PROXY-GUIDED SAMPLING FOR ROBUST PROBABILISTIC MINIMUM BAYES RISK DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimum Bayes Risk (MBR) decoding yields superior generation quality but suffers from a severe $O(N^2)$ computational bottleneck due to pairwise matrix calculation between candidates and pseudo-references. Probabilistic MBR (PMBR) mitigates this by randomly sampling several candidate–pseudo-reference pairs and utilizing Alternating Least Squares (ALS) to approximate the full pairwise matrix. However, randomly sampling the pairs may select uninformative data and starve the ALS of critical pairs. We introduce Bootstrap PMBR (B-PMBR), a framework that utilizes proxy-guided sampling to address this inefficiency. By leveraging model log-probabilities as a proxy score, B-PMBR selects the most influential pairs for evaluation by the utility function. Our extensive experiments demonstrate that B-PMBR consistently exceeds PMBR, especially at low reference sizes, and remains highly competitive with standard MBR. This competitiveness is driven by B-PMBR's ability to mitigate utility metric overfitting, which degrades generalization quality in standard MBR. Finally, ablation studies validate that utilizing model log-probabilities and normalizing the sampled utility scores prior to ALS are critical stabilizing steps for creating a robust, low-error approximation of the ground-truth matrix.

1 INTRODUCTION

Minimum Bayes Risk (MBR) decoding (Kumar & Byrne, 2004) was proposed to mitigate a fundamental limitation of the Maximum a posteriori (MAP) decoding: the most probable output sequence is not always the highest-quality output. Multiple studies (Eikema & Aziz, 2020; Freitag et al., 2022) have demonstrated that MBR decoding yields superior generation quality compared to MAP decoding. However, MBR suffers from a severe computational bottleneck. Because it requires calculating utility scores between every candidate against a pool of pseudo-references, the decoding process incurs an $O(N^2)$ computational cost.

Recent studies have explored methods to improve the efficiency of MBR decoding. One impactful approach is Probabilistic MBR (PMBR) (Trabelsi et al., 2024), which drastically reduces pairwise calculations. PMBR operates by uniformly subsampling a small fraction of candidate–pseudo-reference pairs, calculating their utility scores, and approximating the remaining pairwise matrix using Alternating Least Squares (ALS) factorization (Bell & Koren, 2007).

Despite its success, we argue that the uniform random subsampling employed in PMBR is suboptimal. Uniform random sampling treats all data points equally, wasting constrained computational budget on uninformative, average points. In matrix approximation, highly influential outliers best define population boundaries and latent relationships. Random sampling frequently misses these pairs, starving the ALS algorithm of data crucial for accurate reconstruction. Consequently, uniform random sampling requires a significantly higher sampling budget to adequately approximate the full distribution.

To mitigate this inefficiency, we propose a proxy-guided subsampling framework to replace random sampling in PMBR as depicted in Figure 1. We utilize the model's aggregated token log probabilities, which are generated during the initial decoding phase at zero additional computational cost, as a proxy score to identify influential pairs. We demonstrate that replacing random sampling with our proxy-guided sampling yields better generation quality, without meaningfully increasing inference time. Specifically, our main contributions are as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

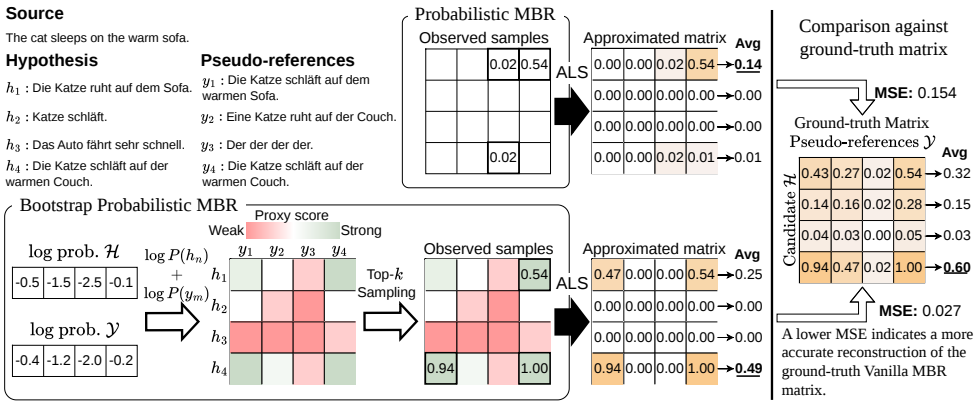

Figure 1: Overview of our proposed methodology. Top-k is depicted here for illustrative purposes.

- We propose Bootstrap PMBR (B-PMBR), a proxy-guided subsampling framework leveraging model log-probabilities to select influential candidate–pseudo-reference pairs for improved ALS matrix reconstruction.
- We demonstrate that B-PMBR consistently outperforms PMBR and performs competitively against MBR across diverse utility functions, evaluation metrics, and reference sizes.
- We show B-PMBR naturally mitigates MBR's utility metric overfitting, establishing it as an efficient alternative to mitigation techniques like SVD-MBR (Soetedjo et al., 2026) and EnsembleMBR (Kovacs et al., 2024).
- Our ablation studies confirm that model log-probabilities outperform lightweight metrics as proxy scorers, and that standardizing sampled scores prior to ALS factorization ensures robust matrix reconstruction across varying parameters.

2 Background and Related Works

MBR decoding. MBR decoding comprises four main components: a candidate set $\mathcal{H}$, a pseudo-reference set $\mathcal{Y}$, a utility function, and approximation estimation (Eikema & Aziz, 2020). Let N and M denote the candidate set and pseudo-references set size, respectively. An existing evaluation metric is used as a utility function $u(h_i, y_j)$ to score each candidate against every pseudo-reference to construct an $N \times M$ pairwise utility matrix. The expected utility $\hat{u}$ for a given candidate h_i is approximated using Monte Carlo estimation by averaging its scores across all pseudo-references:

$$\hat{u}(h_i) = \frac{1}{M} \sum_{y_j \in \mathcal{Y}} u(h_i, y_j). \tag{1}$$

The algorithm selects the candidate h^* that maximizes this expected utility as the final output $h^* = \arg\max_{h_i \in \mathcal{H}} \hat{u}(h_i)$. In this pipeline, the majority of computational resources are consumed by calculating the pairwise scores, requiring $N \times M$ expensive utility calls (Eikema & Aziz, 2022).

PMBR decoding. Due to the expensive resource required for utility function calls in MBR, multiple studies (Eikema & Aziz, 2022) have attempted to optimize this specific bottleneck. PMBR (Trabelsi et al., 2024) reduces this cost by uniformly subsampling a small fraction of candidate–pseudo-reference pairs. Formally, given a budget $1/r$, PMBR constructs a sparse observation mask Ω consisting of randomly selected pair indices, where the total number of sampled pairs is constrained to $|\Omega| = (N \times M)/r$. The utility function $u(h_i, y_j)$ is calculated only for the pairs within this subset, while non-sampled pairs are treated as unobserved data. Relying on the finding that the MBR matrix $\hat{U}$ is inherently low-rank, PMBR approximates the unobserved scores using the ALS algorithm. It learns latent factors to reconstruct the matrix, yielding $\hat{U} \approx XY^\top$, which is then used to select the winning candidate. While efficient, this random masking introduces a trade-off between computational cost and reconstruction accuracy. Natsumi et al. (2025) attempted to mitigate this, utilizing a distilled version of the utility metric to efficiently evaluate more sample pairs.

Efficiency over randomness. In machine learning, subsampling is primarily used to reduce the computational cost of approximating the characteristics of large datasets. However, previous studies demonstrate that uniform random sampling is suboptimal. Wang et al. (2018) argued that regardless of the subsample size, random subsampling results in high sampling error, proposing instead to systematically select the most informative data points. Similarly, Ting & Brochu (2018) demonstrated that to accurately represent a population's boundaries, highly influential outlier points should be selected over random samples, which often waste the computational budget on uninformative, average data points. In the context of PMBR, random sampling risks missing influential pairs, subsequently starving the ALS algorithm of the crucial information required for accurate matrix reconstruction.

3 BOOTSTRAP PMBR DECODING

We propose an approach to improve the subsample selection stage prior to matrix approximation in PMBR as depicted in Figure 1. Instead of relying on random sampling, which often wastes computational budget on uninformative data for the ALS, we introduce Bootstrap Probabilistic MBR (B-PMBR). This framework leverages low-cost proxy scores to guide the subsampling of candidate–pseudo-reference pairs. The B-PMBR pipeline comprises three core components: a proxy scorer to identify influential pairs, a selection algorithm, and the standardization of the sampled pairs' utility function scores prior to matrix completion.

3.1 PROXY SCORER

To identify influential candidate–pseudo-reference pairs without invoking expensive evaluation metrics, we define a proxy scorer as a lightweight approximation metric. Specifically, we utilize the model's aggregated token log-probabilities to construct a pairwise proxy score matrix. Because these log-probabilities are computed by the model during the decoding phase, this signal is retrieved at almost zero additional computational cost.

Given a candidate set $\mathcal{H} = \{h_1, \ldots, h_n\}$ of size N and a pseudo-reference set $\mathcal{Y} = \{y_1, \ldots, y_m\}$ of size M, we define the aggregated token log-probabilities of each candidate and pseudo-reference as $\log P(h_n)$ and $\log P(y_m)$, respectively. Crucially, we use length-normalized log-probabilities to ensure that the proxy accurately reflects sequence quality rather than a bias against longer generations (Jinnai et al., 2024). We formulate the proxy score as an estimation of the joint probability of the candidate and the pseudo-reference occurring together. In log-space, this is mathematically achieved through the addition of their respective length-normalized log-probabilities:

$$Proxy(h_n, y_m) = \log P(h_n) + \log P(y_m). \tag{2}$$

This additive formulation creates a dense $N \times M$ proxy matrix that acts as an efficient model-based quality signal, penalizing pairs where either the candidate or the reference represents a low-confidence generation. Furthermore, in MBR settings where the candidate and pseudo-reference sets are identical (i.e., $\mathcal{H} = \mathcal{Y}$), evaluating a sequence against itself guarantees the maximum utility score when evaluated by conventional evaluation metrics. To prevent the self-pairs from being selected, we mask the diagonal of the proxy matrix by setting $Proxy(h_n, y_n) = -\infty$. We further explore the impact of diagonal masking on PMBR in Appendix E

3.2 SUBSAMPLING ALGORITHMS

Following PMBR, we limit our computational budget to a pre-defined sample size k. To replace naive uniform random sampling, we explore three distinct yet simple subsampling algorithms:

Top-k. This approach deterministically selects the k indices with the highest proxy scores. While it targets high-scoring pairs, it severely shrinks sample variance and risks ignoring mediocre rows, which can lead to zero observations for certain hypotheses and destabilize the matrix factorization.

Edge-k. Inspired by optimal subdata selection principles, this method selects the extreme boundaries of the distribution to maximize efficiency. We choose the top $k/2$ and bottom $k/2$ proxy pairs.

Gumbel-Top-k. This method implements the Gumbel-Top-k trick (Kool et al., 2019) to achieve stochastic sampling without replacement that is proportional to the exponentiated proxy scores. First, we isolate non-masked pairs and then generate uniform noise $u \sim \text{Uniform}(0, 1)$, which is

then transformed into Gumbel noise via the equation $g = -\log(-\log(u + \epsilon))$. The $\epsilon = 10^{-9}$ prevents errors that would corrupt the tensor with NaNs or infinite values. After perturbing the proxy scores with this noise, the algorithm performs a Top-k selection on the noisy distribution.

3.3 Standardization before ALS

Once the subsampling algorithm successfully identifies the subset of influential candidate–pseudo-reference pairs, we evaluate them using the utility function. The remaining uncalculated pairs are initialized to 0, similar to PMBR. However, drawing on established practices from recommendation systems (Zhou et al., 2008), applying matrix factorization directly to raw scores can sometimes distort latent factor learning or slow down convergence. To improve the stability of the ALS process, we apply standardization to the sampled utility scores. Specifically, we apply z-score normalization element-wise to the computed utility scores. This standardizes the distribution scale and improves the model's ability to accurately approximate the uncalculated pairwise scores. Furthermore, Appendix D shows that integrating the standardization into PMBR yields performance improvements.

4 Experiments

4.1 Experimental Setup

Dataset and model. We conducted our experiments using the WMT-22 English-to-German (En→De) test set (Kocmi et al., 2022), a standard benchmark utilized in recent MBR decoding literature (Trabelsi et al., 2024; Soetedjo et al., 2026). We used the M2M100 model (418M parameters) to generate output sets in multiple sizes: 4, 8, 16, 32, 64, 128, and 256 using ϵ-sampling (Hewitt et al., 2022) with $\epsilon = 0.02$ (Freitag et al., 2023). We fixed the candidate set size at $N = 256$ and used all generated sizes as pseudo-references.

Utility function. We employed 5 different metrics as utility functions: BLEU (Papineni et al., 2002), chrF (Popović, 2015), COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020). This allows us to analyze our method across multiple utility spaces, specifically contrasting the behavior of lexical metrics against neural metrics.

Baseline. As baselines, we utilized standard MBR and PMBR under identical candidate sizes, pseudo-reference sizes, and utility functions. For PMBR, we fixed the parameters such that the rank $d = 8$, regularization $\lambda = 0.1$, the reduction factor $r = 8.0$, and ran up to 10 ALS iterations. To ensure a fair comparison against B-PMBR, all parameters remained identical except for the sampling budget k, which was constrained to $(1/r) \times N \times M$ for all B-PMBR settings.

Evaluation. In addition to the specific metrics used as utility functions, we employed XCOMET (Guerreiro et al., 2024) and COMETKiwi (Rei et al., 2022) to measure translation quality. For readability and to measure generalized quality, we combined all the evaluation metrics (i.e., BLEU, chrF, COMET, BLEURT, BERTScore, XCOMET, and COMETKiwi) into a unified metric $\bar{Z}_{other}$ as suggested by Soetedjo et al. (2026), where each metric is standardized using a z-score and averaged. Beyond translation quality, we assessed computational cost by measuring the total execution time of the MBR decoding process (excluding candidate generation time).

Hardware and implementation. To ensure comparable computation costs, all processes were executed on a Tesla V100S-PCIE-32GB GPU using 16 cores of an Intel Xeon W-3223 CPU @ 3.50GHz. All baseline and proposed methods were implemented using the modular `mbrs` (Deguchi et al., 2024) library to guarantee standardized and reproducible evaluations.

4.2 Result 1: B-PMBR exceeds PMBR and competitive to MBR

Figure 2 shows the results in three evaluation metrics: the metric being used as a utility function, XCOMET, and $\bar{Z}_{other}$. Across all evaluated utility functions, B-PMBR consistently and significantly outperforms the PMBR baseline and remains competitive with standard MBR, regardless of the sampling strategy employed (see Appendix B for paired bootstrap test results). This demonstrates that utilizing the model's log-probabilities as a zero-cost proxy successfully guides the algorithm toward high-utility candidate pairs, preserving downstream quality while operating on a strictly reduced computational budget.

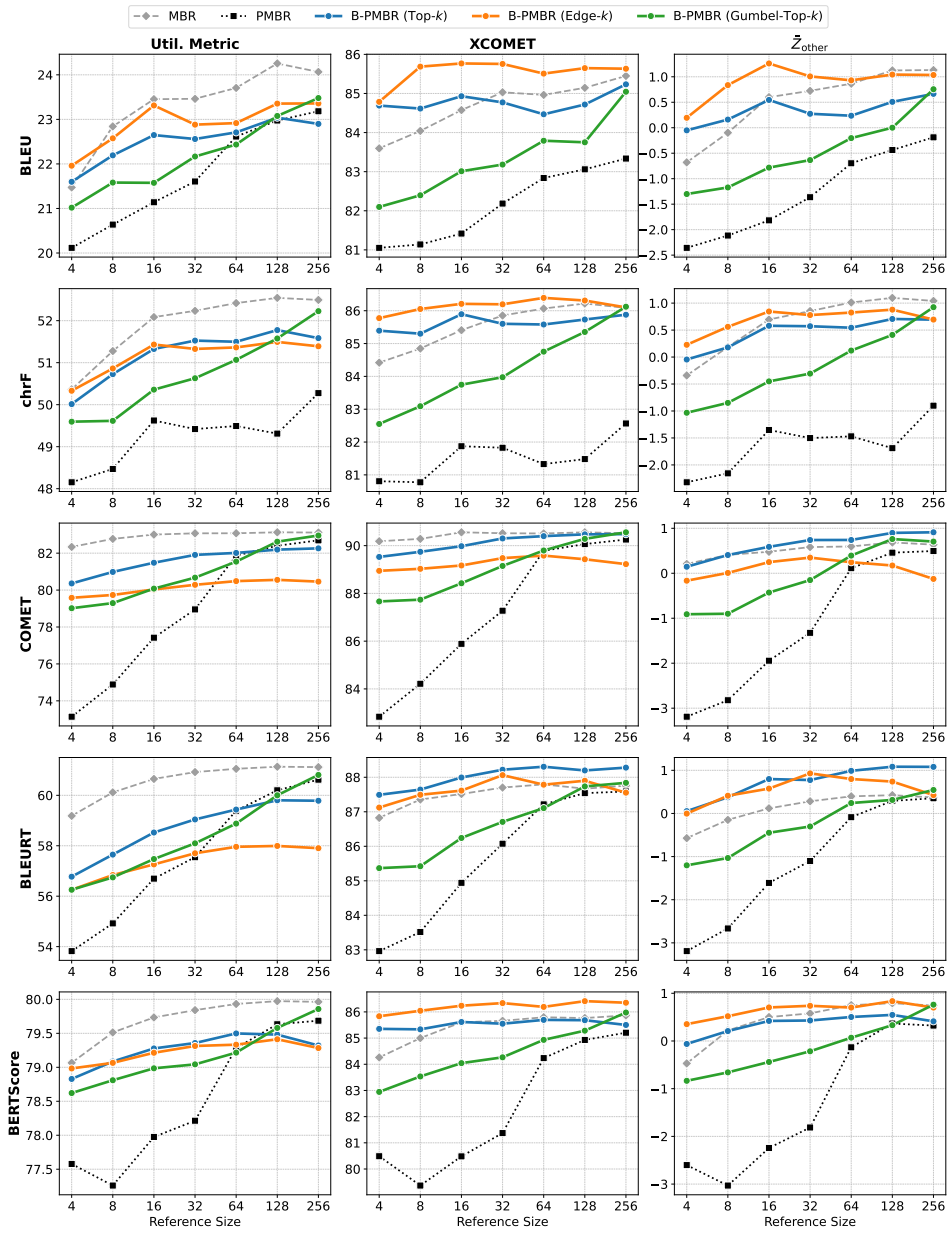

Figure 2: Performance of various B-PMBR against MBR and PMBR on the WMT-22 En→De dataset. Each row in the grid represents a different target utility function, while each column corresponds to the evaluated metrics. Specifically, **Util. Metric** denotes the evaluation result measured using the exact same metric as the utility function for that respective row.

A key limitation of PMBR's random sampling is its reliance on large reference sets, which causes low initial quality that only improves with scale. In contrast, B-PMBR exhibits remarkable stability. By targeting the most informative matrix subsets, B-PMBR achieves near-peak performance even at minimal reference budgets (e.g., 4 or 8) and maintains this high-quality trajectory across scaling.

The proposed framework demonstrates robust adaptability across different scoring spaces, yielding strong results whether optimized on lexical metrics (BLEU, chrF) or neural metrics (BLEURT, COMET, BERTScore). B-PMBR performs exceptionally well when BLEURT is utilized as the target utility function. In this specific configuration, the majority of B-PMBR configurations exceed

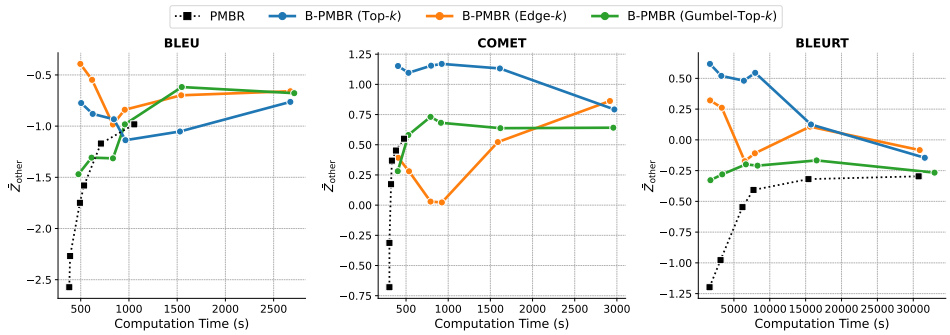

Figure 3: Pareto frontier illustrating the trade-off between $\bar{Z}_{other}$ and total computation time on the WMT-22 En→De dataset. Each subplot represents the utility function being used. The curves track the performance of PMBR against B-PMBR configurations across varying computational budgets (r for PMBR and k for B-PMBR). Configurations that arch toward the top-left of each subplot indicate the optimal balance of high evaluation performance and low computational cost.

the generation quality of the MBR baseline across $\bar{Z}_{other}$. Furthermore, these performance trends generalize to the WMT-22 De→En dataset (see Appendix F).

4.3 Result 2: B-PMBR optimizes the trade-off between quality and cost

To thoroughly assess the cost-to-value performance of our framework, we isolate the evaluation to the largest matrix, fixing both the candidate and reference set sizes at 256. We sweep the computational budget across 6 different scales by adapting the PMBR reduction factor $r \in \{40, 20, 10, 8, 4, 2\}$. We then plot the total computation time of the decoding process against the generalized quality score ($\bar{Z}_{other}$) to form a Pareto frontier, as illustrated in Figure 3.

The Pareto curves demonstrate that B-PMBR yields a drastically more efficient frontier than the random sampling of PMBR. Strikingly, most B-PMBR configurations operating at the lowest reduction factor equivalent ($r = 40$) achieve a higher $\bar{Z}_{other}$ than PMBR configurations utilizing even the largest budget. As expected, scaling the sampling budget up (by decreasing the reduction factor r) increases the total execution time required for B-PMBR to finish the decoding process. However, the utility scoring step continues to dominate the total time (further details are provided in Appendix C).

A critical observation from the frontier is that B-PMBR is highly sensitive to the sampling budget k and the specific utility function being optimized. When utilizing continuous neural metrics like COMET or BLEURT, increasing the sampling budget beyond a certain point actually decreases the generalized score, likely due to more noisy samples being included. Conversely, when optimizing a lexical metric like BLEU, the generalized performance often recovers or increases as the sampling scale grows. Consequently, there is no universally optimal k across all configurations; the ideal budget and sampling strategy must be tailored specifically to the utility function being applied.

4.4 Result 3: B-PMBR mitigates utility metric overfitting

MBR decoding is susceptible to a known overfitting problem: by maximizing the utility function, the algorithm degrade the overall quality (Kovacs et al., 2024; Soetedjo et al., 2026). As observed in Figure 2, although B-PMBR improves $\bar{Z}_{other}$ compared to standard MBR, it consistently scores worse on the utility metric (Util. Metric). To better understand this observation, we evaluate our method against two existing overfitting mitigation techniques: SVD-MBR (Soetedjo et al., 2026) and EnsembleMBR Kovacs et al. (2024). For SVD-MBR, we configure the denoising with top-$k = 2$. For EnsembleMBR, we ensemble all utility functions using the `rankAvg` setting; for each metric, this approach assigns a rank to each of the 128 samples (where 0 is best and 127 is worst) and selects the sample that minimizes the average rank across all metrics.

Table 1 denote our proposed method against both techniques. We treat EnsembleMBR as the theoretical upper bound. When evaluating the generalized quality ($\bar{Z}_{other}$), B-PMBR in most instances

Table 1: Performance of B-PMBR configurations against EnsembleMBR and SVD-MBR on the WMT-22 En→De dataset across different reference sizes (M). The table reports the utility metric score (Util) alongside the generalized quality ($\bar{Z}_{other}$). The ↓ and ↑ symbols indicate that lower and higher scores are better, respectively. **Bold** indicates the best performance between SVD-MBR and B-PMBR for each column, excluding EnsembleMBR.

Util. Function	Method	M=4		M=32		M=256	
		Util↓	$\bar{Z}_{other}$ ↑	Util↓	$\bar{Z}_{other}$ ↑	Util↓	$\bar{Z}_{other}$ ↑
BLEU	EnsembleMBR	22.63	+2.68	23.91	+3.79	24.12	+4.03
	SVD-MBR	21.10	−1.64	23.35	+0.01	23.90	+0.37
	PMBR	**20.12**	−3.24	**21.61**	−2.13	23.18	−0.83
	B-PMBR (Top-k)	21.60	−0.67	22.56	−0.31	**22.90**	+0.11
	B-PMBR (Edge-k)	21.96	**−0.41**	22.88	**+0.50**	23.36	**+0.52**
	B-PMBR (Gumbel Top-k)	21.02	−2.08	22.17	−1.32	23.48	+0.21
chrF	EnsembleMBR	51.00	+1.48	52.07	+2.56	52.30	+2.78
	SVD-MBR	50.27	−1.61	52.11	**+0.16**	52.50	**+0.50**
	PMBR	**48.15**	−4.14	**49.42**	−3.05	**50.28**	−2.25
	B-PMBR (Top-k)	50.01	−1.05	51.53	−0.25	51.58	−0.08
	B-PMBR (Edge-k)	50.33	**−0.69**	51.33	+0.04	51.39	−0.07
	B-PMBR (Gumbel Top-k)	49.60	−2.42	50.63	−1.44	52.23	+0.22
COMET	EnsembleMBR	78.39	−1.33	79.48	+1.00	79.69	+1.49
	SVD-MBR	82.31	**−1.15**	83.17	+0.21	83.28	+0.26
	PMBR	**73.14**	−11.82	**78.95**	−6.00	82.69	−0.32
	B-PMBR (Top-k)	80.36	−1.38	81.91	**+0.47**	82.26	**+1.00**
	B-PMBR (Edge-k)	79.58	−2.35	80.29	−0.75	**80.46**	−2.24
	B-PMBR (Gumbel Top-k)	79.02	−4.70	80.67	−2.33	82.96	+0.35
BLEURT	EnsembleMBR	56.40	+1.58	57.57	+3.45	57.77	+3.85
	SVD-MBR	58.98	−1.71	60.75	+0.16	61.06	+0.38
	PMBR	**53.82**	−7.40	**57.55**	−2.96	60.62	+0.05
	B-PMBR (Top-k)	56.77	**−0.56**	59.04	+0.94	59.78	**+1.56**
	B-PMBR (Edge-k)	56.25	−0.71	57.70	**+1.22**	**57.90**	+0.20
	B-PMBR (Gumbel Top-k)	56.26	−3.24	58.10	−1.33	60.81	+0.46
BERTScore	EnsembleMBR	79.24	+1.58	79.62	+2.68	79.72	+2.90
	SVD-MBR	78.92	−1.61	79.70	+0.06	79.85	**+0.40**
	PMBR	**77.58**	−4.86	**78.21**	−3.65	79.69	−0.35
	B-PMBR (Top-k)	78.83	−0.95	79.36	−0.18	79.32	−0.21
	B-PMBR (Edge-k)	78.98	**−0.31**	79.31	**+0.30**	**79.28**	+0.24
	B-PMBR (Gumbel Top-k)	78.62	−2.14	79.04	−1.19	79.86	+0.33

outperforms SVD-MBR. Crucially, SVD-MBR requires the full pairwise matrix to be calculated before it can apply its singular value decomposition denoising, while B-PMBR strictly evaluates only a small sample of the utility matrix. When comparing $\bar{Z}_{other}$ against EnsembleMBR, B-PMBR still exhibits a notable performance gap across multiple configurations. However, EnsembleMBR is computationally prohibitive for standard use, as it requires calculating a completely different full pairwise matrix for every single utility function included in the ensemble. Given this immense overhead, we consider B-PMBR to be an efficient and practical alternative that captures much of the overfitting-mitigation benefit at a fraction of the cost.

Two additional observations emerge from the results. First, across all configurations, increasing the reference size (M) consistently improves both the Util and $\bar{Z}_{other}$ performance. Second, although the targeted utility score (Util) is lower for B-PMBR compared to EnsembleMBR, this reduction does not automatically guarantee better generalized quality. This highlights that simply minimizing the targeted metric is not a path to better generalization, further demonstrating the necessity of carefully balancing the utility score with the generalized score.

Table 2: Performance of different proxy scorers in multiple sampling algorithms against the PMBR. The reported values denote the generalized quality ($\bar{Z}_{other}$). **Bold** and underlined values denote the best and second-best performances within each column, respectively.

		Util. Function				
Sampling Type	**Proxy Scorer**	**BLEU**	**chrF**	**COMET**	**BLEURT**	**BERTScore**
PMBR		0.22	-0.51	0.46	0.06	0.60
Top-k	Norm. Log-Prob.	0.90	0.95	**1.47**	**2.18**	0.73
	BLEU	0.17	0.38	0.13	0.12	0.11
	chrF	-1.19	-0.75	0.34	-0.35	-0.97
Edge-k	Norm. Log-Prob.	**1.16**	0.96	-1.39	0.22	1.11
	BLEU	0.06	0.17	-1.22	-1.01	-0.11
	chrF	0.10	-0.24	-0.82	-0.92	-0.23
Gumbel-Top-k	Norm. Log-Prob.	0.93	**1.14**	1.06	0.66	**1.15**
	BLEU	-0.33	0.04	0.24	-0.30	-0.39
	chrF	-2.03	-2.13	-0.27	-0.67	-2.01

5 Ablation Study

In this section, we conduct ablation studies to isolate and evaluate the individual contributions of the core mechanisms introduced in our Bootstrap PMBR (B-PMBR) framework. Specifically, we investigate two critical design choices: the effectiveness of utilizing model log-probabilities as a proxy scorer compared to other lightweight evaluation metrics, and the impact of applying z-score standardization to the sampled utility scores prior to the ALS factorization step.

5.1 Alternative proxy scorer

While length-normalized model log-probabilities provide an effective zero-cost proxy, they are inaccessible when evaluating outputs from black-box language models. To accommodate these restricted scenarios, we evaluated two lightweight, metric-based proxy alternatives: BLEU and chrF. The results are detailed in Table 2, where we isolated the evaluation to a fixed candidate and pseudo-reference size of 256. To ensure a fair comparison against the PMBR, we restricted the sampling budget (k) to 12.5% of the total matrix space, which is mathematically equivalent to the 8.0 reduction factor (r) utilized in PMBR.

The empirical results demonstrate that using model log-probability as proxy is overwhelmingly superior. Across every evaluated subsampling algorithm, the log-probability proxy consistently secures the best or second-best $\bar{Z}_{other}$ compared to both BLEU and chrF. Furthermore, it maintains strong, positive downstream performance regardless of the specific utility function being optimized, whereas the metric-based proxies frequently degrade the generalized quality into negative values.

It is worth noting that in a few instances, utilizing BLEU or chrF as a proxy can yield marginally better performance than the random sampling of the PMBR. However, the practical viability of this approach is fundamentally flawed by its computational overhead. Unlike log-probabilities, which are extracted during the initial generation phase, employing an external metric like BLEU or chrF requires the explicit calculation of a full pairwise matrix just to establish the proxy scores. Because this preliminary step demands substantial computation before the actual subsampling even begins, the total execution cost will not be any better than simply utilizing PMBR.

5.2 Impact of normalization

To thoroughly evaluate the impact of normalization, we measure the matrix approximation accuracy by computing the Mean Squared Error (MSE) between the ALS-completed matrix and the ground-truth MBR matrix. Because ALS factorization shrinks the variance of the reconstructed values, directly comparing a z-score standardized matrix to a raw metric matrix would yield heavily biased results. Therefore, for the normalized B-PMBR configurations, we retrieve the mean and standard

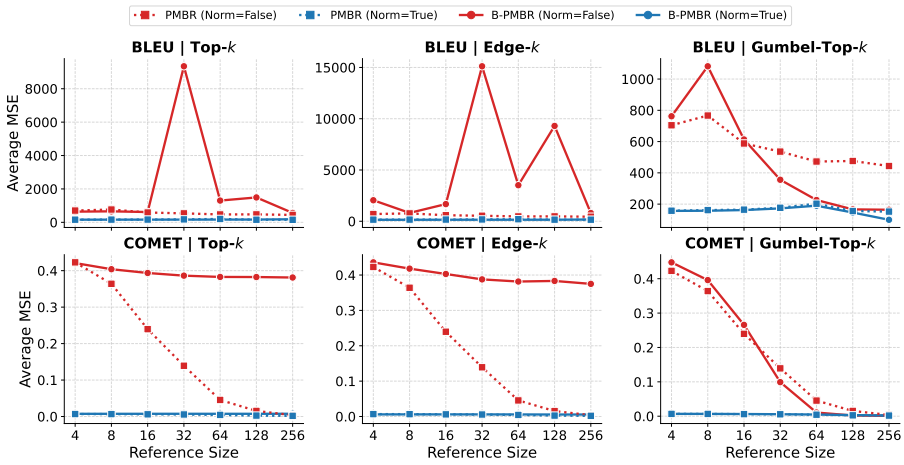

Figure 4: The impact of applying z-score normalization to the sampled utility scores prior to ALS factorization. The figure illustrates the matrix reconstruction error (MSE) against the ground-truth MBR matrix across varying reference counts. Each row represents the utility function being optimized, while each column corresponds to a specific B-PMBR sampling variant.

deviation of the ground-truth matrix to inverse-transform and scale the ALS-completed matrix back to the raw metric space prior to calculating the final MSE. We also calculate the MSE of PMBR in two conditions: normalized and unnormalized, to serve as our baseline comparison.

The results across multiple reference sizes are illustrated in Figure 4. Applying z-score normalization universally stabilizes the matrix factorization. The normalized B-PMBR and PMBR consistently maintain a low MSE that remains close to 0, completely independent of the reference count being evaluated. This sharply contrasts with the unnormalized variants, which suffers from high initial error and only scales down to a smaller MSE as the reference size becomes sufficiently larger. Specifically for the unnormalized B-PMBR, it consistently exhibit the largest reconstruction errors compared to PMBR, failing to accurately reconstruct the matrix regardless of the evaluated scale.

This instability in the unnormalized configurations becomes even more apparent when analyzing the behavior across different sampling types and utility functions. The unnormalized Top-k and Edge-k demonstrate similarly erratic and massive error spikes, while Gumbel-Top-k follows a downward-sloping trend similar to PMBR. From the perspective of the utility functions, the reconstruction of the normalized variants are robust across both lexical (BLEU) and neural (COMET) metric. Without this standardization, the ALS algorithm struggles to properly learn the latent factors, resulting in unpredictable reconstruction losses.

6 Conclusion

In this work, we addressed the computational bottleneck of Minimum Bayes Risk (MBR) decoding by replacing the inefficient uniform random sampling of Probabilistic MBR (PMBR) with Bootstrap PMBR (B-PMBR). B-PMBR utilizes proxy-guided subsampling, leveraging length-normalized model log-probabilities to select the most influential candidate–pseudo-reference pairs for evaluation by the computationally expensive primary utility function.

Evaluations across diverse reference sizes, utility functions, and metrics demonstrate that B-PMBR consistently outperforms PMBR and remains highly competitive with MBR, achieving near-optimal generation quality even at low reference sizes. This competitiveness stems from B-PMBR naturally mitigating the utility metric overfitting that degrades generalized translation quality in MBR, yielding an exceptional cost-to-value ratio where minimal sampling budgets outperform much larger PMBR configurations. Finally, ablation studies confirm that model log-probabilities are a superior proxy compared to lightweight evaluation metrics, and that applying z-score normalization prior to Alternating Least Squares (ALS) factorization is critical for ensuring a stable, low-error approximation of the ground-truth matrix.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods and providing feedback on experiment designs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing research methodology, assisting with translation, supporting qualitative and thematic data analysis, or interpreting results, and generating synthetic data sets, formulating mathematical claims, cleaning and reformatting datasets, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for assisting in the writing or wording of the content to improve readability and sourcing/searching for information. We have reviewed all AI-assisted work. Specifically, we thoroughly tested all AI-implemented methods for correctness, critically evaluated all AI-provided feedback on experiment designs before adoption, manually verified AI-sourced information against peer-reviewed literature, and reviewed all AI-assisted writing to ensure accuracy and maintain our original voice. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study fully complies with the ICLR Code of Ethics and addresses all relevant ethical considerations. All resources used in this work are publicly available and properly licensed, with no concerns regarding licensing. The study does not involve or produce any harmful content. Given these points, we affirm that this work raises no ethical concerns.

REFERENCES

Robert M. Bell and Yehuda Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 43–52, 2007. doi: 10.1109/ICDM.2007.90.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mbrs: A library for minimum Bayes risk decoding. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 351–362, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-demo.37. URL https://aclanthology.org/2024.emnlp-demo.37/.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398/.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.754. URL https://aclanthology.org/2022.emnlp-main.754/.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 07 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00491. URL https://doi.org/10.1162/tacl_a_00491.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9198–9209, Singapore, December 2023. Association

for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.617. URL https://aclanthology.org/2023.findings-emnlp.617/.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. Model-based minimum bayes risk decoding for text generation, 2024. URL https://arxiv.org/abs/2311.05263.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.1. URL https://aclanthology.org/2022.wmt-1.1/.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu (eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kool19a.html.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. Mitigating metric bias in minimum Bayes risk decoding. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1063–1094, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.109. URL https://aclanthology.org/2024.wmt-1.109/.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022/.

Koki Natsumi, Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Agreement-constrained probabilistic minimum Bayes risk decoding. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 484–493, Mumbai, India, December

2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-299-2. doi: 10.18653/v1/2025.ijcnlp-short.39. URL https://aclanthology.org/2025.ijcnlp-short.39/.

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 (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.60. URL https://aclanthology.org/2022.wmt-1.60/.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Riza Setiawan Soetedjo, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Overfitting mitigation via singular value decomposition in minimum bayes risk decoding, 2026. URL https://arxiv.org/abs/2609.01135.

Daniel Ting and Eric Brochu. Optimal subsampling with influence functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/57c0531e13f40b91b3b0f1a30b529a1d-Paper.pdf.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 54714–54733. Curran Associates, Inc., 2024. doi: 10.52202/079017-1734. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/626ab938fe19200324b368f5ee816868-Paper-Conference.pdf.

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2018.

ISSN 1537-274X. doi: 10.1080/01621459.2017.1408468. URL http://dx.doi.org/10.1080/01621459.2017.1408468.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In Rudolf Fleischer and Jinhui Xu (eds.), *Algorithmic Aspects in Information and Management*, pp. 337–348, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-68880-8.

A Limitations and Future Works

Evaluation scope. Our empirical evaluation is conducted on the WMT-22 English-to-German test set within the machine translation (MT) task. We selected this specific language pair as a representative dataset to clearly demonstrate our results. Because our primary contribution is establishing the effectiveness of our proposed proxy-guided subsampling method, we deliberately focused our exploration on evaluating our proposed components in comparison to standard PMBR and other MBR variations. Consequently, it remains unverified how well this method transfers to other language pairs and other standard MBR tasks, such as abstractive text summarization and image captioning.

Exploration of generation methods. Additionally, we did not explore alternative generation methods, testing our framework exclusively under ϵ-sampling. We intentionally constrained our exploration to the MBR decoding process that occurs after the candidates and pseudo-references are generated, as this is where the main computational bottleneck of MBR is localized. Furthermore, as suggested by previous studies (Freitag et al., 2023), we maintained the generation configuration to ϵ-sampling because it is established as the optimal generation strategy for MBR in MT tasks. Future work could investigate the behavior of our proposed method when candidates are generated using recent alternative sampling strategies, or when evaluating disjoint candidate and pseudo-reference sets.

B Statistical Significance Test

To establish the statistical significance of our results, we conducted a paired bootstrap resampling (Koehn, 2004) test utilizing 10,000 iterations. The evaluation focuses on the generalized quality metric, $\bar{Z}_{other}$, measured on the WMT-22 En→De dataset. We kept the reduction factor r to 8 for both PMBR and B-PMBR variants, consistent with the result in Figure 2. We set the p-value < 0.05, which is visualized in Figure 5 where significant improvements are marked in green and non-significant results in red.

When comparing B-PMBR against MBR, our framework achieves statistically significant improvements on several occasions. This is especially evident for the Top-k and Edge-k sampling algorithms, which frequently yield significantly better generalized performance than MBR on specific utility functions. On the other hand, when evaluated against PMBR operating under the identical computational budget, our proposed B-PMBR method is predominantly and significantly better across almost all testing configurations. The only exceptions to this superiority occur at higher reference sizes (i.e. 64, 128, and 256) for a few utility functions, where the performance gap between the two methods narrows and is no longer statistically significant.

C Computation Time Distribution

§4.3 demonstrates that while our proposed methods consistently yield superior generation quality, they inherently incur a higher total computation time than PMBR. To better understand this overhead, we break down the time required for each decoding stage, averaged across the multiple reduction factor budgets tested.

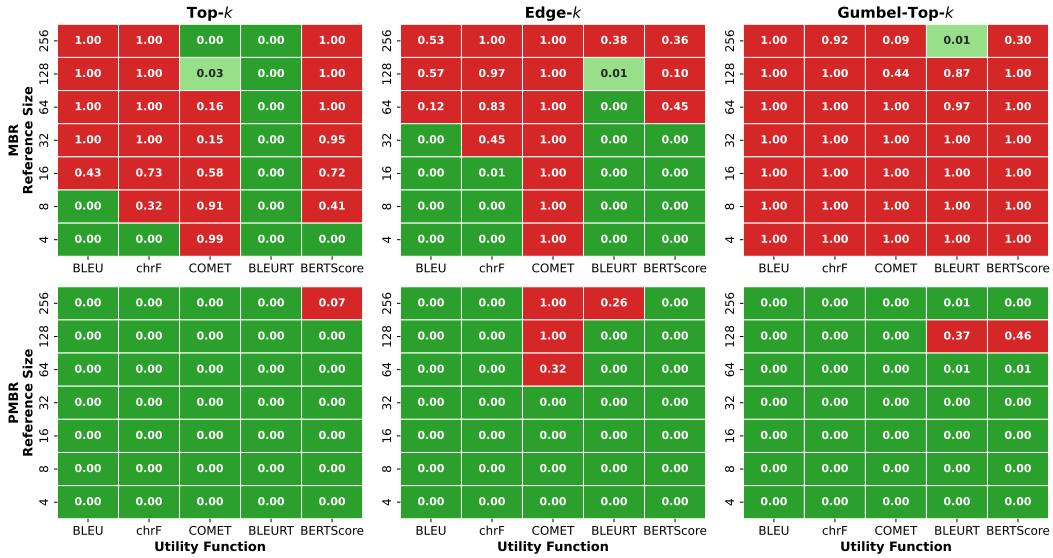

Figure 5: Statistical significance heatmaps comparing B-PMBR sampling variants against MBR (top) and PMBR (bottom) using $\bar{Z}_{other}$ on the WMT-22 En→De dataset. The values within each cell report the exact p-values; green ($p < 0.05$) demonstrate that the B-PMBR variant yields a statistically significant performance improvement over the respective baseline, while red indicate no significant difference.

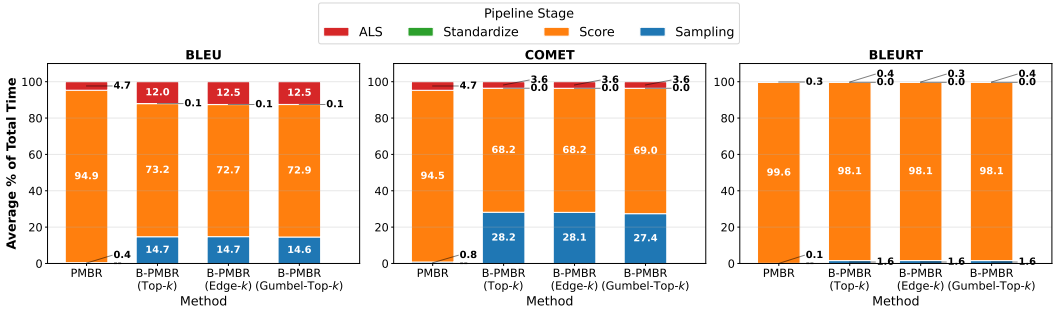

Figure 6: Computation time distribution across the decoding pipeline stages (Sampling, Score, Standardize, and ALS) for PMBR and B-PMBR variants on the WMT-22 En→De dataset. The percentages are averaged across multiple budgets in § 4.3. Each subplot corresponds to a specific utility function. Notably, unlike PMBR, B-PMBR incorporates a standardization step before ALS.

As shown in Figure 6, the proposed proxy-guided sampling stage, which leverages the model's generation log-probabilities, never surpasses the computationally intensive utility scoring step. At most, our targeted sampling algorithms account for approximately 28% of the relative computation time (as observed with COMET). This high proportion is primarily an artifact of the utility function being cacheable and highly optimized, which reduces the time required to compute the scores and consequently inflates the sampling phase's percentage. Conversely, when utilizing a highly expensive neural metric like BLEURT, utility scoring severely dominates the pipeline (accounting for over 98% of the total time), rendering the algorithmic overhead of our proxy sampling process practically negligible at less than 2%.

Furthermore, the addition of our z-score standardization step prior to ALS factorization contributes an inconsequential fraction of the execution time, remaining at or below 0.1% across all tested configurations. Thus, we can conclude that our proposed sampling method introduces a manageable overhead that is eclipsed by the bottleneck of metric evaluations, proving **B-PMBR to be an exceptionally cost-effective trade-off for its performance gains**.

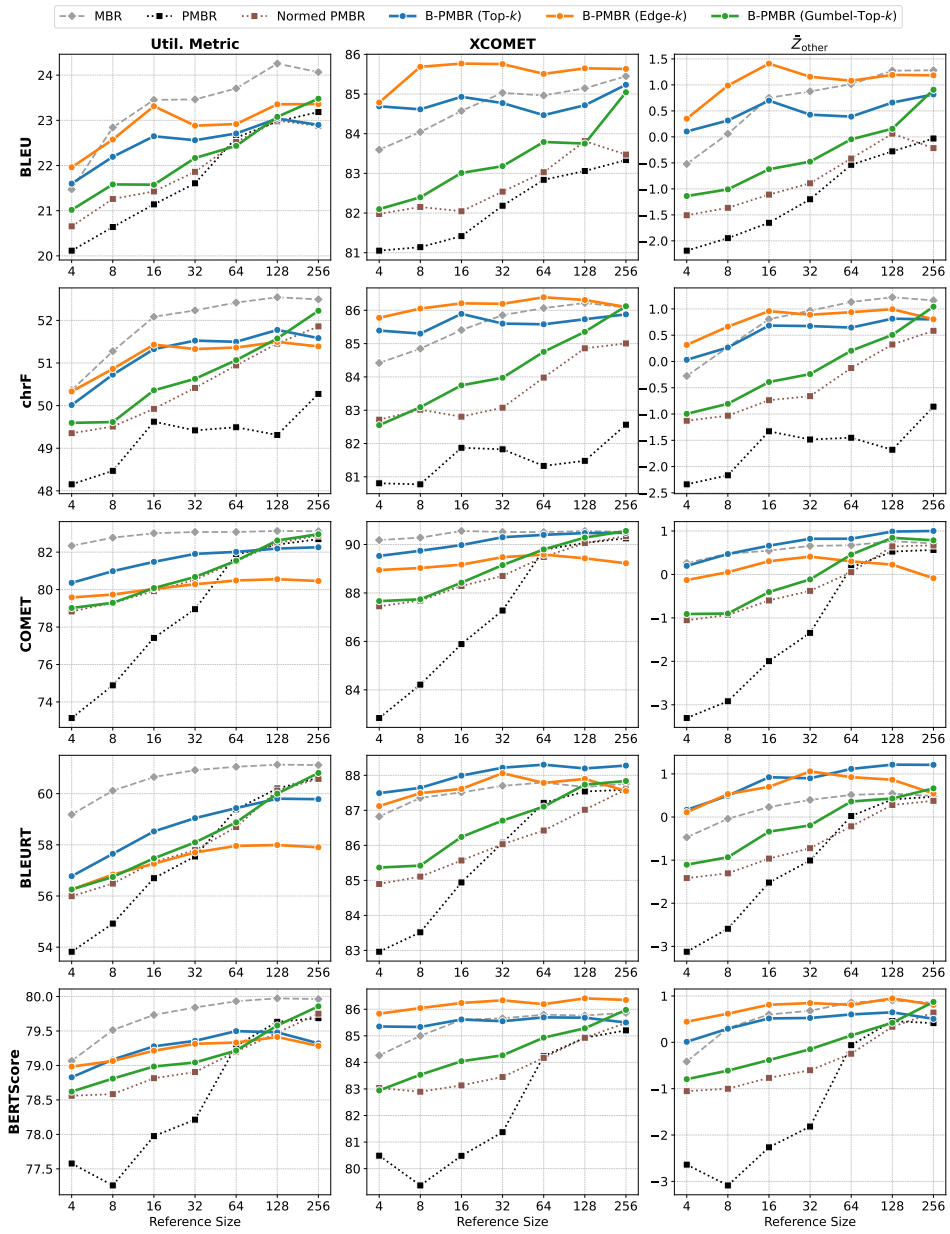

Figure 7: Performance of various B-PMBR against MBR, PMBR, and Normalized PMBR on the WMT-22 En→De dataset. Each row in the grid represents a different target utility function, while each column corresponds to the evaluated metrics. Specifically, **Util. Metric** denotes the evaluation result measured using the exact same metric as the utility function for that respective row.

D STANDARDIZATION BEFORE ALS ALSO IMPROVES PMBR

§5.2 demonstrates that applying z-score normalization to the sampled pairwise scores before ALS factorization achieves a consistently low Mean Squared Error (MSE) against the ground-truth MBR matrix. To understand how this approximation accuracy translates to actual quality, we evaluate the performance of PMBR with and without this standardization step. As illustrated in Figure 7, integrating z-score standardization directly into the PMBR pipeline (Normed PMBR) yields substantial performance improvements over the unnormalized PMBR, particularly when constrained to small reference budgets (4, 8, 16, and 32).

Table 3: Impact of diagonal masking on $\bar{Z}_{other}$ across multiple reduction factors and utility functions. The candidate and pseudo-reference sets are identical ($N = M = 256$). **Bold** and underlined values indicate the highest and second-highest scores, respectively, for a given reduction factor budget.

		Reduction Factor (r)					
Util.Function	**Method**	**40**	**20**	**10**	**8**	**4**	**2**
BLEU	PMBR	-1.46	-1.27	-1.00	-0.92	-0.73	-0.56
	Masked PMBR	-0.95	-0.61	-0.44	-0.37	-0.23	-0.05
	B-PMBR (Top-k)	0.03	-0.02	**-0.05**	-0.14	-0.07	0.06
	B-PMBR (Edge-k)	**0.23**	**0.12**	-0.09	**0.01**	**0.13**	**0.13**
	B-PMBR (Gumbel Top-k)	-0.36	-0.26	-0.25	-0.07	0.11	0.11
chrF	PMBR	-1.61	-1.41	-1.15	-1.07	-0.69	-0.37
	Masked PMBR	-1.11	-0.79	-0.56	-0.54	-0.20	0.14
	B-PMBR (Top-k)	0.25	0.28	0.17	0.17	0.20	0.28
	B-PMBR (Edge-k)	**0.44**	**0.35**	**0.22**	0.16	0.41	0.45
	B-PMBR (Gumbel Top-k)	-0.08	0.07	0.11	**0.30**	**0.44**	**0.46**
COMET	PMBR	-1.02	-0.46	-0.13	-0.06	0.02	0.07
	Masked PMBR	-0.57	0.01	0.38	0.39	0.50	0.60
	B-PMBR (Top-k)	**0.93**	**1.01**	**1.04**	**1.10**	**0.93**	0.70
	B-PMBR (Edge-k)	0.63	0.53	0.17	-0.05	0.57	**0.78**
	B-PMBR (Gumbel Top-k)	0.31	0.63	0.73	0.67	0.61	0.66
BLEURT	PMBR	-1.02	-0.72	-0.49	-0.45	-0.37	-0.34
	Masked PMBR	-0.38	-0.04	0.10	0.12	0.17	0.25
	B-PMBR (Top-k)	0.69	**0.75**	**0.72**	**0.75**	**0.44**	0.25
	B-PMBR (Edge-k)	**0.86**	0.63	0.44	0.27	0.35	**0.33**
	B-PMBR (Gumbel Top-k)	0.21	0.25	0.14	0.23	0.30	0.21
BERTScore	PMBR	-1.32	-0.40	0.03	0.11	0.29	0.33
	Masked PMBR	-1.21	-0.34	0.13	0.15	0.23	0.36
	B-PMBR (Top-k)	0.22	0.32	0.04	-0.03	-0.51	-0.11
	B-PMBR (Edge-k)	**0.33**	**0.34**	**0.16**	0.15	0.43	**0.57**
	B-PMBR (Gumbel Top-k)	-0.22	-0.00	0.05	**0.41**	**0.44**	0.51

For lexical utility functions, this performance gain is maintained across the entire reference sizes. In these metric spaces, Normed PMBR consistently outperforms PMBR and closely mimics the performance trajectory of B-PMBR variants, specifically Gumbel-Top-k. This indicates that applying standardization to the sampled utility scores stabilizes the ALS latent factor learning sufficiently to act as a standalone enhancement for PMBR, proving highly beneficial even in the absence of proxy-guided subsampling.

However,this behavior diverges when optimizing for neural metrics. While Normed PMBR maintains a clear advantage over PMBR at low reference budgets, PMBR improves at a faster rate as the reference size increases (i.e. 64, 128, and 256), eventually overtaking Normed PMBR. This scaling divergence suggests that while z-score standardization prevents ALS factorization failures in data-starved environments, forcibly expanding the variance of naturally compressed neural metrics may slightly hinder the algorithm's ability to approximate the points when abundant observation data is available.

E Diagonal Masking improves PMBR

To evaluate the impact of masking self-pairs from the selection pool, we compare PMBR against a Masked PMBR variant. The motivation of masking is to prevent the algorithm from wasting its computational budget on uninformative self-comparisons, which inherently yield maximum utility scores. For this experiment, both the candidate and pseudo-reference sets are identical and fixed at a

size of 256. We tested the configurations across six different reduction factors: 40, 20, 10, 8, 4, and 2. Because uniform random sampling creates volatility, the results for both PMBR and Masked PMBR are averaged across 5 different random seeds (0, 5, 23, 42, 78) to ensure a trustworthy baseline. Conversely, Bootstrap PMBR (B-PMBR) is kept at seed 0, as its proxy-guided pair selection is deterministic instead of random.

Comparing PMBR and its masked variant in Table 3, it is clear that diagonal masking mostly improves the overall performance of PMBR. This improvement consistently scales with a lower reduction factor (which provides a higher evaluation budget) and remains robust regardless of the specific utility function being used. However, despite this baseline enhancement, there is still a large performance gap when compared to the B-PMBR variants. This performance gap is especially prominent on lower budgets, with the only notable exception occurring when utilizing BERTScore at a reduction factor of 10, where Masked PMBR achieves the second-best score and manages to briefly outperform several B-PMBR configurations.

F Performance on WMT22-De→En

To validate the robustness of our framework across different language directions, we evaluated B-PMBR on the WMT-22 German-to-English (De → En) dataset. We kept all experimental settings and hyperparameters identical to §4. The detailed results across the three evaluation metrics (Util. Metric, XCOMET, and $\bar{Z}_{other}$) are illustrated in Figure 8.

Mirroring the findings from the main experiment, B-PMBR consistently and significantly outperforms the PMBR baseline across all evaluated utility functions. The PMBR configuration exhibits a severe dependency on large reference sets, beginning with poor generation quality at small sizes (e.g., 4 or 8 references) and only gradually climbing as the reference size scales. In stark contrast, the proxy-guided subsampling of B-PMBR demonstrates remarkable stability. By deterministically targeting the most informative subsets, B-PMBR achieves near-peak performance even at extremely low reference budgets and maintains this high-quality trajectory across the entire scaling axis.

Furthermore, B-PMBR proves highly adaptable across both lexical (BLEU, chrF) and neural (COMET, BLEURT, BERTScore) metrics. In this specific dataset, B-PMBR variants mostly do not exceed the generation quality of the MBR across the generalized $\bar{Z}_{other}$ metric, albeit still competitive in most cases. This confirms that the benefits of proxy-guided sampling, as well as its natural mitigation of metric overfitting, generalize across different language directions.

G Full Performance Result

This section presents the comprehensive performance evaluation tables for both the WMT-22 En→De and De→En datasets. We use the results from §4 and Appendix F. The detailed tables report the resulting downstream generation quality across all evaluation metrics and the generalized quality metric ($\bar{Z}_{other}$).

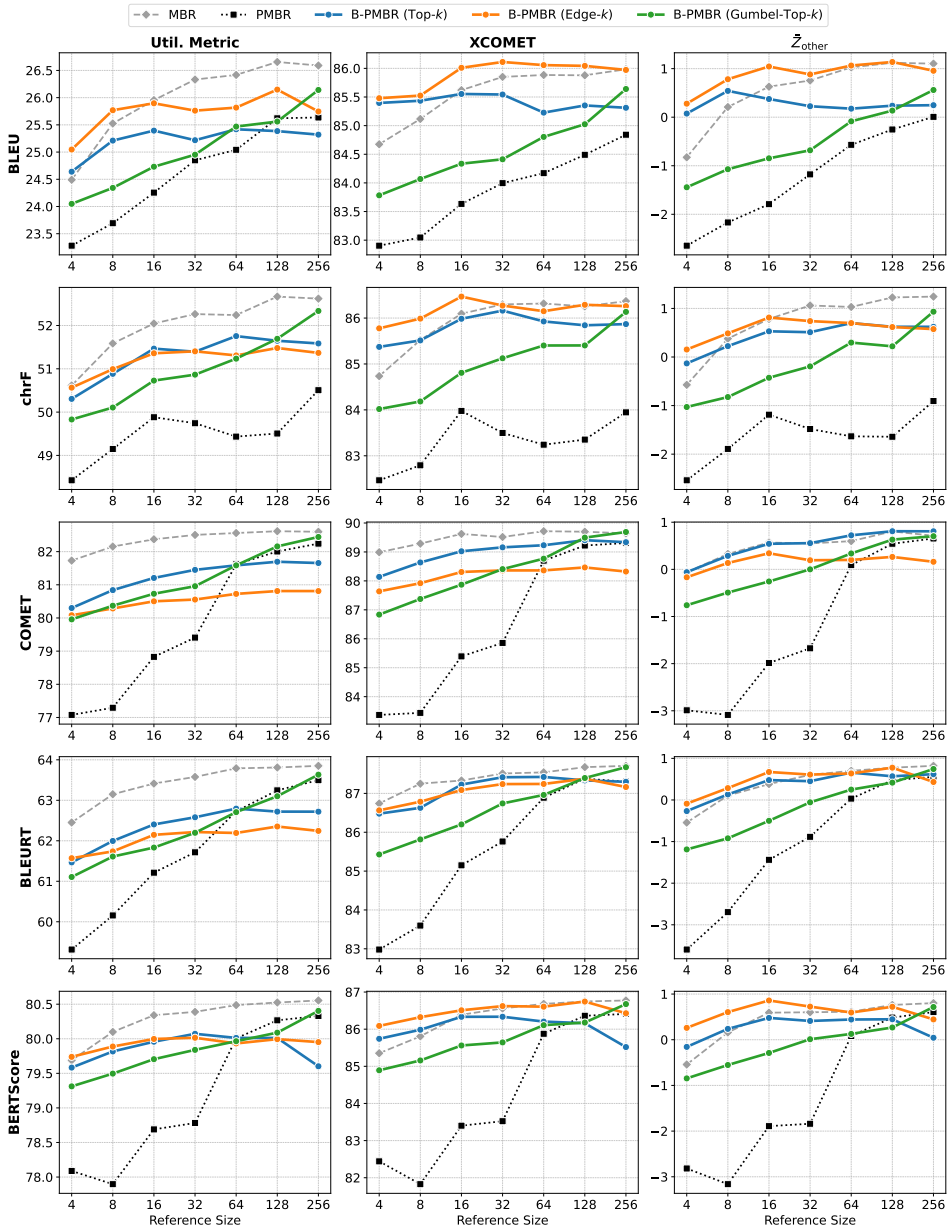

Figure 8: Performance of various B-PMBR against MBR, PMBR, and Normalized PMBR on the WMT-22 De→En dataset. Each row in the grid represents a different target utility function, while each column corresponds to the evaluated metrics. Specifically, **Util. Metric** denotes the evaluation result measured using the exact same metric as the utility function for that respective row.

Table 4: Performance evaluation across metrics ($N = 256$, $M = 4$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	21.47	48.99	72.74	52.10	78.25	83.60	71.69	+0.15
	PMBR	20.12	47.73	70.59	51.03	77.69	81.05	69.42	−1.47
	B-PMBR (Top-k)	21.60	49.25	73.62	52.76	78.36	84.69	72.63	+0.75
	B-PMBR (Edge-k)	**21.96**	**49.54**	**73.91**	**52.98**	**78.47**	**84.79**	**72.85**	**+1.00**
	B-PMBR (Gumbel-Top-k)	21.02	48.77	71.71	51.92	78.10	82.10	70.62	−0.42
chrF	MBR	21.21	**50.36**	73.63	53.33	78.54	84.42	72.63	+0.35
	PMBR	19.42	48.14	70.40	51.05	77.75	80.82	69.25	−1.58
	B-PMBR (Top-k)	21.52	50.01	74.17	53.53	78.54	85.39	73.38	+0.63
	B-PMBR (Edge-k)	**21.89**	50.33	**74.69**	**53.81**	**78.66**	**85.77**	**73.66**	**+0.90**
	B-PMBR (Gumbel-Top-k)	20.90	49.60	72.24	52.56	78.34	82.55	71.38	−0.31
COMET	MBR	21.88	50.22	**82.34**	**56.46**	78.80	**90.18**	**78.41**	**+0.68**
	PMBR	18.42	46.78	73.14	51.58	77.35	82.84	71.10	−1.68
	B-PMBR (Top-k)	**22.31**	**50.60**	80.36	55.72	**78.95**	89.53	77.64	+0.66
	B-PMBR (Edge-k)	22.21	50.25	79.58	55.10	78.77	88.95	77.06	+0.44
	B-PMBR (Gumbel-Top-k)	20.67	49.17	79.02	54.63	78.38	87.66	76.15	−0.10
BLEURT	MBR	20.52	49.07	**76.34**	**59.19**	78.51	86.82	74.38	+0.29
	PMBR	18.61	47.14	71.77	53.82	77.54	82.96	70.69	−1.63
	B-PMBR (Top-k)	21.72	50.03	76.33	56.77	78.75	**87.49**	**74.98**	**+0.76**
	B-PMBR (Edge-k)	**21.94**	**50.19**	75.92	56.25	**78.76**	87.12	74.79	+0.73
	B-PMBR (Gumbel-Top-k)	20.73	49.00	74.39	56.26	78.38	85.37	73.15	−0.15
BERTScore	MBR	21.16	49.65	73.23	53.02	**79.07**	84.26	72.23	+0.22
	PMBR	18.55	46.75	69.42	50.46	77.58	80.49	68.54	−1.64
	B-PMBR (Top-k)	21.49	49.70	74.21	53.53	78.83	85.35	73.16	+0.57
	B-PMBR (Edge-k)	**22.20**	**50.27**	**74.97**	**53.95**	78.98	**85.83**	**73.95**	**+0.93**
	B-PMBR (Gumbel-Top-k)	21.04	49.20	72.54	52.77	78.62	82.94	71.53	−0.09

Table 5: Performance evaluation across metrics ($N = 256$, $M = 8$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**22.84**	49.81	73.19	52.74	78.44	84.04	72.14	+0.33
	PMBR	20.62	48.11	70.87	51.08	77.82	81.14	69.60	−1.41
	B-PMBR (Top-k)	22.19	49.67	73.83	52.94	78.47	84.61	72.72	+0.54
	B-PMBR (Edge-k)	22.57	**50.23**	**74.57**	**53.56**	**78.64**	**85.69**	**73.55**	**+1.12**
	B-PMBR (Gumbel-Top-k)	21.58	48.90	71.81	51.93	78.15	82.40	70.77	−0.58
chrF	MBR	22.11	**51.28**	74.39	53.84	78.78	84.85	73.47	+0.55
	PMBR	19.64	48.48	70.72	51.14	77.88	80.71	69.54	−1.57
	B-PMBR (Top-k)	22.13	50.73	74.48	53.71	78.66	85.30	73.59	+0.54
	B-PMBR (Edge-k)	**22.54**	50.86	**75.07**	**54.05**	**78.82**	**86.05**	**74.23**	**+0.88**
	B-PMBR (Gumbel-Top-k)	21.16	49.61	72.75	52.64	78.33	83.09	71.74	−0.39
COMET	MBR	22.17	50.68	**82.77**	**56.83**	78.87	**90.28**	**78.65**	**+0.72**
	PMBR	18.39	46.89	74.89	52.04	77.46	84.22	72.49	−1.63
	B-PMBR (Top-k)	**22.85**	**50.97**	80.98	56.23	**79.00**	89.74	78.09	+0.72
	B-PMBR (Edge-k)	22.59	50.59	79.73	55.42	78.80	89.03	77.30	+0.43
	B-PMBR (Gumbel-Top-k)	20.61	49.20	79.29	54.64	78.40	87.74	76.15	−0.23
BLEURT	MBR	21.26	49.65	76.44	**60.12**	78.62	87.36	74.94	+0.35
	PMBR	19.12	47.64	72.52	54.92	77.78	83.52	71.36	−1.58
	B-PMBR (Top-k)	22.24	50.39	**76.79**	57.65	78.86	**87.64**	**75.35**	+0.76
	B-PMBR (Edge-k)	**22.55**	**50.64**	76.47	56.84	**78.91**	87.49	75.22	**+0.79**
	B-PMBR (Gumbel-Top-k)	20.96	49.06	75.00	56.74	78.38	85.42	73.44	−0.32
BERTScore	MBR	22.31	50.54	74.38	54.12	**79.51**	85.00	73.26	+0.53
	PMBR	17.97	46.07	68.98	49.90	77.26	79.37	68.06	−1.70
	B-PMBR (Top-k)	22.12	50.39	74.45	53.88	79.09	85.33	73.52	+0.52
	B-PMBR (Edge-k)	**22.62**	**50.61**	**75.03**	**54.15**	79.07	**86.04**	**74.14**	**+0.73**
	B-PMBR (Gumbel-Top-k)	21.17	49.48	72.76	52.97	78.81	83.53	71.73	−0.08

Table 6: Performance evaluation across metrics ($N = 256$, $M = 16$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**23.45**	50.62	74.25	53.24	78.70	84.57	72.79	+0.51
	PMBR	21.14	48.38	71.30	51.33	77.90	81.42	70.12	−1.43
	B-PMBR (Top-k)	22.65	50.22	74.33	53.22	78.59	84.93	73.06	+0.47
	B-PMBR (Edge-k)	23.31	**50.90**	**75.11**	**53.83**	**78.85**	**85.77**	**73.84**	**+1.04**
	B-PMBR (Gumbel-Top-k)	21.58	49.14	72.55	52.26	78.24	83.01	71.18	−0.60
chrF	MBR	22.97	**52.08**	75.11	54.33	**79.05**	85.41	74.00	+0.67
	PMBR	20.60	49.62	71.86	52.07	78.26	81.87	70.80	−1.49
	B-PMBR (Top-k)	22.32	51.33	75.06	54.35	78.91	85.89	74.05	+0.54
	B-PMBR (Edge-k)	**22.99**	51.43	**75.64**	**54.41**	78.95	**86.21**	**74.54**	**+0.82**
	B-PMBR (Gumbel-Top-k)	21.49	50.36	73.25	53.17	78.57	83.75	72.21	−0.54
COMET	MBR	22.15	50.69	**83.01**	**56.87**	78.94	**90.56**	**78.83**	+0.67
	PMBR	19.19	47.74	77.42	53.43	77.83	85.89	74.67	−1.62
	B-PMBR (Top-k)	**23.01**	**51.19**	81.48	56.54	**79.14**	89.98	78.34	**+0.74**
	B-PMBR (Edge-k)	22.84	51.00	80.03	55.74	79.03	89.17	77.44	+0.41
	B-PMBR (Gumbel-Top-k)	21.41	49.79	80.08	55.36	78.58	88.42	76.93	−0.20
BLEURT	MBR	21.56	50.10	76.78	**60.65**	78.79	87.51	75.00	+0.24
	PMBR	19.98	48.49	74.25	56.70	78.22	84.94	72.71	−1.54
	B-PMBR (Top-k)	**22.68**	**50.90**	**77.32**	58.52	**79.09**	**87.99**	**75.76**	**+0.94**
	B-PMBR (Edge-k)	22.60	50.86	76.74	57.26	79.00	87.61	75.49	+0.70
	B-PMBR (Gumbel-Top-k)	21.44	49.71	75.74	57.47	78.66	86.24	74.02	−0.34
BERTScore	MBR	22.85	**51.11**	74.64	54.35	**79.73**	85.60	73.55	+0.59
	PMBR	19.03	47.52	70.13	51.02	77.97	80.48	69.08	−1.66
	B-PMBR (Top-k)	22.41	50.70	74.88	54.17	79.28	85.62	73.80	+0.52
	B-PMBR (Edge-k)	**22.90**	51.04	**75.34**	**54.39**	79.21	**86.24**	**74.28**	**+0.75**
	B-PMBR (Gumbel-Top-k)	21.44	49.70	73.21	53.03	78.99	84.04	72.32	−0.19

Table 7: Performance evaluation across metrics ($N = 256$, $M = 32$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**23.46**	**50.59**	74.60	53.20	78.67	85.03	73.20	+0.74
	PMBR	21.61	48.79	71.69	51.63	78.06	82.18	70.69	−1.38
	B-PMBR (Top-k)	22.56	49.97	74.06	52.88	78.48	84.77	72.82	+0.27
	B-PMBR (Edge-k)	22.88	50.49	**74.80**	**53.60**	**78.69**	**85.76**	**73.90**	**+1.01**
	B-PMBR (Gumbel-Top-k)	22.17	49.48	72.75	52.21	78.23	83.18	71.54	−0.64
chrF	MBR	**22.96**	**52.23**	75.42	**54.48**	**79.10**	85.85	74.43	**+0.77**
	PMBR	20.42	49.42	71.64	51.88	78.14	81.83	70.65	−1.58
	B-PMBR (Top-k)	22.63	51.53	74.97	54.34	78.86	85.60	74.07	+0.49
	B-PMBR (Edge-k)	22.80	51.33	**75.59**	54.28	78.88	**86.20**	**74.72**	+0.70
	B-PMBR (Gumbel-Top-k)	21.63	50.63	73.55	53.26	78.63	83.97	72.52	−0.38
COMET	MBR	22.44	50.94	**83.07**	**57.11**	78.96	**90.52**	**78.84**	+0.68
	PMBR	19.77	48.44	78.95	54.32	78.09	87.28	75.91	−1.59
	B-PMBR (Top-k)	**23.27**	**51.45**	81.91	56.68	**79.15**	90.30	78.68	**+0.81**
	B-PMBR (Edge-k)	22.93	51.10	80.29	55.87	79.04	89.48	77.77	+0.31
	B-PMBR (Gumbel-Top-k)	21.67	49.99	80.67	55.92	78.67	89.15	77.44	−0.22
BLEURT	MBR	21.64	50.19	77.06	**60.92**	78.87	87.70	75.28	+0.23
	PMBR	20.24	48.76	75.06	57.55	78.27	86.07	73.81	−1.44
	B-PMBR (Top-k)	22.55	50.76	**77.32**	59.04	79.02	**88.22**	**75.92**	+0.79
	B-PMBR (Edge-k)	**23.11**	**51.20**	77.23	57.70	**79.13**	88.06	75.74	**+0.93**
	B-PMBR (Gumbel-Top-k)	21.29	49.79	76.06	58.10	78.67	86.71	74.42	−0.51
BERTScore	MBR	**22.95**	**51.28**	74.74	**54.46**	**79.84**	85.66	73.70	+0.61
	PMBR	19.59	48.03	70.84	51.58	78.21	81.37	69.76	−1.67
	B-PMBR (Top-k)	22.67	50.92	74.68	54.05	79.36	85.55	73.73	+0.46
	B-PMBR (Edge-k)	22.92	51.07	**75.44**	54.34	79.31	**86.34**	**74.43**	**+0.75**
	B-PMBR (Gumbel-Top-k)	21.68	49.95	73.75	53.44	79.04	84.27	72.64	−0.15

Table 8: Performance evaluation across metrics ($N = 256$, $M = 64$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**23.71**	**50.66**	74.79	53.37	**78.77**	84.96	73.33	+0.93
	PMBR	22.61	49.55	72.51	52.15	78.34	82.84	71.17	−1.29
	B-PMBR (Top-k)	22.71	50.01	73.74	52.93	78.53	84.47	72.73	+0.01
	B-PMBR (Edge-k)	22.92	50.35	**74.87**	**53.54**	78.70	**85.51**	**73.75**	**+0.97**
	B-PMBR (Gumbel-Top-k)	22.43	49.67	73.16	52.66	78.39	83.79	72.15	−0.62
chrF	MBR	**23.28**	**52.42**	75.60	**54.68**	**79.16**	86.07	74.55	**+0.81**
	PMBR	20.61	49.49	71.77	51.91	78.25	81.33	70.49	−1.67
	B-PMBR (Top-k)	22.55	51.50	75.10	54.05	78.88	85.58	74.21	+0.33
	B-PMBR (Edge-k)	22.80	51.36	**75.62**	54.28	78.91	**86.39**	**74.87**	+0.61
	B-PMBR (Gumbel-Top-k)	22.24	51.07	74.16	53.77	78.73	84.75	73.23	−0.09
COMET	MBR	22.68	50.93	**83.08**	**56.99**	78.93	**90.51**	**78.93**	+0.66
	PMBR	21.89	50.26	81.87	56.30	78.76	89.78	78.11	−0.98
	B-PMBR (Top-k)	**23.17**	**51.36**	82.01	56.76	**79.16**	90.40	78.72	**+1.06**
	B-PMBR (Edge-k)	22.40	50.83	80.49	55.97	78.99	89.58	77.73	−0.64
	B-PMBR (Gumbel-Top-k)	22.64	50.69	81.54	56.50	78.99	89.80	78.29	−0.11
BLEURT	MBR	21.75	50.41	77.24	**61.05**	78.90	87.79	75.40	−0.09
	PMBR	21.40	49.83	76.45	59.39	78.70	87.22	74.89	−1.18
	B-PMBR (Top-k)	22.75	51.05	**77.69**	59.44	79.13	**88.31**	**76.15**	**+1.17**
	B-PMBR (Edge-k)	**22.85**	**51.06**	77.05	57.96	**79.14**	87.79	75.63	+0.65
	B-PMBR (Gumbel-Top-k)	22.22	50.33	76.67	58.88	78.88	87.11	75.05	−0.55
BERTScore	MBR	**23.24**	**51.46**	74.94	**54.79**	**79.93**	85.80	74.04	**+0.92**
	PMBR	22.06	50.28	73.80	53.43	79.28	84.24	72.59	−1.25
	B-PMBR (Top-k)	22.61	50.88	74.87	54.31	79.50	85.69	73.90	+0.30
	B-PMBR (Edge-k)	22.87	51.02	**75.39**	54.38	79.33	**86.19**	**74.32**	+0.79
	B-PMBR (Gumbel-Top-k)	22.18	50.30	74.18	53.66	79.22	84.93	73.09	−0.77

Table 9: Performance evaluation across metrics ($N = 256$, $M = 128$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**24.26**	**51.23**	**74.93**	**53.56**	**78.86**	85.14	73.43	**+1.01**
	PMBR	22.97	50.03	72.83	52.34	78.41	83.06	71.30	−1.28
	B-PMBR (Top-k)	23.04	50.33	74.36	53.15	78.58	84.72	72.91	+0.07
	B-PMBR (Edge-k)	23.35	50.84	74.74	53.43	78.80	**85.65**	**73.60**	+0.85
	B-PMBR (Gumbel-Top-k)	23.08	50.16	73.34	52.80	78.52	83.75	72.05	−0.65
chrF	MBR	**23.33**	**52.54**	**75.78**	**54.70**	**79.22**	86.22	**74.72**	**+0.72**
	PMBR	20.40	49.31	71.36	51.55	78.07	81.48	70.25	−1.74
	B-PMBR (Top-k)	22.62	51.78	75.43	54.35	78.97	85.73	74.35	+0.37
	B-PMBR (Edge-k)	22.73	51.49	75.65	54.43	79.06	**86.31**	74.64	+0.53
	B-PMBR (Gumbel-Top-k)	22.26	51.58	74.79	54.14	78.90	85.35	73.64	+0.11
COMET	MBR	22.71	51.08	**83.13**	**57.11**	79.05	**90.56**	**78.97**	+0.29
	PMBR	22.34	50.86	82.40	56.69	79.01	90.06	78.57	−0.43
	B-PMBR (Top-k)	**23.46**	**51.69**	82.19	56.99	**79.24**	90.47	78.83	**+0.96**
	B-PMBR (Edge-k)	22.54	50.60	80.55	55.80	78.94	89.43	77.68	−1.34
	B-PMBR (Gumbel-Top-k)	23.07	51.30	82.62	57.04	79.19	90.28	78.77	+0.52
BLEURT	MBR	21.90	50.47	77.20	**61.13**	78.94	87.67	75.35	−0.36
	PMBR	21.87	50.31	76.95	60.20	78.89	87.54	75.14	−0.83
	B-PMBR (Top-k)	**23.06**	**51.27**	**77.57**	59.80	**79.26**	**88.20**	**76.00**	**+1.50**
	B-PMBR (Edge-k)	22.74	50.98	77.01	57.99	79.05	87.90	75.68	+0.37
	B-PMBR (Gumbel-Top-k)	21.89	50.27	76.99	60.00	78.86	87.73	75.20	−0.68
BERTScore	MBR	**23.43**	**51.55**	75.08	**54.80**	**79.97**	85.77	73.98	+0.85
	PMBR	23.05	51.17	74.24	54.16	79.63	84.93	73.15	−0.72
	B-PMBR (Top-k)	22.81	50.96	74.95	54.29	79.48	85.68	73.97	−0.13
	B-PMBR (Edge-k)	23.16	51.32	**75.53**	54.47	79.41	**86.41**	**74.47**	**+0.90**
	B-PMBR (Gumbel-Top-k)	22.40	50.78	74.57	54.16	79.58	85.28	73.46	−0.90

Table 10: Performance evaluation across metrics ($N = 256$, $M = 256$) on WMT22 En→De. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**24.07**	**51.02**	74.93	**53.64**	**78.82**	85.45	73.58	**+0.91**
	PMBR	23.18	50.19	72.92	52.54	78.56	83.34	71.63	−1.48
	B-PMBR (Top-k)	22.90	50.36	74.50	53.39	78.54	85.23	73.24	−0.14
	B-PMBR (Edge-k)	23.36	50.62	**75.06**	53.60	78.67	**85.63**	**73.86**	+0.57
	B-PMBR (Gumbel-Top-k)	23.48	50.59	74.36	53.47	78.68	85.04	73.19	+0.14
chrF	MBR	**23.25**	**52.49**	**75.71**	**54.73**	**79.17**	86.09	**74.64**	**+0.69**
	PMBR	21.23	50.28	72.53	52.64	78.40	82.57	71.51	−1.73
	B-PMBR (Top-k)	22.67	51.58	75.41	54.30	78.89	85.88	74.49	+0.25
	B-PMBR (Edge-k)	22.71	51.39	75.33	54.14	78.89	86.10	74.54	+0.25
	B-PMBR (Gumbel-Top-k)	22.95	52.23	75.61	54.69	79.08	**86.12**	74.49	+0.54
COMET	MBR	22.67	50.99	**83.12**	**57.10**	78.98	90.52	78.94	+0.28
	PMBR	22.29	50.76	82.69	56.97	78.99	90.26	78.69	−0.06
	B-PMBR (Top-k)	**23.51**	**51.68**	82.26	56.97	**79.25**	90.48	78.91	**+0.89**
	B-PMBR (Edge-k)	21.83	50.14	80.46	55.45	78.73	89.23	77.61	−1.57
	B-PMBR (Gumbel-Top-k)	22.64	51.15	82.96	57.08	79.08	**90.56**	**79.10**	+0.45
BLEURT	MBR	21.79	50.36	77.15	**61.12**	78.91	87.76	75.31	−0.53
	PMBR	21.76	50.38	77.11	60.62	78.95	87.59	75.24	−0.62
	B-PMBR (Top-k)	**22.94**	**51.24**	**77.74**	59.78	**79.18**	**88.28**	**76.18**	**+1.64**
	B-PMBR (Edge-k)	22.20	50.62	76.87	57.90	78.86	87.55	75.51	−0.50
	B-PMBR (Gumbel-Top-k)	22.01	50.57	77.34	60.81	79.00	87.84	75.50	+0.01
BERTScore	MBR	**23.20**	**51.48**	75.05	54.69	**79.96**	85.87	73.96	+0.64
	PMBR	22.61	50.86	74.41	54.12	79.69	85.20	73.28	−1.09
	B-PMBR (Top-k)	22.62	50.73	74.81	54.03	79.32	85.50	73.79	−0.72
	B-PMBR (Edge-k)	22.81	50.89	**75.53**	54.28	79.28	**86.35**	**74.42**	+0.46
	B-PMBR (Gumbel-Top-k)	23.13	51.35	75.17	**54.77**	79.86	85.98	74.08	**+0.71**

Table 11: Performance evaluation across metrics ($N = 256$, $M = 4$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	24.49	49.24	77.43	59.94	79.06	84.67	74.19	+0.07
	PMBR	23.28	48.25	76.33	59.12	78.29	82.90	73.12	−1.45
	B-PMBR (Top-k)	24.64	49.76	77.91	60.44	79.32	85.40	74.83	+0.83
	B-PMBR (Edge-k)	**25.05**	**49.90**	**78.08**	**60.53**	**79.37**	**85.48**	**74.92**	**+1.00**
	B-PMBR (Gumbel-Top-k)	24.05	48.82	77.01	59.74	78.80	83.78	73.85	−0.45
chrF	MBR	24.59	**50.61**	77.57	60.51	79.17	84.73	74.58	+0.24
	PMBR	22.70	48.43	76.19	59.09	78.14	82.47	73.02	−1.61
	B-PMBR (Top-k)	24.68	50.31	78.11	60.70	79.37	85.37	75.12	+0.64
	B-PMBR (Edge-k)	**25.26**	50.56	**78.29**	**60.81**	**79.53**	**85.78**	**75.25**	**+0.92**
	B-PMBR (Gumbel-Top-k)	24.15	49.83	77.29	60.14	78.96	84.02	74.16	−0.19
COMET	MBR	24.64	49.91	**81.73**	**61.73**	**80.27**	**88.99**	**77.47**	+0.57
	PMBR	21.60	47.10	77.08	59.00	78.11	83.37	73.52	−1.72
	B-PMBR (Top-k)	25.26	50.42	80.30	61.64	80.14	88.14	76.81	**+0.59**
	B-PMBR (Edge-k)	**25.31**	**50.46**	80.08	61.42	80.04	87.64	76.55	+0.51
	B-PMBR (Gumbel-Top-k)	24.22	49.64	79.96	60.99	79.62	86.84	76.22	+0.04
BLEURT	MBR	23.84	49.32	**78.60**	**62.45**	79.35	**86.74**	**75.60**	+0.41
	PMBR	21.26	46.92	76.25	59.31	77.91	82.98	73.12	−1.70
	B-PMBR (Top-k)	24.28	49.89	78.56	61.47	79.54	86.48	75.55	+0.60
	B-PMBR (Edge-k)	**24.87**	**50.11**	78.57	61.57	**79.57**	86.56	75.55	**+0.73**
	B-PMBR (Gumbel-Top-k)	23.54	49.08	77.85	61.11	79.10	85.43	74.85	−0.04
BERTScore	MBR	23.91	49.15	77.93	60.15	79.69	85.35	74.88	+0.23
	PMBR	21.48	46.84	75.97	58.54	78.09	82.44	72.76	−1.67
	B-PMBR (Top-k)	24.39	49.57	78.14	60.59	79.58	85.74	75.11	+0.55
	B-PMBR (Edge-k)	**25.14**	**50.14**	**78.35**	**60.80**	**79.74**	**86.09**	**75.43**	**+0.91**
	B-PMBR (Gumbel-Top-k)	23.48	48.70	77.59	60.13	79.31	84.89	74.63	−0.03

Table 12: Performance evaluation across metrics ($N = 256$, $M = 8$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	25.52	50.07	77.98	60.51	79.37	85.12	74.70	+0.44
	PMBR	23.69	48.49	76.65	59.43	78.47	83.05	73.30	−1.46
	B-PMBR (Top-k)	25.21	50.07	78.15	60.68	79.49	85.43	75.11	+0.70
	B-PMBR (Edge-k)	**25.77**	**50.51**	**78.20**	**60.84**	**79.52**	**85.53**	**75.14**	**+0.90**
	B-PMBR (Gumbel-Top-k)	24.34	49.15	77.25	60.01	78.74	84.07	74.15	−0.58
chrF	MBR	25.71	**51.59**	78.31	**61.18**	79.57	85.52	75.28	+0.69
	PMBR	23.35	49.15	76.63	59.62	78.51	82.79	73.40	−1.53
	B-PMBR (Top-k)	25.45	50.88	78.23	60.95	79.53	85.51	75.29	+0.53
	B-PMBR (Edge-k)	**25.78**	50.99	**78.49**	61.06	**79.65**	**85.99**	**75.47**	**+0.79**
	B-PMBR (Gumbel-Top-k)	24.42	50.11	77.43	60.40	78.90	84.18	74.42	−0.49
COMET	MBR	25.26	50.48	**82.15**	**62.13**	**80.43**	**89.29**	**77.81**	**+0.62**
	PMBR	21.06	46.80	77.29	58.87	78.18	83.44	73.77	−1.72
	B-PMBR (Top-k)	25.73	**50.91**	80.84	61.89	80.33	88.64	77.22	+0.58
	B-PMBR (Edge-k)	**26.02**	50.90	80.29	61.63	80.12	87.92	76.82	+0.47
	B-PMBR (Gumbel-Top-k)	24.52	49.73	80.37	61.30	79.89	87.37	76.52	+0.05
BLEURT	MBR	24.60	49.99	**78.94**	**63.15**	79.65	**87.25**	**76.02**	+0.57
	PMBR	22.03	47.87	76.82	60.16	78.36	83.60	73.73	−1.64
	B-PMBR (Top-k)	25.12	50.19	78.68	62.00	79.62	86.63	75.92	+0.59
	B-PMBR (Edge-k)	**25.48**	**50.42**	78.66	61.74	**79.70**	86.79	75.92	**+0.71**
	B-PMBR (Gumbel-Top-k)	23.88	49.21	78.05	61.61	79.15	85.82	75.18	−0.24
BERTScore	MBR	24.84	49.96	78.39	60.71	**80.10**	85.81	75.42	+0.46
	PMBR	20.83	46.61	75.64	58.36	77.90	81.83	72.50	−1.71
	B-PMBR (Top-k)	24.93	49.96	78.39	60.95	79.82	85.98	75.41	+0.51
	B-PMBR (Edge-k)	**25.65**	**50.51**	**78.61**	**61.09**	79.89	**86.32**	**75.64**	**+0.75**
	B-PMBR (Gumbel-Top-k)	23.68	48.93	77.95	60.28	79.50	85.16	74.99	−0.01

Table 13: Performance evaluation across metrics ($N = 256$, $M = 16$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**25.95**	50.50	78.16	60.65	79.47	85.62	75.06	+0.64
	PMBR	24.26	48.93	76.77	59.63	78.50	83.63	73.60	−1.42
	B-PMBR (Top-k)	25.39	50.20	77.96	60.59	79.37	85.55	75.02	+0.42
	B-PMBR (Edge-k)	25.90	**50.70**	**78.34**	**60.99**	**79.57**	**86.01**	**75.33**	**+0.99**
	B-PMBR (Gumbel-Top-k)	24.73	49.45	77.28	60.09	78.90	84.34	74.17	−0.62
chrF	MBR	**26.11**	**52.05**	78.60	**61.52**	79.72	86.09	75.65	+0.78
	PMBR	23.95	49.88	77.25	60.08	78.80	83.98	74.06	−1.46
	B-PMBR (Top-k)	25.67	51.47	78.47	61.18	79.65	85.98	75.61	+0.48
	B-PMBR (Edge-k)	25.89	51.36	**78.75**	61.36	**79.78**	**86.47**	**75.85**	**+0.80**
	B-PMBR (Gumbel-Top-k)	24.77	50.73	77.77	60.68	79.19	84.81	74.57	−0.60
COMET	MBR	25.50	50.94	**82.37**	**62.33**	**80.56**	**89.62**	**78.02**	**+0.68**
	PMBR	22.44	48.03	78.83	59.93	78.85	85.39	75.09	−1.67
	B-PMBR (Top-k)	26.16	**51.26**	81.21	62.06	80.47	89.02	77.50	+0.64
	B-PMBR (Edge-k)	**26.19**	51.16	80.50	61.85	80.27	88.31	77.04	+0.44
	B-PMBR (Gumbel-Top-k)	24.67	50.13	80.73	61.57	79.97	87.87	76.86	−0.09
BLEURT	MBR	25.04	50.40	**79.14**	**63.41**	79.64	**87.33**	**76.17**	+0.53
	PMBR	23.31	48.83	77.76	61.21	78.88	85.15	74.76	−1.51
	B-PMBR (Top-k)	25.36	50.56	79.02	62.40	79.76	87.23	76.10	+0.63
	B-PMBR (Edge-k)	**25.93**	**51.09**	78.99	62.15	**79.80**	87.09	75.96	**+0.82**
	B-PMBR (Gumbel-Top-k)	24.18	49.65	78.44	61.83	79.32	86.20	75.45	−0.46
BERTScore	MBR	25.22	50.54	**78.78**	61.07	**80.34**	86.39	75.69	+0.58
	PMBR	22.26	47.68	76.86	59.26	78.69	83.40	73.67	−1.64
	B-PMBR (Top-k)	25.15	50.35	78.54	61.14	79.96	86.33	75.58	+0.47
	B-PMBR (Edge-k)	**26.04**	**51.05**	78.71	**61.25**	80.00	**86.51**	**75.74**	**+0.80**
	B-PMBR (Gumbel-Top-k)	23.92	49.41	78.17	60.50	79.70	85.56	75.07	−0.21

Table 14: Performance evaluation across metrics ($N = 256$, $M = 32$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**26.33**	**50.65**	78.23	60.72	79.49	85.85	75.09	+0.84
	PMBR	24.85	49.43	77.07	59.83	78.82	84.00	73.89	−1.29
	B-PMBR (Top-k)	25.22	50.02	77.99	60.44	79.31	85.54	74.99	+0.23
	B-PMBR (Edge-k)	25.76	50.46	**78.24**	**60.87**	**79.58**	**86.11**	**75.29**	**+0.96**
	B-PMBR (Gumbel-Top-k)	24.95	49.64	77.48	60.04	78.94	84.41	74.28	−0.75
chrF	MBR	**26.36**	**52.26**	**78.86**	**61.62**	**79.88**	**86.30**	**75.87**	**+0.92**
	PMBR	23.66	49.74	77.03	59.77	78.69	83.50	73.90	−1.59
	B-PMBR (Top-k)	25.67	51.39	78.47	61.16	79.63	86.16	75.58	+0.38
	B-PMBR (Edge-k)	26.07	51.40	78.57	61.30	79.75	86.27	75.74	+0.60
	B-PMBR (Gumbel-Top-k)	25.21	50.86	77.87	60.72	79.25	85.12	74.90	−0.32
COMET	MBR	25.40	50.74	**82.50**	**62.32**	**80.63**	**89.52**	**78.10**	**+0.68**
	PMBR	22.65	48.26	79.41	60.23	79.15	85.86	75.55	−1.69
	B-PMBR (Top-k)	**25.92**	**51.13**	81.45	62.13	80.53	89.16	77.71	+0.67
	B-PMBR (Edge-k)	25.75	50.95	80.56	61.74	80.21	88.36	77.04	+0.27
	B-PMBR (Gumbel-Top-k)	25.15	50.39	80.96	61.70	80.17	88.41	77.13	+0.08
BLEURT	MBR	25.34	50.68	**79.29**	**63.58**	79.73	**87.51**	**76.25**	**+0.70**
	PMBR	23.78	49.26	78.20	61.72	79.21	85.76	75.08	−1.58
	B-PMBR (Top-k)	25.02	50.48	79.07	62.58	79.78	87.42	76.23	+0.48
	B-PMBR (Edge-k)	**25.74**	**50.86**	79.01	62.22	**79.79**	87.24	76.05	+0.70
	B-PMBR (Gumbel-Top-k)	24.64	50.05	78.69	62.20	79.51	86.74	75.85	−0.31
BERTScore	MBR	25.41	50.48	**78.74**	61.02	**80.39**	86.56	75.75	+0.59
	PMBR	22.24	47.72	76.82	59.30	78.78	83.52	73.85	−1.73
	B-PMBR (Top-k)	25.09	50.16	78.66	60.84	80.07	86.33	75.71	+0.41
	B-PMBR (Edge-k)	**25.69**	**50.61**	78.70	**61.17**	80.01	**86.62**	**75.85**	**+0.70**
	B-PMBR (Gumbel-Top-k)	24.54	49.69	78.35	60.61	79.84	85.64	75.37	+0.03

Table 15: Performance evaluation across metrics ($N = 256$, $M = 64$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**26.42**	**50.81**	78.39	60.84	**79.67**	85.88	75.14	+0.98
	PMBR	25.04	49.62	77.54	60.19	79.05	84.17	74.20	−1.23
	B-PMBR (Top-k)	25.42	50.02	77.89	60.54	79.32	85.23	74.81	−0.21
	B-PMBR (Edge-k)	25.82	50.69	**78.45**	**60.94**	79.57	**86.06**	**75.34**	**+1.02**
	B-PMBR (Gumbel-Top-k)	25.47	50.14	77.77	60.30	79.22	84.80	74.51	−0.56
chrF	MBR	**26.27**	**52.24**	**78.85**	**61.59**	**79.93**	**86.32**	**75.79**	**+0.76**
	PMBR	23.43	49.43	76.93	59.77	78.60	83.24	73.72	−1.73
	B-PMBR (Top-k)	26.08	51.75	78.50	61.35	79.74	85.93	75.59	+0.45
	B-PMBR (Edge-k)	25.79	51.31	78.59	61.27	79.79	86.15	75.72	+0.45
	B-PMBR (Gumbel-Top-k)	25.42	51.23	78.32	61.15	79.57	85.40	75.19	+0.07
COMET	MBR	25.54	50.84	**82.56**	**62.36**	**80.61**	**89.72**	**78.13**	+0.69
	PMBR	24.95	50.37	81.62	61.84	80.26	88.70	77.51	−0.95
	B-PMBR (Top-k)	**26.28**	**51.35**	81.59	62.34	80.60	89.23	77.74	**+1.05**
	B-PMBR (Edge-k)	25.75	50.76	80.73	61.73	80.30	88.36	77.11	−0.63
	B-PMBR (Gumbel-Top-k)	25.47	50.76	81.59	62.02	80.44	88.77	77.52	−0.16
BLEURT	MBR	25.43	50.67	**79.34**	**63.79**	**79.84**	**87.54**	**76.30**	**+0.84**
	PMBR	24.61	49.94	78.86	62.73	79.58	86.89	75.98	−1.27
	B-PMBR (Top-k)	25.52	50.76	79.15	62.79	79.83	87.42	76.24	+0.63
	B-PMBR (Edge-k)	**25.65**	**50.78**	79.01	62.19	79.83	87.24	76.21	+0.52
	B-PMBR (Gumbel-Top-k)	25.06	50.40	78.92	62.71	79.64	86.96	75.92	−0.71
BERTScore	MBR	25.24	50.39	**78.84**	**61.09**	**80.49**	**86.68**	**75.80**	**+0.98**
	PMBR	24.69	49.64	78.40	60.75	80.01	85.88	75.38	−1.08
	B-PMBR (Top-k)	25.33	50.38	78.57	60.91	80.01	86.20	75.52	+0.18
	B-PMBR (Edge-k)	**25.41**	**50.56**	78.64	61.05	79.93	86.61	75.73	+0.79
	B-PMBR (Gumbel-Top-k)	24.57	49.77	78.52	60.80	79.96	86.11	75.36	−0.88

Table 16: Performance evaluation across metrics ($N = 256$, $M = 128$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**26.66**	**51.04**	78.37	60.93	79.65	85.88	75.17	+1.01
	PMBR	25.62	50.14	77.62	60.32	79.16	84.49	74.25	−1.13
	B-PMBR (Top-k)	25.39	50.13	77.90	60.55	79.32	85.35	74.86	−0.38
	B-PMBR (Edge-k)	26.15	50.81	**78.41**	**60.94**	**79.66**	**86.04**	**75.35**	**+1.03**
	B-PMBR (Gumbel-Top-k)	25.56	50.10	77.92	60.47	79.24	85.02	74.79	−0.53
chrF	MBR	**26.91**	**52.67**	**78.87**	**61.75**	**79.98**	86.25	**75.80**	**+0.92**
	PMBR	23.66	49.51	76.89	59.73	78.61	83.35	73.58	−1.68
	B-PMBR (Top-k)	25.84	51.65	78.48	61.33	79.71	85.84	75.56	+0.38
	B-PMBR (Edge-k)	26.01	51.48	78.47	61.24	79.69	**86.29**	75.57	+0.37
	B-PMBR (Gumbel-Top-k)	25.54	51.69	78.17	61.16	79.47	85.40	75.11	+0.01
COMET	MBR	26.03	51.19	**82.62**	**62.54**	**80.72**	**89.70**	**78.11**	+0.65
	PMBR	25.83	50.95	82.01	62.20	80.54	89.22	77.79	−0.34
	B-PMBR (Top-k)	**26.47**	**51.50**	81.70	62.36	80.66	89.41	77.78	**+0.89**
	B-PMBR (Edge-k)	25.93	50.98	80.81	61.78	80.22	88.46	77.20	−1.21
	B-PMBR (Gumbel-Top-k)	25.90	51.07	82.16	62.25	80.59	89.49	77.95	+0.01
BLEURT	MBR	25.51	50.84	**79.39**	**63.81**	79.84	**87.68**	**76.34**	**+0.98**
	PMBR	25.08	50.54	79.11	63.25	79.81	87.38	76.13	−0.63
	B-PMBR (Top-k)	25.34	50.64	79.09	62.72	79.82	87.33	76.21	−0.17
	B-PMBR (Edge-k)	**25.82**	**50.97**	79.12	62.35	**79.90**	87.38	76.20	+0.70
	B-PMBR (Gumbel-Top-k)	25.02	50.43	79.11	63.10	79.72	87.40	76.19	−0.87
BERTScore	MBR	25.57	50.63	**78.84**	**61.22**	**80.52**	**86.75**	**75.85**	**+1.04**
	PMBR	25.23	50.19	78.66	61.04	80.27	86.36	75.67	−0.19
	B-PMBR (Top-k)	25.38	50.44	78.52	60.97	80.01	86.16	75.45	−0.44
	B-PMBR (Edge-k)	**25.61**	**50.75**	78.65	61.15	79.99	86.74	75.85	+0.79
	B-PMBR (Gumbel-Top-k)	24.91	50.02	78.46	60.86	80.09	86.18	75.53	−1.20

Table 17: Performance evaluation across metrics ($N = 256$, $M = 256$) on WMT22 De→En. **Bold** and Underline indicate the best and second-best performance, respectively. Abbreviations: BL: BLEU, CF: chrF, CM: COMET, BT: BLEURT, BS: BERTScore, XC: XCOMET, CK: COMETKiwi, $\bar{Z}_{other}$: Generalized Quality.

Util	Method	BL	CF	CM	BT	BS	XC	CK	$\bar{Z}_{other}$
BLEU	MBR	**26.59**	**50.98**	**78.40**	60.86	**79.62**	**85.98**	75.27	**+1.11**
	PMBR	25.64	50.25	77.71	60.48	79.27	84.84	74.45	−1.17
	B-PMBR (Top-k)	25.32	50.16	77.91	60.56	79.31	85.31	74.85	−0.71
	B-PMBR (Edge-k)	25.75	50.63	78.28	**60.90**	79.57	85.97	**75.30**	+0.80
	B-PMBR (Gumbel-Top-k)	26.14	50.54	78.06	60.66	79.45	85.64	74.92	−0.03
chrF	MBR	**26.73**	**52.62**	**78.92**	**61.81**	**79.96**	**86.37**	**75.92**	**+0.90**
	PMBR	24.27	50.51	77.44	60.30	78.95	83.95	74.22	−1.69
	B-PMBR (Top-k)	25.94	51.59	78.47	61.32	79.65	85.87	75.61	+0.15
	B-PMBR (Edge-k)	25.69	51.37	78.48	61.28	79.70	86.26	75.57	+0.10
	B-PMBR (Gumbel-Top-k)	26.28	52.34	78.69	61.54	79.85	86.13	75.77	+0.53
COMET	MBR	25.85	51.09	**82.60**	**62.45**	80.64	89.64	**78.12**	+0.29
	PMBR	25.86	51.06	82.24	62.42	80.57	89.31	77.96	+0.11
	B-PMBR (Top-k)	**26.50**	**51.43**	81.66	62.39	80.65	89.34	77.80	**+0.83**
	B-PMBR (Edge-k)	25.61	50.85	80.81	61.75	80.21	88.32	77.05	−1.48
	B-PMBR (Gumbel-Top-k)	25.81	51.08	82.44	62.41	**80.68**	**89.69**	78.08	+0.25
BLEURT	MBR	25.54	**50.80**	**79.42**	**63.85**	**79.87**	**87.71**	**76.42**	**+1.08**
	PMBR	25.32	50.61	79.14	63.50	79.77	87.30	76.22	−0.57
	B-PMBR (Top-k)	**25.65**	50.73	79.06	62.72	79.79	87.30	76.19	+0.10
	B-PMBR (Edge-k)	25.28	50.68	78.96	62.24	79.70	87.16	76.07	−1.07
	B-PMBR (Gumbel-Top-k)	25.51	50.66	79.37	63.63	79.84	87.67	76.39	+0.45
BERTScore	MBR	**25.58**	**50.78**	**78.89**	61.25	**80.55**	**86.77**	**75.81**	**+1.00**
	PMBR	25.36	50.36	78.76	61.14	80.33	86.41	75.72	+0.18
	B-PMBR (Top-k)	25.10	50.16	77.97	60.71	79.60	85.52	75.06	−1.52
	B-PMBR (Edge-k)	25.36	50.35	78.49	60.95	79.95	86.43	75.57	−0.24
	B-PMBR (Gumbel-Top-k)	25.40	50.57	78.84	**61.26**	80.40	86.68	75.79	+0.58