# RoSF-BO: Robust Surrogate Fitting via Low-Noise Data Augmentation for Bayesian Optimization

**Anonymous authors**
Paper under double-blind review

## Abstract

Bayesian optimization (BO) relies on surrogate models fitted to limited observations, making its performance particularly sensitive to observation noise. Existing noise-handling approaches mainly focus on surrogate modeling or acquisition design, while observation-level noise mitigation remains relatively underexplored. To address this issue, we propose RoSF-BO, an observation-level data augmentation framework that improves the training data used for surrogate fitting without additional expensive evaluations or discarding original observations. RoSF-BO exploits local functional relationships among evaluated solutions to identify observations with complementary deviations and generates additional training samples with reduced noise interference through residual-cancellation interpolation. We further provide a theoretical analysis to explain the noise-reduction effect of the generated samples. Extensive experiments on static datasets, noisy single-objective BO, and noisy multi-objective BO demonstrate that RoSF-BO consistently improves data quality and translates these improvements into better optimization performance across different BO methods and observation distributions. Overall, RoSF-BO provides a new observation-level perspective for noise handling in Bayesian optimization and complements existing surrogate- and acquisition-level approaches while preserving the standard BO acquisition and evaluation procedures.

## 1 Introduction

Bayesian optimization (BO) has become a widely used framework for optimizing expensive black-box problems, where each evaluation may involve substantial computational, experimental, or financial costs Balandat et al. (2020); Hvarfner et al. (2024); Salgia et al. (2024). By iteratively constructing a probabilistic model of the objective and selecting candidates through an acquisition function, BO supports efficient search under a limited evaluation budget. In practice, stochastic simulations, measurement errors, and environmental fluctuations often produce noisy observations Daulton et al. (2021); Zhou et al. (2024); Jun & Kim (2024). Such noise can obscure the objective landscape, misdirect candidate selection, and reduce both reliability and sample efficiency.

Recent studies mainly handle noise in Bayesian optimization in two ways. One is to modify the acquisition function so that the uncertainty caused by noisy observations is considered when selecting new points for evaluation Zhou et al. (2024); Jun & Kim (2024); Siska et al. (2026). The other is to improve the surrogate model or its fitting procedure so that it becomes more robust to noisy data Hvarfner et al. (2023); Altamirano et al. (2024). Although effective, these approaches do not directly mitigate noise at the data level before surrogate fitting, and often rely on specific surrogate models, fitting procedures, or assumptions about the underlying noise characteristics Deshpande et al. (2024); Do & Zhang (2025).

This motivates handling noise directly at the observation level before surrogate fitting. Existing approaches at this level mainly rely on repeated evaluations, smoothing, or filtering and trimming Binois et al. (2019); Andrade & Takeda (2023); Reifenstein et al. (2024). However, repeated evaluations consume additional evaluation budgets, smoothing may obscure local landscape structures, and filtering or trimming can discard potentially useful observations. These limitations are particularly



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.

restrictive in expensive black-box optimization, where evaluations are scarce and even noisy observations may contain valuable information Binois et al. (2018a); Altamirano et al. (2024). Therefore, it is desirable to extract more reliable information from existing observations without additional evaluations Lu et al. (2024), discarding original data, or imposing strong assumptions on the noise characteristics Ament et al. (2024).

To this end, we propose RoSF-BO, an observation-level data augmentation framework for robust surrogate fitting in noisy Bayesian optimization. RoSF-BO first partitions the evaluated solutions into multiple subspaces and estimates local functional relationships between neighboring regions. Based on these relationships, it identifies pairs of observations whose local prediction residuals tend to cancel and generates additional low-noise training samples by jointly interpolating their inputs and observed values. The generated samples are combined with the original observations only for surrogate fitting, without requiring additional expensive evaluations or modifying the acquisition and evaluation procedures.

Our contributions are threefold. First, we introduce RoSF-BO, an observation-level data augmentation approach for noisy Bayesian optimization that improves the training data before surrogate fitting without requiring additional expensive evaluations or discarding the original observations. Second, we provide a theoretical analysis of the effectiveness of RoSF-BO, clarifying the mechanism behind low-noise sample generation. Third, we conduct extensive experiments on static data augmentation, single-objective and multi-objective Bayesian optimization, and engineering design problems, demonstrating the effectiveness of RoSF-BO across different noisy optimization settings. Overall, this work provides a new observation-level perspective for noise handling in Bayesian optimization and broadens existing approaches beyond surrogate modeling and acquisition design.

## 2 Related Work

**Bayesian Optimization under Noisy Observations.** Noise affects BO through both posterior estimation and the decision induced by that posterior Huang et al. (2006); Picheny et al. (2013); Desimpel et al. (2026). If the observed incumbent is treated as exact, improvement-based criteria may repeatedly exploit points that appear favorable only because of measurement error. Noise-aware acquisitions instead account for uncertainty in latent objective values Scott et al. (2011). NEI marginalizes improvement over uncertain values at evaluated points Letham et al. (2019); qNEHVI extends this principle to parallel noisy multi-objective optimization Daulton et al. (2021); and corrected EI retains covariance with the noisy incumbent under heteroscedastic observations Zhou et al. (2024). Noise-adaptive confidence sets adapt predictive uncertainty to an unknown noise level Jun & Kim (2024), while calibrated and conformal BO recalibrate predictive uncertainty under adaptive, non-i.i.d. feedback Deshpande et al. (2024); Caldas Ramos et al. (2026). Although effective, these approaches mainly handle noise during candidate selection and do not directly improve the observations used for surrogate fitting.

A complementary line focuses on improving the surrogate model or its fitting procedure. Heteroscedastic GPs jointly estimate the latent objective and input-dependent variance Binois et al. (2018a); Goldberg et al. (1997); Lázaro-Gredilla & Titsias (2011). Robust and conjugate GP regression uses generalized Bayesian updating while retaining closed-form conditioning Altamirano et al. (2024), whereas trimmed marginal likelihood adaptively selects a fitting subset to limit outlier influence Andrade & Takeda (2023). Self-correcting BO adapts GP hyperparameters during optimization Hvarfner et al. (2023), and corruption-tolerant GP bandits modify confidence or elimination mechanisms under adversarial perturbations Bogunovic et al. (2022). Although effective in their target regimes, these approaches typically couple noise handling with a specific surrogate formulation, fitting procedure, or decision mechanism. RoSF-BO instead operates at the observation level by augmenting the data before surrogate fitting, providing a complementary direction to existing surrogate- and acquisition-level approaches.

**Noise Handling Methods.** At the data-processing level, replication and averaging reduce observation noise by repeatedly evaluating the same solution, while sequential designs determine whether to replicate an existing input or explore a new one Binois et al. (2019). Although statistically effective, repeated evaluations directly consume the limited evaluation budget. Smoothing-based methods instead borrow information from neighboring observations, with recent anisotropic schemes adapting the smoothing strength to local landscape characteristics Reifenstein et al. (2024). Their effective-

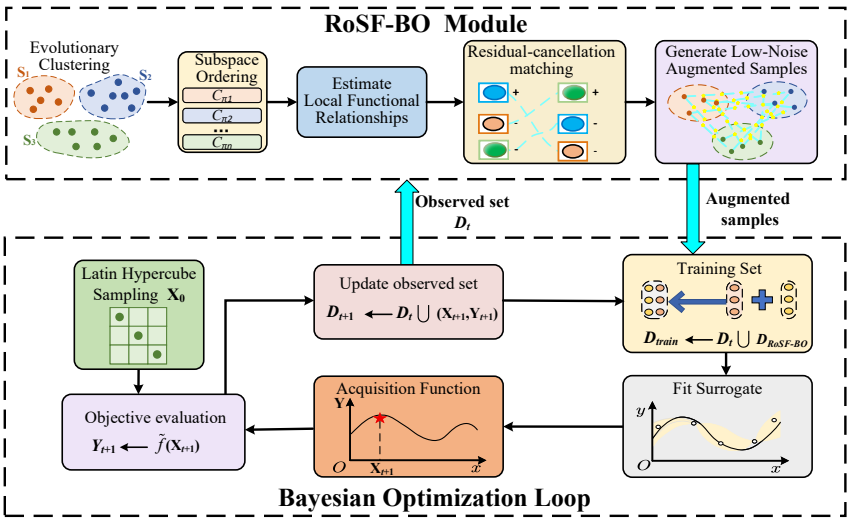


Figure 1: The overall framework of RoSF-BO within the standard Bayesian optimization pipeline.

ness, however, depends on the selected neighborhood or smoothing strength and may obscure local landscape structures.

Related robust strategies reduce the influence of observations considered unreliable Martinez-Cantin et al. (2018). Robust generalized-Bayes GP regression limits the effect of contaminated observations Altamirano et al. (2024), trimmed likelihood selects a fitting subset Andrade & Takeda (2023), and robust phased elimination tolerates budgeted reward corruption Bogunovic et al. (2022). While effective under severe noise or contamination, suppressing or removing observations may also discard potentially useful information, which is particularly undesirable when expensive evaluations are scarce. Despite these efforts, directly improving observed data before surrogate fitting remains relatively underexplored in BO. RoSF-BO addresses this gap by augmenting the original observations with additional training samples, without extra expensive evaluations or changes to the acquisition and evaluation procedures.

## 3 Methodology

### 3.1 Framework

RoSF-BO is incorporated into the standard Bayesian optimization pipeline as an observation-level data augmentation module. As shown in Fig. 1 and Algorithm 1, the optimization starts with an initial set of evaluated samples generated by Latin hypercube sampling (LHS). Given the current noisy observations, RoSF-BO generates additional training samples with reduced noise interference and combines them with the original observations to construct the dataset used for surrogate fitting. The surrogate model is then fitted on the augmented dataset, and the acquisition function is optimized to select the next candidate for expensive evaluation. The newly obtained observation is added only to the original observation set, after which RoSF-BO reconstructs the augmented training dataset for the next iteration. This process continues until the evaluation budget is exhausted. RoSF-BO therefore operates only at the training-data interface, without modifying the acquisition or expensive evaluation procedure.

Within this framework, RoSF-BO consists of three main stages. First, an evolutionary clustering strategy partitions the current evaluated solutions into multiple subspaces, which are subsequently ordered according to their spatial relationships, as detailed in Section 3.2. Second, local functional relationships are estimated between adjacent subspaces using the available observations. Third, these relationships are used to identify observation pairs with complementary prediction residuals and generate additional training samples through joint interpolation. The generated samples are used together with the original observations only for surrogate fitting and are not treated as new expensive evaluations.

**Algorithm 1** RoSF-BO within Bayesian optimization

**Require:** Search domain $\mathcal{X}$; objective function with noisy observations $\tilde{f}(\cdot)$; initial design size $n_0$; total evaluation budget $B$; acquisition function $\alpha$
**Ensure:** Evaluated observation set

$\mathbf{X}_0 \leftarrow \mathrm{LHS}(n_0, \mathcal{X})$
$\mathbf{Y}_0 \leftarrow \tilde{f}(\mathbf{X}_0)$
$D_0 \leftarrow (\mathbf{X}_0, \mathbf{Y}_0)$
$t \leftarrow 0$
**while** $|D_t| < B$ **do**
  $D_{\mathrm{RoSF-BO}} \leftarrow \text{RoSF-BO}(D_t)$
  $D_{\mathrm{train}} \leftarrow D_t \cup D_{\mathrm{RoSF-BO}}$
  $\hat{f}(\cdot) \leftarrow \textsc{FitSurrogate}(D_{\mathrm{train}})$
  $\mathbf{X} \leftarrow \arg\max \alpha(\mathcal{X}; \hat{f}(\cdot))$
  $\mathbf{Y} \leftarrow \tilde{f}(\mathbf{X})$
  $D_{t+1} \leftarrow D_t \cup (\mathbf{X}, \mathbf{Y})$
  $t \leftarrow t + 1$
**end while**
**return** $D_t$

### 3.2 Evolutionary Subspace Partitioning and Ordering

RoSF-BO first organizes the evaluated solutions into spatially coherent subspaces to support local functional modeling. Since the relationship between decision variables and objective values may vary across different regions of the search space, directly modeling all evaluated samples together may overlook such local variations. RoSF-BO therefore partitions the current evaluated solutions according to their spatial distribution and subsequently establishes neighboring relationships among the resulting subspaces. At BO iteration $t$, let $D_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_t}$ denote the current evaluated dataset, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is an evaluated solution and $y_i \in \mathbb{R}$ is its corresponding objective value. Given a predefined number of subspaces $K$, the evaluated solutions are partitioned into $K$ spatial regions, which are then ordered to determine the adjacent subspace pairs used for subsequent local functional modeling and sample augmentation.

**Evolutionary Clustering.** For a given $K$, RoSF-BO searches for $K$ representative solutions from the current evaluated dataset to construct the subspaces. Each candidate is encoded by a set of $K$ distinct indices $S = \{s_1, \ldots, s_K\}$, where $s_k \in \{1, \ldots, n_t\}$, and each index $s_k$ directly specifies an evaluated solution $\mathbf{x}_{s_k}$ as a subspace representative. Given $S$, each evaluated solution is assigned to its closest representative, resulting in $K$ subspaces. The quality of $S$ is evaluated by the total distance of all evaluated solutions to their corresponding closest representatives:

$$F(S) = \sum_{i=1}^{n_t} \min_{k \in \{1,\ldots,K\}} d(\mathbf{x}_i, \mathbf{x}_{s_k}), \tag{1}$$

A smaller $F(S)$ indicates that the selected representatives provide a more compact partition of the evaluated solutions, which facilitates subsequent local functional modeling. We employ an evolutionary strategy to search for the representative index set $S$ Robles-Berumen et al. (2024); Chenan & Tsutsumida (2025). The initial population is generated by randomly selecting $K$ distinct indices from the $n_t$ evaluated solutions. During each generation, candidate index sets are updated through selection, crossover, and mutation, while candidates with smaller values of $F(S)$ are preferentially retained. If duplicate indices appear after the evolutionary operations, they are replaced by unused indices to ensure that each candidate contains $K$ distinct representative solutions. After a predefined number of generations, the candidate with the smallest $F(S)$ is selected to determine the final subspace partition $\{C_1, \ldots, C_K\}$. This index-based representation restricts the clustering search to representative solutions selected from the current evaluated set, instead of directly optimizing arbitrary cluster centers in the continuous decision space. Accordingly, the search space satisfies

$$\Omega_{\mathrm{index}} \subset \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_t}\}^K \subset \mathcal{X}^K, \qquad |\Omega_{\mathrm{index}}| = \binom{n_t}{K}. \tag{2}$$

Therefore, the continuous search over $K$ cluster centers is restricted to a finite search over evaluated representatives, while keeping each subspace anchored to observed regions.

**Subspace Ordering.** After obtaining the $K$ subspaces, RoSF-BO further orders them to determine neighboring relationships for subsequent local functional modeling. For each subspace $C_k$, its center $\mathbf{c}_k$ is calculated from the solutions contained in the subspace. Based on the distances between these centers, a closed traveling salesperson problem (TSP) Croes (1958) is constructed to obtain an ordering of the subspaces. The resulting ordered subspaces are denoted by

$$\mathcal{C}^{\mathrm{ord}} = (C_{\pi_1}, C_{\pi_2}, \ldots, C_{\pi_K}), \tag{3}$$

where $\pi = (\pi_1, \ldots, \pi_K)$ denotes the ordering obtained by the TSP. Two consecutive subspaces are regarded as neighboring regions. Since a closed tour is used, the last subspace is also connected to the first one, i.e., $C_{\pi_{K+1}} = C_{\pi_1}$. These neighboring subspace pairs are subsequently used for cross-subspace functional modeling and sample augmentation.

### 3.3 Cross-Subspace Functional Modeling

Based on the ordered subspaces, RoSF-BO estimates local objective relationships between neighboring regions and uses them to construct augmented training samples. For the $r$-th pair of neighboring subspaces, we define

$$B_r = C_{\pi_r} \cup C_{\pi_{r+1}}, \qquad r = 1, \ldots, K, \tag{4}$$

where $C_{\pi_{K+1}} = C_{\pi_1}$. Each $B_r$ therefore contains the evaluated solutions from two spatially neighboring subspaces and serves as the local region for subsequent functional modeling and sample generation. For each neighboring region $B_r$, RoSF-BO fits a local regression model using the evaluated solutions and their objective values. The fitted model, denoted by $\hat{f}_r(\mathbf{x})$, approximates the local relationship between the decision variables and objective value. For each evaluated solution $\mathbf{x}_i \in B_r$, its local prediction residual is defined as

$$e_{r,i} = \hat{f}_r(\mathbf{x}_i) - y_i, \qquad \mathbf{x}_i \in B_r. \tag{5}$$

The residual reflects the deviation of the observed objective value from the local functional relationship and is subsequently used to identify observation pairs whose deviations can compensate for each other. The local prediction residual provides useful information about the observation noise. For an evaluated solution $\mathbf{x}_i \in B_r$, let $f(\mathbf{x}_i)$ denote its noise-free objective value and $\epsilon_i$ denote the observation noise, such that $y_i = f(\mathbf{x}_i) + \epsilon_i$. The residual can then be written as $e_{r,i} = d_r(\mathbf{x}_i) - \epsilon_i$, where $d_r(\mathbf{x}_i) = \hat{f}_r(\mathbf{x}_i) - f(\mathbf{x}_i)$ denotes the local model discrepancy. Therefore, when the local model provides a reasonable approximation, the residual mainly reflects the deviation introduced by observation noise.

### 3.4 Residual-Cancellation Matching and Low-Noise Data Augmentation

For two neighboring subspaces $C_{\pi_r}$ and $C_{\pi_{r+1}}$, we define the matching cost between observations $i \in C_{\pi_r}$ and $j \in C_{\pi_{r+1}}$ by evaluating their residual cancellation at multiple uniformly spaced interpolation positions:

$$g_r(i,j) = \frac{1}{H} \sum_{q=1}^{H} \left| \left(1 - \frac{q}{H+1}\right) e_{r,i} + \frac{q}{H+1} e_{r,j} \right|, \tag{6}$$

where $H$ denotes the number of interpolation positions used for residual-cancellation evaluation. A smaller $g_r(i,j)$ indicates stronger residual cancellation across the interpolation interval. In all experiments, we set $H = 10$; further discussion on this design is provided in **Appendix A.3**. RoSF-BO first applies the Hungarian algorithm Kuhn (1955) to obtain a minimum-cost one-to-one matching between the two subspaces. When their numbers of observations are unequal, each unmatched observation in the larger subspace is further paired with its minimum-cost counterpart in the other subspace. Thus, every observation participates in at least one pair for each neighboring-subspace relation involving its subspace. Since the subspaces form a closed tour, each observation is involved in at least two cross-subspace pairs overall, corresponding to the two neighboring subspaces. The resulting pair set for the $r$-th neighboring-subspace relation is denoted by $M_r$.

**Algorithm 2** Procedure of RoSF-BO

**Require:** Evaluated dataset $D_t$; number of subspaces $K$; interpolation density $\eta$
**Ensure:** Augmented dataset $D_{\text{RoSF-BO}}$

```
{C_{π_1}, ..., C_{π_K}} ← PartitionAndOrder(D_t, K)
C_{π_{K+1}} ← C_{π_1}
D_RoSF-BO ← ∅
for r = 1, ..., K do
    f̂_r ← FitLocalModel(C_{π_r}, C_{π_{r+1}})
    Compute residuals e_{r,i} = f̂_r(x_i) − y_i
    M_r ← ResidualCancellationPair(C_{π_r}, C_{π_{r+1}})
    for all (i, j) ∈ M_r do
        n_{r,ij} ← ⌈η d(x_i, x_j)⌉
        Generate n_{r,ij} samples by jointly interpolating (x_i, y_i) and (x_j, y_j)
        Add the generated samples to D_RoSF-BO
    end for
end for
return D_RoSF-BO
```

After obtaining the matched pairs from the neighboring subspaces, each pair defines an interpolation bridge for sample generation. For a matched pair $(i, j) \in M_r$ from the $r$-th neighboring subspace pair, let $\ell_{r,ij}$ denote the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in the decision space. RoSF-BO uses a parameter $\eta$ to control the interpolation density, where $\eta$ specifies the number of generated samples per unit distance. The number of samples assigned to each interpolation bridge is determined by $\lceil \eta\, \ell_{r,ij} \rceil$, longer bridges receive more generated samples while each matched pair generates at least one sample. RoSF-BO then generates the assigned samples by linearly interpolating both the decision variables and objective values. Let $n_{r,ij} = \lceil \eta \ell_{r,ij} \rceil$ denote the number of generated samples on the corresponding interpolation bridge. For the $h$-th sample, $h = 1, \ldots, n_{r,ij}$, the interpolation coefficient is defined as $\tau_h = h/(n_{r,ij} + 1)$, and the generated sample is obtained as

$$\tilde{\mathbf{x}}_{r,ij,h} = (1 - \tau_h)\mathbf{x}_i + \tau_h \mathbf{x}_j, \qquad \tilde{y}_{r,ij,h} = (1 - \tau_h) y_i + \tau_h y_j. \tag{7}$$

Since $0 < \tau_h < 1$, the original endpoints are excluded and only intermediate samples are generated. All generated samples from the matched pairs are collected to form the augmented dataset $D_{\text{RoSF-BO}}$. This dataset is combined with the original evaluated observations only when fitting the surrogate model. The generated samples are not treated as new function evaluations and are not added to the original observation set. Therefore, RoSF-BO improves the training data used by the surrogate without changing the actual evaluation budget or the standard BO iteration process. The complete procedure of RoSF-BO is summarized in Algorithm 2. In **Appendix A**, we further provide a theoretical analysis of RoSF-BO, explaining how residual-cancellation interpolation contributes to low-noise sample generation. **Notably**, RoSF-BO is also applicable to multi-objective noisy BO. For each objective, the above procedure is performed independently, and the generated samples are used to augment the training data of the corresponding surrogate model.

## 4 Experiments

In this section, we discuss the following research questions. **RQ1**: How effectively does RoSF-BO reduce noise interference in noisy observation sets? **RQ2**: How sensitive is RoSF-BO to the hyperparameters $K$ and $\eta$? **RQ3**: Does the sampling distribution of the observations affect the performance of RoSF-BO? **RQ4**: How effective is RoSF-BO for noisy single-objective Bayesian optimization? **RQ5**: How effective is RoSF-BO for noisy multi-objective Bayesian optimization? Below, we first introduce the experimental settings and then address RQ1–RQ5.

### 4.1 Experimental Setup

**Experimental Setup.** We use linear regression for local functional modeling, as it provides a simple and stable approximation with limited observations. Following the mixed-noise settings in Guo et al. (2023), we construct noisy observations under different noise intensities, with the detailed

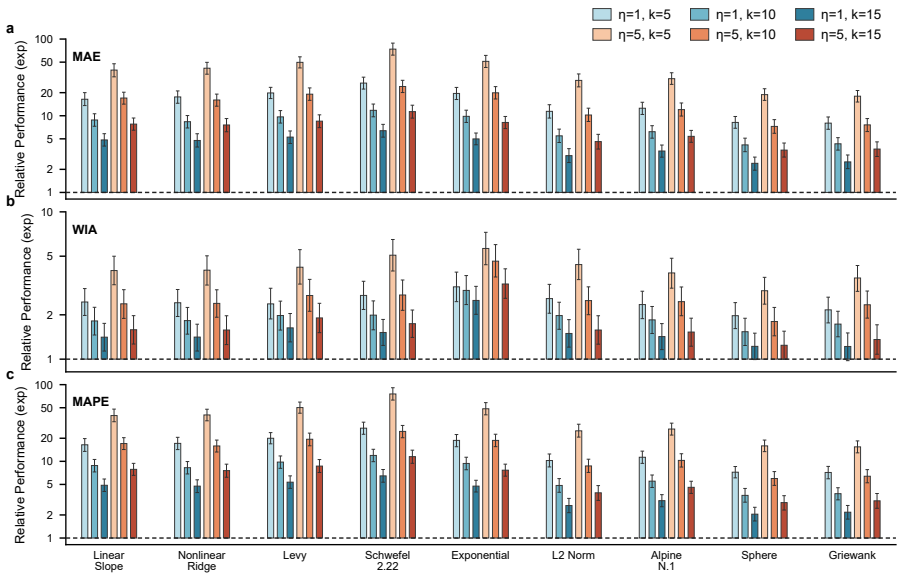


Figure 2: Static-data improvement of RoSF-BO under LHS sampling.

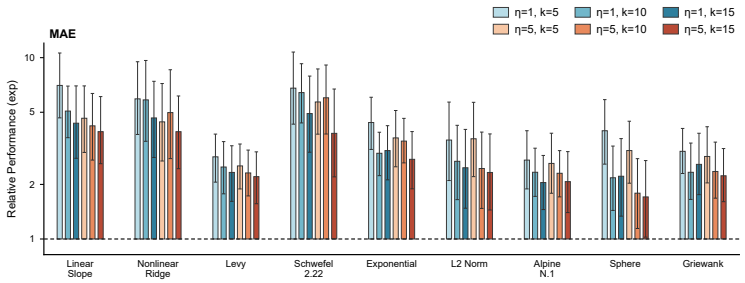


Figure 3: Static-data improvement of RoSF-BO under normally distributed sampling.

noise distributions and parameter settings provided in **Appendix B.1.** For evolutionary subspace partitioning, we use a genetic algorithm with a population size of 50 and 50 generations to search for the representative index set $S$. For subspace ordering, the closed TSP over the subspace centers is solved using a 2-opt heuristic Englert et al. (2014) to obtain the neighboring-subspace sequence.

**Benchmarks.** For the static-data and single-objective experiments, we consider 9 benchmark functions: Linear Slope, Nonlinear Ridge, Levy, Schwefel 2.22, Exponential, $L_2$-Norm, Alpine No. 1, Sphere, and Griewank **?**. For multi-objective optimization, we use ZDT1–ZDT3 Zitzler et al. (2000), DTLZ1–DTLZ3 and DTLZ5–DTLZ7 Deb et al. (2002).

### 4.2 Static Dataset Evaluation (RQ1-RQ3)

To answer RQ1–RQ3, we first evaluate RoSF-BO on static noisy datasets to isolate its observation-level augmentation effect from the sequential BO process. For each of the nine benchmark functions, we consider the 10-dimensional setting and generate 100 observations. Each setting is independently repeated 30 times. To investigate the sensitivity to the two main hyperparameters, we evaluate $K \in \{5, 10, 15\}$ and $\eta \in \{1, 5\}$, resulting in 6 parameter combinations. The static experiments use MAE, WIA, and MAPE to evaluate how well the noisy or augmented objective values match their noise-free counterparts, where lower MAE and MAPE and higher WIA indicate better agreement. For intuitive comparison across different metrics, we adopt a Z-score normalization design Ma et al. (2023), using the original noisy dataset as the baseline. The normalized scores are direction-adjusted such that a larger Z-score consistently indicates better agreement, and are then exponentiated to quantify the relative improvement. Detailed calculations are provided in **Appendix B.2**.

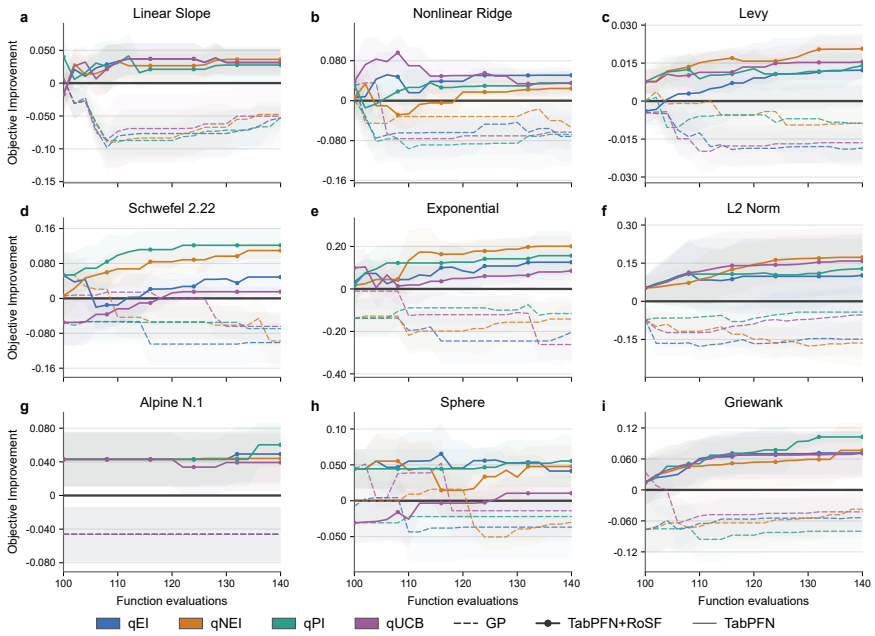


Figure 4: Single-Objective Optimization Results with Different Surrogates and BO Methods. Colors distinguish the BO methods, while solid and dashed curves denote TabPFN+RoSF-BO and GP, respectively. TabPFN combined with each BO method serves as the baseline and is represented only by the zero reference line without an additional color.

As shown in Fig. 2, RoSF-BO improves all three metrics under every evaluated hyperparameter combination, supporting its effectiveness for RQ1. For RQ2, $K = 5$ and $\eta = 5$ achieve the strongest overall results under LHS sampling, while the consistent gains across other settings indicate relatively stable sensitivity to $K$ and $\eta$. For RQ3, we repeat the analysis using normally distributed observations. As shown in Fig. 3, the preferred combinations of $K$ and $\eta$ differ from those under LHS sampling, indicating that the observation distribution affects hyperparameter preference. Additional static-data results across different dimensions are provided in **Appendix B.4**.

### 4.3 Single-Objective Bayesian Optimization (RQ4)

To answer RQ4, we evaluate RoSF-BO on nine 10-dimensional noisy single-objective problems with qNEI Letham et al. (2019); Balandat et al. (2020),qEI Balandat et al. (2020), qUCB, and qPI Wilson et al. (2018). Since the interpolated samples alter the empirical training distribution and introduce dependencies among samples, we combine RoSF-BO with TabPFN Hollmann et al. (2025); Yu et al. (2026); Du et al. (2026); Jin et al. (2026) rather than a standard GP, while retaining GP as a conventional baseline. We report the Best-Objective Improvement relative to TabPFN, defined as

$$\Delta f_t^{(m)} = f\left(\mathbf{x}_{t,\text{best}}^{\text{TabPFN}}\right) - f\left(\mathbf{x}_{t,\text{best}}^{(m)}\right), \tag{8}$$

where $\mathbf{x}_{t,\text{best}}^{(m)}$ is the best solution found by method $m$ up to step $t$. A positive value indicates better performance than TabPFN. After 100 initial samples are generated by LHS, we initialize $K = 5$ and $\eta = 5$ and update them every 10 evaluations, using an elbow criterion for $K$ and hold-out validation for $\eta$; details are provided in **Appendix B.3**. As shown in Fig. 4, the zero line denotes TabPFN with BO methods, solid curves represent TabPFN+RoSF-BO, and dashed curves represent GP, with different colors indicating the four acquisition functions. Curves and shaded regions show the mean and one standard deviation over 15 random seeds, respectively. TabPFN+RoSF-BO achieves positive final improvement in all 36 problem–acquisition combinations, showing consistent benefits across the evaluated single-objective settings.

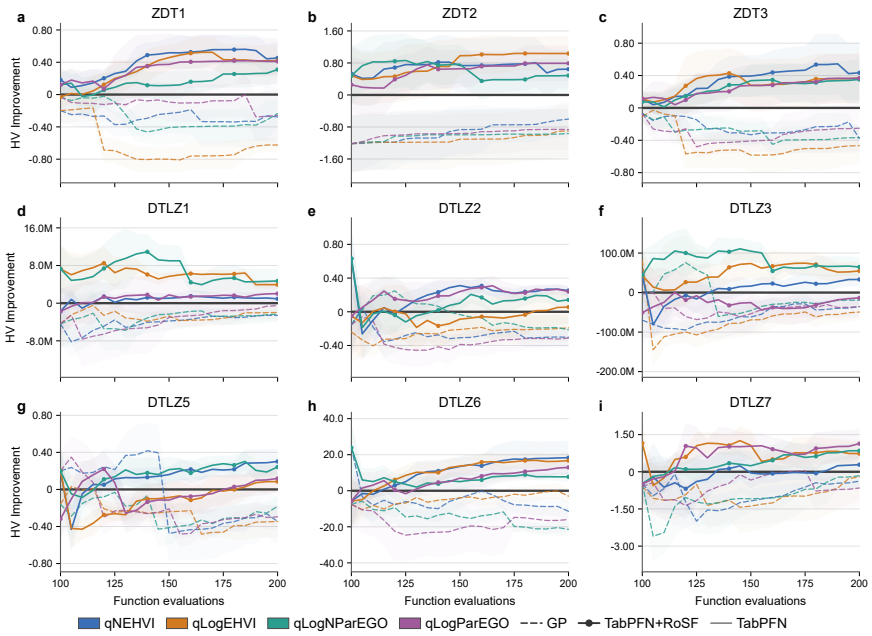


Figure 5: Multi-objective Optimization Results with Different Surrogates and BO Methods.

### 4.4 Multi-Objective Bayesian Optimization (RQ5)

To answer RQ5, we evaluate RoSF-BO on nine 10-dimensional noisy multi-objective problems with qNEHVI Daulton et al. (2021), qLogEHVI Daulton et al. (2020); Ament et al. (2023), qLogNParEGO Knowles (2006); Ament et al. (2023), and qLogParEGO Knowles (2006); Ament et al. (2023). As in the single-objective experiments, RoSF-BO is combined with TabPFN, while GP is retained as a conventional baseline. We report the HV Zitzler & Thiele (1999) Improvement relative to TabPFN:

$$\Delta HV_t^{(m)} = HV_t^{(m)} - HV_t^{\mathrm{TabPFN}}, \tag{9}$$

where a positive value indicates better performance than TabPFN under the corresponding BO method. As shown in Fig. 5, the zero line denotes TabPFN with the corresponding BO method, solid curves represent TabPFN+RoSF-BO, and dashed curves represent GP. TabPFN+RoSF-BO achieves positive final HV Improvement in all problem–BO combinations, while GP remains below the TabPFN baseline, demonstrating broadly consistent improvements across different multi-objective BO methods. These results further confirm that the observation-level augmentation strategy remains effective when extending from single-objective to multi-objective optimization. Higher-dimensional single- and multi-objective BO results are provided in **Appendix B.5**, with practical engineering applications in **Appendix B.6**.

## 5 Conclusion

This work proposes RoSF-BO from an observation-level perspective for noise handling in Bayesian optimization. Experimental results show that RoSF-BO can effectively improve data quality under noisy observations and further enhance both single-objective and multi-objective optimization across different BO methods. While the augmented training distribution introduced by RoSF-BO may not be equally suitable for all surrogate models, RoSF-BO itself remains an independent observation-level module that can be readily integrated into existing BO pipelines without modifying their acquisition or evaluation procedures, preserving a plug-and-play interface at the framework level. Overall, this work provides a new observation-level perspective for noise handling in Bayesian optimization and broadens existing approaches beyond surrogate modeling and acquisition design. Future work will focus on more adaptive hyperparameter determination to reduce the computational overhead during sequential optimization.

### AI USE STATEMENT

In this work, we used generative AI tools only to assist with language polishing and improving the clarity of the manuscript. All AI-assisted text was carefully reviewed and revised by the authors. We take full responsibility for the final content of this work.

## REFERENCES


Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and conjugate gaussian process regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1155–1185, 2024.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in neural information processing systems*, volume 36, pp. 20577–20612. Curran Associates, Inc., 2023.

Sebastian Ament, Elizabeth Santorella, David Eriksson, Ben Letham, Maximilian Balandat, and Eytan Bakshy. Robust gaussian processes via relevance pursuit. *Advances in Neural Information Processing Systems*, 37:61700–61734, 2024.

Daniel Andrade and Akiko Takeda. Robust gaussian process regression with the trimmed marginal likelihood. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 67–76, 2023.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mickael Binois, Robert B. Gramacy, and Michael Ludkovski. Practical heteroskedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018a.

Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018b.

Mickael Binois, Jiangeng Huang, Robert B. Gramacy, and Michael Ludkovski. Replication or exploration? sequential design for stochastic simulation experiments. *Technometrics*, 61(1):7–23, 2019.

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mayk Caldas Ramos, Shane S Michtavy, Andrew D White, and Marc D Porosoff. Bayesian optimization of catalysis with in-context learning. *ACS Central Science*, 12(5):599–615, 2026.

Huang Chenan and Narumasa Tsutsumida. A scalable k-medoids clustering via whale optimization algorithm. *Array*, pp. 100599, 2025.

Georges A Croes. A method for solving traveling-salesman problems. *Operations research*, 6(6): 791–812, 1958.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in neural information processing systems*, volume 33, pp. 9851–9864. Curran Associates, Inc., 2020.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2187–2200, 2021.

Kalyanmoy Deb, Lothar Thiele, Marco Laumanns, and Eckart Zitzler. Scalable multi-objective optimization test problems. In *Proceedings of the 2002 congress on evolutionary computation. CEC'02 (Cat. No. 02TH8600)*, volume 1, pp. 825–830, 2002.

Shachi Deshpande, Charles Marx, and Volodymyr Kuleshov. Online calibrated and conformal prediction improves bayesian optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1450–1458, 2024.

Stefan Desimpel, Matthieu Dorbec, Kevin M Van Geem, and Christian V Stevens. Bayesian optimization for chemical reactions. *Chemical Society Reviews*, 55(5):2731–2775, 2026.

Bach Do and Ruda Zhang. Multifidelity bayesian optimization: A review. *AIAA Journal*, 63(6): 2286–2322, 2025.

Yukun Du, Haiyue Yu, Xiaotong Xie, Yan Zheng, Lixin Zhan, Yudong Du, Chongshuang Hu, Boxuan Wang, and Jiang Jiang. Meta-black-box optimization with bi-space landscape analysis and dual-control mechanism for saea. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36891–36899, 2026.

Matthias Englert, Heiko Röglin, and Berthold Vöcking. Worst case and probabilistic analysis of the 2-opt algorithm for the tsp. *Algorithmica*, 68(1):190–264, 2014.

Paul Goldberg, Christopher Williams, and Christopher Bishop. Regression with input-dependent noise: A gaussian process treatment. In M. Jordan, M. Kearns, and S. Solla (eds.), *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

Yaqing Guo, Wenjian Wang, and Xuejun Wang. A robust linear regression feature selection method for data sets with unknown noise. *IEEE Transactions on Knowledge and Data Engineering*, 35 (1):31–44, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

D. Huang, T. T. Allen, W. I. Notz, and N. Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34(3):441–466, mar 2006.

Carl Hvarfner, Erik Orm Hellsten, Frank Hutter, and Luigi Nardi. Self-correcting bayesian optimization through bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla bayesian optimization performs great in high dimensions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20793–20817, 2024.

Xiao Jin, Yongxiong Wang, Haobo Liu, Yudong Du, and Yukun Du. Meta-black-box optimization with ensemble surrogate modeling for robustness–accuracy trade-off within saea. *Swarm and Evolutionary Computation*, 106:102425, 2026.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to bayesian optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22643–22671, 2024.

J. Knowles. Parego: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

Yongfan Lu, Bingdong Li, and Aimin Zhou. Are you concerned about limited function evaluations: Data-augmented pareto set learning for expensive multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14202–14210, 2024.

Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic gaussian process regression. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 841–848. Omnipress, 2011.

Zeyuan Ma, Hongshu Guo, Jiacheng Chen, Zhenrui Li, Guojun Peng, Yue-Jiao Gong, Yining Ma, and Zhiguang Cao. Metabox: A benchmark platform for meta-black-box optimization with reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10775–10795. Curran Associates, Inc., 2023.

Ruben Martinez-Cantin, Kevin Tee, and Michael McCourt. Practical bayesian optimization in the presence of outliers. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1722–1731. PMLR, 09–11 Apr 2018.

Nobuo Namura. Single and multi-objective optimization benchmark problems focusing on human-powered aircraft design. In *International Conference on Evolutionary Multi-Criterion Optimization*, pp. 195–210. Springer, 2025.

Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, September 2013.

Sam Reifenstein, Timothee Leleu, and Yoshihisa Yamamoto. Dynamic anisotropic smoothing for noisy derivative-free optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42332–42364, 2024.

Hermes Robles-Berumen, Amelia Zafra, and Sebastián Ventura. A survey of genetic algorithms for clustering: Taxonomy and empirical analysis. *Swarm and Evolutionary Computation*, 91:101720, 2024.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. Random exploration in bayesian optimization: Order-optimal regret and computational efficiency. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43112–43141, 2024.

Warren Scott, Peter Frazier, and Warren Powell. The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM Journal on Optimization*, 21(3):996–1026, 2011.

Maximilian Siska, Emma Pajak, Katrin Rosenthal, Antonio del Rio Chanona, Eric von Lieres, and Laura M Helleckes. A guide to bayesian optimization in bioprocess engineering. *Biotechnology and bioengineering*, 123(4):805–830, 2026.

James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for Bayesian optimization. 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.

Rosen Yu, Cyril Picard, and Faez Ahmed. GIT-BO: High-dimensional bayesian optimization with tabular foundation models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International conference on learning representations*, 2026.

Han Zhou, Xingchen Ma, and Matthew B. Blaschko. A corrected expected improvement acquisition function under noisy observations. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pp. 1747–1762, 2024.

E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation*, 8(2):173–195, 2000.

# A THEORETICAL ANALYSIS OF RESIDUAL-CANCELLATION INTERPOLATION

This section of appendix provides a concise theoretical analysis of the residual-cancellation interpolation used in RoSF-BO. We focus on three aspects:

1. We show that when the local functional model reasonably captures the relationship between neighboring subspaces, residual cancellation can reduce the noise magnitude of the generated samples (see **Appendix A.1**).
2. We analyze the case of imperfect residual cancellation and show that interpolation does not cause uncontrolled noise amplification (see **Appendix A.2**).
3. We further discuss the multi-point residual-cancellation matching criterion and its relationship with the interpolation process (see **Appendix A.3**)

Since multi-objective problems are handled independently for each objective in RoSF-BO, the following analysis considers the scalar-output case for clarity.

## A.1 EFFECTIVENESS OF RESIDUAL-CANCELLATION INTERPOLATION

Consider a matched pair $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$, where $y_u = f(\mathbf{x}_u) + \epsilon_u$, $u \in \{i, j\}$. For an interpolation coefficient $0 < \tau < 1$, RoSF-BO generates $\tilde{\mathbf{x}}_\tau = (1-\tau)\mathbf{x}_i + \tau\mathbf{x}_j$ and $\tilde{y}_\tau = (1-\tau)y_i + \tau y_j$. The effective noise of the generated sample can therefore be written as

$$\begin{aligned} \tilde{\epsilon}_\tau &= \tilde{y}_\tau - f(\tilde{\mathbf{x}}_\tau) = (1-\tau)\epsilon_i + \tau\epsilon_j + \delta_\tau, \\ \delta_\tau &= (1-\tau)f(\mathbf{x}_i) + \tau f(\mathbf{x}_j) - f(\tilde{\mathbf{x}}_\tau), \end{aligned} \tag{10}$$

where $\delta_\tau$ represents the local interpolation error of the noise-free objective function. To quantify the noise reduction, define the weighted endpoint noise level and the expected cancellation gain as

$$\begin{aligned} B_\tau &= (1-\tau)\mathbb{E}|\epsilon_i| + \tau\mathbb{E}|\epsilon_j|, \\ C_\tau &= B_\tau - \mathbb{E}\left|(1-\tau)\epsilon_i + \tau\epsilon_j\right|. \end{aligned} \tag{11}$$

By the triangle inequality, $C_\tau \geq 0$, confirming that the linearly combined endpoint-noise term cannot exceed the weighted endpoint noise level. Any reduction corresponds to a positive cancellation gain, thereby supporting the residual-cancellation mechanism used in RoSF-BO. Using Eq. 10, the expected noise magnitude of the generated sample satisfies

$$\begin{aligned} \mathbb{E}|\tilde{\epsilon}_\tau| &\leq \mathbb{E}\left|(1-\tau)\epsilon_i + \tau\epsilon_j\right| + \mathbb{E}|\delta_\tau|, \\ \therefore \quad B_\tau - \mathbb{E}|\tilde{\epsilon}_\tau| &\geq C_\tau - \mathbb{E}|\delta_\tau|. \end{aligned} \tag{12}$$

Eq 12 gives a lower bound on the expected noise reduction achieved by interpolation. It shows that the improvement is determined by the cancellation gain $C_\tau$ and the local interpolation error. In particular, whenever $C_\tau > \mathbb{E}|\delta_\tau|$, the generated sample has a smaller expected noise magnitude than the weighted endpoint noise level.

The residual-cancellation matching in RoSF-BO is designed to increase this cancellation effect. Since $e_{r,u} = d_r(\mathbf{x}_u) - \epsilon_u$, a sufficiently accurate local model gives $e_{r,u} \approx -\epsilon_u$. Therefore, the linearly combined residual $|(1-\tau)e_{r,i} + \tau e_{r,j}|$ approximately reflects the magnitude of the corresponding combined noise component $|(1-\tau)\epsilon_i + \tau\epsilon_j|$. Accordingly, the matching criterion in Eq. 6 favors observation pairs that exhibit stronger residual cancellation across multiple interpolation positions, thereby promoting the cancellation effect characterized by $C_\tau$. **Importantly**, the above result does not require homoscedastic noise, independence between $\epsilon_i$ and $\epsilon_j$, or any specific parametric noise distribution. The two observations may follow different and dependent noise distributions; only finite first absolute moments are required for the expectations above to exist.

## A.2 ROBUSTNESS TO IMPERFECT RESIDUAL CANCELLATION

The residuals estimated from the local model may not always perfectly reflect the directions of the true observation noise. We therefore further consider the case where residual cancellation is imperfect or even absent.

From Eq. 10, regardless of whether the two noise components cancel, the triangle inequality gives

$$\begin{aligned}\mathbb{E}|\tilde{\epsilon}_\tau| &\le \mathbb{E}\left|(1-\tau)\epsilon_i + \tau\epsilon_j\right| + \mathbb{E}|\delta_\tau| \\ &\le (1-\tau)\mathbb{E}|\epsilon_i| + \tau\mathbb{E}|\epsilon_j| + \mathbb{E}|\delta_\tau| \\ &= B_\tau + \mathbb{E}|\delta_\tau|.\end{aligned} \tag{13}$$

Therefore, even when residual cancellation is ineffective, the expected noise magnitude of the generated sample cannot exceed the weighted endpoint noise level by more than the local interpolation error. In particular, when the noise-free objective is locally close to linear such that $\mathbb{E}|\delta_\tau|$ is small, imperfect residual cancellation only reduces the additional cancellation gain rather than causing uncontrolled noise amplification.

Together with the result in Appendix A.1, residual cancellation can therefore be viewed as an additional noise-reduction mechanism: effective cancellation provides a positive gain $C_\tau$, whereas imperfect cancellation mainly causes this gain to diminish. The interpolation itself remains bounded by Eq. 13. This result likewise requires neither homoscedasticity, independence between the endpoint noises, nor any specific noise distribution.

### A.3 Multi-Point Residual-Cancellation Matching

To characterize residual cancellation over the entire interpolation interval, we consider the continuous matching cost

$$g_r^{\mathrm{cont}}(i,j) = \int_0^1 \left|(1-\tau)e_{r,i} + \tau e_{r,j}\right| \, d\tau. \tag{14}$$

Eq. 6 approximates this quantity using $H$ uniformly spaced interpolation positions. When $H = 1$, it reduces to

$$g_r(i,j) = \frac{1}{2}\left|e_{r,i} + e_{r,j}\right|, \tag{15}$$

which corresponds to midpoint-only residual cancellation. Therefore, using $H > 1$ provides a more representative characterization of cancellation over the interpolation interval. We keep $H$ fixed and independent of the augmentation density $\eta$, where $H$ is used only for matching and $\eta$ determines the subsequent number of augmented samples. In all experiments, we set $H = 10$ to provide a practical multi-point approximation while facilitating efficient computation and avoiding unnecessary matching complexity.

## B Additional Discussion

### B.1 Detailed Noise Settings

We adopt a heterogeneous noise setting throughout the experiments. Noise is added only to the evaluated objective values, while the decision variables remain noise-free. Rather than using fixed absolute noise amplitudes, the noise magnitude is scaled according to the absolute value of the corresponding noise-free objective value. Specifically, each observation is randomly assigned one of the three noise components summarized in Table 1.

Table 1: Heterogeneous noise setting used in the experiments.

| Noise component | Probability | Noise setting |
|---|---|---|
| High Gaussian | 0.20 | Normal$(0,\ (0.30\|y\|)^2)$ |
| Uniform | 0.30 | Uniform$(-0.15\|y\|,\ 0.15\|y\|)$ |
| Low Gaussian | 0.50 | Normal$(0,\ (0.03\|y\|)^2)$ |

Here, $y$ denotes the corresponding noise-free objective value. Therefore, the noise intensity varies with the magnitude of the objective value, avoiding the use of a fixed absolute noise scale across problems with different response ranges.

### B.2 Relative Improvement

To enable intuitive comparison across different evaluation metrics, we normalize the static-data results using the original noisy observations as the reference. For a given benchmark and metric, let $v_r^{\mathrm{ori}}$ and $v_r^{\mathrm{aug}}$ denote the metric values before and after RoSF-BO augmentation in the $r$-th independent run, respectively. The mean and standard deviation of the original noisy results are calculated as $\mu_{\mathrm{ori}}$ and $\sigma_{\mathrm{ori}}$. The direction-adjusted standardized score of the augmented result is defined as

$$z_r = s_m \frac{v_r^{\mathrm{aug}} - \mu_{\mathrm{ori}}}{\sigma_{\mathrm{ori}}}, \qquad s_m = \begin{cases} -1, & m \in \{\mathrm{MAE}, \mathrm{MAPE}\}, \\ 1, & m = \mathrm{WIA}, \end{cases} \tag{16}$$

where $s_m$ ensures that a larger score consistently indicates better performance for all metrics. The scores are averaged over the independent runs and exponentiated to obtain the reported relative improvement:

$$S = \exp\left(\frac{1}{R}\sum_{r=1}^{R} z_r\right). \tag{17}$$

Under this definition, the original noisy observations have an average standardized score of zero and therefore correspond to $S = 1$. Hence, $S > 1$ indicates improvement after RoSF-BO augmentation, whereas $S < 1$ indicates degradation. The exponential transformation is monotonic and maps the standardized scores to a positive scale centered at the baseline value of one, facilitating visual comparison across different metrics.

### B.3 Elbow-based Hyperparameter Selection

After the initial LHS sampling, we initialize RoSF-BO with $K = 5$ and $\eta = 5$. During sequential optimization, the hyperparameters are updated every 10 new expensive evaluations. For $K$, we evaluate the clustering objective in Eq. 1 under different numbers of subspaces and select the elbow point beyond which increasing $K$ yields only marginal improvement. For $\eta$, we adopt a hold-out validation strategy. The current observations are randomly divided into training and validation sets with a ratio of 7:3. We evaluate $\eta \in \{1, 5, 10\}$ by constructing the augmented training set for each candidate value and measuring the corresponding surrogate prediction error on the validation set. The value of $\eta$ with the lowest validation error is selected. The resulting $K$ and $\eta$ are then used for the following 10 evaluations before being updated again.

### B.4 Static-Data Improvement of RoSF-BO across Different Dimensions

**Table 2** reports the complete static-data results for 10-, 20-, and 30-dimensional settings across all nine combinations of $K \in \{5, 10, 15\}$ and $\eta \in \{1, 5, 10\}$. The dimensionality of each problem is indicated after its name using the notation D10, D20, and D30, respectively. For each benchmark and dimension, each entry denotes the mean over 30 independent runs. Statistical comparisons are conducted against the corresponding results before RoSF-BO augmentation using a two-sided paired Wilcoxon signed-rank test at the 0.05 significance level without multiple-comparison correction. For each dataset, the three consecutive rows report MAE, WIA, and MAPE, respectively. Lower MAE and MAPE values indicate better performance, whereas higher WIA values are preferred. The symbols $+$, $-$, and $\approx$ denote a significant improvement, a significant deterioration, and no significant difference relative to the corresponding pre-augmentation result, respectively. The triplets in the final three rows are reported in the order (MAE, WIA, MAPE).

Table 2: Static-Data Results of RoSF-BO across Different Dimensions and Hyperparameter Settings.

| Dataset | Before | After RoSF-BO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\eta = 1$ | | | $\eta = 5$ | | | $\eta = 10$ | | |
| | | $K = 5$ | $K = 10$ | $K = 15$ | $K = 5$ | $K = 10$ | $K = 15$ | $K = 5$ | $K = 10$ | $K = 15$ |
| F1 Linear Slope D10 | 0.1459 | 0.1157+ | 0.1221+ | 0.1277+ | 0.1075+ | 0.1154+ | 0.1233+ | 0.1068+ | 0.1148+ | 0.1229+ |
| | 0.5692 | 0.6096+ | 0.5959+ | 0.5872+ | 0.6281+ | 0.6074+ | 0.5902+ | 0.6298+ | 0.6085+ | 0.5908+ |
| | 9.7966 | 7.7543+ | 8.1933+ | 8.5554+ | 7.2043+ | 7.7432+ | 8.2525+ | 7.1563+ | 7.7038+ | 8.2303+ |
| F1 Linear Slope D20 | 0.0810 | 0.0649+ | 0.0687+ | 0.0722+ | 0.0596+ | 0.0647+ | 0.0693+ | 0.0591+ | 0.0643+ | 0.0691+ |
| | 0.5878 | 0.6296+ | 0.6176+ | 0.6027+ | 0.6526+ | 0.6296+ | 0.6095+ | 0.6547+ | 0.6307+ | 0.6100+ |
| | 5.4153 | 4.3362+ | 4.5873+ | 4.8247+ | 3.9803+ | 4.3205+ | 4.6272+ | 3.9508+ | 4.2969+ | 4.6105+ |
| F1 Linear Slope D30 | 0.0814 | 0.0642+ | 0.0680+ | 0.0719+ | 0.0586+ | 0.0640+ | 0.0689+ | 0.0581+ | 0.0636+ | 0.0686+ |
| | 0.5042 | 0.5532+ | 0.5355+ | 0.5224+ | 0.5827+ | 0.5511+ | 0.5291+ | 0.5859+ | 0.5520+ | 0.5298+ |
| | 5.4366 | 4.2903+ | 4.5421+ | 4.8045+ | 3.9109+ | 4.2699+ | 4.6022+ | 3.8776+ | 4.2459+ | 4.5845+ |
| F2 Nonlinear Ridge D10 | 0.1429 | 0.1125+ | 0.1203+ | 0.1260+ | 0.1043+ | 0.1141+ | 0.1219+ | 0.1034+ | 0.1136+ | 0.1217+ |
| | 0.6515 | 0.6855+ | 0.6773+ | 0.6677+ | 0.7039+ | 0.6853+ | 0.6698+ | 0.7060+ | 0.6864+ | 0.6698+ |
| | 9.2915 | 7.3083+ | 7.8003+ | 8.1711+ | 6.7765+ | 7.3866+ | 7.8985+ | 6.7157+ | 7.3530+ | 7.8787+ |
| F2 Nonlinear Ridge D20 | 0.0804 | 0.0639+ | 0.0687+ | 0.0717+ | 0.0589+ | 0.0650+ | 0.0688+ | 0.0584+ | 0.0647+ | 0.0686+ |
| | 0.6550 | 0.6952+ | 0.6822+ | 0.6692+ | 0.7170+ | 0.6939+ | 0.6754+ | 0.7189+ | 0.6951+ | 0.6762+ |
| | 5.1902 | 4.1197+ | 4.4258+ | 4.6218+ | 3.7928+ | 4.1863+ | 4.4300+ | 3.7585+ | 4.1680+ | 4.4160+ |
| F2 Nonlinear Ridge D30 | 0.0829 | 0.0654+ | 0.0697+ | 0.0733+ | 0.0598+ | 0.0653+ | 0.0704+ | 0.0593+ | 0.0650+ | 0.0700+ |
| | 0.5506 | 0.5959+ | 0.5799+ | 0.5675+ | 0.6240+ | 0.5959+ | 0.5737+ | 0.6271+ | 0.5972+ | 0.5744+ |
| | 5.3380 | 4.2065+ | 4.4788+ | 4.7149+ | 3.8467+ | 4.1987+ | 4.5227+ | 3.8104+ | 4.1745+ | 4.5003+ |
| F3 Levy D10 | 0.0815 | 0.0659+ | 0.0704+ | 0.0732+ | 0.0614+ | 0.0672+ | 0.0716+ | 0.0610+ | 0.0669+ | 0.0717+ |
| | 0.2530 | 0.2814+ | 0.2771+ | 0.2687+ | 0.2965+ | 0.2851+ | 0.2692+ | 0.2983+ | 0.2855+ | 0.2690+ |
| | 7.3840 | 5.9719+ | 6.3803+ | 6.6313+ | 5.5639+ | 6.0890+ | 6.4823+ | 5.5222+ | 6.0598+ | 6.4868+ |
| F3 Levy D20 | 0.0820 | 0.0652+ | 0.0692+ | 0.0728+ | 0.0598+ | 0.0652+ | 0.0698+ | 0.0592+ | 0.0649+ | 0.0695+ |
| | 0.1306 | 0.1492+ | 0.1450+ | 0.1403+ | 0.1618+ | 0.1524+ | 0.1443+ | 0.1634+ | 0.1529+ | 0.1447+ |
| | 7.5093 | 5.9605+ | 6.3244+ | 6.6503+ | 5.4687+ | 5.9604+ | 6.3759+ | 5.4117+ | 5.9311+ | 6.3506+ |
| F3 Levy D30 | 0.1470 | 0.1131+ | 0.1206+ | 0.1276+ | 0.1026+ | 0.1127+ | 0.1217+ | 0.1015+ | 0.1120+ | 0.1212+ |
| | 0.0636 | 0.0737+ | 0.0714+ | 0.0697+ | 0.0809+ | 0.0752+ | 0.0720+ | 0.0819+ | 0.0755+ | 0.0722+ |
| | 13.4985 | 10.3719+ | 11.0593+ | 11.7027+ | 9.4047+ | 10.3264+ | 11.1565+ | 9.3007+ | 10.2621+ | 11.1081+ |
| F4 Schwefel 2.22 D10 | 0.0834 | 0.0679+ | 0.0726+ | 0.0750+ | 0.0632+ | 0.0693+ | 0.0729+ | 0.0627+ | 0.0690+ | 0.0728+ |
| | 0.6217 | 0.6495+ | 0.6409+ | 0.6355+ | 0.6618+ | 0.6469+ | 0.6373+ | 0.6624+ | 0.6473+ | 0.6370+ |
| | 5.7560 | 4.6838+ | 5.0031+ | 5.1721+ | 4.3540+ | 4.7739+ | 5.0217+ | 4.3202+ | 4.7559+ | 5.0170+ |
| F4 Schwefel 2.22 D20 | 0.1461 | 0.1125+ | 0.1197+ | 0.1263+ | 0.1025+ | 0.1124+ | 0.1204+ | 0.1015+ | 0.1117+ | 0.1199+ |
| | 0.3366 | 0.3740+ | 0.3630+ | 0.3514+ | 0.3986+ | 0.3760+ | 0.3572+ | 0.4012+ | 0.3772+ | 0.3574+ |
| | 9.9162 | 7.6332+ | 8.1236+ | 8.5673+ | 6.9556+ | 7.6233+ | 8.1654+ | 6.8867+ | 7.5771+ | 8.1338+ |
| F4 Schwefel 2.22 D30 | 0.0828 | 0.0645+ | 0.0691+ | 0.0726+ | 0.0586+ | 0.0650+ | 0.0691+ | 0.0579+ | 0.0647+ | 0.0687+ |
| | 0.4060 | 0.4546+ | 0.4392+ | 0.4259+ | 0.4873+ | 0.4551+ | 0.4332+ | 0.4914+ | 0.4565+ | 0.4343+ |
| | 5.5815 | 4.3501+ | 4.6564+ | 4.8947+ | 3.9511+ | 4.3848+ | 4.6596+ | 3.9049+ | 4.3611+ | 4.6362+ |
| F5 Exponential D10 | 0.0829 | 0.0723+ | 0.0752+ | 0.0782+ | 0.0685+ | 0.0724+ | 0.0769+ | 0.0681+ | 0.0720+ | 0.0767+ |
| | 0.6641 | 0.6958+ | 0.6892+ | 0.6769+ | 0.7110+ | 0.7015+ | 0.6788+ | 0.7122+ | 0.7031+ | 0.6795+ |
| | 4.6316 | 4.0729+ | 4.2405+ | 4.4086+ | 3.8706+ | 4.0922+ | 4.3454+ | 3.8531+ | 4.0742+ | 4.3353+ |
| F5 Exponential D20 | 0.1483 | 0.1148+ | 0.1226+ | 0.1295+ | 0.1046+ | 0.1153+ | 0.1241+ | 0.1034+ | 0.1146+ | 0.1236+ |
| | 0.1615 | 0.1968+ | 0.1940+ | 0.1902+ | 0.2169+ | 0.2078+ | 0.1987+ | 0.2197+ | 0.2093+ | 0.1997+ |
| | 7.5845 | 5.8828+ | 6.2855+ | 6.6436+ | 5.3644+ | 5.9165+ | 6.3746+ | 5.3079+ | 5.8851+ | 6.3497+ |
| F5 Exponential D30 | 0.0835 | 0.0660+ | 0.0697+ | 0.0738+ | 0.0602+ | 0.0655+ | 0.0706+ | 0.0596+ | 0.0650+ | 0.0703+ |
| | 0.0503 | 0.0677+ | 0.0678+ | 0.0661+ | 0.0767+ | 0.0751+ | 0.0706+ | 0.0779+ | 0.0760+ | 0.0711+ |
| | 4.1940 | 3.3178+ | 3.5075+ | 3.7100+ | 3.0284+ | 3.2933+ | 3.5504+ | 2.9980+ | 3.2710+ | 3.5374+ |
| F6 L2 Norm D10 | 0.1506 | 0.1214+ | 0.1294+ | 0.1353+ | 0.1128+ | 0.1241+ | 0.1327+ | 0.1121+ | 0.1238+ | 0.1324+ |
| | 0.4265 | 0.4645+ | 0.4565+ | 0.4470+ | 0.4834+ | 0.4629+ | 0.4483+ | 0.4848+ | 0.4629+ | 0.4487+ |
| | 9.6320 | 7.8163+ | 8.3345+ | 8.7146+ | 7.2873+ | 8.0147+ | 8.5715+ | 7.2442+ | 8.0006+ | 8.5523+ |
| F6 L2 Norm D20 | 0.0823 | 0.0698+ | 0.0736+ | 0.0765+ | 0.0646+ | 0.0700+ | 0.0740+ | 0.0641+ | 0.0697+ | 0.0738+ |
| | 0.4701 | 0.5106+ | 0.4999+ | 0.4834+ | 0.5352+ | 0.5134+ | 0.4853+ | 0.5373+ | 0.5143+ | 0.4853+ |
| | 5.2435 | 4.4768+ | 4.7295+ | 4.9112+ | 4.1570+ | 4.5081+ | 4.7701+ | 4.1287+ | 4.4899+ | 4.7575+ |

Continued on the next page

Table 2 (continued)

| Dataset | Before | After RoSF-BO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\eta = 1$ | | | $\eta = 5$ | | | $\eta = 10$ | | |
| | | $K = 5$ | $K = 10$ | $K = 15$ | $K = 5$ | $K = 10$ | $K = 15$ | $K = 5$ | $K = 10$ | $K = 15$ |
| | 0.0830 | 0.0702+ | 0.0742+ | 0.0779+ | 0.0649+ | 0.0707+ | 0.0752+ | 0.0645+ | 0.0704+ | 0.0749+ |
| F6 L2 Norm D30 | 0.3657 | 0.4100+ | 0.3956+ | 0.3845+ | 0.4351+ | 0.4060+ | 0.3877+ | 0.4378+ | 0.4069+ | 0.3877+ |
| | 5.2704 | 4.4922+ | 4.7559+ | 4.9974+ | 4.1698+ | 4.5458+ | 4.8396+ | 4.1446+ | 4.5266+ | 4.8232+ |
| | 0.1486 | 0.1204+ | 0.1274+ | 0.1327+ | 0.1127+ | 0.1212+ | 0.1298+ | 0.1121+ | 0.1208+ | 0.1298+ |
| F7 Alpine N.1 D10 | 0.5000 | 0.5305+ | 0.5246+ | 0.5125+ | 0.5447+ | 0.5334+ | 0.5094+ | 0.5462+ | 0.5340+ | 0.5082≈ |
| | 10.8893 | 8.8877+ | 9.4061+ | 9.8056+ | 8.3543+ | 8.9802+ | 9.6131+ | 8.3082+ | 8.9505+ | 9.6125+ |
| | 0.0825 | 0.0688+ | 0.0726+ | 0.0758+ | 0.0636+ | 0.0689+ | 0.0731+ | 0.0631+ | 0.0685+ | 0.0728+ |
| F7 Alpine N.1 D20 | 0.5293 | 0.5692+ | 0.5561+ | 0.5467+ | 0.5919+ | 0.5686+ | 0.5522+ | 0.5938+ | 0.5692+ | 0.5528+ |
| | 6.0292 | 5.0628+ | 5.3562+ | 5.5904+ | 4.7022+ | 5.0946+ | 5.4094+ | 4.6676+ | 5.0682+ | 5.3935+ |
| | 0.0825 | 0.0680+ | 0.0723+ | 0.0760+ | 0.0624+ | 0.0682+ | 0.0730+ | 0.0619+ | 0.0679+ | 0.0728+ |
| F7 Alpine N.1 D30 | 0.4394 | 0.4835+ | 0.4711+ | 0.4568+ | 0.5119+ | 0.4875+ | 0.4625+ | 0.5150+ | 0.4892+ | 0.4627+ |
| | 6.0197 | 4.9956+ | 5.3166+ | 5.5909+ | 4.6030+ | 5.0359+ | 5.3945+ | 4.5664+ | 5.0124+ | 5.3767+ |
| | 0.1495 | 0.1223+ | 0.1293+ | 0.1334+ | 0.1138+ | 0.1238+ | 0.1299+ | 0.1130+ | 0.1234+ | 0.1296+ |
| F8 Sphere D10 | 0.4897 | 0.5232+ | 0.5084+ | 0.5049+ | 0.5387+ | 0.5132+ | 0.5042+ | 0.5400+ | 0.5129+ | 0.5033+ |
| | 11.2672 | 9.2995+ | 9.8468+ | 10.1565+ | 8.6977+ | 9.4734+ | 9.9291+ | 8.6402+ | 9.4487+ | 9.9126+ |
| | 0.0821 | 0.0707+ | 0.0750+ | 0.0784+ | 0.0656+ | 0.0719+ | 0.0759+ | 0.0652+ | 0.0716+ | 0.0757+ |
| F8 Sphere D20 | 0.5141 | 0.5494+ | 0.5374+ | 0.5192≈ | 0.5713+ | 0.5448+ | 0.5214≈ | 0.5731+ | 0.5457+ | 0.5211≈ |
| | 6.1717 | 5.3731+ | 5.7083+ | 5.9696+ | 5.0109+ | 5.4973+ | 5.8045+ | 4.9834+ | 5.4759+ | 5.7889+ |
| | 0.0823 | 0.0702+ | 0.0742+ | 0.0781+ | 0.0651+ | 0.0704+ | 0.0755+ | 0.0646+ | 0.0700+ | 0.0753+ |
| F8 Sphere D30 | 0.4407 | 0.4815+ | 0.4672+ | 0.4533+ | 0.5059+ | 0.4793+ | 0.4542+ | 0.5078+ | 0.4806+ | 0.4540+ |
| | 6.1793 | 5.3312+ | 5.6354+ | 5.9424+ | 4.9616+ | 5.3750+ | 5.7714+ | 4.9307+ | 5.3481+ | 5.7531+ |
| | 0.0835 | 0.0740+ | 0.0780+ | 0.0798+ | 0.0705+ | 0.0756+ | 0.0785+ | 0.0701+ | 0.0754+ | 0.0784+ |
| F9 Griewank D10 | 0.7542 | 0.7745+ | 0.7646+ | 0.7565≈ | 0.7846+ | 0.7702+ | 0.7579≈ | 0.7859+ | 0.7709+ | 0.7574≈ |
| | 6.0516 | 5.4247+ | 5.7109+ | 5.8537+ | 5.1837+ | 5.5571+ | 5.7737+ | 5.1555+ | 5.5429+ | 5.7697+ |
| | 0.1508 | 0.1194+ | 0.1272+ | 0.1341+ | 0.1093+ | 0.1201+ | 0.1289+ | 0.1083+ | 0.1195+ | 0.1285+ |
| F9 Griewank D20 | 0.5552 | 0.5959+ | 0.5862+ | 0.5683+ | 0.6223+ | 0.6012+ | 0.5734+ | 0.6247+ | 0.6021+ | 0.5738+ |
| | 10.4458 | 8.3303+ | 8.8918+ | 9.3825+ | 7.6597+ | 8.4341+ | 9.0529+ | 7.5923+ | 8.3933+ | 9.0250+ |
| | 0.0815 | 0.0711+ | 0.0742+ | 0.0779+ | 0.0660+ | 0.0705+ | 0.0754+ | 0.0655+ | 0.0702+ | 0.0752+ |
| F9 Griewank D30 | 0.7024 | 0.7338+ | 0.7257+ | 0.7101+ | 0.7560+ | 0.7405+ | 0.7161+ | 0.7582+ | 0.7417+ | 0.7166+ |
| | 5.5003 | 4.8538+ | 5.0680+ | 5.3241+ | 4.5244+ | 4.8376+ | 5.1752+ | 4.4954+ | 4.8229+ | 5.1595+ |
| + | | (27,27,27) | (27,27,27) | (27,25,27) | (27,27,27) | (27,27,27) | (27,25,27) | (27,27,27) | (27,27,27) | (27,24,27) |
| − | | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) | (0,0,0) |
| ≈ | | (0,0,0) | (0,0,0) | (0,2,0) | (0,0,0) | (0,0,0) | (0,2,0) | (0,0,0) | (0,0,0) | (0,3,0) |

### B.5 Additional 20-Dimensional BO Results

This section reports supplementary BO results for $D = 20$, following the same acquisition functions and reporting procedures as the 10-dimensional experiments in the main text. Each run is initialized with 200 LHS samples. Fig. 6 presents the Objective Improvement results on nine 20-dimensional single-objective benchmark functions, while Fig. 7 reports the corresponding Hypervolume Improvement results on nine 20-dimensional multi-objective benchmark problems. The results further demonstrate that RoSF-BO maintains consistent performance improvements in higher-dimensional settings.

### B.6 Human-Powered Aircraft Design Experiment

To further examine the applicability of RoSF-BO to an engineering-oriented optimization problem, we conducted additional experiments on the human-powered aircraft (HPA) aerostructural design benchmark Namura (2025). The benchmark describes the aerodynamic and structural design of a human-powered aircraft wing using governing equations from aerodynamics and material mechanics. We used the unconstrained penalty-based versions of six Level-0 problems, HPA101–HPA601, with the number of wing segments fixed to four. All design variables were normalized to $[0, 1]^D$.

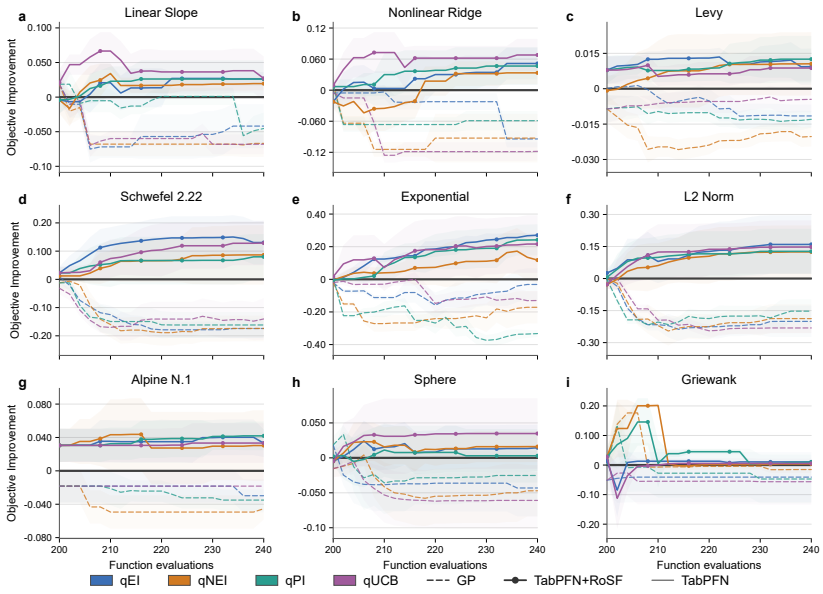


Figure 6: Single-Objective BO Results on 20-Dimensional Benchmark Problems.

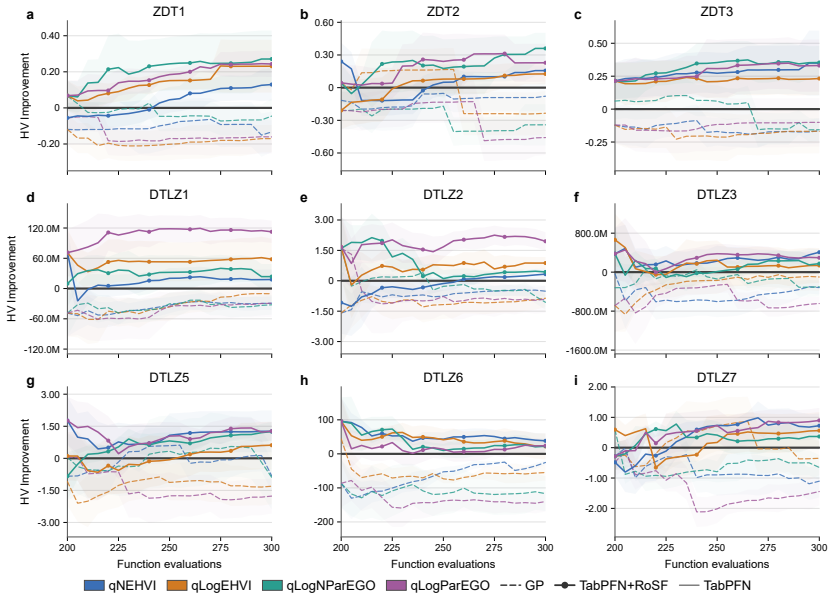


Figure 7: Multi-Objective BO Results on 20-Dimensional Benchmark Problems.

The six problems cover optimization tasks with one to six objectives. HPA101 is a 17-dimensional single-objective problem, whereas HPA201–HPA601 are 18-dimensional multi-objective problems. Their objectives involve aerodynamic power or drag, cruise speed, wing twist, aerodynamic efficiency, structural weight, and wingspan. The basic experimental configurations are summarized in Table 3. As shown in Fig. 8, TabPFN+RoSF-BO produced a positive final mean improvement in 24 of the 24 problem–acquisition combinations. These results further demonstrate that RoSF-BO remains effective in practical engineering optimization problems.

Table 3: Experimental configurations for the HPA aerostructural design problems. Here, $D$ and $m$ denote the numbers of design variables and objectives, respectively; $n_0$ is the initial design size; $q$ is the batch size; and $B$ is the total number of expensive evaluations.

| Problem | $D$ | $m$ | $n_0$ | $q$ | $B$ | Performance measure |
|---|---|---|---|---|---|---|
| HPA101 | 17 | 1 | 170 | 1 | 220 | Best objective |
| HPA201 | 18 | 2 | 180 | 5 | 280 | Hypervolume |
| HPA301 | 18 | 3 | 180 | 5 | 280 | Hypervolume |
| HPA401 | 18 | 4 | 180 | 5 | 280 | Hypervolume |
| HPA501 | 18 | 5 | 180 | 5 | 280 | Hypervolume |
| HPA601 | 18 | 6 | 180 | 5 | 280 | Hypervolume |

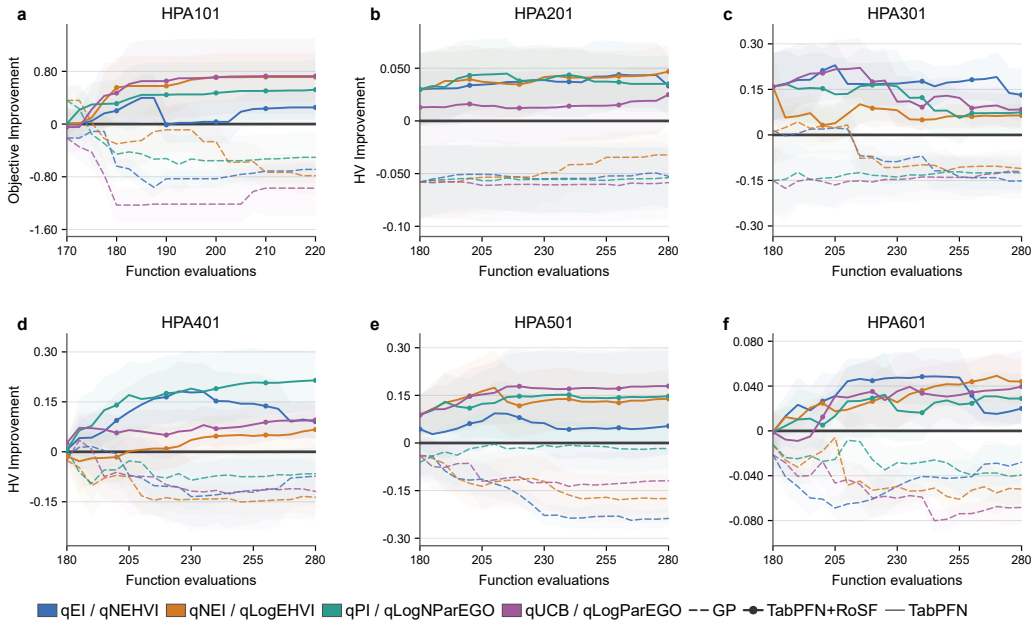


Figure 8: Optimization improvements on six Level-0 HPA aerostructural design problems.

### B.7 Ablation on Pairing and Data-Processing Strategies

To further investigate whether the improvement of RoSF-BO can be explained by interpolation alone or by generic pairing strategies, we compare it with several alternative data-processing strategies on the static noisy datasets. For a fair comparison, we first record the total number of observation pairs and augmented samples generated by RoSF-BO for each dataset. Random Interpolation randomly constructs the same number of observation pairs, while Nearest-Neighbor Interpolation computes pairwise distances among all observations and selects the same number of pairs with the smallest distances. For both interpolation-based baselines, the total number of generated samples is further controlled to be identical to that of RoSF-BO, thereby eliminating the influence of different augmentation sizes. We additionally include an Outlier Removal strategy, where observations are identified as outliers using the interquartile range (IQR) criterion and removed if their objective values fall outside $[Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}]$. Following the evaluation protocol in **Section 4.2**, the original noisy dataset is used as the reference baseline, and the relative performance of each data-processing strategy is calculated based on its improvement over the corresponding original-data metric. The comparison results are shown in Fig. 9. RoSF-BO achieves the best overall performance across the evaluated metrics, indicating that its improvement cannot be explained solely by interpolation-based augmentation or generic pairing strategies. Instead, explicitly constructing observation pairs with complementary residual behavior plays an important role in generating higher-quality augmented data.

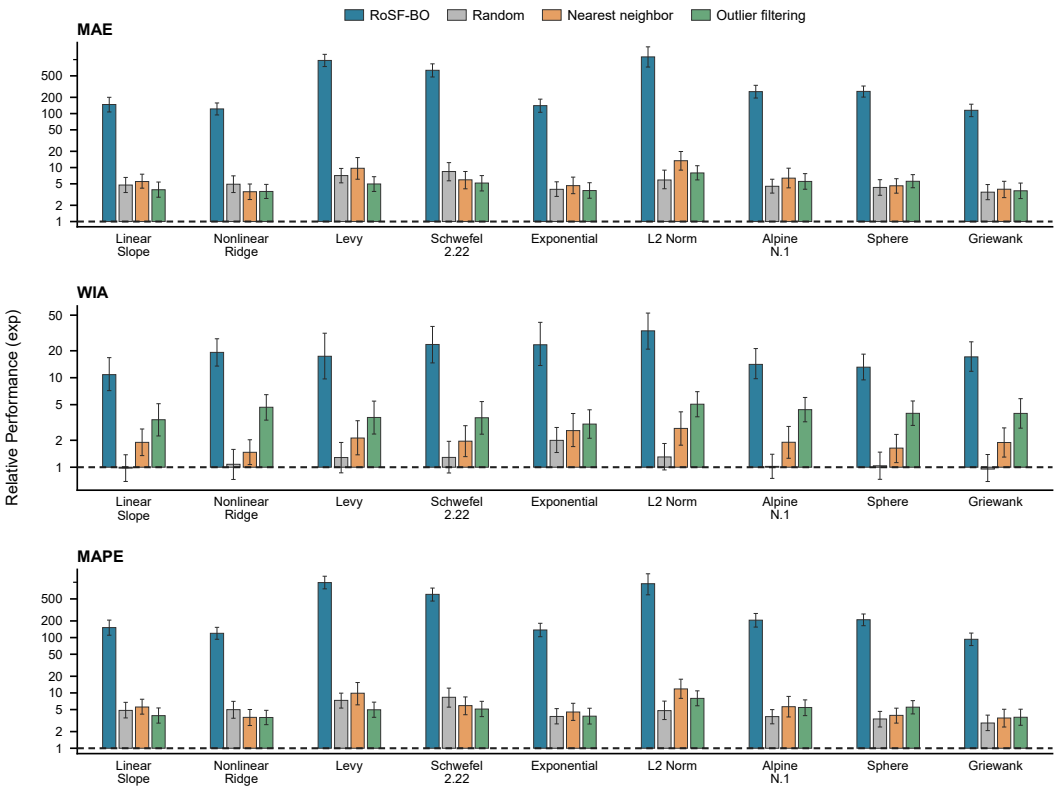


Figure 9: Relative Performance of Different Data-Processing Strategies on Static Noisy Datasets.

### B.8 Comparison of Data-Processing and Surrogate Strategies for Noisy Bayesian Optimization

To further evaluate the effectiveness of RoSF-BO in practical Bayesian optimization, we conduct additional comparisons on the six HPA benchmark problems. We compare RoSF-BO with Random Interpolation, Nearest-Neighbor Interpolation, and Outlier Removal, all using TabPFN as the surrogate model, and additionally include a noise-aware GP surrogate Binois et al. (2018b) as a model-level baseline. Multiple acquisition functions are considered, and the final optimization performance of each method is reported relative to the corresponding TabPFN-based BO baseline (see Fig 10–15). The results show that RoSF-BO achieves positive improvement across all six HPA problems, while the other compared methods generally fail to provide consistent benefits and often lead to degraded optimization performance. These results indicate that improving Bayesian optimization under noisy observations requires more than generic data modification or noise-aware modeling. In contrast, RoSF-BO consistently benefits the downstream optimization process through residual-cancellation-based observation augmentation. This comparison further suggests that directly processing noisy observations before surrogate fitting provides a promising and complementary direction for improving Bayesian optimization under observation noise.

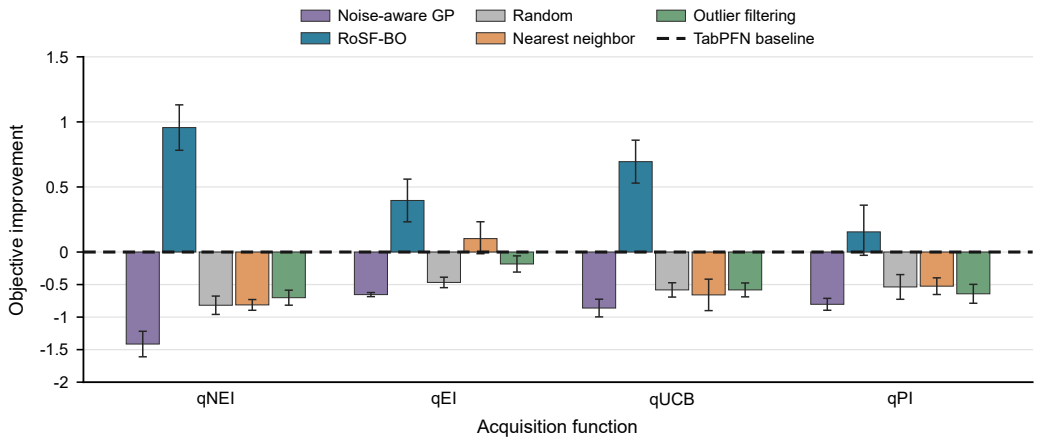


Figure 10: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA101.

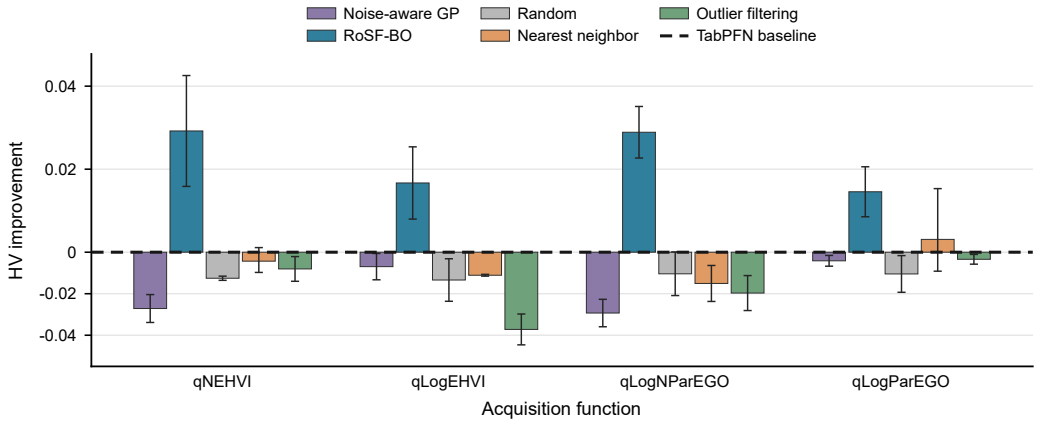


Figure 11: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA201.

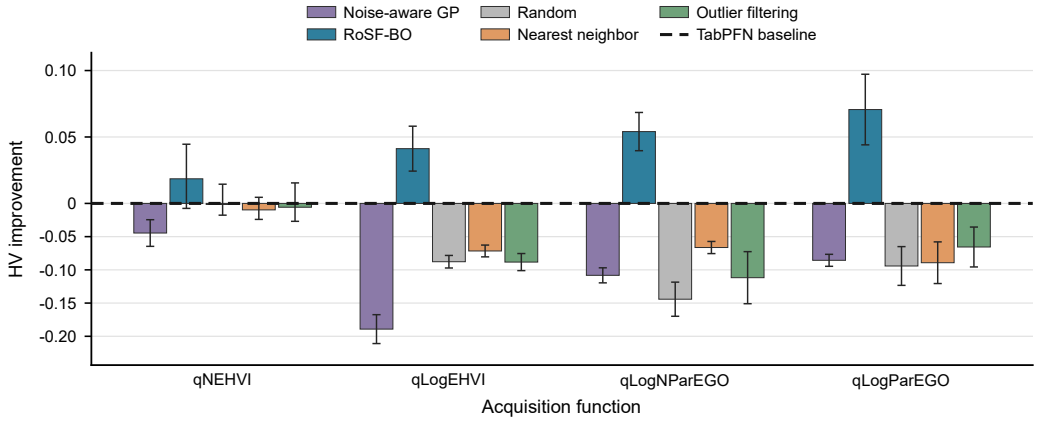


Figure 12: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA301.

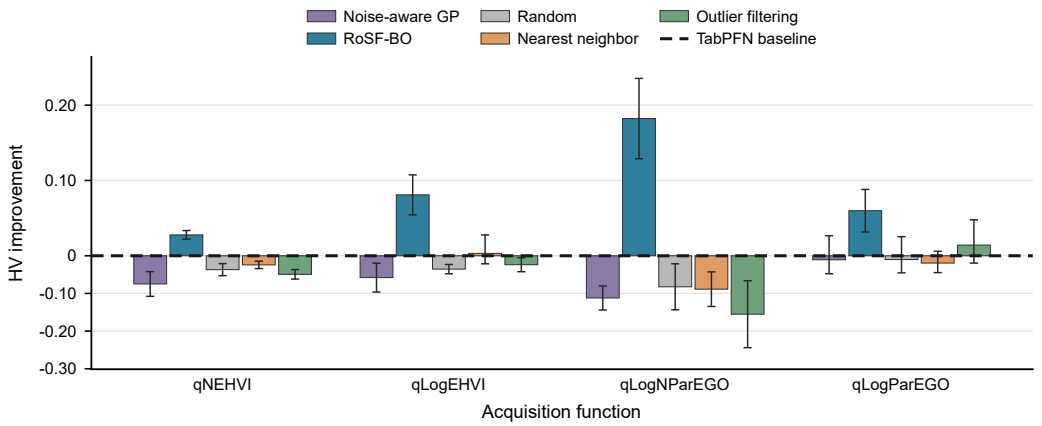


Figure 13: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA401.

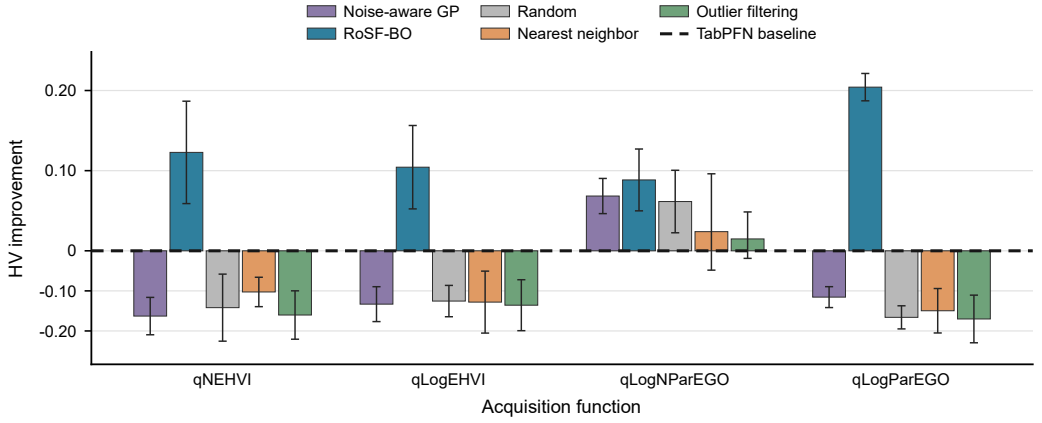


Figure 14: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA501.

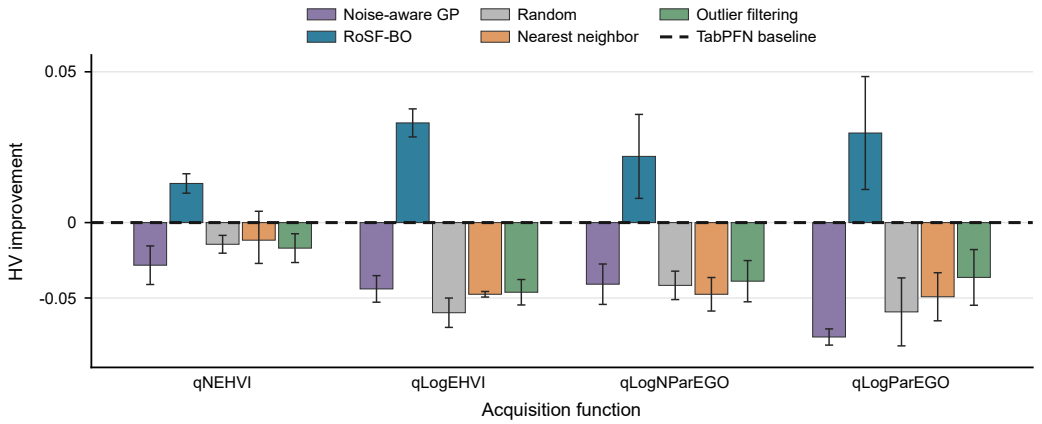


Figure 15: Optimization Performance of Different Noise-Handling and Data-Processing Strategies on HPA601.