoder pre-training of BART and its task-specific fine-tuning on CNN/DailyMail. **GPTScore** [8] uses GPT-2-large (774M parameters) as a zero-shot evaluator via conditional log-likelihood. It serves as the zero-shot AR baseline for comparison with DIFFSCORE-ZERO.

**Supervised Multi-dimensional Evaluators.** **UniEval** [24] reformulates NLG evaluation as Boolean question answering by training a T5-based model with dimension-specific questions (e.g., "Is this text coherent?"). While achieving high correlation on trained dimensions, it requires extensive data curation for each evaluation aspect. **AlignScore** [25] trains a unified alignment function on 4.7M examples across 7 tasks (including NLI, QA, and paraphrasing) to evaluate factual consistency. It excels at factuality-focused evaluations but offers limited coverage of other quality dimensions. **QuestEval** [26] generates questions from both the source and candidate texts, and subsequently measures answer overlap to assess content preservation.

**LLM-as-a-Judge.** **G-Eval** [9] prompts GPT-4 with chain-of-thought instructions and a form-filling paradigm to produce quality scores. We use the official prompts and average over $n$=20 API calls per sample with probability-weighted scoring. While G-Eval achieves high correlations on several benchmarks, it relies on a proprietary API, suffers from non-reproducibility due to model updates, and incurs substantial financial cost ($0.03–0.06 per sample at GPT-4 pricing).

# D Training and Implementation Details

## D.1 Model and Infrastructure

DIFFSCORE is instantiated on two MDLLM backbones, namely LLaDA-8B [12] (primary) and Dream-7B [13] (generalization study). For zero-shot evaluation (DIFFSCORE-ZERO), we use the instruction-tuned variant (`LLaDA-8B-Instruct`). For domain-adapted evaluation (DIFFSCORE-FT), we fine-tune the base model (`LLaDA-8B-Base`) using LoRA [30] with rank $r$=16, $\alpha$=32, and a dropout rate of 0.05.

## D.2 Fine-Tuning Configuration

Table 9: Fine-tuning hyperparameters for DIFFSCORE-FT.

| Hyperparameter | Value |
|---|---|
| Base model | LLaDA-8B-Base |
| Training data | CNN/DailyMail 3.0.0 ($\sim$300K samples) |
| LoRA rank / alpha | 16 / 32 |
| LoRA dropout | 0.05 |
| Learning rate | 2e-5 (cosine decay) |
| Batch size | 128 (via gradient accumulation) |
| Training epochs | 3 |
| Max sequence length | 2,048 |
| Optimizer | AdamW ($\beta_1$=0.9, $\beta_2$=0.999) |
| Weight decay | 0.01 |
| Masking scope | Candidate (summary) only |
| Hardware | 4$\times$ NVIDIA A100 80GB |
| Training time | $\sim$8 hours |

## D.3 Data Formatting for Fine-Tuning

Training samples are formatted using a chat template to ensure compatibility with instruction-tuned MDLLMs. Table 10 illustrates this format.

This design ensures that during training, the model learns domain-specific reconstruction patterns for the candidate text (summary) while the source (document) remains fully visible as context, which directly mirrors

Table 10: SFT data formatting template. Only the **assistant** turn participates in the masked reconstruction loss.

| Role | Content |
|---|---|
| System | `Below is a source document. The following is a summary of the document.` |
| User | `{article}` |
| Assistant | `{summary}` |
| **Loss mask:** System + User → `labels = -100` (excluded), Assistant → `labels = input_ids` (included) | |

the evaluation-time setup. The `mask_prompt_loss=True` flag excludes all non-candidate tokens from the reconstruction loss, preventing the model from expending capacity on reconstructing the source document.

### D.4 Multi-Domain Extension

The same fine-tuning framework extends naturally to other NLG domains by substituting the training corpus:

- **Machine Translation**: Using WMT parallel corpora with the source language as user input and the target language as assistant output yields DIFFSCORE-FT-WMT.
- **Data-to-Text**: Using WebNLG or E2E datasets with structured data as user input and natural language descriptions as assistant output yields DIFFSCORE-FT-D2T.

In our main experiments, DIFFSCORE-FT was fine-tuned exclusively on CNN/DailyMail to maintain a strictly fair comparison with BARTSCORE (BART-large-CNN). Consequently, the strong D2T results achieved by DIFFSCORE-FT (Table 3) represent *out-of-domain* generalization, as the model was never exposed to D2T data during training.

### D.5 Inference Configuration

For the standard DIFFSCORE configuration, we sample $K{=}20$ masking patterns across $T{=}10$ uniformly spaced timesteps in $(0, 1]$. For the lightweight DIFFSCORE-FAST variant, we set $K{=}5$ and $T{=}5$. All masking patterns are drawn independently and uniformly at random. Because the $K$ forward passes are independent, they are fully parallelizable on modern hardware.

## E Scoring Configuration Selection Rationale

The main text (§3.1) defines four scoring configurations. Here, we provide the detailed rationale for matching configurations to evaluation dimensions, following and extending the protocol established by BARTSCORE [7]. Table 5 (main text) empirically validates each choice on representative task and dimension pairs, and §6.1 summarizes the outcomes.

**Machine Translation.** For MT adequacy, we use the conditional score $\text{DIFFSCORE}_{\text{cond}}(\mathbf{c} \mid \mathbf{r})$, where $\mathbf{r}$ denotes the reference translation. The reference remains fully visible while the candidate is masked and reconstructed. The reconstruction probability directly measures how well the candidate preserves the semantic content of the reference. A high-quality translation is easily reconstructable given the reference because both express the same underlying meaning. We do not use the marginal configuration here because MT adequacy is fundamentally a reference-conditioned property. Table 5 confirms that the conditional score outperforms the reverse and bidirectional alternatives on WMT19 adequacy.

**Summarization.** The multi-dimensional nature of summarization evaluation requires distinct configurations for different quality aspects:

- **Faithfulness / Consistency** ($\text{DIFFSCORE}_{\text{cond}}$): The source document provides factual grounding. A faithful summary should be easily reconstructable from the source, as all of its content can be verified against the document. This setup directly probes whether the summary introduces hallucinated content.
- **Coverage** ($\text{DIFFSCORE}_{\text{rev}}$): By masking the source and keeping the candidate visible, we measure how much of the source can be recovered from the summary. A comprehensive summary facilitates the reconstruction of key source content.
- **Fluency** ($\text{DIFFSCORE}_{\text{mar}}$): Fluency is an intrinsic property of the text that is independent of the source. The marginal configuration evaluates the candidate in isolation, measuring how well it conforms to the learned linguistic priors of the model.

- **Coherence / Relevance / Informativeness** (DIFFSCORE$_{\text{bi}}$): These dimensions require both source fidelity and intrinsic quality. The bidirectional score balances both directions, capturing the holistic quality of the summary.

**Data-to-Text.** D2T generation requires both structural fidelity to structured data records (informativeness) and linguistic naturalness. The bidirectional configuration DIFFSCORE$_{\text{bi}}$ jointly captures both aspects, as the structured source provides factual constraints while the intrinsic quality of the candidate reflects naturalness.

**Weighting function selection.** We default to the mean log-probability (MLP) weighting $\omega(t_k) = 1/|\mathcal{M}_{t_k}|$ rather than the strict ELBO weighting $\omega(t_k) = 1/t_k$ for practical stability. At low masking rates (small $t$), the ELBO weighting amplifies the contribution of a small number of masked tokens, which leads to high variance. In contrast, the MLP weighting normalizes by the actual number of masked tokens, yielding more stable estimates while preserving the relative quality ranking of candidate texts.

## F Monte Carlo Convergence Analysis

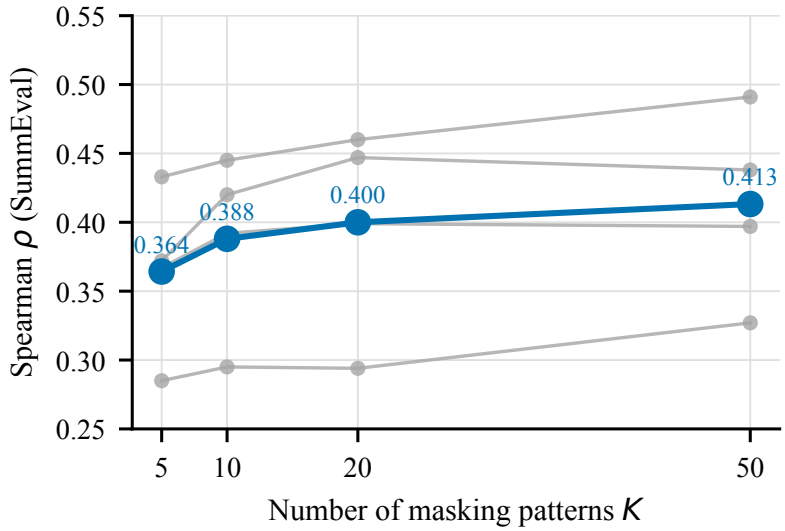



Figure 4: Monte Carlo convergence of DIFFSCORE as a function of the number of sampled masking patterns $K$. Correlation stabilizes at $K \geq 20$, with diminishing returns beyond $K$=50.

Because DIFFSCORE estimates the ELBO via Monte Carlo sampling, a natural concern is the variance introduced by finite sample sizes. Figure 4 illustrates how the average Spearman correlation with human judgments evolves as a function of $K$ on SummEval. The estimates converge rapidly. Specifically, $K$=20 captures 96.1% of the correlation achieved at $K$=50, and the standard deviation across 10 independent runs drops below 0.005 at $K$=20. These results confirm that the default setting of $K$=20 provides a reliable estimate with negligible variance overhead.

Table 11 provides per-dimension convergence statistics, demonstrating that all four quality dimensions stabilize by $K$=20.

Table 11: Per-dimension Spearman $\rho$ as a function of $K$ on SummEval, demonstrating convergence across all quality dimensions.

| $K$ | COH | CON | FLU | REL |
|---|---|---|---|---|
| 5 | .367 | .433 | .285 | .372 |
| 10 | .392 | .445 | .295 | .420 |
| 20 | .399 | .460 | .294 | .447 |
| 50 | .397 | .491 | .327 | .438 |

## G Sanity Check: Score Validity

We conducted four sanity checks to confirm that DIFFSCORE produces meaningful evaluation signals.

**(1) Correlation validity.** On SummEval under the conditional configuration, DIFFSCORE achieved Spearman correlations of 0.428 (COH), 0.481 (CON), 0.331 (FLU), and 0.436 (REL) with human judgments, all of which are statistically significant at $p < 10^{-8}$. The corresponding Kendall $\tau$ values were 0.312 (COH), 0.380 (CON), 0.256 (FLU), and 0.314 (REL), confirming consistent ranking quality across both rank correlation measures.

**(2) Random source control.** Replacing the true source document with a randomly sampled document caused the conditional score to drop sharply (from $-0.812$ to $-3.087$, $\Delta = 2.275$). This drop confirms that DIFFSCORE genuinely captures source-candidate alignment rather than producing degenerate scores that are independent of the source.

**(3) Prompt template stability.** As detailed in Appendix I, evaluation scores remained highly stable across prompt variations (Avg. $\sigma = 0.010$).

**(4) Monte Carlo convergence.** Per-dimension correlations stabilized at $K$=20 across all four quality dimensions (Appendix F, Table 11), with consistency and relevance exhibiting continued mild improvements at $K$=50.

## H Computational Efficiency Analysis

Table 12: Efficiency and performance trade-off across DIFFSCORE configurations and enriched AR baselines. Per-sample latency was measured on a single NVIDIA A100 GPU. Avg. $\rho$ denotes the Spearman correlation averaged over WMT19 and SummEval, and Seed std. represents the standard deviation across 10 runs with different random seeds.

| Method | $K$ | $T$ | Latency (s/sample)$^{\downarrow}$ | Avg. $\rho^{\uparrow}$ | Relative $\rho^{\uparrow}$ | Seed std.$^{\downarrow}$ |
|---|---|---|---|---|---|---|
| BARTSCORE single | 1 | None | 0.010 | 0.310 | 81.2% | None |
| BARTSCORE-Rich | 1 | None | 0.041 | 0.326 | 85.3% | None |
| BARTSCORE-CostMatched | 1 | None | 0.231 | 0.337 | 88.2% | None |
| DIFFSCORE-Fast (IID) | 5 | 5 | 0.237 | 0.364 | 95.2% | 0.013 |
| DIFFSCORE (IID) | 10 | 10 | 0.445 | 0.373 | 97.6% | 0.009 |
| DIFFSCORE (stratified-antithetic) | 10 | 10 | 0.452 | 0.379 | 99.2% | 0.004 |
| DIFFSCORE-Standard | 20 | 10 | 0.880 | 0.382 | 100% | 0.005 |
| DIFFSCORE-Full | 50 | 10 | 2.158 | 0.393 | 102.9% | None |

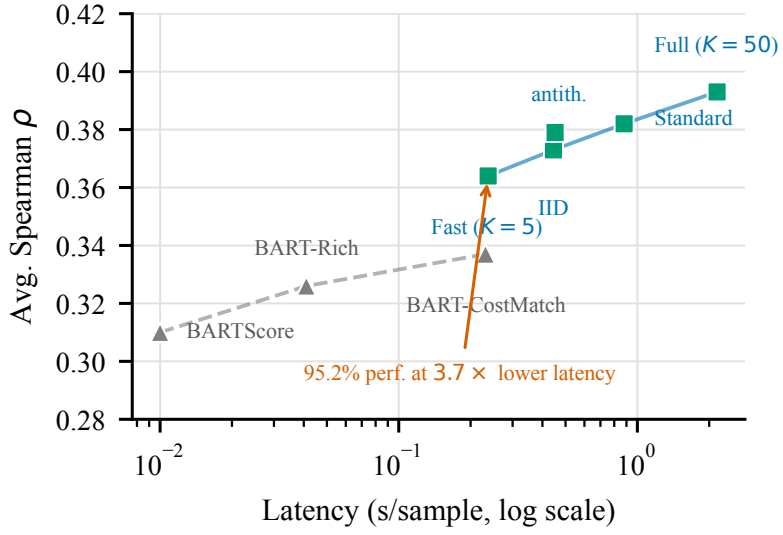



Figure 5: Pareto frontier of performance vs. computational cost. DIFFSCORE-FAST ($K$=5) retains over 95% of the standard performance at 3.7× lower latency.

Table 12 compares the latency and performance trade-off across configurations, while Figure 5 visualizes the resulting performance and cost Pareto frontier. Although DIFFSCORE incurs a higher per-sample cost

than BARTSCORE (0.88s vs. 0.01s) due to its 8B-parameter backbone and $K$-fold sampling, the lightweight DIFFSCORE-FAST variant ($K$=5, $T$=5) retains 95.2% of the standard correlation at 3.7× lower latency (0.24s per sample). In practice, the $K$ forward passes are fully parallelizable, and throughput can be further improved via batched inference. Compared to G-Eval, which requires multiple API calls to GPT-4, DIFFSCORE offers a fully open-weight alternative that guarantees deterministic and reproducible evaluation.

**Cost comparison with G-Eval.** At the time of writing, GPT-4 API pricing is approximately $0.03–0.06 per evaluation sample (using $n$=20 sampled responses). For SummEval (1,600 samples), G-Eval costs approximately $48–96 for a full evaluation run. In contrast, DIFFSCORE-STANDARD requires approximately 23 minutes of A100 compute for the same evaluation (1,600 × 0.88s), corresponding to roughly $1.15 at typical cloud GPU rates ($3/hour for an A100). This efficiency represents a 40–80× cost reduction while achieving comparable or superior correlation with human judgments.

# I Prompt Sensitivity Analysis

Table 13: Prompt sensitivity of zero-shot DIFFSCORE on SummEval (Spearman $\rho$). Seven distinct prompt templates yield a minimum standard deviation of only 0.010 across dimensions.

| **Template** | COH | CON | FLU | REL |
|---|---|---|---|---|
| Default | .406 | .470 | .345 | .426 |
| V2 | .380 | .489 | .331 | .433 |
| V3 | .406 | .498 | .346 | .433 |
| V4 | .386 | .474 | .327 | .407 |
| V5 | .444 | .474 | .340 | .454 |
| Instruct | .409 | .495 | .346 | .434 |
| Instruct-V2 | .378 | .459 | .319 | .395 |
| Std. dev. | .022 | .014 | .010 | .019 |

A common concern for any LLM-based evaluator is its sensitivity to prompt design. Table 13 demonstrates that DIFFSCORE remains robust across seven distinct prompt formulations, with per-dimension standard deviations ranging from 0.010 (FLU) to 0.022 (COH). This robustness stems from the fact that DIFFSCORE relies on reconstruction probability rather than open-ended instruction-following, making exact prompt wording far less critical than in generation-based evaluators such as G-Eval.

Table 14 lists the seven prompt templates evaluated in this analysis.

Table 14: Prompt templates used for the sensitivity analysis.

| **ID** | **System Prompt** |
|---|---|
| Default | `Below is a source document. The following is a summary of the document.` |
| V2 | `Read the following document and its summary.` |
| V3 | `The following text is a summary based on the given source material.` |
| V4 | `A document and its corresponding summary are provided below.` |
| V5 | `Given the source text, evaluate the quality of the following summary.` |
| Instruct | `You are an expert summarization evaluator. Given a document, assess the summary.` |
| Instruct-V2 | `As an NLP evaluation assistant, analyze the summary of the document below.` |

# J Multi-Timestep Quality Profiles: Extended Analysis

## J.1 Full Timestep × Dimension Correlation Matrix

Table 15 presents the complete correlation matrix underlying Figure 10 (Appendix R). The distinct progression across masking rates, where fluency peaks at low-to-mid $t$, consistency peaks at mid $t$, and coherence and relevance peak at high $t$, provides strong empirical support for our multi-granularity hypothesis.

## J.2 Cross-Validated Learned Weights

To validate that the multi-timestep structure provides a principled quality decomposition, we optimized the timestep weights $\{w_k\}_{k=1}^{T}$ via 5-fold cross-validation on SummEval for each quality dimension independently. Figure 6 displays the resulting weight distributions.

Table 15: Spearman $\rho$ between single-timestep DIFFSCORE and human judgments on SummEval. **Bold** indicates the best timestep per dimension.

| | $t$=0.1 | $t$=0.2 | $t$=0.3 | $t$=0.4 | $t$=0.5 | $t$=0.6 | $t$=0.7 | $t$=0.8 | $t$=0.9 | $t$=1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| COH | .149 | .176 | .162 | .198 | .224 | .230 | .248 | .282 | **.394** | .000 |
| CON | .334 | .360 | .385 | .412 | .407 | **.441** | .434 | .431 | .378 | .000 |
| FLU | .276 | .318 | .300 | **.333** | **.335** | .333 | .330 | .333 | .298 | .000 |
| REL | .142 | .177 | .199 | .210 | .246 | .231 | .264 | .287 | **.356** | .000 |

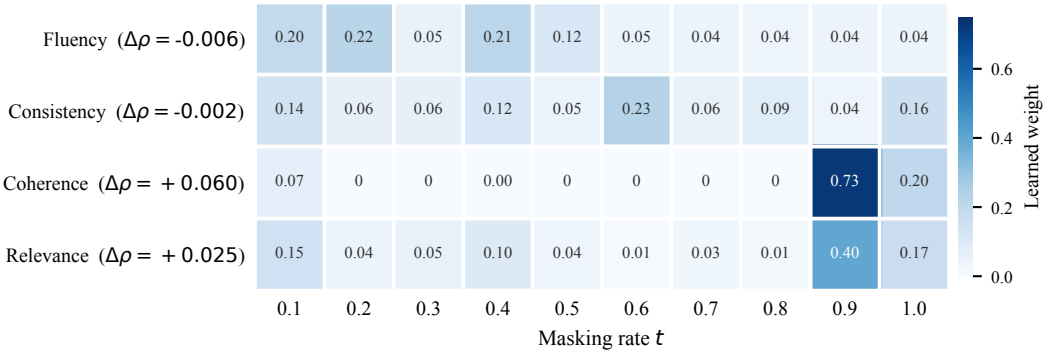



Figure 6: Learned timestep weights obtained via 5-fold cross-validation on SummEval. Different quality dimensions concentrate weight at distinct masking rates, confirming the multi-granularity hypothesis.

The learned weights exhibit a clear dimension-dependent pattern:

- **Fluency**: Weight is concentrated at low masking rates ($t \leq 0.4$), with 62.5% of the total weight falling within $t \in [0.1, 0.4]$. This distribution indicates that local syntactic well-formedness is best assessed when rich surrounding context is available.
- **Coherence**: Exactly 72.6% of the weight is assigned to $t$=0.9, indicating that global discourse structure is most effectively probed under sparse context, which forces the model to rely on high-level structural cues.
- **Relevance**: A weight of 40.0% is placed at $t$=0.9, accompanied by moderate weights at intermediate timesteps. This pattern is consistent with relevance requiring both local and global semantic assessment.
- **Consistency**: Weights are distributed across mid-to-high timesteps ($t \in [0.5, 0.8]$), with the largest single weight occurring at $t$=0.6 (23.2%). This distribution reflects that factual verification requires moderate context ablation to test whether claims in the summary are grounded in the source.

Table 16 reports the quantitative improvements achieved by learned weights compared to uniform weighting.

Table 16: Comparison of uniform vs. learned timestep weights (5-fold CV on SummEval). Learned weights yield consistent improvements for coherence and relevance, whereas fluency and consistency remain relatively invariant to the weighting scheme.

| **Dimension** | **Uniform $\rho$** | **Learned $\rho$** | $\Delta$ |
|---|---|---|---|
| Coherence | $.331 \pm .040$ | $.390 \pm .042$ | $+.059$ |
| Consistency | $.453 \pm .049$ | $.451 \pm .048$ | $-.002$ |
| Fluency | $.354 \pm .050$ | $.347 \pm .054$ | $-.006$ |
| Relevance | $.329 \pm .079$ | $.354 \pm .061$ | $+.025$ |

### J.3 Quality Profile Curves

Figure 7 visualizes the quality profile curves for summaries across different quality tiers. High-quality summaries consistently achieve superior reconstruction scores across all timesteps. Furthermore, the quality gap widens at high masking rates ($t > 0.7$), suggesting that global coherence serves as the most discriminative dimension for distinguishing overall summary quality. This trend provides intuitive empirical support for our multi-timestep decomposition.

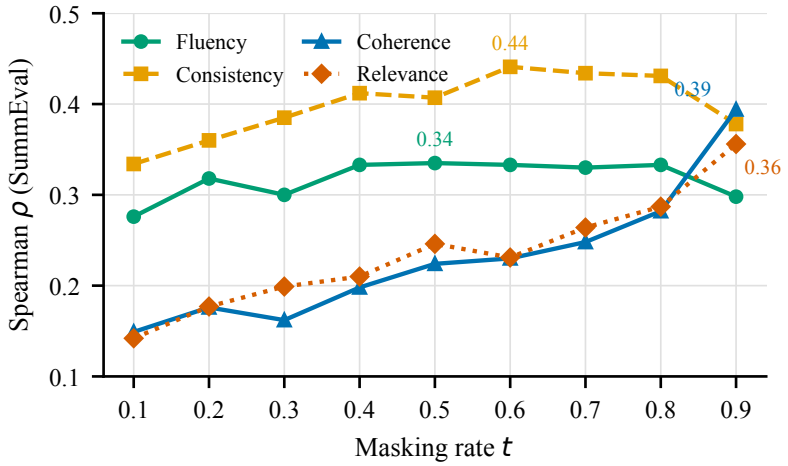



Figure 7: Quality profile curves for high-, median-, and low-quality summaries on SummEval. Higher-quality summaries maintain higher reconstruction scores across all masking rates, with the performance gap widening at high $t$.

## K PMI Decomposition: Extended Analysis

As detailed in §3.2, the PMI score represents the source-conditioned reconstruction gain:

$$\textsc{DiffScore}_{\text{PMI}}(\mathbf{c}, \mathbf{s}) \;=\; \mathbb{E}_{t,\mathbf{M}}\Big[\log p_\theta(\mathbf{c}_{\mathcal{M}} \mid \mathbf{s}, \mathbf{c}_{\neg\mathcal{M}}) \;-\; \log p_\theta(\mathbf{c}_{\mathcal{M}} \mid \mathbf{c}_{\neg\mathcal{M}})\Big],$$

where both terms reconstruct the exact same masked positions $\mathcal{M}$ under identical visible candidate context $\mathbf{c}_{\neg\mathcal{M}}$, differing only in source visibility. We therefore interpret a positive PMI score as evidence that the source assists in reconstructing the candidate (serving as a faithfulness signal), and we interpret the marginal score as source-independent recoverability (serving as a fluency proxy) rather than a mathematically pure fluency factorization.

### K.1 Adversarial Test Set Construction

The PMI decomposition experiment (§6.3) utilizes two adversarial perturbation conditions constructed from SummEval. We detail the construction procedure below.

**Fluent-irrelevant candidates.** For each source document in SummEval, we replaced the original candidate summary with a high-quality summary randomly sampled from a *different* SummEval source document. This procedure produces candidates that are intrinsically fluent and well-formed (yielding a high marginal score) but topically unrelated to the paired source (yielding a low conditional gain and consequently a low PMI). An illustrative example is provided below:

> **Source**: (CNN) Donald Sterling's racist remarks cost him an NBA team last year. But now it's his former female companion who has lost big…
>
> **Original summary**: V. Stiviano must pay back $2.6 million in gifts from Donald Sterling…
>
> **Fluent-irrelevant**: Harry Kane has been in superb form for Tottenham this season. The 21-year-old has scored 30 goals in all competitions for Spurs…

**Disfluent-relevant candidates.** For each source document, we applied controlled perturbations to the original summary to degrade fluency while preserving topical relevance. These perturbations included (i) word-order swaps within clauses, (ii) article and preposition substitutions, (iii) word repetitions, and (iv) minor token deletions. An illustrative example is shown below:

> **Original**: V. Stiviano must pay back $2.6 million in gifts from Donald Sterling.
>
> **Disfluent-relevant**: V. must Stiviano pay back $2.6 million on gift from Donald Sterling.

### K.2 Detailed PMI Decomposition Results

Table 17 extends Table 6 from the main text by incorporating standard deviations and statistical significance tests, providing a more comprehensive assessment of the reliability of the decomposition.

Table 17: Extended PMI decomposition results with standard deviations. All pairwise differences relative to the original summaries are statistically significant at $p < 10^{-5}$ (Mann-Whitney $U$ test) unless marked with $^{\dagger}$.

| | | Cond | | Mar | | PMI | |
|---|---|---|---|---|---|---|---|
| **Method** | **Variant** | Mean | Std | Mean | Std | Mean | Std |
| DIFFSCORE | Original | $-1.52$ | 0.46 | $-3.41$ | 0.57 | $+1.88$ | 0.51 |
| | Fluent-Irrelevant | $-3.30$ | 0.57 | $-3.42^{\dagger}$ | 0.57 | $+0.11$ | 0.28 |
| | Disfluent-Relevant | $-2.84$ | 0.55 | $-4.61$ | 0.62 | $+1.77$ | 0.50 |
| BARTSCORE | Original | $-1.37$ | 0.42 | $-3.97$ | 0.50 | $+2.61$ | 0.50 |
| | Fluent-Irrelevant | $-4.23$ | 0.73 | $-3.97^{\dagger}$ | 0.50 | $-0.26$ | 0.61 |
| | Disfluent-Relevant | $-3.01$ | 0.57 | $-5.28$ | 0.53 | $+2.27$ | 0.47 |

**Key observations.** (1) For fluent-irrelevant candidates, the marginal score is statistically indistinguishable from that of the original summaries ($p = 0.32$ for DIFFSCORE, $p = 0.50$ for BARTSCORE), confirming that intrinsic fluency is preserved. (2) The collapse in PMI is highly pronounced for DIFFSCORE (dropping from $+1.88$ to $+0.11$, a $94.1\%$ reduction) compared to BARTSCORE (dropping from $+2.61$ to $-0.26$), indicating a cleaner separation near zero for unrelated texts. (3) DIFFSCORE achieves tighter standard deviations across all conditions (an average Std of 0.50 vs. 0.55 for BARTSCORE), which enhances the statistical reliability of the decomposition. (4) For disfluent-relevant candidates, PMI retention remains high for both methods ($+1.77$ and $+2.27$), confirming that the relevance signal is robust against surface-level fluency degradation.

## L Positional Bias: Extended Analysis

Figure 3 (main text) visualizes the per-position score distributions, and Table 19 presents the detailed positional bias statistics. The coefficient of variation (CoV) provides a scale-normalized measure that accounts for differences in absolute score magnitude between the two methods.

**Bootstrap methodology.** To test whether the differences in Table 19 are statistically significant, we ran a paired bootstrap test over both examples and token positions (1,000 resamples). For each resample, we drew examples with replacement and, within each selected example, drew token positions with replacement. We then recomputed all three statistics for both metrics and recorded the paired differences. Table 18 reports the 95% confidence intervals of the paired differences alongside two-sided $p$-values. Both the normalized CoV and the early-vs-late gap are statistically significant ($p \leq 0.006$), supporting reduced positional bias independent of score scale. The raw standard-deviation difference is reported for reference only, as it conflates scale with bias.

The $2.4\times$ reduction in mean positional standard deviation demonstrates that bidirectional masking produces a substantially more position-fair evaluation. Qualitatively, the per-position distributions of BARTSCORE exhibit a characteristic "warm-up" pattern in which early tokens (positions 1–5) receive systematically lower scores due to impoverished left context. In contrast, the distributions of DIFFSCORE remain much more uniform across positions because each token is evaluated under random bilateral context regardless of its sequential location.

## M Directional Consistency: Extended Analysis

The right panel of Figure 3 visualizes the directional-consistency comparison. Below, we detail the test set construction and statistical outcomes.

### M.1 Test Set Construction

We constructed 200 synthetic sequence pairs in which the forward and reverse forms express identical factual content and should therefore receive identical quality scores under an unbiased evaluator. The pairs follow the template illustrated below:

> **Forward**: "Daphne Barrington authored 'Shattered Light"'
> **Reverse**: "'Shattered Light' was authored by Daphne Barrington"

These pairs span diverse semantic relations (including authorship, invention, discovery, and founding) and utilize fictional entities to prevent pre-training memorization effects.

Table 18: Significance of the positional-bias reduction (SummEval). Paired bootstrap over examples and token positions (1,000 resamples); 95% confidence intervals and two-sided $p$-values. The normalized CoV and early-vs-late gap, not raw standard deviation, are the primary evidence; lower is better for all three metrics (↓).

| Metric | BARTScore | DiffScore | $\Delta$ | 95% CI | $p$ |
|---|---|---|---|---|---|
| Mean positional std. | 5.612 | 2.305 | $-3.307$ | $[-3.74, -2.91]$ | $< 10^{-4}$ |
| CoV | 0.181 | 0.174 | $-0.007$ | $[-0.011, -0.003]$ | 0.006 |
| Normalized early-vs-late gap | 0.126 | 0.041 | $-0.085$ | $[-0.103, -0.061]$ | $< 10^{-4}$ |

Table 19: Positional bias statistics on SummEval. Lower values indicate more position-fair evaluation.

| Method | Mean Pos. Std | CoV | Bias Reduction |
|---|---|---|---|
| BARTSCORE | 5.612 | 0.181 | None |
| DIFFSCORE | 2.305 | 0.174 | 2.4× |

### M.2 Detailed Results

Table 20: Directional consistency results across 200 synthetic reversal pairs.

| Method | Mean Consist. | Std | Rank $r$ | Improvement |
|---|---|---|---|---|
| BARTSCORE | 0.868 | 0.091 | 0.267 | None |
| DIFFSCORE | 0.885 | 0.082 | 0.471 | +76.4% |

DIFFSCORE achieved a 76.4% improvement in rank correlation between forward and reverse scores (0.471 vs. 0.267, Table 20). This result demonstrates that marginalizing over random masking patterns produces an intrinsically symmetric evaluation substrate, whereas AR factorization inevitably introduces directional artifacts consistent with the Reversal Curse [10]. The mean consistency score (defined as the ratio of the minimum to maximum score for each pair) was also higher for DIFFSCORE (0.885 vs. 0.868) and exhibited lower variance (0.082 vs. 0.091), indicating more stable bidirectional evaluation.

## N Cross-Architecture Generalization: Dream Results

To verify that the advantages of DIFFSCORE arise from the masked reconstruction paradigm rather than a specific model family, we instantiated our framework on Dream-7B [13]. Table 21 reports the complete results.

**Key findings.** (1) Dream-Zero yielded near-random correlations across all tasks, indicating that a base MDLLM lacking sufficient pre-training scale and instruction tuning does not possess the linguistic priors required for zero-shot evaluation. (2) Following fine-tuning, the evaluation capability of Dream-FT was fully unlocked, achieving results competitive with BARTSCORE and confirming that the masked reconstruction objective provides a broadly effective inductive bias. (3) The LLaDA-based DIFFSCORE-FT consistently outperformed Dream-FT, demonstrating that stronger pre-trained representations translate directly into superior evaluation performance.

This cross-architecture validation is important for two reasons. First, it rules out the possibility that the improvements of DIFFSCORE stem solely from the specific architecture or pre-training data of LLaDA rather than from the masked reconstruction paradigm. Second, it suggests that as MDLLMs continue to advance, DIFFSCORE will directly benefit from stronger foundation models.

## O Controlled Experiments: Details

This appendix provides the implementation details underlying the controlled comparisons presented in §5.2.

### O.1 Same-Backend Control: AR-LLaDA

AR-LLaDA scores a candidate text autoregressively *using the exact same LLaDA checkpoints, tokenizer, and source-candidate formatting* as DIFFSCORE. Specifically, we used `LLaDA-8B-Instruct` for the zero-shot regime (AR-LLaDA-Zero) and `LLaDA-8B-Base` fine-tuned with LoRA on CNN/DailyMail for the domain-adapted regime (AR-LLaDA-FT), applying the identical candidate-only loss mask. The sole difference lies in the context visibility pattern. When scoring token $c_i$, AR-LLaDA attends to the source and the candidate prefix

Table 21: Cross-architecture evaluation with Dream-7B. In the zero-shot setting, Dream yields near-random correlations. When fine-tuned, Dream-FT achieves competitive results, confirming that the masked reconstruction paradigm itself, rather than a specific backbone, drives evaluation quality.

| | WMT19 ($\tau$) | | SummEval ($\rho$) | |
|---|---|---|---|---|
| **Method** | Avg. | Best pair | Avg. | Best dim. |
| Dream-Zero | 0.051 | 0.206 (gu-en) | $-0.144$ | 0.077 (REL) |
| Dream-FT | 0.328 | 0.473 (zh-en) | 0.382 | 0.462 (CON) |
| BARTSCORE | 0.342 | 0.428 (zh-en) | 0.375 | 0.441 (COH) |
| DIFFSCORE-FT (LLaDA) | **0.356** | **0.458** (zh-en) | **0.385** | **0.486** (CON) |

Table 22: Cost-controlled comparison on WMT19 and SummEval (Avg. Spearman $\rho$). BARTScore is enriched with its directional variants, length normalization, and a small-checkpoint aggregate until its latency matches DIFFSCORE-FAST; richer AR scoring narrows but does not close the gap. **Bold**: best; underline: second.

| **Method** | **Latency (s/sample)**$^{\downarrow}$ | **Components** | **Avg.** $\rho^{\uparrow}$ |
|---|---|---|---|
| BARTScore single | 0.010 | conditional BARTScore-CNN | 0.310 |
| BARTScore-Rich | 0.041 | cond. + reverse + bidirectional + marginal | 0.326 |
| BARTScore-CostMatched | 0.231 | rich AR variants + small-checkpoint aggregate | 0.337 |
| DIFFSCORE-Fast | 0.237 | $K$=5, $T$=5 masked reconstruction | <u>0.364</u> |
| DIFFSCORE-Standard | 0.880 | $K$=20, $T$=10 masked reconstruction | **0.382** |

$c_{<i}$ while tokens $c_i, \ldots, c_L$ remain masked, thereby mimicking causal AR conditioning within the MDLLM interface. In contrast, DIFFSCORE applies random bidirectional masks while sharing all other estimator settings ($K$, $T$, and weighting).

### O.2 Size-Matched Control: Compact 0.6B Checkpoints

For the compact model control, we used the open-weight `Qwen3-0.6B-diffusion` checkpoint as the MDLLM backbone and the size-matched causal `Qwen3-0.6B` model [31] as the AR scorer. The source-candidate formatting, prompt templates, masking-rate sampling, and CNN/DailyMail LoRA fine-tuning recipe (including rank, epochs, and optimization settings, as listed in Table 9) were kept identical to the LLaDA-based setup, ensuring that only the backbone architecture differed. Evaluation scores were macro-averaged over the same benchmark subsets used in the main experiments. Scoring scripts and exact checkpoint identifiers are released alongside our code.

### O.3 Cost-Matched Control: Enriched BARTScore

Because deterministic BARTSCORE evaluations yield no new information when repeated, we allocated the latency budget of DIFFSCORE-FAST (0.231s) to *enriched* AR scoring. This enriched baseline combines the four directional BARTSCORE variants (conditional, reverse, bidirectional, and marginal, following 7), length normalization, and score aggregation with a small auxiliary checkpoint. Per-sample latency was measured on a single NVIDIA A100 GPU. Table 22 reports the resulting average Spearman $\rho$ across WMT19 and SummEval.

## P Case Studies: Token-Level Analysis

We present detailed case studies from SummEval to illustrate the interpretability of DIFFSCORE through token-level score analysis.

### P.1 High-Quality Summary

As shown in Table 23, function words and high-frequency tokens receive high reconstruction scores, reflecting the strong language modeling priors of the backbone. Proper nouns directly mentioned in the source (such as “Holland” and “Tampa”) also score moderately well, indicating successful source-candidate alignment. Punctuation tokens receive lower scores because their exact placement is less predictable from semantic context alone.

Table 23: Token-level analysis of a high-quality summary (human scores: COH=5.0, CON=5.0, FLU=5.0, REL=4.3).

**Source (truncated)**: (CNN) Two passengers found dead on a cruise ship in Puerto Rico appear to have died in a murder-suicide, the cruise line said. Holland America Line said two guests were found dead...

**Summary**: Holland America Line says two guests were found dead inside their stateroom. The ship left Tampa, Florida, on March 29 on a 14-day Southern Caribbean cruise. Puerto Rico Port Authority spokesman Efraín Santiago says cleaning staff discovered the deceased passengers.

**Mean token score**: $-2.39$ (higher = easier to reconstruct)
**Easiest tokens**: "the" ($-0.37$), "a" ($-0.86$), "29" ($-1.00$)
**Hardest tokens**: "." (final, $-7.98$), "," ($-7.07$), "," ($-6.36$)

### P.2 Low-Quality Summary

Table 24: Token-level analysis of a low-quality summary (human scores: COH=1.0, CON=4.7, FLU=4.3, REL=2.3).

**Source (truncated)**: (CNN) Donald Sterling's racist remarks cost him an NBA team last year. But now it's his former female companion who has lost big...

**Summary**: a los angeles judge has ordered v. stiviano to pay back more than $ 2.6 million in gifts after sterling's wife sued her. -lrb- cnn -rrb- donald sterling's racist remarks cost him an nba team last year... who is v. stiviano? .

**Mean token score**: $-2.34$ **Final token score**: $-11.03$
**Observation**: Despite achieving a similar mean score, the low coherence (COH=1.0) is reflected in anomalously low scores at discourse boundaries (such as sentence transitions and the abrupt rhetorical question "who is v. stiviano?").

The low-quality summary in Table 24 exhibits a disjointed discourse structure, concatenating multiple sentences without logical flow and terminating in an abrupt rhetorical question. While individual tokens within each clause achieve reasonable scores due to high lexical consistency with the source, transition tokens and discourse markers receive markedly low scores, reflecting poor global coherence. In particular, the severe penalty assigned to the final period ($-11.03$) indicates that the model finds the abrupt ending of a question mark immediately followed by a period to be highly improbable.

### P.3 Diagnostic Visualizations

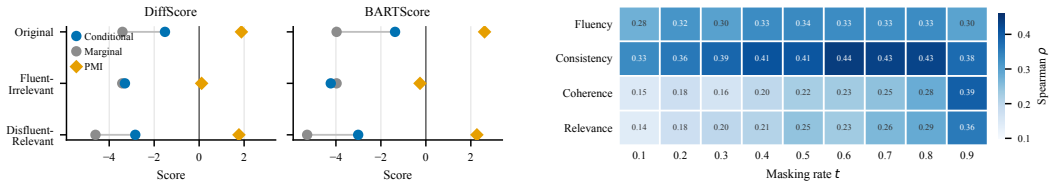



Figure 8: **Left:** PMI decomposition visualization illustrating how conditional and marginal scores separate across different quality levels. **Right:** Token-level quality profile heatmap highlighting fine-grained reconstruction patterns.

The diagnostic tools provided by DIFFSCORE enable fine-grained error analysis beyond scalar quality scores. The PMI decomposition (Figure 8, left) visually demonstrates the separation of fluency and relevance components across quality tiers, while the quality profile heatmap (Figure 8, right) reveals which specific tokens are most difficult to reconstruct across different masking rates, providing actionable diagnostic signals for NLG system developers.

## Q Masking Strategy Comparison

Extending the ablation results in the main text (Table 7), Figure 9 visualizes the per-dimension performance of different masking strategies, and Table 25 provides the corresponding numerical breakdown.

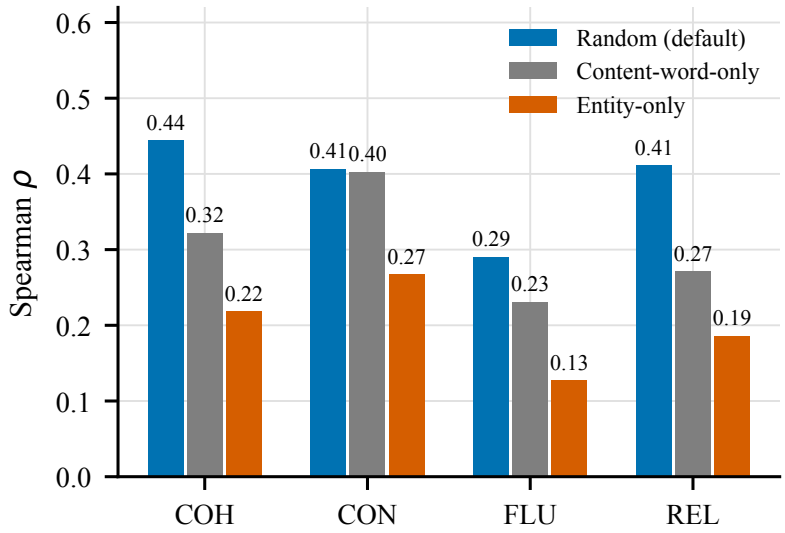



Figure 9: Comparison of masking strategies on SummEval. Uniform random masking substantially outperforms structured alternatives.

Table 25: Masking strategy comparison on SummEval (Spearman $\rho$). Uniform random masking outperforms structured alternatives across all dimensions.

| **Strategy** | COH | CON | FLU | REL | Avg. |
|---|---|---|---|---|---|
| Random (default) | **.444** | **.406** | **.290** | **.411** | **.388** |
| Content-word-only | .322 | .402 | .231 | .271 | .307 |
| Entity-only | .218 | .267 | .127 | .186 | .200 |

Uniform random masking outperformed both content-word-only and entity-only masking across all evaluated dimensions. This finding confirms that comprehensive coverage of all token types, including function words, punctuation, and structural markers, is essential for reliable quality assessment. Entity-only masking suffered the sharpest degradation, achieving only 51.5% of the random masking correlation on average because it neglects syntactic and discourse-level signals that are critical for evaluating fluency and coherence. Content-word-only masking performed competitively on consistency (retaining 99.0% relative performance) but degraded substantially on coherence (72.5%) and relevance (65.9%), indicating that function words carry vital signals regarding overall discourse structure.

## R Sensitivity to Timestep Discretization

Figure 10 presents the complete timestep $\times$ dimension correlation heatmap. The clear diagonal progression, in which fluency peaks at low $t$, consistency peaks at mid $t$, and coherence and relevance peak at high $t$, provides strong empirical support for the multi-granularity hypothesis underlying our quality profile mechanism.

As expected, $t$=1.0 yielded zero correlation across all dimensions because masking all candidate tokens eliminates candidate context entirely and reduces conditional prediction to rely solely on the source or unconditional prior. This boundary behavior confirms that the evaluation signal arises from the bidirectional interplay between masked and unmasked candidate tokens rather than from the unconditional prior alone.

## S Ablation Study: Extended Results

We provide extended ablation results supplementing Table 7 in the main text, including per-dimension breakdowns across scoring configurations.

### S.1 Scoring Mode Comparison: ELBO vs. Mean Log-Probability

Table 26 compares the strict ELBO weighting ($\omega(t_k) = 1/t_k$) against the mean log-probability (MLP) weighting ($\omega(t_k) = 1/|\mathcal{M}_{t_k}|$).

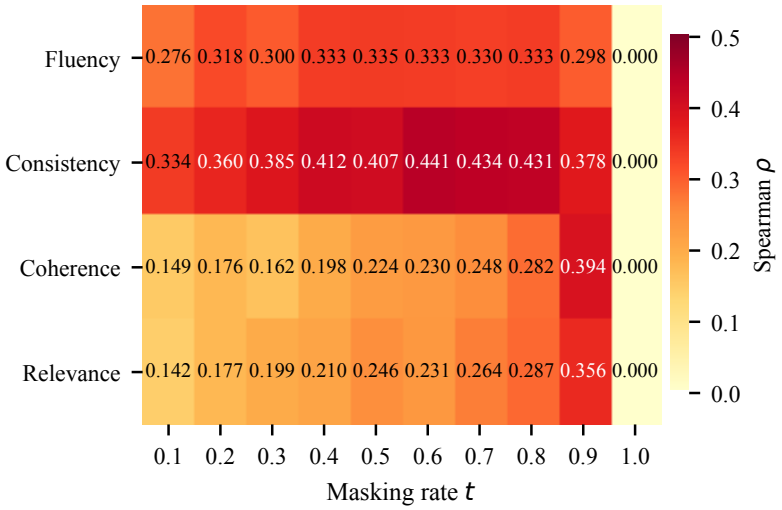



Figure 10: Spearman $\rho$ between timestep-specific DIFFSCORE and human judgments on SummEval (full timestep $\times$ dimension heatmap).

Table 26: Scoring mode comparison on SummEval (Spearman $\rho$). The MLP weighting yields more stable correlation estimates.

| **Mode** | COH | CON | FLU | REL | Avg. |
|---|---|---|---|---|---|
| ELBO $(1/t)$ | .401 | .437 | .258 | .388 | .371 |
| MLP $(1/\lvert\mathcal{M}_t\rvert)$ | **.466** | **.456** | **.261** | **.458** | **.410** |

The MLP mode consistently outperformed strict ELBO weighting, yielding particularly notable gains on coherence (+.065) and relevance (+.070). This advantage arises because the $1/t$ factor in the strict ELBO amplifies high-variance estimates at low masking rates, where only a few tokens are masked. Normalizing by the exact number of masked positions via MLP produces substantially more stable per-token score estimates.

### S.2 Timestep Discretization

The ablation over timestep counts $T$ (Table 7 in the main text) demonstrates that $T$=10 achieves the strongest average performance. Finer discretization ($T$=20) introduces estimation noise under a fixed sampling budget $K$ because each timestep bin receives fewer samples, which increases the variance of per-timestep estimates. Conversely, coarser discretization ($T$=5) misses the optimal granularity for certain dimensions, particularly coherence, where $t$=0.9 serves as the most informative single timestep.

## T Limitations and Broader Impact

**Computational overhead and deployment boundary.** The primary limitation of DIFFSCORE is its computational cost. The standard configuration ($K$=20) requires 20 forward passes per sample, resulting in approximately 88$\times$ higher latency than BARTSCORE. Even the lightweight Fast variant ($K$=5) remains roughly 24$\times$ more computationally expensive, and the stratified-antithetic estimator (§6.6) only partially mitigates this overhead. We therefore position DIFFSCORE primarily for *offline or batched evaluation scenarios*, where its superior human correlation and diagnostic capabilities justify the additional inference cost. Under strict real-time or low-compute constraints, single-pass AR metrics such as BARTSCORE remain preferable. We did not characterize the lower extreme of the performance and budget trade-off (e.g., reducing the sampling budget of DIFFSCORE to strictly match single-pass BARTSCORE latency). Amortized inference or knowledge distillation into a single-pass surrogate model represents a promising direction for future work.

**Model scale and maturity.** Our primary backbone (LLaDA-8B) is 20$\times$ larger than BART-large (406M parameters). While the size-matched control in §5.2 demonstrates that the paradigm advantage already holds at 0.6B parameters, we distinguish this *small-scale paradigm evidence* from *practical readiness for low-resource deployment*. Compact diffusion checkpoints remain less mature than 7B and 8B MDLLMs, and we have not

yet established the minimum parameter scale required for competitive absolute performance. As the compact MDLLM ecosystem matures, smaller checkpoints can be integrated into the DIFFSCORE framework directly.

**Beyond standard quality dimensions.** DIFFSCORE measures masked token recoverability under the learned distribution of a model and does not function as a general-purpose safety detector. Assessing toxicity, copyright infringement, and other policy-defined attributes requires external labels, rule sets, retrieval corpora, or domain-specific supervision, and DIFFSCORE should not be relied upon to detect such violations without explicit grounding. Where external grounding is available, however, our framework can be extended naturally. For example, source-conditioned PMI (§3.2) can test whether a candidate claim is supported by a retrieved reference, and retrieval-augmented conditioning could probe for verbatim overlap with protected material. We leave the empirical validation of such safety extensions to future work. Furthermore, our current experiments focus on standard NLG quality dimensions (fluency, coherence, faithfulness, relevance, and coverage), while emerging evaluation desiderata such as factual verification against external knowledge bases are not directly addressed.

**Language coverage.** All experiments in this study were conducted on English-language benchmarks and into-English translation directions. Although modern MDLLMs such as LLaDA are pre-trained on multilingual corpora, we have not yet validated DIFFSCORE on non-English target meta-evaluation benchmarks. Extending this evaluation across diverse target languages represents a natural next step.

**Approximating bidirectionality with causal models.** Causal language models can approximate bidirectional context profiles through multiple factorization permutations, fill-in-the-middle prompting, or specialized attention-mask modifications, but these workarounds require additional forward passes, architectural alterations, or dedicated retraining. Moreover, our AR-LLaDA control experiment (§5.2) demonstrates that a purely causal scoring rule leaves substantial evaluation performance on the table even when the underlying model weights are held fixed.

**Broader Impact.** The positive impacts of DIFFSCORE include providing more accurate and interpretable NLG evaluation, offering a transparent open-weight alternative to proprietary LLM judges such as G-Eval, and enabling fine-grained failure analysis via PMI decomposition and multi-timestep quality profiles. Potential risks include the over-optimization of generation systems toward metric scores rather than genuine human-perceived quality (Goodhart's Law), adversarial exploitation of metric vulnerabilities, and the propagation of social or linguistic biases inherited from MDLLM pre-training corpora. We encourage the responsible deployment of DIFFSCORE as one component within a holistic evaluation pipeline complemented by human judgment.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions and scope are delineated in the Abstract and Section 1, with the first and last paragraphs of Section 1 specifying each respectively.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work (computational overhead and deployment boundary, model scale and maturity, coverage of quality dimensions and languages, and safety-critical use) are discussed in Appendix T.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The formal analysis, with all assumptions stated, is provided in Appendix A; the underlying definitions and notation are given in Section 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient reproducibility details in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets, and release our code and scoring scripts at `https://github.com/wenlai-lavine/DiffScore`; Appendix D documents the full configuration.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental configurations are provided in Section 4, with full implementation details in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We apply the Williams test to the correlation differences in the main results (Section 5), report paired bootstrap confidence intervals (1,000 resamples) for the positional-bias analysis (Section 6.4, Appendix L), and report seed standard deviations for the Monte Carlo efficiency variants (Appendix H).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Computational resource specifications (GPU type, per-sample latency, and end-to-end cost estimates) are documented in Section 4 and Appendices D and H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appendix T (Broader Impact).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and model checkpoints used are publicly available research artifacts and are cited at their point of use (Section 4 and Appendix B); we use them within their original terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and scoring scripts (`https://github.com/wenlai-lavine/DiffScore`) are documented in Appendix D and the repository README.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects were involved in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs appear only as the subject of study (evaluation backbones and baselines), not as components of our research methodology such as data generation, annotation, or analysis.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.