HOLOCOSM: BEYOND SINGLE-METRIC BENCHMARK COMPRESSION ACROSS MULTIPLE METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark compression reduces evaluation cost by selecting a small set of samples. Existing methods typically guide selection using a single metric, although benchmarks often assess models along multiple dimensions. We find that model rankings are often weakly correlated across metrics, indicating that different metrics capture distinct aspects of model performance. Samples selected for one metric do not reliably preserve scores and rankings on others when used for compressed evaluation. At 80% compression, each of the five single-metric baselines yields higher reconstruction error than equal-size random sampling in over 70% of cross-metric comparisons. To address this challenge, we present HoloCosm, which jointly minimizes normalized score-reconstruction error across source models and metrics while constraining ranking distortion. We evaluate HoloCosm across multiple compression rates on 15 language, multimodal, VLA, and agent benchmark configurations with 85 component metrics. With 80% compression, HoloCosm outperforms Random, the strongest baseline overall: average NMAE drops by 35.3% to 0.236, worst-component NMAE drops by 39.5% to 0.330, and average Spearman correlation rises from 0.884 to 0.915. In addition, HoloCosm remains competitive with existing baselines in single-metric evaluation. We will release our code and experimental results.

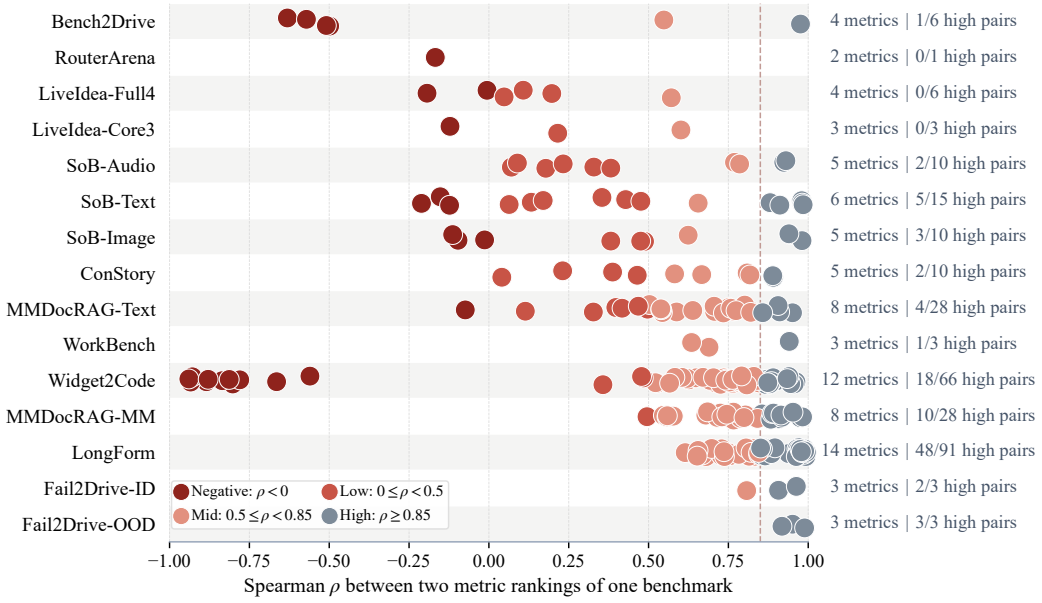

Figure 1: **Model-ranking consistency across metrics.** Each point shows Spearman ρ between rankings induced by two official metrics, all direction-aligned so higher is better. Colors show correlation ranges, the dashed line marks $\rho \geq 0.85$, and right-side annotations give metric counts and high-correlation pairs. Weak and negative correlations are common.

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.

1 INTRODUCTION

As the number of models and the scale of evaluation continue to grow, repeatedly evaluating models on full benchmarks incurs increasing computational and time costs. Benchmark compression reduces this burden by selecting a small set of representative samples to approximate full-benchmark model scores and rankings. Existing benchmark compression methods (**BCMs**), such as AnchorPoints, tinyBenchmarks, mRMR-based compression, SubLIME, and ZipBench, aim to substantially reduce benchmark size while maintaining high evaluation fidelity through representative sample selection for downstream model comparison (Vivek et al., 2024; Maia Polo et al., 2024; Bowyer et al., 2026; Saranathan et al., 2025; Huang et al., 2026). However, these methods typically guide sample selection using a single evaluation metric, whereas modern benchmarks often assess models along multiple dimensions. This exposes a key issue: samples considered representative under one metric may not be equally representative for other metrics.

We therefore examine the consistency of model rankings across evaluation metrics. Across 85 official metrics from 15 benchmark configurations spanning language, multimodal, VLA, and agent tasks, we compute Spearman correlations between metric-induced model rankings within each benchmark. As shown in Fig. 1, many metric pairs exhibit weak ranking agreement, with negative correlations appearing across several benchmarks. Using $\rho \geq 0.85$ as the threshold for high correlation (Granato, 2014), we find that highly correlated pairs are only a small fraction in many benchmarks: Bench2Drive has only one out of six such pairs, while LiveIdea-Full4 (Ruan et al., 2026) has none. These results show that different metrics capture distinct, sometimes conflicting, model-comparison information, raising a concern for compression guided by a single metric.

These cross-metric differences raise a question for single-metric compression: can samples selected to represent one metric also preserve the evaluation information captured by others? To examine this question, we select items using one metric and evaluate them on the remaining metrics of the same benchmark, comparing against equal-size Random sampling at 80% compression. As shown in the left panel of Fig. 2, existing single-metric BCMs perform worse than Random in 70.5%–97.7% of cross-metric score-reconstruction comparisons across five baselines. The right panel further shows that even when the selected samples achieve lower NMAE than Random on the selection metric, this advantage often fails to carry over to other metrics, where reconstruction errors instead exceed those of Random. These results show that a subset representative of one metric does not necessarily preserve the information needed to evaluate others.

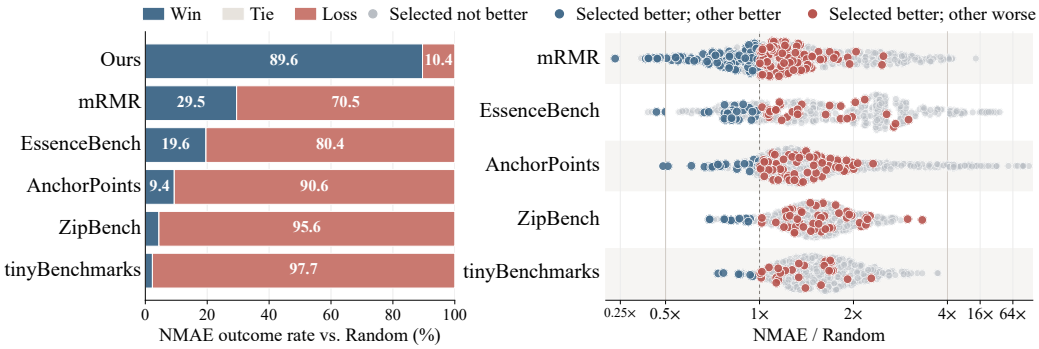

Figure 2: **Single-metric selection fails across metrics. Left:** Cross-metric NMAE win, tie, and loss rates against Random across 15 benchmarks at 80% compression. HoloCosm wins in 89.6% of comparisons. **Right:** NMAE relative to Random on metrics other than the selection metric, with compressed axis tails. Red points show cross-metric failure despite selected-metric gains.

To address this systematic cross-metric failure, we present HoloCosm, which directly optimizes evaluation fidelity across multiple metrics rather than assuming that items selected for one metric will transfer to others. Using full-benchmark evaluation results from source models, HoloCosm minimizes normalized score-reconstruction error across models and metrics while constraining ranking distortion. Size-matched reference selections calibrate the ranking-distortion constraint, a feasibility-first exchange search refines the selected items, and a source-assisted correction esti-

mates full-benchmark scores while preserving within-metric rankings. At 80% compression (20% retention), HoloCosm achieves the best aggregate performance across 15 benchmark configurations and 85 evaluation metrics. Despite being designed for multi-metric evaluation, HoloCosm remains competitive with existing single-metric BCMs in single-metric settings. Our contributions are:

- **Multi-metric inconsistency.** Across 85 official metrics from 15 benchmark configurations, we find widespread weak and negative correlations between metric-induced model rankings, showing that different evaluation dimensions often capture substantially different information across different types of benchmarks.
- **Systematic cross-metric failure.** At 80% compression, existing single-metric BCMs perform worse than Random in 70.5%–97.7% of cross-metric score-reconstruction comparisons, revealing a systematic failure of single-metric-guided compression. To our knowledge, we are the first to systematically study this multi-metric compression problem.
- **Multi-metric benchmark compression.** We introduce HoloCosm to preserve score and ranking fidelity across metrics. It achieves the best overall fidelity across 15 benchmark configurations and 85 metrics at 80% compression while remaining competitive in single-metric evaluation under the same compression setting.

2 Related Work

Benchmark compression. Benchmark compression reduces evaluation cost by estimating full-benchmark performance from fewer items. AnchorPoints, mRMR, and ZipBench select representative items using response correlation, mutual-information-based feature selection, and synthesized-response representations, respectively (Vivek et al., 2024; Bowyer et al., 2026; Huang et al., 2026). tinyBenchmarks and bias-bounded item selection focus on accurate score estimation through IRT-based selection and bias-bounded submodular selection (Maia Polo et al., 2024; Zhuang et al., 2025). More specialized benchmark compression includes MiniLongBench, which learns sample representations and clusters them to prune redundant long-context samples while preserving model rankings (Huang et al., 2025). SubLIME uses predicted rank correlation to guide subset selection, while EssenceBench combines redundancy-aware filtering with fitness-based subset search (Saranathan et al., 2025; Wang et al., 2026). Despite these differences, existing benchmark-compression methods primarily select items for one evaluation metric at a time. Tab. 1 summarizes this distinction. Our work instead uses multiple evaluation metrics for item selection within the same benchmark while preserving both scores and model rankings.

Table 1: **Benchmark-compression methods by selection basis, objective, and metric setting.**

Method	Selection basis	Objective	Metric setting
AnchorPoints	Response correlation	Coverage	Single
mRMR	Mutual information	Score reconstruction	Single
tinyBenchmarks	IRT embeddings	Score estimation	Single
SubLIME	Predicted rank correlation	Ranking preservation	Single
EssenceBench	Redundancy + subset fitness	Score & rank fidelity	Single
MiniLongBench	Learned sample representations	Ranking preservation	Single
ZipBench	Synthesized-response representations	Score & rank fidelity	Single
HoloCosm	Cross-metric fidelity	Score & rank fidelity	**Single / Multi**

Efficient performance estimation. Related work considers multiple benchmarks or exploits cross-metric structure for more efficient performance estimation (Kipnis et al., 2025; Liao et al., 2025). These settings do not directly study multi-metric item selection across multiple evaluation metrics within the same benchmark, which is the focus of our work.

Evaluation reliability. Prior work studies the reliability of reduced evaluations and resulting benchmark rankings (Perlitz et al., 2024; Yauney et al., 2026; Neuhof & Benjamini, 2026). These studies mainly consider reliability under a fixed evaluation criterion. Our work instead asks whether items selected to represent one metric remain reliable for other metrics, evaluating both score reconstruction and ranking preservation consistently across evaluation metrics.

3 METHODOLOGY

HoloCosm formulates multi-metric benchmark compression as fidelity-constrained reconstruction. Each candidate set is characterized by reconstruction accuracy and pairwise-margin distortion across source models and metrics. Size-matched reference sets calibrate a ranking-risk threshold. A feasibility-first exchange search seeks B items with low multi-metric reconstruction error under this constraint, after which source-assisted estimation estimates full-benchmark target scores.

Preliminaries. Let $\mathcal{M} = [M]$ denote the source-model index set, with source models $\{f_m\}_{m\in\mathcal{M}}$, and let $\mathcal{D} = \{x_1, \ldots, x_N\}$ denote the benchmark items. Item-level scores for all source models and items under K evaluation metrics form a tensor $X \in \mathbb{R}^{M\times N\times K}$, where X_{mnk} is the score of model f_m on item x_n under metric k. For metrics obtained by averaging item-level scores, the full-benchmark score and its uncorrected subset estimate are

$$\mu_{mk} = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_{mnk}, \qquad \widehat{\mu}_{mk}(S) = \frac{1}{|S|}\sum\nolimits_{n\in S} X_{mnk}, \tag{1}$$

where $n \in [N]$ indexes benchmark items and $S \subseteq [N]$ indexes the selected items. Collecting these scores gives the full evaluation matrix $\boldsymbol{\mu} \in \mathbb{R}^{M\times K}$ and its subset estimate $\widehat{\boldsymbol{\mu}}(S)$. Given a selection size $B < N$, our goal is to choose a single subset S with $|S| = B$ such that $\widehat{\mu}_{mk}(S) \approx \mu_{mk}$ across source models and metrics. The same selected items are used to estimate every metric. Item selection uses only source-model responses; held-out target models are evaluated after S is fixed.

3.1 MULTI-METRIC FIDELITY PROFILE

For every size-B candidate selection S, HoloCosm characterizes its fidelity by reconstruction risk $R_{\text{mean}}(S)$ and ranking risk $R_{\text{rank}}(S)$. The primary term R_{mean} averages normalized reconstruction errors across source models and metrics. Normalization by the dispersion of full-benchmark source scores makes errors across metric scales comparable, while retaining every metric in each candidate assessment. Concretely, for a source pool $\mathcal{A}$ of q models, let $s_k(\mathcal{A})$ be the standard deviation of the full-benchmark scores for metric k (set to one when numerically zero). We define

$$a_k(S;\mathcal{A}) = \frac{1}{q\, s_k(\mathcal{A})}\sum_{m\in\mathcal{A}} |\widehat{\mu}_{mk}(S) - \mu_{mk}|, \qquad r_{\text{mean}}(S;\mathcal{A}) = \frac{1}{K}\sum_{k=1}^{K} a_k(S;\mathcal{A}). \tag{2}$$

Reconstruction accuracy cannot guarantee faithful model comparisons. For each metric, R_{rank} measures reductions in signed pairwise margins. Let $d_{abk} = \mu_{ak} - \mu_{bk}$, $\widehat{d}_{abk}(S) = \widehat{\mu}_{ak}(S) - \widehat{\mu}_{bk}(S)$, and let $p_k(\mathcal{A})$ be the median nonzero absolute margin for metric k. Let $\mathcal{Q}(\mathcal{A})$ denote the pooled set of valid metric-indexed model pairs, given by $\{(k,a,b) : k \in [K],\ (a,b) \in \mathcal{P}_k(\mathcal{A})\}$. We define the aggregate ranking risk as the mean normalized margin shortfall over this set:

$$r_{\text{rank}}(S;\mathcal{A}) = |\mathcal{Q}(\mathcal{A})|^{-1}\sum\nolimits_{(k,a,b)\in\mathcal{Q}(\mathcal{A})} p_k(\mathcal{A})^{-1}\left[|d_{abk}| - \text{sign}(d_{abk})\widehat{d}_{abk}(S)\right]_+. \tag{3}$$

This loss penalizes both weakened separation and reversed orderings. The final risks average these quantities over leave-one-source-out pools, $R_j(S) = M^{-1}\sum_{m=1}^{M} r_j(S;\mathcal{M}\setminus\{m\})$ for $j \in \{\text{mean}, \text{rank}\}$, with normalization recomputed in each pool. Appendix A.2 gives the zero-scale conventions, valid-pair construction, source-pool averaging, and full definitions.

3.2 REFERENCE-CALIBRATED CONSTRAINED SELECTION

HoloCosm minimizes multi-metric reconstruction error subject to a calibrated constraint on aggregate pairwise-ranking risk. Using an ϵ-constraint formulation (Mesquita-Cunha et al., 2023), it solves the following constrained subset-selection problem:

$$\min_{S\subseteq[N]} R_{\text{mean}}(S) \quad \text{subject to} \quad |S| = B, \quad R_{\text{rank}}(S) \le \epsilon_{\text{rank}}. \tag{4}$$

Here, reconstruction is the primary objective, while ranking risk is a feasibility constraint. We draw 128 uniform size-B subsets $\mathcal{R}$ and set $\epsilon_{\text{rank}} = Q_{0.25}(\{R_{\text{rank}}(S) : S \in \mathcal{R}\})$, the 25th percentile of their ranking risks. Thus, lower reconstruction error cannot compensate for risk above this threshold.

3.3 FEASIBILITY-FIRST EXCHANGE SEARCH

All candidate selections are ordered by one lexicographic rule. Among infeasible candidates, smaller normalized constraint violation takes priority. Every feasible candidate is preferred to every infeasible candidate; within the feasible region, lower R_{mean} determines the ordering.

Starting from the best calibrated candidate, HoloCosm repeatedly evaluates one-for-one exchanges of a selected and an unselected item. An exchange is accepted only when it improves the lexicographic criterion. Every update preserves $|S| = B$. The initial subset is feasible; accepted exchanges preserve feasibility and reduce reconstruction error.

3.4 SOURCE-ASSISTED SCORE ESTIMATION

For each target model, we correct its selected-item mean using source-average scores as auxiliary predictions, following prediction-powered mean estimation (Angelopoulos et al., 2023). The resulting full-benchmark score estimate is

$$\widetilde{\mu}_{gk}(S) = \widehat{\mu}_{gk}(S) + N^{-1}\sum\nolimits_{n=1}^{N} z_{nk} - B^{-1}\sum\nolimits_{n\in S} z_{nk}, \quad z_{nk} = \frac{1}{M}\sum_{m=1}^{M} X_{mnk}. \tag{5}$$

Here, $\widehat{\mu}_{gk}(S) = B^{-1}\sum_{n\in S} Y_{gnk}$ is the observed subset mean of target model g, where Y_{gnk} is its score on item n under metric k. The two sums over z_{nk} are the full-benchmark and selected-item means of the source-average scores, respectively. Their difference is shared by all target models for a given metric, so the correction preserves pairwise score differences and within-metric rankings. Ranking fidelity is determined by item selection, while the correction improves full-score estimation without requiring target responses outside S. Algorithm 1 summarizes the procedure.

Algorithm 1 HoloCosm multi-metric subset selection and score estimation

Require: Source scores X, target models $\mathcal{T}$, subset size B
Ensure: Selected subset S and corrected target score estimates
1: **Calibration:** Draw a bank $\mathcal{R}$ of 128 uniform size-B subsets.
2: Evaluate R_{mean} and R_{rank} for each $S \in \mathcal{R}$.
3: $\epsilon_{\text{rank}} \leftarrow Q_{0.25}(\{R_{\text{rank}}(S) : S \in \mathcal{R}\})$.
4: **Selection:** Initialize S with the best reference subset under the feasibility-first rule.
5: **for** at most 80 refinement rounds **do**
6: Sample up to 64 selected–unselected item swaps and evaluate both risks for each proposal.
7: Let S' be the best proposal under the feasibility-first rule; **break** if it does not improve S.
8: Accept the improving proposal and update the current subset: $S \leftarrow S'$.
9: **Estimation:** Evaluate each target model on S.
10: Apply the source-assisted correction in Eq. 5.

4 EXPERIMENTS

Benchmarks and splits. We evaluate on 15 benchmark configurations with 85 official component metrics, as summarized in Tab. 2. They span six LLM, five multimodal-LLM, three VLA, and one agent configuration, with 12–59 models, 78–8,400 items, and 2–14 metrics each. Most benchmark matrices are reconstructed from public item-level records; the remainder use official evaluation pipelines. For each configuration, we first compute each model's full-benchmark mean on every metric. We orient the metrics so that higher values indicate better performance, min–max normalize each metric across the candidate models, and take their equally weighted mean as each model's score. After ranking models by this score, we select six evenly spaced sources from lowest to highest, using the rest as targets. All methods use the same fixed source–target split.

Baselines. We compare HoloCosm with Random, tinyBenchmarks (Maia Polo et al., 2024), ZipBench (Huang et al., 2026), mRMR with Ridge reconstruction (Bowyer et al., 2026), AnchorPoints (Vivek et al., 2024), and EssenceBench (Wang et al., 2026). Random performs uniform random sampling without replacement. We adapt the baselines to the common six-source-model protocol; implementation details are provided in Appendix B.1.

Table 2: **Benchmark configurations used in the cross-metric transfer analysis.** The 15 configurations contain 85 component metrics across LLM, multimodal, VLA, and agent evaluations.

Type	Benchmark	Task	Models	Items	Metrics
LLM	ConStory (Li et al., 2026)	Story evaluation	33	1,179	5
	LiveIdea-Core3 (Ruan et al., 2026)	Ideation	41	869	3
	LiveIdea-Full4 (Ruan et al., 2026)	Ideation	41	235	4
	LongForm (Paech, 2025)	Writing	59	108	14
	MMDocRAG-Text (Dong et al., 2025)	Retrieval	51	1,955	8
	RouterArena (Lu et al., 2026)	Routing	12	8,400	2
MLLM	MMDocRAG-Multimodal (Dong et al., 2025)	Retrieval	31	1,974	8
	SoB-Text (Singh et al., 2026)	Structured output	35	5,000	6
	SoB-Image (Singh et al., 2026)	Structured output	35	209	5
	SoB-Audio (Singh et al., 2026)	Structured output	34	115	5
	Widget2Code (Zhang et al., 2026)	Code generation	13	288	12
VLA	Bench2Drive (Jia et al., 2024)	Driving	32	220	4
	Fail2Drive-ID (Gerstenecker et al., 2026)	Driving	33	82	3
	Fail2Drive-OOD (Gerstenecker et al., 2026)	Driving	33	78	3
Agent	WorkBench (Styles & Miller, 2026)	Tool use	24	690	3

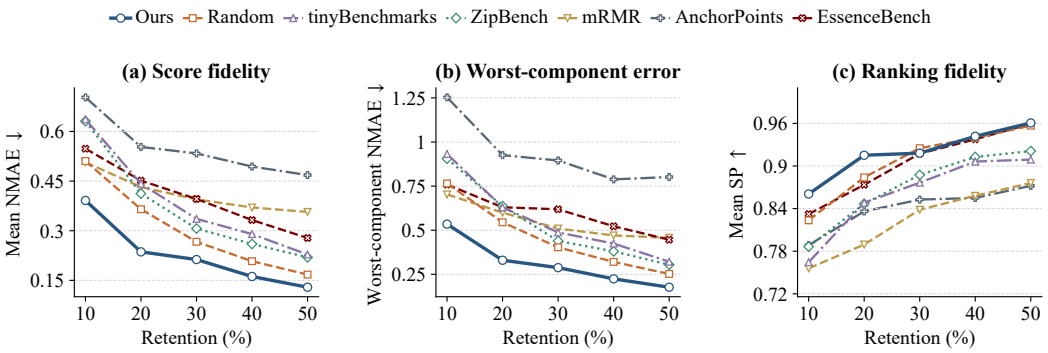

Figure 3: **Multi-metric fidelity across retention ratios.** Mean NMAE, worst-component NMAE, and mean SP are computed across evaluation metrics within each benchmark, averaged over runs, then equally across 15 benchmarks. For each single-target baseline, the target with lowest run-averaged mean NMAE is selected at each benchmark and retention level and used for all panels.

Comparison protocol. We test retention ratios of 10%, 20%, 30%, 40%, and 50%, with 20% as the main setting. For N items at retention ratio ρ, we set the selection size to the nearest integer to ρN, with at least one item. At each ratio, every method selects the same number of unique items. Each single-target baseline selects items using each component metric in turn and evaluates the resulting selection on all components. HoloCosm selects items across all components, whereas Random uniformly samples items and evaluates all components on those items.

Evaluation metrics. We measure score reconstruction using normalized mean absolute error (NMAE) and ranking preservation using Spearman rank correlation (SP) across target models. NMAE normalizes reconstruction error by target-score dispersion, making errors comparable across metrics with different scales. For component k of benchmark b, we define NMAE as

$$\mathrm{NMAE}_{b,k} = \frac{|\mathcal{T}_b|^{-1} \sum_{t \in \mathcal{T}_b} |\widehat{y}_{t,k} - y_{t,k}|}{\sigma_{b,k}}, \qquad \sigma_{b,k} = \mathrm{Std}_{t \in \mathcal{T}_b}(y_{t,k}). \tag{6}$$

Here, $\mathcal{T}_b$ denotes the target models, and $y_{t,k}$ and $\widehat{y}_{t,k}$ denote the full-benchmark and estimated scores, respectively. $\sigma_{b,k}$ is the population standard deviation across target models. SP measures agreement between rankings induced by the full and estimated scores. For each run, we report mean and worst-component values, using maximum NMAE or minimum SP. Results are averaged over runs and macro-averaged across benchmarks. Lower NMAE and higher SP indicate better fidelity.

Table 3: **Multi-metric score and ranking fidelity at 20% retention.** Entries average over K metrics and runs with benchmarks weighted equally; each baseline's best-NMAE target supplies the measures. W/T/L denotes wins, ties, and losses against Random. Bold indicates best results.

		All-metric / random selection				Single-metric selection (best across metrics)									
		Ours		Random		tinyBenchmarks		ZipBench		mRMR		AnchorPoints		EssenceBench	
Benchmark	K	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑
ConStory	5	0.30	0.88	**0.29**	0.88	0.43	0.84	0.45	0.81	0.67	0.69	1.24	0.52	0.44	**0.88**
Bench2Drive	4	**0.27**	0.84	0.39	0.84	0.46	0.78	0.53	0.76	0.37	0.83	0.59	0.79	0.34	**0.85**
Fail2Drive-ID	3	**0.28**	**0.80**	0.58	0.59	0.74	0.79	0.51	0.57	0.37	0.49	0.28	0.71	1.00	0.51
Fail2Drive-OOD	3	**0.39**	**0.86**	0.79	0.79	0.68	0.69	0.73	0.65	0.50	0.51	0.61	0.80	0.58	0.76
LiveIdea-Core3	3	0.10	0.98	0.11	0.98	0.16	0.96	0.16	0.96	0.33	0.97	0.12	0.98	**0.10**	**0.98**
LiveIdea-Full4	4	**0.22**	**0.95**	0.29	0.94	0.39	0.87	0.40	0.88	0.42	0.89	0.27	0.94	0.25	0.94
RouterArena	2	**0.09**	0.97	0.15	**0.99**	0.28	0.91	0.25	0.96	1.02	0.86	0.28	0.97	0.42	0.93
LongForm	14	**0.12**	**0.98**	0.14	0.98	0.17	0.97	0.17	0.97	0.13	0.98	0.15	0.98	0.32	0.96
SoB-Audio	5	0.58	**0.83**	0.70	0.78	0.99	0.65	0.78	0.75	**0.56**	0.75	0.72	0.80	0.58	0.74
SoB-Image	5	**0.51**	**0.77**	1.17	0.65	1.03	0.62	0.92	0.71	0.56	0.66	0.87	0.71	1.41	0.70
SoB-Text	6	**0.16**	**0.96**	0.21	0.96	0.39	0.86	0.39	0.87	0.51	0.88	1.80	0.67	0.46	0.96
MMDocRAG-MM	8	**0.06**	**0.99**	0.07	0.99	0.14	0.98	0.11	0.98	0.10	0.99	0.09	0.99	0.08	0.99
MMDocRAG-Text	8	**0.10**	0.98	0.11	0.98	0.14	0.97	0.17	0.96	0.23	0.97	0.13	0.97	0.11	**0.98**
Widget2Code	12	**0.22**	**0.94**	0.30	0.93	0.42	0.88	0.38	0.90	0.24	0.92	0.33	0.86	0.30	0.94
WorkBench	3	**0.14**	**0.98**	0.17	0.97	0.22	0.95	0.22	0.97	0.48	0.46	0.83	0.84	0.38	0.96
Macro-average	–	**0.24**	**0.92**	0.37	0.88	0.44	0.85	0.41	0.85	0.43	0.79	0.55	0.84	0.45	0.87
W/T/L vs. Random	–	**14/0/1**	**11/0/4**	–	–	2/0/13	1/0/14	3/0/12	1/0/14	7/0/8	1/0/14	4/0/11	5/0/10	5/0/10	8/0/7

Table 4: **Cross-metric fidelity on Bench2Drive at 20% retention.** Underlining marks selection-metric results; Mean averages NMAE, and blue-gray marks each baseline's lowest-mean target.

Method	**Selection metric**	**NMAE ↓**					**SP ↑**			
		Efficiency	**Score**	**Smoothness**	**Success**	**Mean**	**Efficiency**	**Score**	**Smoothness**	**Success**
Ours	All metrics	0.053	0.435	0.159	0.448	0.274	0.782	0.845	0.911	0.835
ZipBench	Efficiency	<u>0.111</u>	0.872	0.343	0.818	0.536	<u>0.758</u>	0.748	0.862	0.787
	Score	0.100	<u>0.831</u>	0.338	0.860	0.532	0.668	<u>0.750</u>	0.869	0.747
	Smoothness	0.174	0.944	<u>0.299</u>	0.973	0.598	0.847	0.783	<u>0.845</u>	0.825
	Success	0.229	0.807	0.315	<u>0.821</u>	0.543	0.689	0.815	0.872	<u>0.726</u>

4.1 Main results

For each single-metric baseline, we report the selection target with the lowest mean NMAE for each benchmark and retention level, and use the same target for all measures.

Results at 20% retention. Tab. 3 reports the main results at 20% retention. HoloCosm achieves the best aggregate fidelity, while Random is the strongest baseline overall. Compared with Random, HoloCosm reduces mean and worst-component NMAE by 35.3% and 39.5%, respectively, while also improving ranking fidelity. In contrast, all single-metric baselines underperform Random in macro-average NMAE and SP, with NMAE ranging from 0.41 to 0.55 and SP from 0.79 to 0.87. This advantage is also broadly consistent across benchmarks: HoloCosm outperforms Random in mean NMAE on 14 of 15 benchmarks and in mean SP on 11 of 15.

Across retention levels. To assess performance beyond the 20% setting, Fig. 3 compares HoloCosm with all baselines from 10% to 50% retention. HoloCosm achieves the lowest mean and worst-component NMAE at every level, while maintaining consistently competitive ranking fidelity throughout the full tested retention range. Both error measures decrease steadily and mean SP rises as more items are retained. The margin is clearest under aggressive compression and persists at higher retention, showing that the gains are not tied to a single operating point.

Cross-metric transfer. Tab. 3 reports each single-metric baseline's results using the selection metric with the lowest mean NMAE. Tab. 4 details ZipBench on Bench2Drive. ZipBench achieves its lowest mean NMAE with Score as the compression target. Switching to Smoothness lowers Smoothness error but increases Efficiency error. HoloCosm jointly lowers all four component errors, avoiding this target-dependent trade-off. Complete results are provided in Appendix D.3.

4.2 Ablation Study

Table 5: **Selection ablation at 20% retention.** All variants use the same source-assisted estimator. Results are three-seed means with benchmarks weighted equally.

Variant	Selection rule	Mean NMAE ↓	Worst NMAE ↓	Mean SP ↑
Random + correction	Uniform sampling	0.2812	0.3828	0.8840
Multi-metric reconstruction	Minimize R_{mean}	0.2404	0.3390	0.9077
HoloCosm	Minimize R_{mean} s.t. $R_{\text{rank}} \leq \epsilon_{\text{rank}}$	**0.2362**	**0.3299**	**0.9151**

As shown in Tab. 5, we isolate the contribution of each selection component at 20% retention across 15 benchmarks, while keeping the source-assisted estimator unchanged. We compare uniform random sampling with source-assisted correction, multi-metric reconstruction without explicit ranking protection, and the full HoloCosm selection with ranking-aware constraints.

Multi-metric reconstruction improves score and ranking fidelity over random sampling with correction, reducing mean NMAE from 0.2812 to 0.2404 and raising mean SP from 0.884 to 0.9077. Adding ranking protection yields gains, reaching 0.2362 mean NMAE, 0.3299 worst-component NMAE, and 0.9151 mean SP. This shows that ranking-aware selection improves not only ordering but also reconstruction fidelity, including the hardest-to-preserve components. Since source-assisted correction preserves model ordering, the SP gains reflect the selection.

4.3 Additional Analyses

(1) Can Simple Multi-Metric Strategies Preserve Benchmark Scores and Rankings?

Table 6: Simple multi-metric compression strategies at 20% retention. Results are averaged equally over 15 configurations; mean and worst-metric errors are computed per configuration.

Method	Strategy	Mean NMAE ↓	Worst NMAE ↓	Mean SP ↑
Random	—	0.3650	0.5450	0.8840
AnchorPoints	Single metric	0.5533	0.9256	0.8361
	AverageMetric	0.6106	1.1302	0.8497
	Vote	0.6288	0.9719	0.9038
K-means	Multi-metric responses	0.3242	0.4288	0.8842
Ours	AverageMetric	0.2822	0.3896	0.8946
	All metrics	**0.2362**	**0.3299**	**0.9151**

To test whether simple multi-metric strategies can preserve benchmark scores and rankings, we compare three strategies on the original metrics at 20% retention.

AverageMetric: We direction-align all metrics, normalize each using source-derived min–max statistics, and average them equally into a scalar selection target. For AnchorPoints and Ours, this scalar serves as the sole selection target, while their original score estimators remain unchanged.

Vote: We run AnchorPoints for each metric to select B items, giving candidates one vote per metric. We select by vote count; boundary ties favor the least-covered metrics, followed by tie-breaking. This yields B distinct items, whose unweighted means estimate target metrics.

K-means: We concatenate each item's source responses after scaling metrics by the standard deviation of source full-benchmark means. Lloyd K-means forms B clusters and retains the item nearest each center. Target scores are cluster-size-weighted averages of the representatives.

Tab. 6 shows uneven gains. Vote improves ranking correlation but increases score error, and random selection remains a competitive baseline. Ours achieves lower error and higher ranking agreement. Compared with K-means, all-metric selection reduces mean and worst NMAE by 27.1% and 23.1%, respectively, while increasing mean SP from 0.8842 to 0.9151. Thus, directly optimizing metric-specific response structure is more effective than generic response-space clustering.

(2) Can HoloCosm Match Strong Baselines on Single-Metric Benchmarks?

Table 7: **Single-metric evaluation at 20% retention.** MAE measures score reconstruction and SP measures ranking preservation. Results are averaged over runs; Macro-average weights the three benchmarks equally. Bold marks the best result in each column.

	MMStar		SimpleQA		SimpleVQA		Macro-average	
Method	MAE ↓	SP ↑	MAE ↓	SP ↑	MAE ↓	SP ↑	MAE ↓	SP ↑
Ours	0.017	0.964	0.009	**1.000**	**0.013**	**1.000**	**0.013**	0.988
Random	**0.014**	**0.988**	**0.009**	**1.000**	0.022	0.992	0.015	**0.993**
tinyBenchmarks	0.032	0.952	0.022	0.988	0.018	0.984	0.024	0.975
ZipBench	0.018	0.964	0.018	0.988	0.031	0.937	0.023	0.963
mRMR	0.067	0.893	0.030	**1.000**	0.022	0.976	0.040	0.956
AnchorPoints	0.125	0.810	0.119	0.881	0.156	0.262	0.134	0.651

We further examine whether joint multi-metric optimization compromises performance when only one metric is evaluated on three single-metric benchmarks: MMStar, SimpleQA, and SimpleVQA (Chen et al., 2024; Wei et al., 2024; Cheng et al., 2025), spanning diverse evaluation settings.

As shown in Tab. 7, at 20% retention HoloCosm achieves the lowest macro-average MAE of 0.013 and a macro-average SP of 0.988, close to Random's 0.993, while outperforming all existing compression baselines. These results show that HoloCosm's effectiveness is not limited to multi-metric evaluation and remains competitive when only a single metric is available.

(3) Can Weaker Source Models Guide Benchmark Compression for Stronger Models?

To evaluate transfer beyond the source pool's capability, we construct a weak-to-strong setting. We rank models by the four-metric average on LiveIdea-Full4, SSIM on Widget2Code, and Correct on WorkBench. Six weaker models guide selection for disjoint stronger models. Fig. 4 shows that ranking fidelity improves with retention. HoloCosm's advantage over Random is most pronounced at low retention on LiveIdea-Full4 and Widget2Code, while the two methods perform similarly on WorkBench. The curves converge as retention increases. These results suggest that weaker source models can guide selection for stronger targets, with gains varying across benchmarks.

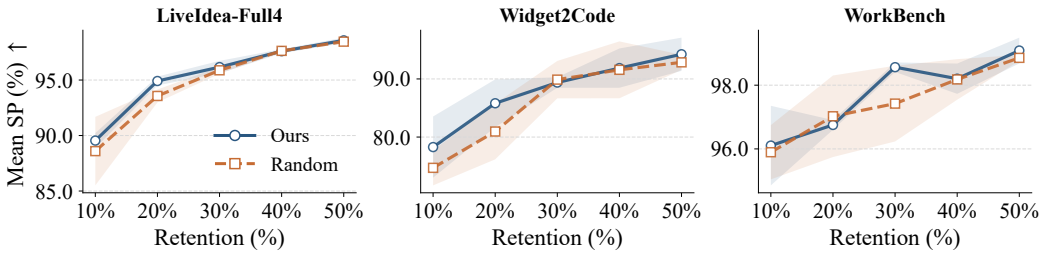

Figure 4: **Weak-to-strong generalization.** Six weaker models guide selection for disjoint stronger models. Curves show mean SP ± standard deviation over three seeds.

5 Conclusion

We uncover a systematic cross-metric failure in benchmark compression: subsets selected for one metric often perform worse than equal-size random sampling on others. To address this, HoloCosm minimizes multi-metric reconstruction error under a ranking-risk constraint. Across 15 benchmark configurations, HoloCosm improves score and ranking fidelity over Random. It remains competitive on single-metric benchmarks and transfers from weaker source models to stronger targets. Benchmark compression for multi-metric evaluation should preserve fidelity across metrics.

Reproducibility Statement

Some benchmark matrices are reconstructed from public item-level records; the remainder are obtained through official evaluation pipelines. The methodology and appendix describe the selection algorithm, score estimator, baseline implementations, and experimental protocols. We will release code, processed sample-level records, and source–target splits to support reproducibility.

AI Use Statement

Generative AI tools were used to debug experimental code, write experimental scripts, and improve the clarity of the manuscript. All AI-assisted code and text were reviewed and verified by the authors, who take full responsibility for the final paper.

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Sam Bowyer, Acyr Locatelli, and Kris Cao. Efficient benchmarking is just feature selection and multiple regression. *arXiv preprint arXiv:2605.25773*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646. IEEE, 2025.

Kuicai Dong, CHANG YUJING, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimomal generation for document question answering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0619. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1a93178950e92fd2e7b7448f7d68fd7d-Paper-Datasets_and_Benchmarks_Track.pdf.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026.

Simon Gerstenecker, Andreas Geiger, and Katrin Renz. Fail2drive: Benchmarking closed-loop driving generalization. *arXiv preprint arXiv:2604.08535*, 2026.

Gregory E Granato. Statistics for stochastic modeling of volume reduction, hydrograph extension, and water-quality treatment by structural stormwater runoff best management practices (bmps). Technical report, US Geological Survey, 2014.

Zhongzhan Huang, Guoming Ling, Shanshan Zhong, Hefeng Wu, and Liang Lin. Minilongbench: the low-cost long context understanding benchmark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11442–11460, 2025.

Zhongzhan Huang, Junxin Li, Guoming Ling, Yupei Lin, Shanshan Zhong, and Hefeng Wu. Zipbench: Low-cost framework for compressing comprehensive benchmarks of large language models, 2026.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. metabench-a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31734–31770, 2025.

Junjie Li, Xinrui Guo, Yuhao Wu, Roy Ka-Wei Lee, Hongzhi Li, and Yutao Xie. Lost in stories: Consistency bugs in long story generation by llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8400–8428, 2026.

Lele Liao, Qile Zhang, Ruofan Wu, and Guanhua Fang. Toward a unified framework for data-efficient evaluation of large language models. *arXiv preprint arXiv:2510.04051*, 2025.

Yifan Louie Lu, Rixin Liu, Jiayi Yuan, Xingqi Cui, Shenrun Zhang, Hongyi Liu, and Jiarong Xing. Routerarena: An open platform for comprehensive comparison of llm routers. In *International Conference on Learning Representations*, volume 2026, pp. 44385–44406, 2026.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Mariana Mesquita-Cunha, José Rui Figueira, and Ana Paula Barbosa-Póvoa. New ϵ-constraint methods for multi-objective integer linear programming: A pareto front representation approach. *European Journal of Operational Research*, 306(1):286–307, 2023.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in llm benchmarks. *arXiv preprint arXiv:2607.16259*, 2026.

S.J. Paech. Longform creative writing benchmark, 2025. URL `https://github.com/EQ-bench/longform-writing-bench`. GitHub repository.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2519–2536, 2024.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Nature communications*, 17(1):3625, 2026.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. Sublime: Subset selection via rank correlation prediction for data-efficient llm evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30572–30593, 2025.

Abhinav Kumar Singh, Harsha Vardhan Khurdula, Yoeven D Khemlani, and Vineet Agarwal. The structured output benchmark: A multi-source benchmark for evaluating structured output quality in large language models. *arXiv preprint arXiv:2604.25359*, 2026.

Olly Styles and Sam Miller. WorkBench Revisited: Workplace agents two years on. *arXiv preprint arXiv:2606.13715*, 2026.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, Xingzhang Ren, Fei Huang, Dayiheng Liu, and Linfeng Zhang. Rethinking llm evaluation: Can we evaluate llms with 200×less data? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 109852–109870, 2026.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Gregory Yauney, Shahzaib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations*, volume 2026, pp. 27389–27412, 2026.

Houston H Zhang, Tao Zhang, Baoze Lin, Yuanqi Xue, Yincheng Zhu, Huan Liu, Li Gu, Linfeng Ye, Ziqiang Wang, Xinxin Zuo, et al. Widget2Code: From visual widgets to UI code via multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20293–20302, 2026.

Yan Zhuang, Junhao Yu, Qi Liu, Yuxuan Sun, Jiatong Li, Zhenya Huang, and Enhong Chen. Efficient benchmarking via bias-bounded subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

APPENDIX CONTENTS

A Methodological Details **14**
- A.1 Problem Formulation and Notation 14
- A.2 Multi-Metric Selection Risks 14
- A.3 Constraint Calibration and Exchange Search 15
- A.4 Source-Assisted Estimation 16
- A.5 Source-to-Target Error Analysis 17

B Experimental Setup **18**
- B.1 Baseline Implementations 18
- B.2 Simple Multi-Metric Strategies 18

C Analysis and Discussion **19**
- C.1 Why Random Sampling Is a Strong Baseline 19
- C.2 When Is One Metric Sufficient? 19
- C.3 Alternative Multi-Metric Strategies 21
- C.4 Source-Model Selection 21
- C.5 Selection and Estimation Ablations 22
- C.6 Evaluation Cost 24

D Detailed Results **26**
- D.1 Aggregate Results 26
- D.2 Per-Benchmark Strategy Comparisons 26
- D.3 Cross-Metric Results 26

A METHODOLOGICAL DETAILS

A.1 PROBLEM FORMULATION AND NOTATION

Let X_{mnk} denote the score of source model m on item n under metric k, and let Y_{gnk} denote the corresponding score of target model g. The benchmark contains $N \geq 2$ items and K metrics. We select $S \subseteq [N]$ with $|S| = B$, where $1 \leq B < N$. For any model, its full-benchmark score and subset estimate are the corresponding arithmetic means over all items and over S, respectively.

For a target model, define

$$\mu_{gk} = \frac{1}{N} \sum_{n=1}^{N} Y_{gnk}, \qquad \widehat{\mu}_{gk}(S) = \frac{1}{B} \sum_{n \in S} Y_{gnk},$$

and its signed estimation error as $\delta_{gk}(S) = \widehat{\mu}_{gk}(S) - \mu_{gk}$. The same definitions apply to source models by replacing Y with X. Recall that $\mathcal{M} = [M]$ is the source-model index set.

Multi-metric compression jointly uses all K metrics to select a subset S of B items. The following sections define the source-based reconstruction objective and ranking-risk constraint used to select this subset.

A.2 MULTI-METRIC SELECTION RISKS

This section defines the multi-metric reconstruction objective and the pairwise ranking risk used by the selector. All quantities are computed from source scores. Let $\mathcal{A} \subseteq \mathcal{M}$ be a source-model pool containing q models.

A.2.1 NORMALIZED MEAN ERROR

For each metric, normalize errors by the population standard deviation of full-benchmark scores across models in $\mathcal{A}$:

$$\sigma_k(\mathcal{A}) = \sqrt{\frac{1}{q} \sum_{m \in \mathcal{A}} \left(\mu_{mk} - \overline{\mu}_{\mathcal{A},k}\right)^2}, \qquad \overline{\mu}_{\mathcal{A},k} = \frac{1}{q} \sum_{m \in \mathcal{A}} \mu_{mk}.$$

Use $s_k(\mathcal{A}) = \sigma_k(\mathcal{A})$ when $\sigma_k(\mathcal{A}) > \eta$, and $s_k(\mathcal{A}) = 1$ otherwise, where η is machine precision. Define the normalized model error and the reconstruction error of each metric as

$$e_{mk}(S; \mathcal{A}) = \frac{|\widehat{\mu}_{mk}(S) - \mu_{mk}|}{s_k(\mathcal{A})}, \qquad a_k(S; \mathcal{A}) = \frac{1}{q} \sum_{m \in \mathcal{A}} e_{mk}(S; \mathcal{A}).$$

The multi-metric mean risk is

$$r_{\text{mean}}(S; \mathcal{A}) = \frac{1}{K} \sum_{k=1}^{K} a_k(S; \mathcal{A}) = \frac{1}{qK} \sum_{m \in \mathcal{A}} \sum_{k=1}^{K} e_{mk}(S; \mathcal{A}).$$

Normalization makes the metric errors comparable, and the outer average gives each metric equal weight. Crucially, all metric terms are evaluated together for each candidate S: a proposed exchange is assessed through its combined effect across metrics. A single-metric reconstruction criterion contains only one component a_k, while the multi-metric criterion includes all K components. Source-pool averaging below preserves this decomposition.

A.2.2 PAIRWISE MARGIN RISK

For each model pair and metric, define $d_{abk} = \mu_{ak} - \mu_{bk}$ and $\widehat{d}_{abk}(S) = \widehat{\mu}_{ak}(S) - \widehat{\mu}_{bk}(S)$. Let

$$\mathcal{P}_k(\mathcal{A}) = \{(a, b) : a, b \in \mathcal{A},\ a < b,\ |d_{abk}| > \eta\}.$$

For nonempty $\mathcal{P}_k(\mathcal{A})$, use the scale $p_k(\mathcal{A}) = \mathrm{median}_{(a,b)\in\mathcal{P}_k(\mathcal{A})} |d_{abk}|$; set $p_k(\mathcal{A}) = 1$ when the pair set is empty. The ranking risk is

$$r_{\mathrm{rank}}(S;\mathcal{A}) = \frac{\displaystyle\sum_{k=1}^{K} \sum_{(a,b)\in\mathcal{P}_k(\mathcal{A})} \frac{\Big[|d_{abk}| - \mathrm{sign}(d_{abk})\widehat{d}_{abk}(S)\Big]_+}{p_k(\mathcal{A})}}{\displaystyle\sum_{k=1}^{K} |\mathcal{P}_k(\mathcal{A})|},$$

where $[u]_+ = \max(u, 0)$. If there are no valid pairs, the risk is zero.

The loss penalizes a reduction in the correctly oriented score gap and increases further when the ordering reverses. Valid pair–metric entries are pooled into one average.

A.2.3 Source-Pool Risk Averaging

HoloCosm evaluates each candidate selection across overlapping source-model pools. With M source models, define $\mathcal{A}_{-m} = \mathcal{M} \setminus \{m\}$ and

$$R_j(S) = \frac{1}{M} \sum_{m=1}^{M} r_j(S;\mathcal{A}_{-m}), \qquad j \in \{\mathrm{mean}, \mathrm{rank}\}. \tag{7}$$

The same candidate S is evaluated in every pool. Normalization scales and pairwise losses are recomputed using the models in that pool. With six source models, this produces six five-model risk evaluations.

This averaging incorporates several source-pool compositions into each candidate assessment. In the implementation, it is enabled when at least four source models are available; with three models, risks are evaluated directly on the full source pool.

A.3 Constraint Calibration and Exchange Search

Constrained selection. The ϵ-constraint approach optimizes a primary objective while constraining another criterion (Mesquita-Cunha et al., 2023). We apply it to multi-metric reconstruction error and pairwise-ranking distortion:

$$\min_{S:\,|S|=B} R_{\mathrm{mean}}(S) \quad \text{subject to} \quad R_{\mathrm{rank}}(S) \le \epsilon_{\mathrm{rank}}.$$

The mean objective combines score reconstruction across metrics, and the ranking constraint controls pairwise margin distortion.

A.3.1 Constraint Calibration

We draw $L = 128$ uniform size-B candidate selections, sampling without replacement within each selection. The ranking threshold is

$$\epsilon_{\mathrm{rank}} = Q_{0.25}\big(\{R_{\mathrm{rank}}(S_\ell)\}_{\ell=1}^{L}\big), \tag{8}$$

where $Q_{0.25}$ uses linear interpolation between ordered observations.

This threshold expresses the desired ranking quality relative to selections of the same size. It remains fixed throughout refinement. The candidate pool also supplies the initial selection.

A.3.2 Exchange Search and Its Properties

Define the relative excess

$$E(v,b) = \begin{cases} \max(0, v/b - 1), & b > \eta, \\ 0, & b \le \eta \text{ and } v \le \eta, \\ v/\eta, & b \le \eta \text{ and } v > \eta. \end{cases}$$

The ranking-constraint violation is $V(S) = E(R_{\text{rank}}(S), \epsilon_{\text{rank}})$. Candidates are compared lexicographically using

$$\Psi(S) = \begin{cases} (0, 0, R_{\text{mean}}(S)), & V(S) = 0, \\ (1, V(S), R_{\text{mean}}(S)), & V(S) > 0. \end{cases} \tag{9}$$

A candidate is numerically feasible when $V(S) = 0$, using tolerance η when $\epsilon_{\text{rank}} \leq \eta$. Feasible candidates are preferred; among feasible candidates, lower mean error is preferred. For infeasible candidates, lower constraint violation takes priority, followed by mean error.

One-for-one refinement. Initialize S with the candidate having the smallest comparison key. At each round, sample up to 64 distinct exchange pairs $(i, j) \in S \times ([N] \setminus S)$ and evaluate $S' = (S \setminus \{i\}) \cup \{j\}$. Choose the proposal with the smallest key and accept it if $\Psi(S') <_{\text{lex}} \Psi(S)$. Stop when no sampled proposal improves the key or after 80 accepted exchanges. The algorithm returns the best selection encountered under the feasibility-first comparison rule. The ranking constraint was satisfied in all 45 runs of the 20% retention experiment across 15 benchmarks and three seeds.

Properties of refinement. The reference-bank quantile guarantees at least one feasible reference subset, so feasibility-first initialization starts from a feasible subset. Each accepted exchange preserves $|S| = B$ and feasibility while strictly decreasing $R_{\text{mean}}(S)$. Strict decrease over the finite collection of size-B subsets prevents revisits and guarantees finite termination.

A.4 Source-Assisted Estimation

Algorithm 1 evaluates selection risks using uncorrected source means and applies source-assisted correction after the subset is fixed.

Residual Estimation Identity After fixing S, we combine the full-benchmark mean of auxiliary predictions with the mean observed target residual on S. This uses the correction principle of prediction-powered mean estimation (Angelopoulos et al., 2023), with source-average scores serving as the auxiliary predictions. Let

$$z_{nk} = \frac{1}{M} \sum_{m=1}^{M} X_{mnk}, \qquad \xi_{gnk} = Y_{gnk} - z_{nk}.$$

Bars denote full-item means, and a subscript S denotes subset means. Then

$$\mu_{gk} = \overline{z}_k + \overline{\xi}_{gk},$$
$$\widetilde{\mu}_{gk}(S) = \widehat{\mu}_{gk}(S) + \overline{z}_k - \overline{z}_{S,k} = \overline{z}_k + \overline{\xi}_{gS,k}.$$

Subtracting the full target mean gives the exact identity

$$\boxed{\widetilde{\mu}_{gk}(S) - \mu_{gk} = \overline{\xi}_{gS,k} - \overline{\xi}_{gk} = \delta_{gk}(S) - \frac{1}{M} \sum_{m=1}^{M} \delta_{mk}(S)}. \tag{10}$$

The adjustment replaces the auxiliary component's subset mean with its known full-benchmark mean. The remaining error is the subset error of the target residual.

Exact Preservation of Model Ordering **Proposition 1.** Source-assisted difference correction preserves every target-model pairwise score difference within each metric.

Proof. For metric k, the correction $c_k = \overline{z}_k - \overline{z}_{S,k}$ is shared by all target models. Hence, for any a, b,

$$\widetilde{\mu}_{ak} - \widetilde{\mu}_{bk} = (\widehat{\mu}_{ak} + c_k) - (\widehat{\mu}_{bk} + c_k) = \widehat{\mu}_{ak} - \widehat{\mu}_{bk}.$$

The correction preserves all within-metric orders and ties. Ranking fidelity is determined by item selection.

A.5 Source-to-Target Error Analysis

A.5.1 From Source Errors to Target Errors

Source-guided selection is useful when source and target models share item-level performance patterns. The following decomposition makes this relationship explicit.

For metric k, write the target score vector as

$$\mathbf{y}_{gk} = a_{gk}\mathbf{1} + \sum_{m=1}^{M} \alpha_{gmk}\mathbf{x}_{mk} + \mathbf{e}_{gk},$$

where $\mathbf{x}_{mk}, \mathbf{y}_{gk} \in \mathbb{R}^N$, and $\mathbf{e}_{gk}$ captures the remaining item-level variation. We use this representation to analyze transfer.

Define $\mathbf{d}_S = \mathbf{1}_S/B - \mathbf{1}/N$. Then $\mathbf{d}_S^\top \mathbf{1} = 0$, and the target error satisfies

$$\delta_{gk}(S) = \sum_{m=1}^{M} \alpha_{gmk}\delta_{mk}(S) + \mathbf{d}_S^\top \mathbf{e}_{gk}.$$

Consequently,

$$\boxed{|\delta_{gk}(S)| \le \sum_{m=1}^{M} |\alpha_{gmk}|\,|\delta_{mk}(S)| + \sqrt{\frac{1}{B} - \frac{1}{N}}\,\|\mathbf{e}_{gk}\|_2}. \tag{11}$$

Proof. Multiply the score decomposition by $\mathbf{d}_S^\top$. The constant term vanishes because $\mathbf{d}_S^\top \mathbf{1} = 0$. Apply the triangle inequality and Cauchy–Schwarz inequality, together with

$$\|\mathbf{d}_S\|_2^2 = B\left(\frac{1}{B} - \frac{1}{N}\right)^2 + (N-B)\frac{1}{N^2} = \frac{1}{B} - \frac{1}{N}.$$

The bound separates target error into a source-error contribution and a residual contribution. When source scores capture the target's item-level variation with moderate coefficients and a small residual, controlling errors across source models directly controls the first contribution. The relation applies separately to each metric, motivating simultaneous control of source reconstruction errors across the full evaluation matrix.

A.5.2 Score Errors and Ranking Preservation

Ranking preservation depends on estimation errors relative to the original score gaps. Consider two models a and b under metric k, with full-benchmark gap $d_{abk} = \mu_{ak} - \mu_{bk} > 0$. Their subset gap is

$$\widehat{d}_{abk}(S) = d_{abk} + \delta_{ak}(S) - \delta_{bk}(S).$$

Hence,

$$|\delta_{ak}(S)| + |\delta_{bk}(S)| < d_{abk} \quad \Longrightarrow \quad \widehat{d}_{abk}(S) > 0.$$

Derivation.

$$\widehat{d}_{abk}(S) \ge d_{abk} - |\delta_{ak}(S)| - |\delta_{bk}(S)| > 0.$$

The sufficient condition depends on the errors of a particular model pair relative to its score gap. A small error averaged over all models and metrics does not ensure this pair-specific condition, especially for closely spaced models. This distinguishes overall score reconstruction from preservation of model comparisons and motivates evaluating signed pairwise margins directly. The ranking risk defined in Appendix A.2.2 penalizes margin shrinkage as well as reversed orderings.

B EXPERIMENTAL SETUP

Tab. 2 summarizes the tasks and model, item, and component counts for the 15 benchmark configurations, which together provide 85 component metrics. Component scores retain their benchmark-specific definitions, and evaluation NMAE uses the held-out target-model scale in Eq. 6.

B.1 BASELINE IMPLEMENTATIONS

Common protocol. All external baselines use the same six source models as HoloCosm. Preprocessing statistics, hyperparameter choices, item selection, and estimator fitting are determined using only source-model information and method-specific item features. Target-model responses are never used during selection or tuning; target models are evaluated only after the selected subset and estimation procedure have been fixed.

tinyBenchmarks. For continuous components, we min–max normalize source scores and choose a binarization threshold to match the binary response mean to the original source-score mean. We fit a five-dimensional 2PL model with hierarchical priors and represent each item by its fitted discrimination and difficulty parameters. Weighted K-means forms B clusters, and linear-sum assignment selects one distinct item per center. Cluster frequencies provide aggregation weights.

ZipBench. ZipBench uses the same preprocessing and IRT configuration as tinyBenchmarks. We augment the six observed source profiles with six synthetic profiles constructed by majority voting over source-model combinations across performance levels. We cluster the fitted item representations into B groups, select each center's nearest item, and weight selected items by cluster frequency.

mRMR. mRMR greedily selects items using the mutual-information quotient, with the full-benchmark source-model mean as the relevance target. We use appropriate mutual-information estimators for binary and continuous components. A separate RidgeCV estimator is fitted for each evaluation component to reconstruct target scores from the selected items.

AnchorPoints. AnchorPoints computes Pearson correlations between item response patterns across source models and applies fasterpam to the distance $1 - \mathrm{correlation}$. The medoids are selected, and each item is assigned to its most-correlated medoid. Cluster sizes provide aggregation weights.

EssenceBench. We use BGE-M3 embeddings and size-constrained filtering, with GA/EBM search when the filtered pool exceeds the required subset size. Selection uses leave-one-source-model-out validation and GAM-based fitness. Target scores are estimated by averaging selected scores.

B.2 SIMPLE MULTI-METRIC STRATEGIES

All strategies use identical source–target splits and 20% retention. Evaluation uses the original metrics. We compute mean and worst-component fidelity per run, average runs within each configuration, and weight the 15 configurations equally. For the Single-metric oracle reference, we report the selection metric with the lowest target-set mean NMAE.

AverageMetric. We align metric directions, apply source-derived min–max normalization, and average metrics equally. Methods use it for selection and retain existing score-estimation procedures.

Vote. We combine the size-B subsets selected by AnchorPoints for individual metrics. Each selected item receives one vote per metric, and the B highest-voted distinct items are retained, with balanced metric coverage used for tie-breaking. Their unweighted means estimate all target metrics.

K-means. Let s_k be the standard deviation, across source models, of their full-benchmark means on metric k; numerically zero scales are set to one. For each item, we concatenate source responses normalized by s_k:

$$v_n = \mathrm{vec}([X_{mnk}/s_k]_{m,k}), \qquad \widehat{\mu}_{gk} = \sum_{j=1}^{B} \frac{|C_j|}{N} Y_{g,s_j,k}.$$

We use Lloyd K-means with B clusters and select the actual item nearest each cluster center, enforcing distinct representatives when needed. Cluster proportions provide the weights above, and target models evaluate only the B representatives. Detailed comparisons are reported in Appendix D.2.

C Analysis and Discussion

C.1 Why Random Sampling Is a Strong Baseline

Uniform sampling estimates full-benchmark means without requiring source-model information. Its competitiveness is also highlighted by Yauney et al. (2026), who examine the reliability of micro-benchmarks in preserving pairwise model comparisons.

For a subset S sampled uniformly without replacement from all size-B subsets, the uncorrected subset mean satisfies

$$\mathbb{E}_S[\widehat{\mu}_{gk}(S)] = \mu_{gk}, \qquad \mathrm{Var}_S[\widehat{\mu}_{gk}(S)] = \frac{1 - B/N}{B}\, v_{gk},$$

where $v_{gk} = (N-1)^{-1} \sum_{n=1}^{N} (Y_{gnk} - \mu_{gk})^2$ is the finite-population variance of the target model's item scores.

The estimator is unbiased, so its mean squared error equals this variance and decreases as B increases. These properties help explain why Random is a strong baseline for score reconstruction. In benchmark compression, however, one subset is fixed and reused across model evaluations, making the quality of that particular subset consequential.

Empirical variability across random seeds. We evaluate Bench2Drive, LongForm, and SoB-Image at retention ratios of 10%, 20%, 30%, 40%, and 50%. For each benchmark–retention setting, we fix the source–target split and repeat item selection with 10 random seeds. Performance is first averaged across component metrics within each run, and the reported means and standard deviations are computed across the 10 seeds.

Fig. 5 shows how evaluation fidelity varies across random seeds at each retention level. Thus, strong average performance can coexist with substantial variation in the quality of individual random selections. Source-guided selection addresses this practical need by using observed model responses to identify items that preserve evaluation fidelity across metrics.

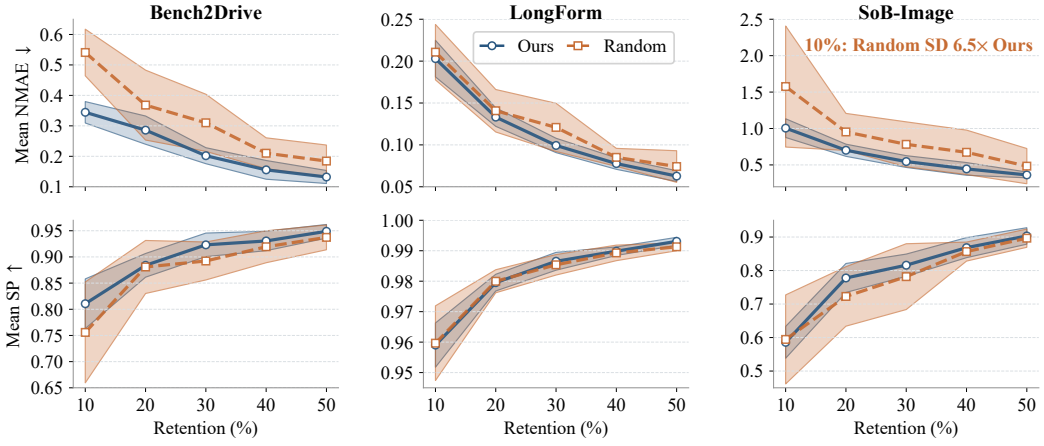

Figure 5: **Random-selection variability across 10 seeds.** Results cover Bench2Drive, LongForm, and SoB-Image at 10%, 20%, 30%, 40%, and 50% retention under fixed model splits. Curves show ten-seed means; shaded bands indicate one standard deviation. Lower NMAE and higher SP indicate better fidelity.

C.2 When Is One Metric Sufficient?

DeepResearch Bench (Du et al., 2026) provides a high-correlation counterpart to Fig. 1. All ten metric pairs have model-ranking Spearman correlations of 0.92–0.99 (median 0.95), while Overall Score correlates with every component at 0.96–0.99. We further test whether this alignment makes

Table 8: **Cross-metric fidelity on DeepResearch at 20% retention.** Each mRMR and AnchorPoints row uses one component for selection and evaluates the selected items on all five metrics. HoloCosm uses all metrics, while Random uses no guiding metric. Entries report NMAE ↓ / SP ↑, averaged over three seeds.

Method	Selection metric	Comp.	Insight	Inst. Follow.	Read.	Overall	Avg.
HoloCosm	All metrics	0.12 / 0.99	0.10 / 0.99	0.21 / 0.98	0.10 / 0.98	0.10 / 0.98	**0.13** / 0.98
Random	–	0.10 / 0.99	0.09 / 0.99	0.27 / 0.97	0.15 / 0.99	0.09 / 0.99	0.14 / **0.99**
mRMR	Comprehensiveness	0.13 / 0.98	0.08 / 0.99	0.26 / 0.98	0.22 / 0.98	0.11 / 0.99	0.16 / 0.98
	Insight	0.10 / 0.98	0.08 / 0.99	0.32 / 0.94	0.11 / 0.98	0.09 / 0.98	0.14 / 0.98
	Instruction Following	0.15 / 0.99	0.09 / 0.99	0.30 / 0.94	0.18 / 0.98	0.10 / 1.00	0.16 / 0.98
	Readability	0.11 / 0.98	0.08 / 0.99	0.27 / 0.96	0.13 / 0.96	0.10 / 0.99	0.14 / 0.98
	Overall Score	0.11 / 0.97	0.09 / 0.99	0.22 / 0.96	0.17 / 0.96	0.11 / 0.98	0.14 / 0.97
AnchorPoints	Comprehensiveness	0.14 / 0.98	0.18 / 0.99	0.36 / 0.96	0.19 / 0.98	0.18 / 0.99	0.21 / 0.98
	Insight	0.13 / 0.99	0.09 / 0.99	0.23 / 0.96	0.12 / 0.99	0.11 / 0.99	0.13 / 0.98
	Instruction Following	0.15 / 0.99	0.11 / 0.99	0.22 / 0.96	0.14 / 0.99	0.13 / 0.99	0.15 / 0.98
	Readability	0.12 / 0.99	0.17 / 0.98	0.31 / 0.95	0.11 / 0.99	0.13 / 0.98	0.17 / 0.98
	Overall Score	0.12 / 0.98	0.12 / 0.99	0.28 / 0.96	0.12 / 0.99	0.11 / 0.98	0.15 / 0.98

single-metric selection transferable. At 20% retention, each component is used in turn as the sole selection metric for mRMR and AnchorPoints, and the selected items are evaluated on all five metrics. We additionally evaluate HoloCosm and Random at the same retention level.

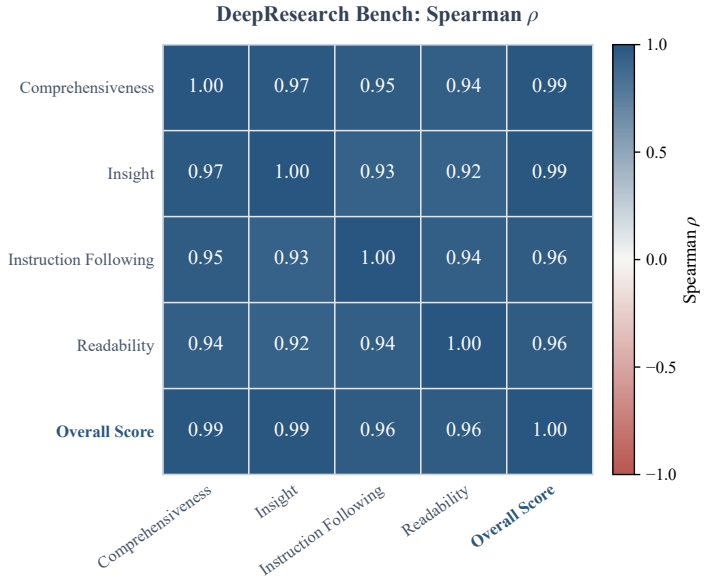

Figure 6: **A high-correlation control on DeepResearch Bench.** Pairwise model-ranking Spearman correlations among four component dimensions and Overall Score. Overall Score is highlighted.

Regardless of the component used for selection, mRMR achieves an average SP of 0.9714–0.9829 and AnchorPoints achieves 0.9786–0.9842 across all evaluation metrics. Their average NMAE ranges are 0.1400–0.1645 and 0.1348–0.2087, respectively. HoloCosm's multi-metric selection reaches 0.1254/0.9817 in average NMAE/SP. These results show that when component rankings are highly aligned, items selected using any single metric already transfer well to the remaining metrics.

Table 9: Aggregate-guided compression at 20% retention on five existing configurations. Official Overall includes the restricted LiveIdea-Core3 average described in the text.

Method	Strategy	Mean NMAE	Worst NMAE	Mean SP	Worst SP
AnchorPoints	AverageMetric	0.2955	0.4285	0.9137	0.8613
AnchorPoints	Official Overall	0.3781	0.6402	0.9142	0.8707
Ours	All metrics	0.1573	0.2420	0.9561	0.9006
Ours	AverageMetric	0.1809	0.3055	0.9530	0.8942
Ours	Official Overall	0.1729	0.2679	0.9534	0.9009
Random	Unguided	0.1747	0.2813	0.9549	0.9199
ZipBench	AverageMetric	0.2802	0.4487	0.9271	0.8735
ZipBench	Official Overall	0.2667	0.4074	0.9152	0.8478
mRMR	AverageMetric	0.3859	0.4718	0.8519	0.7727
mRMR	Official Overall	0.3772	0.4801	0.8191	0.7134
tinyBenchmarks	AverageMetric	0.2656	0.4009	0.9224	0.8544
tinyBenchmarks	Official Overall	0.2788	0.4380	0.9220	0.8717

Table 10: Aggregate-guided compression at 20% retention on eight configurations with common coverage. Official Overall includes the restricted LiveIdea-Core3 average described in the text.

Method	Strategy	Mean NMAE	Worst NMAE	Mean SP	Worst SP
AnchorPoints	AverageMetric	0.3951	0.5538	0.8756	0.8200
AnchorPoints	Official Overall	0.5118	0.7924	0.8454	0.7867
Ours	All metrics	0.2156	0.3050	0.9105	0.8582
Ours	AverageMetric	0.2711	0.3878	0.8845	0.8282
Random	Unguided	0.3295	0.4558	0.8748	0.8215
ZipBench	AverageMetric	0.4429	0.6315	0.8030	0.7382
ZipBench	Official Overall	0.4082	0.5670	0.8167	0.7392
mRMR	AverageMetric	0.4052	0.4886	0.7616	0.6692
mRMR	Official Overall	0.4519	0.6511	0.7345	0.5891
tinyBenchmarks	AverageMetric	0.4244	0.5924	0.8360	0.7539
tinyBenchmarks	Official Overall	0.4245	0.5850	0.8179	0.7416

C.3 Alternative Multi-Metric Strategies

We compare AverageMetric, Vote, and K-means at 20% retention. Tab. 6 summarizes the comparison; Appendix D.2 reports the per-benchmark results.

C.3.1 Benchmark-Defined Overall Scores

We examine whether benchmark-level aggregate scores provide effective guidance for preserving component metrics, using AverageMetric as the reference. All selected subsets are evaluated on the original component metrics at 20% retention.

Tab. 9 compares aggregation strategies on ConStory, LiveIdea-Core3/Full4, and MMDocRAG Text/Multimodal. LiveIdea-Core3 uses the average of its three included metrics, and ConStory uses CED responses with full-data normalization. Tab. 10 extends the baseline comparison to Bench2Drive and Fail2Drive-ID/OOD, using driving_score and HM, respectively. Results are averaged over the configurations included in each comparison.

Benchmark-level aggregate guidance does not consistently improve component fidelity. HoloCosm achieves the lowest mean NMAE in both comparisons.

C.4 Source-Model Selection

C.4.1 Number of Source Models

We vary the number of source models from 3 to 15 on LiveIdea-Full4 at 20% retention, using the same ten target models throughout. The remaining 31 models form the source candidate pool. We rank these candidates by their average metric-wise min–max-normalized full-benchmark scores and

Table 11: **Random-6 versus diverse source models on SoB-Audio at 20% retention.** Results are averaged over three seeds.

	NMAE ↓		SP ↑	
Metric	Random-6	HoloCosm	Random-6	HoloCosm
Hierarchical path recall	0.6434	**0.5412**	0.7500	**0.8389**
Leaf-value EM	0.9410	**0.6525**	0.5333	**0.7944**
Path-set F1	0.7366	**0.5841**	0.6389	**0.8444**
Schema compliance	0.7012	**0.5230**	0.7608	**0.8507**
Value-token F1	0.5520	**0.4463**	0.7667	**0.9333**
Mean	0.7148	**0.5494**	0.6899	**0.8524**

select 15 evenly spaced representatives. For each source count m, we choose m evenly spaced representatives, including both endpoints, and use the same selection across seeds. Results are averaged over three random seeds.

Fig. 7 shows that six source models achieve the highest SP among the tested source counts. With six sources, NMAE is 0.2065 and SP is 0.9535, compared with 0.2115 and 0.9475 with fifteen sources. These results indicate that a compact source cohort can support accurate multi-metric compression in this setting.

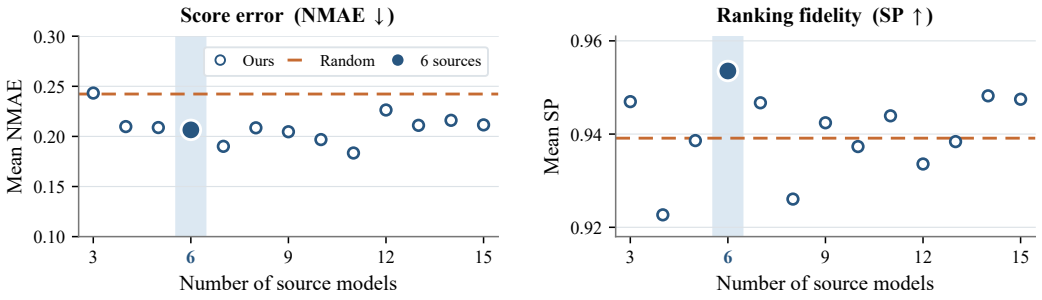

Figure 7: **Source-model count sensitivity on LiveIdea-Full4.** Results at 20% retention are averaged over three random seeds. All configurations use the same target models. Lower NMAE and higher SP indicate better fidelity.

C.4.2 Source Selection Strategy

We compare two strategies for selecting six source models on SoB-Audio at 20% retention. Our strategy selects models at evenly spaced positions in the source candidate pool's performance ranking. *Random-6* samples six models uniformly without replacement from the same pool. Both strategies are evaluated on the same nine target models, with results averaged over three seeds.

Tab. 11 shows that selecting source models across the performance ranking improves both NMAE and SP on all five component metrics. Mean NMAE decreases from 0.7148 to 0.5494, a reduction of 23.1%, while mean SP increases from 0.6899 to 0.8524.

C.5 Selection and Estimation Ablations

We compare two additional variants at 20% retention across all 15 configurations, using the same source–target splits, seeds, and source-assisted estimator as in Tab. 5.

Effect of swap refinement. *Without swap refinement* uses the same 128 reference subsets and ranking threshold as HoloCosm, selects the feasible reference subset with the lowest reconstruction error, and stops without exchanging items. This isolates the contribution of refinement from

Table 12: **Swap refinement and reconstruction-objective ablations at 20% retention.** All variants use matched seeds and the same source-assisted estimator. Results are averaged equally over 15 configurations.

Variant	Mean NMAE ↓	Worst NMAE ↓	Mean SP ↑
Without swap refinement	0.2500	0.3544	0.9130
Corrected-residual objective	0.2817	0.3802	0.8983
HoloCosm	**0.2362**	**0.3299**	**0.9151**

reference-bank selection. As shown in Tab. 12, refinement further reduces Mean NMAE by 5.53%, with a smaller increase in Mean SP.

Reconstruction objective. *Corrected-residual objective* replaces only the mean reconstruction loss, retaining the source-pool folds, metric scales, ranking constraint, and search budget. Within each fold, each source model is treated as a pseudo-target and corrected using the other models in that fold. For a fold F of size L, let $\delta_{mk}(S)$ be model m's uncorrected subset error. The corrected error is

$$e_{mk}^{(-m)}(S) = \delta_{mk}(S) - \frac{1}{L-1} \sum_{j \in F \setminus \{m\}} \delta_{jk}(S) = \frac{L}{L-1} \left(\delta_{mk}(S) - \overline{\delta}_{F,k}(S) \right).$$

We average its normalized absolute value across models, metrics, and the existing folds. This alternative increases mean NMAE by 19.29% and lowers mean SP relative to HoloCosm.

Tab. 13 isolates multi-metric selection and pairwise-ranking protection on Fail2Drive-ID. *Single-metric selection* uses HM, DR, and SR in turn to guide selection and averages the results over guiding metrics. *Without ranking constraint* removes only the pairwise-ranking constraint. Both ablations degrade all three metrics: multi-metric selection reduces average NMAE by 35.1% over single-metric selection, while removing the constraint changes NMAE/SP from 0.2762/0.7964 to 0.3573/0.6594.

Table 13: **Ablation on Fail2Drive-ID at 20% retention.** Three-seed NMAE ↓ / SP ↑ means; Avg. averages evaluation metrics. Single-metric results also average guiding metrics.

Variant	HM	DR	SR	Avg.
Ours	**0.2940 / 0.7845**	**0.2246 / 0.8345**	**0.3101 / 0.7703**	**0.2762 / 0.7964**
Single-metric selection	0.4507 / 0.6663	0.3309 / 0.6968	0.4960 / 0.6636	0.4259 / 0.6756
Without ranking constraint	0.3710 / 0.6429	0.3099 / 0.6714	0.3911 / 0.6640	0.3573 / 0.6594

Tabs. 14 and 15 resolve this comparison at the component level on two three-metric benchmarks. On WorkBench, Ours reduces NMAE on all three metrics, with the largest reduction on Exact Match (16.5%). On LiveIdea-Core3, the reduction is concentrated on Feasibility (16.7%), while Clarity and Originality change by less than 3%.

Table 14: **Per-metric constrained-selection ablation on WorkBench.** Results at 20% retention are three-run means. Reconstruction and Ours use the same source-assisted correction. NMAE reduction is $(\mathrm{Recon} - \mathrm{Ours})/\mathrm{Recon}$; SP gain is $\mathrm{Ours} - \mathrm{Recon}$. Positive values favor Ours.

	NMAE ↓			SP ↑		
Metric	Recon	Ours	Reduction	Recon	Ours	Gain
Correct	0.1384	**0.1288**	6.9%	0.9790	**0.9855**	+0.0065
Exact Match	0.1561	**0.1304**	16.5%	0.9814	**0.9837**	+0.0022
Safety	0.1696	**0.1578**	6.9%	0.9592	**0.9614**	+0.0022

Table 15: **Per-metric constrained-selection ablation on LiveIdea-Core3.** Results at 20% retention are three-run means. Reconstruction and Ours use the same source-assisted correction. NMAE reduction is $(\mathrm{Recon} - \mathrm{Ours})/\mathrm{Recon}$; SP gain is $\mathrm{Ours} - \mathrm{Recon}$. Positive values favor Ours.

	NMAE ↓			SP ↑		
Metric	Recon	Ours	Reduction	Recon	Ours	Gain
Clarity	**0.1054**	0.1074	−1.9%	0.9725	**0.9738**	+0.0012
Feasibility	0.1564	**0.1303**	16.7%	0.9641	**0.9771**	+0.0130
Originality	**0.0690**	0.0707	−2.6%	**0.9947**	**0.9947**	0.0000

C.5.1 Comparison under a Common Estimator

We compare HoloCosm with AverageMetric-guided ZipBench under the same uniform source-assisted estimator at 20% retention across all 15 configurations. We also report ZipBench's original weighted estimator. The ZipBench variants use the same selected items, and all estimates are evaluated on the original component metrics.

Table 16: **Common-estimator comparison with AverageMetric-guided ZipBench at 20% retention.** Results average three runs per method and weight the 15 benchmarks equally. All ZipBench variants use the same selected items.

Estimator / method	Mean NMAE ↓	Worst NMAE ↓	Mean SP ↑
ZipBench: original weighted	0.4920	0.7376	0.8145
ZipBench: uniform + correction	0.3423	0.4720	0.8705
HoloCosm	**0.2362**	**0.3299**	**0.9151**

Tab. 16 shows that, under the same uniform source-corrected estimator, HoloCosm-selected subsets achieve 31.0% lower mean NMAE and 30.1% lower worst-component NMAE than AverageMetric-guided ZipBench subsets. Mean NMAE is lower on 13 of the 15 configurations.

C.5.2 Sensitivity to the Ranking-Constraint Quantile

We evaluate ranking-constraint calibration quantiles $q \in \{0.10, 0.25, 0.50\}$ on Fail2Drive-ID, SoB-Audio, and LiveIdea-Core3 at 20% retention. Tab. 17 reports results averaged over three seeds.

Table 17: **Sensitivity to the ranking-constraint quantile at 20% retention.** Each entry reports mean/worst-component NMAE / mean SP.

Setting	Fail2Drive-ID	SoB-Audio	LiveIdea-Core3
$q = 0.10$	0.3996 / 0.4671 / 0.5782	0.6236 / 0.7902 / 0.7960	0.1028 / 0.1336 / 0.9819
$q = 0.25$ (default)	**0.2762 / 0.3219 / 0.7964**	**0.5798 / 0.7520 / 0.8293**	**0.1028 / 0.1336 / 0.9819**
$q = 0.50$	0.3573 / 0.3922 / 0.6594	0.5798 / 0.7520 / 0.8293	0.1102 / 0.1564 / 0.9771

The default $q = 0.25$ achieves the best overall fidelity across the three benchmarks.

C.6 Evaluation Cost

C.6.1 Evaluation Size

For metric-wise selection, the nominal per-metric selection size need not equal the final evaluation cost. Let $S_k(B)$ denote the selection for component k. The union $S_{\mathrm{union}}(B) = \bigcup_{k=1}^{K} S_k(B)$ contains between B and $\min(N, KB)$ unique items. For each benchmark, method, retention level, and seed, the evaluation cost is the number of distinct items in this union, with items shared by several component selections counted only once.

We report the inflation factor $I_{\text{union}} = |S_{\text{union}}(B)|/B$. Tab. 18 gives representative tinyBenchmarks selection sizes and absolute union sizes. For example, on LongForm a per-component selection size of 22 items produces a union of 101 items, or 4.591 times the nominal selection size.

Table 18: **Representative actual union sizes for tinyBenchmarks.** At 20% nominal retention, independent component-wise selections can require substantially more than B unique evaluations.

Benchmark	B	$\lvert S_{\text{union}}\rvert$	I_{union}
LongForm	22	101	4.591
Widget2Code	58	263	4.534
MMDocRAG-MM	395	1,624	4.111
SoB-Text	1,000	3,320	3.320
RouterArena	1,680	3,020	1.798

C.6.2 Offline Item-Selection Time

At 20% retention, item selection averages 0.77 seconds across all 45 benchmark–seed runs. Tab. 19 gives results for three larger benchmarks; the longest individual run is 3.71 seconds on RouterArena.

Table 19: **Offline item-selection time at 20% retention.** Times are means and one standard deviation over three seeds.

Benchmark	Items	Metrics	Selection size	Time (s)
RouterArena	8,400	2	1,680	2.53 ± 1.12
SoB-Text	5,000	6	1,000	2.38 ± 0.28
MMDocRAG-Multimodal	1,974	8	395	1.32 ± 0.17

D Detailed Results

D.1 Aggregate Results

Tab. 20 reports the aggregate fidelity measures. We additionally report Kendall's τ_b, which accounts for tied ranks (Kendall, 1945). For each single-target method and benchmark, every official component is used as a selection target, and the selected items are evaluated on all components. Component means and extrema are computed for each selection before averaging over repetitions. Each baseline retains the target with the lowest mean NMAE and reports all fidelity measures for that target. The final summary weights the 15 benchmarks equally. HoloCosm and Random each perform one selection per run rather than metric-specific selection.

Table 20: **Aggregate fidelity across 15 benchmarks at 20% retention.** Mean and worst-component fidelity are aggregated within benchmarks before averaging benchmarks equally. Each scalar method uses its best selection target by mean NMAE, with the same target used for all fidelity measures. Bold and underline indicate the best and second-best values in each column.

	NMAE ↓		SP ↑		Kendall's τ_b ↑
Method	Mean	Worst	Mean	Worst	Mean
HoloCosm	**0.2362**	**0.3299**	**0.9151**	**0.8624**	**0.8173**
Random	<u>0.3650</u>	<u>0.5450</u>	<u>0.8840</u>	<u>0.8302</u>	<u>0.7859</u>
tinyBenchmarks	0.4421	0.6297	0.8483	0.7545	0.7252
ZipBench	0.4123	0.6364	0.8467	0.7646	0.7288
mRMR	0.4321	0.5997	0.7894	0.5952	0.6734
AnchorPoints	0.5533	0.9256	0.8361	0.7508	0.7196
EssenceBench	0.4515	0.6302	0.8736	0.7898	0.7643

D.2 Per-Benchmark Strategy Comparisons

Tabs. 21 and 22 expand Tab. 6. All values are averages over available runs within each configuration; Average weights configurations equally.

Tab. 23 reports additional baseline results using AverageMetric for item selection at 20% retention.

D.3 Cross-Metric Results

Tabs. 24, 25, and 26 report cross-metric fidelity at 20% retention. Baselines select using the indicated metric; underlining marks diagonal entries.

The following tables report the remaining benchmark-level cross-metric results at 20% retention. Entries are NMAE / SP, averaged over three seeds.

Table 21: Per-configuration simple-strategy results at 20% retention: Random and AnchorPoints. NMAE and SP are means over metrics.

Configuration	Random		AnchorPoints: Single		AnchorPoints: AverageMetric		AnchorPoints: Vote	
	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑
Widget2Code	0.2979	0.9335	0.3280	0.8628	0.3148	0.9167	0.4707	0.8911
ConStory	0.2938	0.8813	1.2431	0.5209	0.8375	0.7038	0.3320	0.8606
Bench2Drive	0.3881	0.8418	0.5910	0.7899	0.5200	0.8120	0.9476	0.8800
Fail2Drive-ID	0.5805	0.5918	0.2839	0.7051	0.4309	0.7524	0.8388	0.8696
Fail2Drive-OOD	0.7939	0.7903	0.6111	0.7968	0.7323	0.8720	0.7760	0.7794
LiveIdea-Core3	0.1075	0.9834	0.1207	0.9811	0.1182	0.9733	0.1107	0.9827
LiveIdea-Full4	0.2888	0.9359	0.2693	0.9410	0.3016	0.9225	0.2539	0.9490
LongForm	0.1400	0.9814	0.1459	0.9801	0.1483	0.9780	0.1546	0.9738
MMDocRAG-Multimodal	0.0729	0.9915	0.0864	0.9882	0.0886	0.9908	0.1494	0.9883
MMDocRAG-Text	0.1106	0.9824	0.1259	0.9740	0.1318	0.9780	0.3201	0.9583
RouterArena	0.1464	0.9905	0.2783	0.9714	0.5413	0.9048	0.8428	0.9857
SoB-Audio	0.7048	0.7791	0.7179	0.8040	1.3456	0.6978	1.3263	0.7372
SoB-Image	1.1731	0.6484	0.8664	0.7090	1.0741	0.7303	1.8610	0.7579
SoB-Text	0.2104	0.9598	1.8020	0.6720	1.8754	0.7646	0.8672	0.9657
WorkBench	0.1664	0.9683	0.8299	0.8448	0.6982	0.7484	0.1812	0.9775
Average	0.3650	0.8840	0.5533	0.8361	0.6106	0.8497	0.6288	0.9038

Table 22: Per-configuration simple-strategy results at 20% retention: K-means and Ours. NMAE and SP are means over metrics.

Configuration	K-means		Ours: AverageMetric		Ours: All metrics	
	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑
Widget2Code	0.2231	0.9196	0.2086	0.9276	0.2178	0.9444
ConStory	0.5427	0.7896	0.4015	0.8618	0.2972	0.8754
Bench2Drive	0.2938	0.8346	0.3234	0.8677	0.2736	0.8432
Fail2Drive-ID	0.2370	0.7865	0.3761	0.7031	0.2762	0.7964
Fail2Drive-OOD	0.4327	0.7447	0.5646	0.7401	0.3889	0.8642
LiveIdea-Core3	0.1425	0.9738	0.1056	0.9811	0.1028	0.9819
LiveIdea-Full4	0.3216	0.9335	0.2414	0.9471	0.2247	0.9481
LongForm	0.1474	0.9744	0.1355	0.9763	0.1172	0.9839
MMDocRAG-Multimodal	0.0756	0.9903	0.0658	0.9917	0.0645	0.9929
MMDocRAG-Text	0.1411	0.9702	0.0902	0.9832	0.0972	0.9821
RouterArena	0.2161	0.9524	0.1207	0.9810	0.0928	0.9714
SoB-Audio	0.6312	0.8668	0.5812	0.7927	0.5798	0.8293
SoB-Image	0.7921	0.6235	0.7079	0.7217	0.5145	0.7723
SoB-Text	0.3489	0.9415	0.1649	0.9692	0.1560	0.9642
WorkBench	0.3171	0.9620	0.1463	0.9753	0.1390	0.9769
Average	0.3242	0.8842	0.2822	0.8946	0.2362	0.9151

Table 23: Additional AverageMetric-guided results at 20% retention. Ours: All metrics uses the full metric set. NMAE and SP are metric means; Average weights 15 configurations equally.

Configuration	tinyBenchmarks		ZipBench		mRMR		AnchorPoints		Ours: AverageMetric		Ours: All metrics	
	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑	NMAE ↓	SP ↑
Widget2Code	0.4120	0.8958	0.3892	0.8562	0.2476	0.9167	0.3148	0.9167	0.2086	0.9276	0.2178	0.9444
ConStory	0.4377	0.8259	0.5151	0.8100	0.7738	0.4386	0.8375	0.7038	0.4015	0.8618	0.2972	0.8754
Bench2Drive	0.7194	0.7078	0.6091	0.7980	0.5039	0.7149	0.5200	0.8120	0.3234	0.8677	0.2736	0.8432
Fail2Drive-ID	0.6051	0.7966	0.6112	0.5426	0.3388	0.5753	0.4309	0.7524	0.3761	0.7031	0.2762	0.7964
Fail2Drive-OOD	0.7430	0.5717	0.9218	0.4478	0.4691	0.5431	0.7323	0.8720	0.5646	0.7401	0.3889	0.8642
LiveIdea-Core3	0.2125	0.9534	0.2012	0.9672	0.3742	0.9657	0.1182	0.9733	0.1056	0.9811	0.1028	0.9819
LiveIdea-Full4	0.3656	0.8915	0.4070	0.9139	0.4173	0.8896	0.3016	0.9225	0.2414	0.9471	0.2247	0.9481
LongForm	0.1878	0.9641	0.2170	0.9623	0.1727	0.9735	0.1483	0.9780	0.1355	0.9763	0.1172	0.9839
MMDocRAG-Multimodal	0.1340	0.9798	0.1110	0.9813	0.1147	0.9917	0.0886	0.9908	0.0658	0.9917	0.0645	0.9929
MMDocRAG-Text	0.1781	0.9612	0.1665	0.9631	0.2495	0.9736	0.1318	0.9780	0.0902	0.9832	0.0972	0.9821
RouterArena	0.3973	0.9714	0.3411	0.9048	1.0962	0.8857	0.5413	0.9048	0.1207	0.9810	0.0928	0.9714
SoB-Audio	1.0299	0.6290	1.1876	0.6782	0.5518	0.7420	1.3456	0.6978	0.5812	0.7927	0.5798	0.8293
SoB-Image	1.1644	0.6768	1.0201	0.5886	0.7144	0.2875	1.0741	0.7303	0.7079	0.7217	0.5145	0.7723
SoB-Text	0.4390	0.8792	0.4560	0.8417	0.5707	0.8218	1.8754	0.7646	0.1649	0.9692	0.1560	0.9642
WorkBench	0.2717	0.9527	0.2256	0.9612	0.3894	0.8390	0.6982	0.7484	0.1463	0.9753	0.1390	0.9769
Average	0.4865	0.8438	0.4920	0.8145	0.4656	0.7706	0.6106	0.8497	0.2822	0.8946	0.2362	0.9151

Table 24: **Cross-metric fidelity on Fail2Drive-ID and Fail2Drive-OOD at 20% retention.** Ours and Random do not use metric-specific selection. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Fail2Drive-ID								
Method	**Selection metric**	**NMAE ↓**				**SP ↑**		
		HM	**DR**	**SR**	**Mean**	**HM**	**DR**	**SR**
Ours	All metrics	0.294	0.225	0.310	0.276	0.785	0.835	0.770
Random	None	0.611	0.488	0.643	0.580	0.571	0.646	0.558
tinyBenchmarks	HM	0.804	0.702	0.820	0.775	0.482	0.608	0.296
	DR	0.766	0.681	0.814	0.754	0.477	0.583	0.506
	SR	0.785	0.602	0.837	0.741	0.848	0.883	0.653
ZipBench	HM	0.544	0.396	0.592	0.510	0.544	0.596	0.565
	DR	0.800	0.583	0.842	0.742	0.630	0.687	0.580
	SR	0.766	0.668	0.809	0.748	0.717	0.776	0.671
mRMR	HM	0.439	0.383	0.440	0.420	0.593	0.550	0.357
	DR	0.403	0.296	0.405	0.368	0.507	0.614	0.345
	SR	0.424	0.453	0.457	0.445	0.561	0.496	0.591
AnchorPoints	HM	0.290	0.235	0.327	0.284	0.705	0.684	0.726
	DR	0.698	0.696	0.704	0.699	0.551	0.604	0.602
	SR	0.914	0.874	0.969	0.919	0.699	0.712	0.728
EssenceBench	HM	1.113	0.997	1.105	1.072	0.518	0.496	0.615
	DR	1.031	0.930	1.027	0.996	0.432	0.550	0.551
	SR	1.294	1.056	1.312	1.220	0.543	0.596	0.648
Fail2Drive-OOD								
Method	**Selection metric**	**NMAE ↓**				**SP ↑**		
		HM	**DR**	**SR**	**Mean**	**HM**	**DR**	**SR**
Ours	All metrics	0.421	0.295	0.451	0.389	0.865	0.869	0.858
Random	None	0.922	0.501	0.958	0.794	0.777	0.865	0.728
tinyBenchmarks	HM	0.687	0.637	0.707	0.677	0.702	0.693	0.664
	DR	0.979	0.528	1.107	0.872	0.358	0.675	0.152
	SR	0.889	0.660	0.968	0.839	0.544	0.544	0.544
ZipBench	HM	0.930	0.766	0.973	0.889	0.635	0.683	0.573
	DR	0.830	0.532	0.874	0.745	0.492	0.651	0.539
	SR	0.784	0.612	0.808	0.735	0.635	0.731	0.586
mRMR	HM	1.853	0.389	0.590	0.944	0.129	0.539	0.675
	DR	0.710	0.352	0.440	0.501	0.632	0.482	0.424
	SR	0.657	0.611	0.709	0.659	0.407	0.396	0.369
AnchorPoints	HM	0.710	0.465	0.716	0.630	0.816	0.799	0.753
	DR	0.633	0.496	0.704	0.611	0.795	0.743	0.852
	SR	0.916	0.633	0.987	0.845	0.710	0.711	0.706
EssenceBench	HM	0.723	0.539	0.758	0.673	0.704	0.768	0.784
	DR	0.723	0.539	0.758	0.673	0.704	0.768	0.784
	SR	0.614	0.476	0.652	0.581	0.725	0.746	0.796

Table 25: **Cross-metric fidelity on LiveIdea-Core3 and WorkBench at 20% retention.** Ours and Random do not use metric-specific selection. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

LiveIdea-Core3								
Method	**Selection metric**	**NMAE ↓**				**SP ↑**		
		Originality	**Feasibility**	**Clarity**	**Mean**	**Originality**	**Feasibility**	**Clarity**
Ours	All metrics	0.071	0.130	0.107	0.103	0.995	0.977	0.974
Random	None	0.071	0.149	0.103	0.108	0.995	0.976	0.979
tinyBenchmarks	Originality	0.103	0.382	0.208	0.231	0.990	0.963	0.956
	Feasibility	0.117	0.187	0.184	0.162	0.985	0.952	0.946
	Clarity	0.122	0.274	0.179	0.191	0.987	0.965	0.941
ZipBench	Originality	0.114	0.228	0.171	0.171	0.988	0.949	0.954
	Feasibility	0.138	0.334	0.154	0.208	0.986	0.942	0.949
	Clarity	0.111	0.228	0.155	0.165	0.988	0.948	0.956
mRMR	Originality	0.136	0.442	0.499	0.359	0.994	0.946	0.950
	Feasibility	0.096	0.486	0.514	0.365	0.996	0.960	0.929
	Clarity	0.117	0.414	0.474	0.335	0.994	0.958	0.954
AnchorPoints	Originality	0.091	0.151	0.120	0.121	0.994	0.976	0.974
	Feasibility	0.097	0.177	0.142	0.139	0.995	0.969	0.968
	Clarity	0.088	0.212	0.145	0.149	0.992	0.976	0.968
EssenceBench	Originality	0.075	0.152	0.086	0.104	0.996	0.986	0.983
	Feasibility	0.059	0.161	0.085	0.102	0.997	0.983	0.976
	Clarity	0.073	0.113	0.107	0.098	0.996	0.988	0.967

WorkBench								
Method	**Selection metric**	**NMAE ↓**				**SP ↑**		
		Correct	**Safety**	**Exact match**	**Mean**	**Correct**	**Safety**	**Exact match**
Ours	All metrics	0.129	0.158	0.130	0.139	0.986	0.961	0.984
Random	None	0.138	0.179	0.183	0.166	0.983	0.940	0.981
tinyBenchmarks	Correct	0.179	0.273	0.227	0.226	0.977	0.939	0.973
	Safety	0.189	0.239	0.218	0.215	0.971	0.904	0.971
	Exact match	0.221	0.310	0.217	0.249	0.982	0.942	0.974
ZipBench	Correct	0.183	0.268	0.217	0.223	0.986	0.954	0.963
	Safety	0.263	0.302	0.270	0.278	0.966	0.942	0.961
	Exact match	0.211	0.258	0.224	0.231	0.972	0.961	0.962
mRMR	Correct	0.323	0.765	0.351	0.480	0.913	-0.421	0.874
	Safety	0.323	0.765	0.351	0.480	0.913	-0.421	0.874
	Exact match	0.323	0.765	0.351	0.480	0.913	-0.421	0.874
AnchorPoints	Correct	0.749	0.940	0.801	0.830	0.881	0.847	0.807
	Safety	1.559	2.399	1.521	1.826	0.810	0.729	0.680
	Exact match	0.726	0.933	0.861	0.840	0.891	0.825	0.843
EssenceBench	Correct	0.875	1.012	0.796	0.894	0.977	0.975	0.951
	Safety	1.607	2.388	1.086	1.694	0.989	0.954	0.925
	Exact match	0.464	0.321	0.344	0.376	0.976	0.965	0.951

Table 26: **Cross-metric fidelity on benchmarks with 4 metrics at 20% retention.** Each baseline selects using the indicated metric and is evaluated on every component. Underlining marks diagonal entries. Ours and Random do not use metric-specific selection. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Bench2Drive

Method	Selection metric	NMAE ↓					SP ↑			
		Efficiency	Score	Smoothness	Success	Mean	Efficiency	Score	Smoothness	Success
Ours	All metrics	0.053	0.435	0.159	0.448	0.274	0.782	0.845	0.911	0.835
Random	None	0.139	0.631	0.253	0.529	0.388	0.733	0.857	0.933	0.844
tinyBenchmarks	Efficiency	0.077	0.714	0.339	0.706	0.459	0.766	0.770	0.824	0.741
	Score	0.143	0.782	0.346	0.929	0.550	0.845	0.755	0.824	0.714
	Smoothness	0.123	1.176	0.255	1.174	0.682	0.658	0.806	0.836	0.656
	Success	0.121	0.905	0.256	0.912	0.549	0.725	0.610	0.866	0.521
ZipBench	Efficiency	0.111	0.872	0.343	0.818	0.536	0.758	0.748	0.862	0.787
	Score	0.100	0.831	0.338	0.860	0.532	0.668	0.750	0.869	0.747
	Smoothness	0.174	0.944	0.299	0.973	0.598	0.847	0.783	0.845	0.825
	Success	0.229	0.807	0.315	0.821	0.543	0.689	0.815	0.872	0.726
mRMR	Efficiency	0.243	0.685	0.503	0.558	0.497	0.581	0.611	0.910	0.542
	Score	0.205	0.378	0.516	0.361	0.365	0.660	0.871	0.939	0.863
	Smoothness	0.171	0.444	0.516	0.427	0.390	0.617	0.721	0.896	0.840
	Success	0.271	0.501	0.632	0.440	0.461	0.553	0.786	0.863	0.763
AnchorPoints	Efficiency	0.075	0.816	0.523	0.951	0.591	0.647	0.899	0.788	0.826
	Score	0.446	1.913	0.734	2.066	1.290	0.672	0.631	0.849	0.641
	Smoothness	0.202	1.150	0.343	1.027	0.680	0.775	0.731	0.741	0.809
	Success	0.509	2.046	0.954	2.477	1.496	0.591	0.644	0.801	0.660
EssenceBench	Efficiency	0.068	0.610	0.110	0.559	0.337	0.685	0.918	0.921	0.892
	Score	0.107	1.190	0.403	1.478	0.795	0.810	0.921	0.957	0.910
	Smoothness	0.070	0.674	0.157	0.674	0.394	0.617	0.846	0.817	0.918
	Success	0.134	0.863	0.160	1.195	0.588	0.821	0.907	0.975	0.920

LiveIdea-Full4

Method	Selection metric	NMAE ↓					SP ↑			
		Originality	Feasibility	Clarity	Fluency	Mean	Originality	Feasibility	Clarity	Fluency
Ours	All metrics	0.142	0.234	0.178	0.345	0.225	0.990	0.952	0.944	0.906
Random	None	0.135	0.241	0.227	0.552	0.289	0.988	0.949	0.923	0.883
tinyBenchmarks	Originality	0.222	0.440	0.348	0.670	0.420	0.965	0.869	0.870	0.741
	Feasibility	0.190	0.371	0.354	0.631	0.387	0.969	0.861	0.899	0.770
	Clarity	0.192	0.494	0.319	0.585	0.398	0.972	0.859	0.896	0.799
	Fluency	0.226	0.461	0.386	0.548	0.405	0.966	0.880	0.876	0.770
ZipBench	Originality	0.164	0.429	0.356	0.687	0.409	0.975	0.888	0.894	0.827
	Feasibility	0.229	0.369	0.323	0.669	0.398	0.967	0.890	0.901	0.776
	Clarity	0.231	0.470	0.330	0.576	0.402	0.966	0.896	0.895	0.817
	Fluency	0.240	0.415	0.334	0.609	0.400	0.970	0.891	0.878	0.801
mRMR	Originality	0.181	0.418	0.539	0.527	0.416	0.981	0.911	0.865	0.793
	Feasibility	0.163	0.489	0.591	0.434	0.419	0.987	0.848	0.705	0.873
	Clarity	0.189	0.495	0.573	0.514	0.443	0.987	0.846	0.787	0.887
	Fluency	0.168	0.449	0.576	0.577	0.442	0.979	0.918	0.830	0.773
AnchorPoints	Originality	0.169	0.306	0.248	0.476	0.300	0.974	0.923	0.947	0.888
	Feasibility	0.193	0.318	0.275	0.377	0.290	0.985	0.942	0.947	0.916
	Clarity	0.129	0.261	0.229	0.459	0.269	0.989	0.947	0.938	0.889
	Fluency	0.141	0.343	0.257	0.544	0.321	0.982	0.928	0.932	0.847
EssenceBench	Originality	0.173	0.284	0.213	0.386	0.264	0.993	0.920	0.934	0.922
	Feasibility	0.146	0.230	0.211	0.426	0.253	0.988	0.945	0.947	0.896
	Clarity	0.159	0.227	0.204	0.426	0.254	0.985	0.951	0.955	0.867
	Fluency	0.133	0.477	0.220	0.447	0.319	0.987	0.933	0.952	0.909

Table 27: **Cross-metric fidelity on RouterArena at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	**Acc.**	**Cost**	**Mean**
Ours	All metrics	0.11/0.96	0.07/0.98	0.09
Random	None	0.19/1.00	0.10/0.98	0.15
tinyBenchmarks	Acc.	0.43/0.89	0.14/0.94	0.28
	Cost	0.55/0.90	0.16/0.98	0.35
ZipBench	Acc.	0.37/0.96	0.12/0.96	0.25
	Cost	0.43/0.92	0.11/0.98	0.27
mRMR	Acc.	0.68/0.83	1.36/0.89	1.02
	Cost	0.62/0.77	1.48/0.77	1.05
AnchorPoints	Acc.	9.44/0.46	1.13/0.81	5.29
	Cost	0.43/0.95	0.13/0.99	0.28
EssenceBench	Acc.	6.39/1.00	0.48/0.94	3.43
	Cost	0.45/0.92	0.38/0.94	0.42

Table 28: **Cross-metric fidelity on ConStory at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	**Char.**	**Fact.**	**Narr.**	**Time.**	**World**	**Mean**
Ours	All metrics	0.51/0.74	0.20/0.97	0.27/0.79	0.23/0.96	0.28/0.92	0.30
Random	None	0.46/0.79	0.23/0.96	0.24/0.79	0.19/0.97	0.36/0.90	0.29
tinyBenchmarks	Char.	0.75/0.67	0.31/0.93	0.50/0.71	0.27/0.93	0.55/0.71	0.48
	Fact.	0.89/0.72	0.44/0.87	0.60/0.77	0.31/0.92	0.71/0.74	0.59
	Narr.	0.71/0.74	0.33/0.87	0.38/0.77	0.25/0.96	0.50/0.85	0.43
	Time.	1.04/0.64	0.38/0.88	0.43/0.78	0.34/0.91	0.62/0.79	0.56
	World	0.56/0.79	0.38/0.91	0.43/0.76	0.29/0.90	0.62/0.80	0.45
ZipBench	Char.	0.92/0.70	0.28/0.91	0.40/0.73	0.26/0.93	0.63/0.79	0.50
	Fact.	0.89/0.58	0.38/0.88	0.44/0.56	0.45/0.89	0.82/0.75	0.59
	Narr.	0.77/0.73	0.27/0.90	0.37/0.63	0.28/0.91	0.59/0.86	0.45
	Time.	0.73/0.63	0.29/0.94	0.46/0.70	0.30/0.92	0.54/0.81	0.47
	World	0.74/0.67	0.31/0.92	0.43/0.62	0.33/0.93	0.58/0.81	0.48
mRMR	Char.	0.79/0.03	0.69/0.87	0.65/0.15	0.72/0.90	0.75/0.50	0.72
	Fact.	0.79/0.17	0.78/0.72	0.59/0.53	0.66/0.94	0.59/0.83	0.68
	Narr.	0.85/0.03	0.68/0.90	0.59/0.82	0.62/0.92	0.61/0.75	0.67
	Time.	0.85/0.26	0.66/0.89	0.65/-0.16	0.85/0.31	0.79/0.38	0.76
	World	0.80/0.25	0.74/0.81	0.58/0.33	0.56/0.92	0.83/0.63	0.70
AnchorPoints	Char.	4.61/0.36	2.44/0.38	2.52/0.34	1.99/0.36	3.77/0.14	3.07
	Fact.	1.61/0.39	1.22/0.60	1.18/0.50	0.95/0.63	1.25/0.49	1.24
	Narr.	4.68/0.40	2.39/0.33	1.31/-0.05	2.64/0.48	1.33/0.11	2.47
	Time.	3.74/0.52	1.32/0.50	0.77/0.53	1.25/0.68	1.89/0.54	1.80
	World	1.52/0.46	1.78/0.37	0.76/0.38	1.83/0.61	1.48/0.35	1.47
EssenceBench	Char.	0.54/0.91	0.21/0.96	0.42/0.65	0.25/0.97	0.81/0.93	0.45
	Fact.	0.73/0.91	0.28/0.96	0.22/0.81	0.30/0.96	0.79/0.95	0.47
	Narr.	0.58/0.90	0.21/0.96	0.40/0.66	0.24/0.97	0.79/0.93	0.44
	Time.	0.67/0.80	0.23/0.96	0.27/0.78	0.39/0.97	0.72/0.95	0.45
	World	0.60/0.90	0.22/0.96	0.37/0.71	0.28/0.97	0.85/0.93	0.46

Table 29: **Cross-metric fidelity on SoB-Audio at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	**Schema**	**LeafEM**	**ValueF1**	**PathRec**	**PathF1**	**Mean**
Ours	All metrics	0.55/0.77	0.73/0.83	0.54/0.92	0.49/0.84	0.58/0.79	0.58
Random	None	0.56/0.73	1.14/0.75	0.76/0.86	0.50/0.82	0.57/0.73	0.70
tinyBenchmarks	Schema	0.82/0.70	2.02/0.65	1.45/0.75	0.90/0.73	0.96/0.63	1.23
	LeafEM	0.89/0.64	1.18/0.59	0.93/0.73	0.98/0.69	0.97/0.58	0.99
	ValueF1	0.88/0.66	1.53/0.57	0.97/0.79	0.76/0.67	0.94/0.55	1.02
	PathRec	0.71/0.74	2.37/0.73	1.68/0.78	0.75/0.73	0.89/0.70	1.28
	PathF1	0.89/0.58	1.60/0.50	0.91/0.62	0.84/0.51	0.95/0.49	1.04
ZipBench	Schema	0.61/0.79	1.45/0.58	0.62/0.79	0.55/0.82	0.64/0.75	0.78
	LeafEM	0.78/0.68	1.44/0.70	0.94/0.81	0.80/0.69	0.87/0.66	0.96
	ValueF1	0.71/0.76	2.15/0.47	1.06/0.72	0.71/0.80	0.84/0.70	1.10
	PathRec	0.65/0.73	1.50/0.51	1.10/0.72	0.66/0.82	0.70/0.66	0.92
	PathF1	0.75/0.76	1.98/0.58	1.15/0.77	0.69/0.80	0.81/0.75	1.07
mRMR	Schema	0.51/0.83	0.78/0.43	0.75/0.60	0.49/0.82	0.56/0.79	0.62
	LeafEM	0.72/0.66	0.66/0.79	0.56/0.82	0.84/0.50	0.99/0.16	0.75
	ValueF1	0.53/0.65	0.78/0.76	0.46/0.87	0.43/0.80	0.58/0.64	0.56
	PathRec	0.53/0.74	0.84/0.67	0.46/0.89	0.43/0.78	0.53/0.66	0.56
	PathF1	0.89/0.62	0.84/0.60	0.61/0.68	0.69/0.71	0.96/0.48	0.80
AnchorPoints	Schema	2.28/0.76	3.55/0.32	3.23/0.46	2.53/0.64	2.67/0.60	2.85
	LeafEM	0.56/0.81	1.40/0.68	0.60/0.89	0.49/0.86	0.55/0.78	0.72
	ValueF1	1.03/0.76	1.17/0.66	0.83/0.72	0.89/0.75	1.00/0.62	0.98
	PathRec	1.30/0.67	2.21/0.46	2.15/0.50	1.35/0.76	1.50/0.62	1.70
	PathF1	1.72/0.64	2.65/0.48	1.76/0.41	1.52/0.70	1.82/0.66	1.89
EssenceBench	Schema	1.83/0.94	2.24/0.47	2.39/0.75	1.82/0.82	2.01/0.88	2.06
	LeafEM	0.47/0.85	0.83/0.49	0.61/0.76	0.44/0.86	0.54/0.77	0.58
	ValueF1	0.59/0.82	1.55/0.74	0.88/0.80	0.52/0.82	0.61/0.76	0.83
	PathRec	0.73/0.91	1.58/0.37	1.10/0.71	0.94/0.88	0.84/0.83	1.04
	PathF1	0.48/0.89	0.78/0.74	0.61/0.87	0.56/0.88	0.55/0.83	0.60

Table 30: **Cross-metric fidelity on SoB-Image at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	**PathRec**	**LeafEM**	**PathF1**	**Schema**	**ValueF1**	**Mean**
Ours	All metrics	0.51/0.77	0.38/0.88	0.56/0.74	0.54/0.80	0.58/0.67	0.51
Random	None	1.01/0.68	1.20/0.77	1.10/0.57	1.00/0.62	1.57/0.62	1.17
tinyBenchmarks	PathRec	0.95/0.60	1.37/0.69	1.11/0.54	0.95/0.63	1.90/0.52	1.25
	LeafEM	0.87/0.64	1.19/0.71	0.94/0.51	0.93/0.59	1.20/0.64	1.03
	PathF1	1.16/0.62	1.65/0.70	1.30/0.52	1.10/0.62	1.98/0.59	1.44
	Schema	1.06/0.60	1.15/0.69	1.21/0.48	1.02/0.55	1.37/0.32	1.16
	ValueF1	1.14/0.64	1.22/0.65	1.28/0.52	1.14/0.61	1.50/0.55	1.26
ZipBench	PathRec	1.11/0.64	1.65/0.71	1.25/0.57	1.18/0.58	1.62/0.54	1.36
	LeafEM	1.76/0.62	1.31/0.45	2.05/0.52	1.80/0.57	1.76/0.43	1.73
	PathF1	0.82/0.74	0.83/0.77	0.87/0.67	0.86/0.75	1.21/0.62	0.92
	Schema	1.04/0.72	1.38/0.80	1.22/0.65	1.00/0.69	1.66/0.68	1.26
	ValueF1	1.30/0.61	1.50/0.77	1.45/0.55	1.36/0.56	1.47/0.41	1.42
mRMR	PathRec	0.72/0.56	0.71/0.64	0.99/0.54	0.83/0.74	0.79/0.56	0.81
	LeafEM	0.81/0.35	0.45/0.87	0.68/0.26	0.68/0.56	0.84/0.44	0.69
	PathF1	0.74/0.35	0.50/0.85	0.73/0.25	0.74/0.29	0.60/0.69	0.66
	Schema	0.73/0.51	0.66/0.58	0.88/0.39	0.88/0.42	0.70/0.58	0.77
	ValueF1	0.80/0.27	0.46/0.87	0.48/0.68	0.54/0.72	0.54/0.76	0.56
AnchorPoints	PathRec	2.58/0.65	2.86/0.61	2.82/0.61	2.62/0.66	3.52/0.38	2.88
	LeafEM	0.67/0.78	1.03/0.73	0.77/0.65	0.70/0.70	1.15/0.68	0.87
	PathF1	2.11/0.70	3.23/0.46	2.25/0.66	2.11/0.71	3.72/0.52	2.69
	Schema	6.89/0.43	5.90/0.31	7.45/0.46	6.70/0.50	8.73/0.31	7.13
	ValueF1	1.39/0.81	1.24/0.47	1.64/0.77	1.34/0.80	1.54/0.63	1.43
EssenceBench	PathRec	2.72/0.89	5.52/0.85	2.85/0.82	1.99/0.87	6.84/0.61	3.98
	LeafEM	1.21/0.70	1.80/0.80	1.28/0.73	1.24/0.78	1.53/0.47	1.41
	PathF1	2.58/0.89	6.13/0.64	3.16/0.83	2.06/0.89	7.31/0.71	4.25
	Schema	7.94/0.65	5.63/0.15	8.91/0.71	7.97/0.73	9.16/0.26	7.92
	ValueF1	0.76/0.81	2.54/0.82	0.86/0.74	0.75/0.83	2.26/0.61	1.43

Table 31: **Cross-metric fidelity on SoB-Text at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	**Schema**	**LeafEM**	**ValueF1**	**PathRec**	**PathF1**	**JSONEM**	**Mean**
Ours	All metrics	0.14/0.97	0.16/0.95	0.18/0.96	0.14/0.97	0.14/0.97	0.17/0.97	0.16
Random	None	0.19/0.95	0.21/0.97	0.19/0.96	0.18/0.96	0.18/0.95	0.32/0.96	0.21
tinyBenchmarks	Schema	0.36/0.91	0.45/0.83	0.44/0.83	0.36/0.87	0.39/0.89	0.65/0.90	0.44
	LeafEM	0.37/0.91	0.34/0.90	0.35/0.88	0.36/0.89	0.35/0.89	0.70/0.91	0.41
	ValueF1	0.31/0.92	0.45/0.79	0.46/0.79	0.31/0.93	0.31/0.90	0.50/0.87	0.39
	PathRec	0.35/0.84	0.53/0.86	0.50/0.83	0.38/0.81	0.37/0.78	0.56/0.88	0.45
	PathF1	0.39/0.81	0.71/0.85	0.71/0.87	0.38/0.81	0.43/0.81	0.51/0.86	0.52
	JSONEM	0.37/0.93	0.38/0.90	0.42/0.88	0.38/0.93	0.40/0.91	0.49/0.86	0.41
ZipBench	Schema	0.33/0.85	0.45/0.86	0.57/0.86	0.38/0.87	0.37/0.83	0.35/0.89	0.41
	LeafEM	0.37/0.89	0.41/0.87	0.44/0.88	0.37/0.86	0.38/0.88	0.44/0.89	0.40
	ValueF1	0.45/0.86	0.51/0.89	0.52/0.87	0.45/0.86	0.47/0.85	0.52/0.93	0.49
	PathRec	0.39/0.88	0.53/0.87	0.51/0.87	0.39/0.89	0.45/0.88	0.42/0.90	0.45
	PathF1	0.38/0.85	0.34/0.89	0.43/0.90	0.38/0.85	0.38/0.84	0.45/0.90	0.39
	JSONEM	0.34/0.94	0.46/0.91	0.47/0.91	0.38/0.92	0.36/0.92	0.43/0.89	0.41
mRMR	Schema	0.51/0.88	0.49/0.84	0.54/0.88	0.52/0.88	0.52/0.86	0.46/0.94	0.51
	LeafEM	0.53/0.93	0.55/0.77	0.59/0.75	0.52/0.91	0.52/0.91	0.46/0.96	0.53
	ValueF1	0.62/0.83	0.55/0.76	0.60/0.73	0.63/0.68	0.63/0.69	0.48/0.92	0.58
	PathRec	0.58/0.97	0.52/0.94	0.54/0.94	0.56/0.95	0.56/0.95	0.50/0.93	0.54
	PathF1	0.56/0.95	0.52/0.89	0.54/0.90	0.56/0.95	0.55/0.93	0.51/0.93	0.54
	JSONEM	0.51/0.88	0.49/0.84	0.54/0.88	0.52/0.88	0.52/0.86	0.46/0.94	0.51
AnchorPoints	Schema	16.19/0.72	16.74/0.14	18.93/0.23	16.78/0.66	17.39/0.59	10.12/0.25	16.02
	LeafEM	0.77/0.86	2.38/0.46	1.80/0.57	0.86/0.80	0.88/0.79	4.11/0.56	1.80
	ValueF1	0.98/0.84	2.99/0.58	2.02/0.64	1.10/0.80	1.21/0.80	5.14/0.74	2.24
	PathRec	6.18/0.82	21.45/0.33	24.03/0.40	10.12/0.59	9.42/0.56	9.87/0.48	13.51
	PathF1	2.47/0.84	15.47/0.42	17.68/0.41	3.72/0.63	4.39/0.42	8.42/0.57	8.69
	JSONEM	1.32/0.68	3.82/0.35	4.35/0.41	1.52/0.61	1.76/0.54	5.67/0.42	3.07
EssenceBench	Schema	3.16/0.92	3.65/0.81	4.00/0.93	3.36/0.92	3.58/0.91	1.61/0.95	3.23
	LeafEM	0.37/0.97	2.81/0.95	1.53/0.97	0.39/0.99	0.39/0.96	5.31/0.95	1.80
	ValueF1	1.01/0.97	5.45/0.93	4.40/0.93	1.22/0.98	1.29/0.96	7.26/0.93	3.44
	PathRec	3.22/0.94	5.38/0.78	6.03/0.90	3.95/0.94	4.01/0.92	2.37/0.92	4.16
	PathF1	2.99/0.94	6.13/0.80	6.94/0.93	3.56/0.94	4.01/0.92	2.63/0.94	4.38
	JSONEM	0.16/0.97	0.96/0.96	0.84/0.98	0.17/0.97	0.17/0.96	0.49/0.94	0.46

Table 32: **Cross-metric fidelity on MMDocRAG-MM at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	QuoteP	QuoteR	QuoteF1	Fluency	CitQual	TxtImg	Logic	Fact.	Mean
Ours	All metrics	0.09/0.99	0.05/0.99	0.07/0.99	0.06/1.00	0.06/0.99	0.06/0.99	0.07/1.00	0.06/0.99	0.06
Random	None	0.11/0.99	0.06/1.00	0.07/0.99	0.07/0.99	0.06/0.99	0.05/0.99	0.08/0.99	0.09/0.99	0.07
tinyBenchmarks	QuoteP	0.19/0.97	0.12/0.99	0.15/0.98	0.12/0.97	0.12/0.98	0.11/0.98	0.15/0.97	0.13/0.98	0.14
	QuoteR	0.21/0.95	0.13/0.98	0.16/0.98	0.14/0.98	0.15/0.97	0.15/0.98	0.13/0.98	0.14/0.97	0.15
	QuoteF1	0.15/0.98	0.10/0.99	0.11/0.98	0.13/0.98	0.14/0.98	0.12/0.98	0.18/0.97	0.19/0.98	0.14
	Fluency	0.27/0.93	0.15/0.98	0.18/0.97	0.16/0.96	0.17/0.97	0.14/0.98	0.20/0.98	0.16/0.98	0.18
	CitQual	0.23/0.96	0.13/0.99	0.17/0.98	0.12/0.98	0.12/0.98	0.11/0.98	0.15/0.98	0.16/0.98	0.15
	TxtImg	0.20/0.97	0.12/0.99	0.15/0.97	0.13/0.98	0.13/0.98	0.12/0.98	0.17/0.98	0.17/0.97	0.15
	Logic	0.19/0.98	0.11/0.98	0.13/0.98	0.13/0.97	0.15/0.97	0.11/0.98	0.16/0.98	0.14/0.98	0.14
	Fact.	0.15/0.97	0.17/0.98	0.14/0.98	0.12/0.98	0.13/0.97	0.11/0.99	0.15/0.99	0.13/0.98	0.14
ZipBench	QuoteP	0.15/0.96	0.11/0.99	0.12/0.98	0.10/0.98	0.10/0.98	0.09/0.99	0.11/0.99	0.14/0.98	0.11
	QuoteR	0.15/0.98	0.09/0.99	0.11/0.99	0.09/0.99	0.12/0.98	0.11/0.98	0.15/0.98	0.12/0.98	0.12
	QuoteF1	0.13/0.98	0.09/0.99	0.10/0.99	0.11/0.98	0.14/0.97	0.11/0.98	0.13/0.99	0.11/0.98	0.12
	Fluency	0.17/0.98	0.11/0.99	0.14/0.98	0.15/0.96	0.15/0.98	0.13/0.98	0.18/0.98	0.18/0.98	0.15
	CitQual	0.17/0.98	0.11/0.99	0.12/0.97	0.11/0.97	0.12/0.98	0.10/0.98	0.12/0.98	0.11/0.99	0.12
	TxtImg	0.15/0.97	0.08/0.99	0.11/0.99	0.12/0.97	0.11/0.99	0.09/0.98	0.13/0.99	0.12/0.98	0.11
	Logic	0.17/0.97	0.15/0.99	0.16/0.98	0.10/0.98	0.14/0.98	0.09/0.99	0.11/0.99	0.11/0.98	0.13
	Fact.	0.16/0.97	0.09/0.99	0.12/0.98	0.11/0.98	0.11/0.98	0.08/0.99	0.11/0.99	0.11/0.98	0.11
mRMR	QuoteP	0.20/0.94	0.20/0.99	0.20/0.99	0.12/0.98	0.15/0.99	0.12/0.99	0.14/0.98	0.12/0.99	0.16
	QuoteR	0.26/0.91	0.30/0.98	0.28/0.97	0.15/0.97	0.16/0.99	0.14/0.99	0.12/0.99	0.11/0.98	0.19
	QuoteF1	0.20/0.94	0.21/0.99	0.21/0.98	0.10/0.96	0.13/0.99	0.13/0.99	0.12/0.99	0.12/0.99	0.15
	Fluency	0.25/0.96	0.17/0.99	0.18/0.99	0.12/0.98	0.15/0.99	0.12/0.99	0.14/0.99	0.12/0.99	0.16
	CitQual	0.19/0.99	0.10/1.00	0.12/0.99	0.07/0.99	0.31/0.99	0.17/0.99	0.11/0.99	0.13/0.98	0.15
	TxtImg	0.21/0.99	0.07/1.00	0.12/1.00	0.07/1.00	0.19/0.99	0.18/0.99	0.17/0.99	0.10/0.98	0.14
	Logic	0.14/0.98	0.07/0.99	0.09/0.99	0.13/0.99	0.22/0.99	0.16/0.99	0.21/0.99	0.13/0.99	0.14
	Fact.	0.11/0.98	0.10/1.00	0.09/0.99	0.08/0.98	0.12/0.99	0.08/0.99	0.11/0.99	0.15/0.97	0.10
AnchorPoints	QuoteP	0.21/0.94	0.13/0.98	0.15/0.96	0.14/0.97	0.17/0.97	0.14/0.98	0.19/0.97	0.22/0.97	0.17
	QuoteR	0.40/0.85	0.33/0.96	0.33/0.93	0.24/0.93	0.30/0.95	0.24/0.93	0.31/0.95	0.32/0.95	0.31
	QuoteF1	0.17/0.96	0.14/0.99	0.14/0.97	0.13/0.97	0.14/0.97	0.13/0.98	0.18/0.97	0.19/0.98	0.15
	Fluency	0.26/0.95	0.15/0.98	0.19/0.96	0.17/0.97	0.20/0.95	0.15/0.97	0.20/0.97	0.19/0.97	0.19
	CitQual	0.12/0.98	0.08/0.99	0.09/0.99	0.08/0.99	0.08/0.99	0.07/0.99	0.09/0.99	0.09/0.99	0.09
	TxtImg	0.11/0.99	0.09/0.99	0.10/0.99	0.07/0.99	0.10/0.99	0.09/0.99	0.12/0.99	0.15/0.99	0.10
	Logic	0.12/0.98	0.08/0.99	0.09/0.99	0.09/0.99	0.09/0.99	0.08/0.99	0.10/0.99	0.10/0.99	0.10
	Fact.	0.11/0.99	0.06/0.99	0.08/0.99	0.08/0.99	0.10/0.98	0.08/0.99	0.09/0.99	0.09/0.99	0.09
EssenceBench	QuoteP	0.14/0.98	0.09/0.99	0.11/0.98	0.07/0.99	0.09/0.99	0.06/0.99	0.08/0.99	0.09/0.99	0.09
	QuoteR	0.14/0.98	0.11/0.99	0.12/0.99	0.06/0.99	0.08/0.98	0.06/0.99	0.06/0.99	0.08/0.99	0.09
	QuoteF1	0.22/0.99	0.07/0.99	0.12/0.99	0.08/0.99	0.13/1.00	0.14/0.99	0.17/0.99	0.25/0.99	0.15
	Fluency	0.15/0.99	0.06/1.00	0.10/0.99	0.20/0.98	0.21/0.99	0.19/0.99	0.36/0.98	0.41/0.98	0.21
	CitQual	0.11/0.98	0.06/0.99	0.08/0.99	0.05/0.99	0.10/0.99	0.10/0.99	0.14/0.99	0.19/0.99	0.10
	TxtImg	0.11/0.97	0.08/0.98	0.09/0.99	0.07/0.99	0.09/0.99	0.08/0.99	0.15/0.99	0.18/1.00	0.11
	Logic	0.13/0.98	0.06/0.99	0.06/0.99	0.08/0.98	0.06/0.99	0.05/0.99	0.10/0.98	0.10/0.99	0.08
	Fact.	0.10/0.99	0.07/0.99	0.08/0.99	0.13/0.99	0.18/0.99	0.17/0.99	0.39/0.99	0.53/0.99	0.21

Table 33: **Cross-metric fidelity on MMDocRAG-Text at 20% retention.** Each cell reports NMAE / SP. Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	QuoteP	QuoteR	QuoteF1	Fluency	CitQual	TxtImg	Logic	Fact.	Mean
Ours	All metrics	0.09/0.99	0.08/0.99	0.10/0.98	0.08/0.98	0.10/0.98	0.10/0.98	0.11/0.98	0.12/0.97	0.10
Random	None	0.12/0.99	0.09/0.99	0.13/0.98	0.09/0.98	0.11/0.98	0.11/0.97	0.12/0.98	0.12/0.98	0.11
tinyBenchmarks	QuoteP	0.13/0.98	0.13/0.98	0.15/0.95	0.13/0.97	0.16/0.96	0.15/0.96	0.14/0.97	0.17/0.95	0.14
	QuoteR	0.16/0.97	0.15/0.98	0.19/0.95	0.16/0.97	0.19/0.96	0.17/0.96	0.21/0.95	0.29/0.94	0.19
	QuoteF1	0.13/0.98	0.12/0.98	0.14/0.97	0.12/0.97	0.14/0.97	0.14/0.96	0.14/0.98	0.20/0.95	0.14
	Fluency	0.19/0.98	0.16/0.98	0.24/0.96	0.15/0.96	0.17/0.97	0.19/0.96	0.23/0.94	0.30/0.92	0.20
	CitQual	0.15/0.97	0.16/0.97	0.21/0.96	0.15/0.96	0.16/0.95	0.16/0.95	0.17/0.97	0.23/0.94	0.17
	TxtImg	0.18/0.98	0.13/0.98	0.19/0.95	0.15/0.96	0.19/0.95	0.19/0.95	0.20/0.96	0.23/0.95	0.18
	Logic	0.18/0.97	0.16/0.98	0.22/0.96	0.14/0.96	0.17/0.96	0.15/0.95	0.17/0.96	0.19/0.96	0.17
	Fact.	0.16/0.97	0.17/0.98	0.21/0.97	0.13/0.96	0.16/0.96	0.14/0.96	0.15/0.97	0.16/0.96	0.16
ZipBench	QuoteP	0.14/0.96	0.17/0.97	0.19/0.95	0.15/0.97	0.19/0.95	0.16/0.94	0.20/0.95	0.25/0.95	0.18
	QuoteR	0.13/0.97	0.13/0.98	0.16/0.97	0.16/0.97	0.20/0.95	0.17/0.96	0.21/0.96	0.27/0.94	0.18
	QuoteF1	0.16/0.96	0.17/0.98	0.20/0.95	0.12/0.97	0.17/0.95	0.15/0.95	0.18/0.97	0.20/0.95	0.17
	Fluency	0.27/0.98	0.18/0.98	0.29/0.95	0.14/0.95	0.17/0.96	0.19/0.93	0.19/0.95	0.23/0.94	0.21
	CitQual	0.19/0.98	0.16/0.98	0.23/0.96	0.16/0.97	0.19/0.96	0.21/0.95	0.19/0.96	0.22/0.95	0.19
	TxtImg	0.19/0.97	0.14/0.98	0.19/0.96	0.14/0.96	0.18/0.96	0.19/0.94	0.22/0.94	0.23/0.92	0.18
	Logic	0.20/0.97	0.13/0.98	0.21/0.97	0.16/0.96	0.18/0.94	0.18/0.93	0.19/0.95	0.21/0.94	0.18
	Fact.	0.20/0.97	0.14/0.98	0.20/0.94	0.14/0.98	0.18/0.96	0.16/0.95	0.17/0.96	0.24/0.93	0.18
mRMR	QuoteP	0.25/0.98	0.33/0.98	0.28/0.97	0.16/0.98	0.24/0.97	0.21/0.97	0.23/0.97	0.22/0.95	0.24
	QuoteR	0.28/0.97	0.43/0.98	0.37/0.93	0.17/0.97	0.23/0.97	0.20/0.97	0.21/0.95	0.23/0.95	0.26
	QuoteF1	0.23/0.99	0.32/0.99	0.29/0.97	0.17/0.98	0.24/0.97	0.20/0.97	0.17/0.98	0.23/0.95	0.23
	Fluency	0.28/0.98	0.38/0.98	0.30/0.96	0.24/0.97	0.28/0.96	0.25/0.97	0.24/0.97	0.26/0.95	0.28
	CitQual	0.28/0.97	0.33/0.98	0.28/0.97	0.19/0.98	0.30/0.98	0.21/0.97	0.24/0.97	0.23/0.95	0.26
	TxtImg	0.28/0.97	0.36/0.98	0.32/0.96	0.19/0.99	0.29/0.97	0.28/0.98	0.23/0.98	0.24/0.97	0.27
	Logic	0.28/0.97	0.36/0.99	0.28/0.97	0.19/0.98	0.26/0.98	0.24/0.98	0.25/0.96	0.23/0.95	0.26
	Fact.	0.24/0.97	0.31/0.98	0.29/0.97	0.19/0.97	0.24/0.96	0.22/0.97	0.23/0.95	0.29/0.93	0.25
AnchorPoints	QuoteP	0.12/0.98	0.11/0.99	0.14/0.97	0.11/0.98	0.13/0.98	0.14/0.97	0.15/0.98	0.19/0.97	0.14
	QuoteR	0.48/0.87	0.61/0.92	0.66/0.76	0.38/0.88	0.44/0.82	0.45/0.81	0.48/0.81	0.60/0.78	0.51
	QuoteF1	0.10/0.99	0.14/0.98	0.16/0.97	0.10/0.98	0.14/0.96	0.12/0.98	0.14/0.97	0.19/0.96	0.14
	Fluency	0.26/0.94	0.24/0.97	0.32/0.94	0.20/0.94	0.27/0.92	0.27/0.93	0.38/0.92	0.48/0.89	0.30
	CitQual	0.11/0.98	0.11/0.98	0.13/0.97	0.11/0.98	0.14/0.98	0.15/0.96	0.21/0.96	0.25/0.96	0.15
	TxtImg	0.13/0.98	0.11/0.99	0.14/0.98	0.11/0.97	0.13/0.96	0.11/0.97	0.13/0.97	0.14/0.97	0.13
	Logic	0.13/0.99	0.13/0.99	0.15/0.97	0.11/0.98	0.13/0.97	0.12/0.97	0.14/0.97	0.16/0.97	0.13
	Fact.	0.19/0.98	0.14/0.98	0.22/0.96	0.13/0.97	0.15/0.97	0.15/0.97	0.15/0.97	0.15/0.96	0.16
EssenceBench	QuoteP	0.11/0.98	0.19/0.99	0.14/0.96	0.11/0.98	0.14/0.98	0.15/0.97	0.22/0.97	0.25/0.96	0.16
	QuoteR	0.50/0.97	0.78/0.98	0.90/0.94	0.28/0.96	0.46/0.92	0.48/0.92	0.66/0.90	0.92/0.79	0.62
	QuoteF1	0.15/0.97	0.15/0.99	0.19/0.98	0.20/0.97	0.27/0.96	0.25/0.98	0.44/0.94	0.55/0.88	0.28
	Fluency	0.19/0.98	0.15/0.99	0.24/0.97	0.28/0.97	0.36/0.95	0.38/0.95	0.61/0.92	0.86/0.82	0.38
	CitQual	0.08/0.99	0.08/0.99	0.10/0.98	0.08/0.99	0.12/0.99	0.08/0.98	0.16/0.98	0.21/0.97	0.11
	TxtImg	0.09/0.99	0.11/0.98	0.14/0.97	0.13/0.99	0.16/0.97	0.16/0.98	0.32/0.96	0.38/0.92	0.19
	Logic	0.09/0.99	0.08/0.99	0.10/0.98	0.12/0.99	0.13/0.98	0.09/0.99	0.25/0.97	0.30/0.96	0.15
	Fact.	0.38/0.98	0.29/0.99	0.48/0.98	0.29/0.98	0.39/0.96	0.39/0.94	0.68/0.89	0.97/0.77	0.48

Table 34: **Cross-metric NMAE on Widget2Code at 20% retention.** Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	Margin	Content	Area	TextJac	Contrast	Local	Palette	Vibrancy	Polarity	SSIM	LP	Geom.	Mean
Ours	All metrics	0.27	0.10	0.10	0.10	0.51	0.24	0.32	0.30	0.33	0.19	0.10	0.04	0.22
Random	None	0.35	0.23	0.13	0.12	0.56	0.32	0.82	0.37	0.29	0.21	0.13	0.05	0.30
tinyBenchmarks	Margin	0.49	0.24	0.23	0.18	0.86	0.47	1.08	0.68	1.07	0.50	0.21	0.07	0.51
	Content	0.49	0.20	0.29	0.19	0.95	0.32	1.14	0.40	0.47	0.36	0.20	0.08	0.42
	Area	0.53	0.31	0.27	0.18	0.92	0.47	1.07	0.48	0.57	0.38	0.24	0.08	0.46
	TextJac	0.51	0.28	0.21	0.14	1.05	0.40	0.90	0.88	0.71	0.26	0.25	0.09	0.47
	Contrast	0.47	0.25	0.40	0.24	0.79	0.56	0.89	1.00	0.61	0.36	0.29	0.11	0.50
	Local	0.39	0.25	0.23	0.15	1.18	0.39	0.80	0.79	0.83	0.40	0.30	0.06	0.48
	Palette	0.55	0.35	0.28	0.24	1.34	0.51	0.84	0.97	0.54	0.38	0.22	0.10	0.53
	Vibrancy	0.65	0.49	0.22	0.22	1.04	0.40	0.98	0.51	0.69	0.47	0.26	0.09	0.50
	Polarity	0.54	0.30	0.31	0.20	1.34	0.38	0.97	0.85	0.49	0.30	0.25	0.14	0.51
	SSIM	0.59	0.39	0.29	0.16	1.16	0.39	1.19	1.11	0.66	0.44	0.31	0.08	0.56
	LP	0.51	0.30	0.25	0.21	1.03	0.56	1.11	0.54	0.58	0.35	0.17	0.12	0.48
	Geom.	0.44	0.25	0.25	0.21	1.69	0.50	0.89	0.79	0.48	0.29	0.17	0.07	0.50
ZipBench	Margin	0.44	0.25	0.25	0.29	0.96	0.39	0.98	0.56	0.49	0.23	0.32	0.07	0.43
	Content	0.32	0.21	0.20	0.24	1.67	0.41	0.77	0.48	0.56	0.41	0.16	0.09	0.46
	Area	0.36	0.22	0.23	0.18	0.90	0.40	0.81	0.56	0.81	0.58	0.34	0.07	0.46
	TextJac	0.47	0.30	0.21	0.14	0.97	0.29	0.77	0.73	0.31	0.42	0.42	0.10	0.43
	Contrast	0.48	0.30	0.31	0.31	1.01	0.34	1.06	0.63	0.64	0.44	0.27	0.06	0.49
	Local	0.57	0.33	0.18	0.18	1.23	0.40	0.87	0.72	0.43	0.47	0.22	0.07	0.47
	Palette	0.44	0.30	0.26	0.20	0.84	0.47	0.58	0.54	0.65	0.33	0.20	0.10	0.41
	Vibrancy	0.63	0.46	0.26	0.18	1.50	0.43	0.70	0.62	0.61	0.31	0.26	0.11	0.51
	Polarity	0.24	0.28	0.20	0.26	1.86	0.25	0.93	0.83	0.29	0.36	0.28	0.07	0.49
	SSIM	0.47	0.38	0.22	0.17	0.85	0.32	0.74	0.45	0.52	0.20	0.13	0.09	0.38
	LP	0.37	0.22	0.19	0.22	1.22	0.31	0.60	0.55	0.51	0.36	0.24	0.07	0.41
	Geom.	0.32	0.29	0.27	0.13	1.09	0.40	0.91	0.56	0.37	0.38	0.30	0.06	0.42
mRMR	Margin	0.27	0.31	0.24	0.09	0.58	0.25	0.65	0.34	0.27	0.29	0.12	0.06	0.29
	Content	0.33	0.25	0.26	0.16	0.50	0.20	0.66	0.38	0.39	0.12	0.16	0.08	0.29
	Area	0.31	0.32	0.30	0.14	0.39	0.16	0.72	0.28	0.37	0.28	0.18	0.07	0.29
	TextJac	0.36	0.30	0.29	0.15	0.45	0.30	0.49	0.31	0.40	0.22	0.19	0.06	0.29
	Contrast	0.40	0.19	0.17	0.19	0.55	0.24	0.43	0.20	0.18	0.24	0.13	0.08	0.25
	Local	0.45	0.30	0.17	0.10	0.53	0.32	0.19	0.21	0.19	0.17	0.14	0.07	0.24
	Palette	0.39	0.19	0.31	0.15	0.60	0.25	0.43	0.36	0.27	0.23	0.15	0.05	0.28
	Vibrancy	0.42	0.23	0.13	0.14	0.40	0.12	0.74	0.29	0.32	0.25	0.17	0.03	0.27
	Polarity	0.35	0.22	0.19	0.11	0.63	0.25	0.46	0.29	0.42	0.34	0.16	0.06	0.29
	SSIM	0.33	0.14	0.17	0.13	0.67	0.24	0.47	0.35	0.30	0.16	0.12	0.05	0.26
	LP	0.48	0.15	0.26	0.17	0.71	0.18	0.28	0.41	0.31	0.18	0.16	0.04	0.28
	Geom.	0.38	0.25	0.23	0.15	0.47	0.21	0.37	0.36	0.36	0.18	0.19	0.12	0.27
AnchorPoints	Margin	0.33	0.18	0.25	0.21	1.29	0.32	0.81	0.44	0.29	0.41	0.17	0.05	0.40
	Content	0.30	0.18	0.24	0.21	1.11	0.35	0.50	0.41	0.40	0.20	0.16	0.09	0.35
	Area	0.41	0.28	0.21	0.15	1.21	0.44	0.86	1.03	0.33	0.50	0.19	0.07	0.48
	TextJac	0.41	0.16	0.23	0.25	1.11	0.60	0.72	0.88	0.51	0.46	0.28	0.06	0.47
	Contrast	0.46	0.28	0.28	0.18	0.80	0.43	0.93	0.92	0.63	0.26	0.24	0.09	0.46
	Local	0.50	0.31	0.26	0.20	0.93	0.38	0.65	0.38	0.42	0.23	0.20	0.08	0.38
	Palette	0.54	0.28	0.19	0.15	0.85	0.38	0.43	0.46	0.32	0.16	0.18	0.08	0.33
	Vibrancy	0.44	0.23	0.16	0.10	0.84	0.31	0.64	0.33	0.37	0.24	0.20	0.07	0.33
	Polarity	0.61	0.28	0.23	0.18	1.90	0.43	0.92	0.48	0.52	0.49	0.22	0.07	0.53
	SSIM	0.50	0.33	0.32	0.20	1.62	0.36	0.87	0.65	0.55	0.29	0.25	0.09	0.50
	LP	0.32	0.30	0.21	0.22	1.43	0.27	0.62	0.67	0.38	0.35	0.34	0.07	0.43
	Geom.	0.54	0.38	0.25	0.22	0.89	0.50	1.08	0.56	0.48	0.36	0.22	0.04	0.46
EssenceBench	Margin	0.73	0.39	0.29	0.17	2.44	0.50	0.91	0.49	0.29	0.82	0.22	0.09	0.61
	Content	0.62	0.33	0.14	0.08	2.06	0.27	0.47	0.37	0.35	0.54	0.21	0.04	0.46
	Area	0.47	0.20	0.21	0.16	1.55	0.22	0.41	0.22	0.39	0.34	0.16	0.05	0.36
	TextJac	0.29	0.22	0.20	0.24	0.59	0.27	0.50	0.27	0.16	0.28	0.53	0.05	0.30
	Contrast	0.29	0.27	0.11	0.10	0.50	0.33	0.71	0.63	0.38	0.17	0.23	0.04	0.31
	Local	0.57	0.25	0.15	0.09	1.68	0.30	0.65	0.95	0.52	0.58	0.34	0.06	0.51
	Palette	0.20	0.34	0.15	0.12	0.44	0.21	1.23	1.31	0.38	0.20	0.22	0.06	0.41
	Vibrancy	0.24	0.15	0.25	0.14	0.51	0.26	0.46	1.06	0.48	0.34	0.29	0.06	0.35
	Polarity	0.30	0.24	0.12	0.11	1.07	0.34	0.53	0.25	0.54	0.32	0.16	0.04	0.34
	SSIM	0.89	0.44	0.18	0.09	4.13	0.96	0.74	0.57	0.49	1.75	0.43	0.04	0.89
	LP	0.33	0.20	0.16	0.13	0.73	0.19	0.33	0.73	0.21	0.45	0.43	0.04	0.33
	Geom.	0.47	0.37	0.23	0.19	1.02	0.14	0.31	0.29	0.30	0.28	0.20	0.19	0.33

Table 35: **Cross-metric SP on Widget2Code at 20% retention.** Underlining marks the selection metric; pale blue-gray marks each single-metric baseline's selection target with the lowest mean NMAE.

Method	Selection metric	Margin	Content	Area	TextJac	Contrast	Local	Palette	Vibrancy	Polarity	SSIM	LP	Geom.
Ours	All metrics	0.90	0.98	0.99	0.95	0.86	0.94	0.92	0.93	0.99	0.94	0.96	0.98
Random	None	0.92	0.99	0.95	0.93	0.87	0.98	0.83	0.88	0.99	0.98	0.93	0.96
tinyBenchmarks	Margin	0.71	0.96	0.99	0.86	0.79	0.89	0.89	0.93	0.87	0.92	0.95	0.88
	Content	0.88	0.93	0.95	0.85	0.75	0.94	0.61	0.89	0.93	0.96	0.95	0.94
	Area	0.86	0.82	0.89	0.89	0.73	0.88	0.80	0.87	0.90	0.92	0.95	0.87
	TextJac	0.88	0.85	0.93	0.92	0.86	0.85	0.73	0.76	0.90	0.94	0.96	0.90
	Contrast	0.80	0.92	0.86	0.92	0.70	0.89	0.90	0.88	0.87	0.86	0.94	0.88
	Local	0.86	0.89	0.96	0.90	0.62	0.92	0.80	0.86	0.80	0.96	0.88	0.86
	Palette	0.93	0.95	0.87	0.89	0.64	0.89	0.81	0.81	0.89	0.88	0.99	0.79
	Vibrancy	0.82	0.96	0.98	0.90	0.51	0.94	0.74	0.95	0.73	0.82	0.89	0.89
	Polarity	0.92	0.90	0.98	0.85	0.80	0.89	0.57	0.95	0.85	0.99	0.88	0.90
	SSIM	0.67	0.85	0.89	0.86	0.88	0.94	0.65	0.76	0.73	0.90	0.90	0.89
	LP	0.86	0.89	0.93	0.90	0.79	0.99	0.63	0.83	0.86	0.92	0.92	0.82
	Geom.	0.86	0.93	0.96	0.87	0.83	0.98	0.85	0.93	0.93	0.95	0.96	0.88
ZipBench	Margin	0.94	0.92	0.88	0.79	0.68	0.94	0.89	0.92	0.95	0.93	0.99	0.90
	Content	0.94	0.86	0.98	0.95	0.62	0.92	0.89	0.92	0.87	0.92	0.95	0.94
	Area	0.74	0.81	0.96	0.89	0.85	0.93	0.60	0.93	0.77	0.90	0.89	0.95
	TextJac	0.82	0.94	0.98	0.96	0.86	0.95	0.76	0.93	0.89	0.98	0.90	0.94
	Contrast	0.92	0.87	0.85	0.94	0.62	0.94	0.80	0.80	0.83	0.90	0.93	0.92
	Local	0.81	0.80	0.98	0.89	0.68	0.88	0.73	0.92	0.87	0.98	0.90	0.87
	Palette	0.87	0.95	0.85	0.82	0.74	0.95	0.80	0.87	0.90	0.94	0.93	0.87
	Vibrancy	0.87	0.76	0.89	0.96	0.69	0.89	0.79	0.93	0.88	0.90	0.96	0.86
	Polarity	0.92	0.92	0.99	0.88	0.73	0.92	0.81	0.79	0.88	0.94	0.94	0.76
	SSIM	0.83	0.93	0.88	0.94	0.80	0.94	0.81	0.90	0.95	0.96	0.96	0.92
	LP	0.95	0.82	0.99	0.83	0.92	0.98	0.81	0.92	0.95	0.95	0.94	0.87
	Geom.	0.80	0.85	0.95	0.94	0.70	0.87	0.80	0.81	0.94	0.95	0.90	0.87
mRMR	Margin	0.89	0.82	1.00	0.96	0.61	1.00	0.86	0.89	0.96	0.93	1.00	0.96
	Content	0.93	0.96	0.96	0.89	0.71	0.96	0.93	0.86	0.86	0.96	0.96	0.93
	Area	0.89	0.96	0.96	0.86	0.86	1.00	0.89	0.96	0.89	0.93	1.00	0.96
	TextJac	0.75	0.96	1.00	1.00	0.82	1.00	0.93	0.79	0.89	0.93	0.96	0.93
	Contrast	0.86	0.96	0.96	0.61	0.93	0.96	0.86	0.93	1.00	0.96	0.96	0.71
	Local	0.68	1.00	0.96	0.86	0.82	1.00	0.93	0.96	0.93	0.96	0.96	0.96
	Palette	0.89	0.89	0.93	0.96	0.54	0.96	0.86	0.96	1.00	0.96	0.96	0.96
	Vibrancy	0.89	0.96	1.00	0.89	0.86	1.00	0.68	0.96	0.86	0.93	1.00	0.96
	Polarity	0.93	0.93	0.96	0.96	0.57	0.96	0.86	1.00	0.89	0.86	0.96	0.96
	SSIM	0.93	0.96	1.00	0.96	0.43	1.00	0.93	0.79	0.93	0.86	0.96	0.93
	LP	0.96	1.00	0.96	0.96	0.54	0.89	0.86	0.86	0.89	0.96	1.00	0.96
	Geom.	0.79	0.96	0.96	0.96	0.61	0.96	0.93	0.93	0.96	0.96	0.93	0.96
AnchorPoints	Margin	0.84	0.96	0.96	0.94	0.84	0.99	0.97	0.62	0.85	0.89	1.00	0.88
	Content	0.98	0.94	0.93	0.95	0.70	0.97	0.94	0.97	0.89	0.99	0.96	0.71
	Area	0.97	0.96	0.97	0.93	0.75	0.96	0.90	0.93	0.86	0.99	1.00	0.97
	TextJac	0.90	0.95	0.96	0.83	0.83	0.93	0.86	0.76	0.82	0.80	0.96	0.88
	Contrast	0.78	0.75	0.89	0.87	0.81	0.92	0.90	0.88	0.85	0.96	0.95	0.80
	Local	0.84	0.91	0.95	0.80	0.61	0.90	0.85	0.76	0.85	0.97	0.98	0.92
	Palette	0.88	0.81	0.99	0.97	0.89	0.99	0.90	0.95	0.95	0.99	0.86	0.92
	Vibrancy	0.89	0.83	0.98	0.91	0.77	0.87	0.79	0.84	0.86	0.85	0.91	0.85
	Polarity	0.81	0.91	0.91	0.94	0.85	0.91	0.72	0.85	0.98	0.96	0.95	0.89
	SSIM	0.71	0.85	0.89	0.93	0.93	0.95	0.82	0.86	0.80	0.93	0.98	0.92
	LP	0.81	0.95	0.97	0.95	0.76	0.96	0.81	0.77	0.89	0.96	0.97	0.93
	Geom.	0.84	0.86	0.82	0.86	0.62	0.96	0.75	0.85	0.92	0.93	0.97	0.99
EssenceBench	Margin	0.68	0.61	0.93	0.89	0.82	1.00	0.89	0.93	0.96	0.86	0.89	0.96
	Content	0.75	0.79	0.94	0.96	0.82	1.00	0.88	0.86	0.94	0.95	0.89	0.94
	Area	0.93	0.89	0.93	0.96	0.79	0.96	0.79	0.86	0.93	0.96	0.93	1.00
	TextJac	0.96	0.86	0.93	1.00	0.93	0.96	0.93	0.82	0.96	1.00	1.00	0.93
	Contrast	0.93	0.96	0.96	0.96	0.89	0.96	0.93	0.89	1.00	0.96	0.93	0.96
	Local	0.93	0.96	0.96	0.96	0.64	0.93	0.75	0.82	0.93	0.86	0.96	0.96
	Palette	0.93	1.00	1.00	0.96	0.93	0.93	0.75	0.64	0.96	0.96	1.00	0.96
	Vibrancy	0.96	0.96	0.96	0.86	0.75	0.93	0.77	0.86	0.85	0.93	0.89	0.73
	Polarity	0.96	1.00	1.00	0.93	0.89	0.96	0.93	1.00	0.93	0.89	0.93	1.00
	SSIM	0.54	0.61	0.96	1.00	0.96	1.00	0.75	0.75	0.96	0.93	0.96	0.93
	LP	0.86	0.89	0.93	0.96	0.93	1.00	0.93	0.86	0.96	1.00	0.96	1.00
	Geom.	0.93	0.82	0.93	0.96	0.93	0.93	0.96	0.93	0.93	0.93	0.96	0.96

Table 36: **Cross-metric NMAE on LongForm at 20% retention.** Underlining marks the selection metric. The Mean column averages NMAE across metrics; pale blue-gray marks each single-metric baseline's lowest-mean selection target.

Method	Selection metric	Amat.	Profile	Coher.	Plot	Engage	Prompt	Plan	Poetry	Nuance	Purple	Tell	Uncreat.	Dialogue	Tone	Mean
Ours	All metrics	0.11	0.11	0.12	0.11	0.10	0.12	0.11	0.16	0.10	0.14	0.11	0.13	0.11	0.10	0.12
Random	None	0.12	0.13	0.14	0.13	0.13	0.14	0.15	0.20	0.13	0.18	0.12	0.13	0.13	0.13	0.14
tinyBenchmarks	Amat.	0.17	0.18	0.22	0.22	0.16	0.23	0.20	0.27	0.16	0.24	0.17	0.22	0.17	0.18	0.20
	Profile	0.17	0.19	0.19	0.20	0.18	0.19	0.21	0.26	0.17	0.24	0.18	0.17	0.18	0.20	0.19
	Coher.	0.19	0.17	0.23	0.20	0.18	0.20	0.22	0.33	0.18	0.35	0.18	0.22	0.19	0.19	0.22
	Plot	0.18	0.16	0.17	0.17	0.16	0.18	0.18	0.21	0.16	0.22	0.17	0.18	0.17	0.17	0.18
	Engage	0.14	0.18	0.17	0.18	0.16	0.20	0.19	0.21	0.16	0.20	0.16	0.16	0.17	0.16	0.17
	Prompt	0.24	0.21	0.26	0.23	0.21	0.27	0.24	0.32	0.21	0.29	0.22	0.27	0.24	0.25	0.25
	Plan	0.22	0.22	0.24	0.22	0.21	0.25	0.24	0.27	0.21	0.25	0.23	0.23	0.23	0.23	0.23
	Poetry	0.17	0.16	0.19	0.18	0.15	0.19	0.15	0.29	0.15	0.26	0.16	0.19	0.17	0.16	0.18
	Nuance	0.16	0.17	0.18	0.17	0.17	0.20	0.17	0.21	0.15	0.21	0.17	0.19	0.19	0.18	0.18
	Purple	0.20	0.19	0.24	0.20	0.19	0.25	0.24	0.29	0.20	0.29	0.19	0.24	0.20	0.22	0.22
	Tell	0.22	0.22	0.24	0.23	0.20	0.24	0.25	0.28	0.21	0.28	0.22	0.24	0.21	0.20	0.23
	Uncreat.	0.14	0.16	0.16	0.16	0.14	0.19	0.19	0.22	0.13	0.18	0.15	0.18	0.14	0.16	0.17
	Dialogue	0.16	0.18	0.21	0.18	0.16	0.22	0.20	0.23	0.17	0.22	0.15	0.18	0.18	0.18	0.19
	Tone	0.16	0.15	0.18	0.19	0.16	0.19	0.17	0.27	0.16	0.23	0.16	0.20	0.18	0.18	0.18
ZipBench	Amat.	0.22	0.21	0.24	0.25	0.18	0.27	0.24	0.35	0.19	0.35	0.21	0.25	0.21	0.21	0.24
	Profile	0.22	0.19	0.22	0.19	0.18	0.19	0.21	0.32	0.21	0.32	0.21	0.25	0.23	0.20	0.23
	Coher.	0.19	0.19	0.22	0.20	0.18	0.22	0.21	0.26	0.19	0.25	0.17	0.21	0.18	0.19	0.20
	Plot	0.24	0.22	0.22	0.25	0.21	0.24	0.26	0.33	0.22	0.31	0.25	0.26	0.24	0.25	0.25
	Engage	0.15	0.18	0.20	0.19	0.16	0.22	0.19	0.24	0.16	0.21	0.17	0.18	0.17	0.17	0.19
	Prompt	0.18	0.19	0.22	0.18	0.16	0.20	0.21	0.26	0.19	0.24	0.18	0.20	0.20	0.19	0.20
	Plan	0.19	0.20	0.21	0.20	0.19	0.20	0.23	0.22	0.19	0.21	0.18	0.20	0.19	0.21	0.20
	Poetry	0.24	0.22	0.25	0.25	0.20	0.28	0.25	0.35	0.21	0.30	0.23	0.26	0.23	0.22	0.25
	Nuance	0.19	0.16	0.19	0.18	0.18	0.22	0.18	0.26	0.16	0.22	0.20	0.20	0.17	0.19	0.19
	Purple	0.17	0.17	0.19	0.19	0.17	0.22	0.22	0.28	0.15	0.27	0.16	0.20	0.17	0.18	0.20
	Tell	0.14	0.15	0.19	0.16	0.17	0.19	0.17	0.25	0.15	0.22	0.15	0.18	0.16	0.16	0.17
	Uncreat.	0.21	0.20	0.26	0.25	0.23	0.31	0.21	0.29	0.21	0.28	0.20	0.22	0.20	0.25	0.24
	Dialogue	0.18	0.17	0.19	0.19	0.18	0.23	0.18	0.27	0.17	0.23	0.19	0.21	0.19	0.19	0.20
	Tone	0.17	0.17	0.21	0.19	0.15	0.20	0.18	0.27	0.16	0.24	0.16	0.19	0.17	0.17	0.19
mRMR	Amat.	0.16	0.16	0.15	0.11	0.13	0.16	0.13	0.20	0.13	0.16	0.16	0.19	0.14	0.17	0.15
	Profile	0.15	0.13	0.19	0.18	0.14	0.16	0.16	0.22	0.14	0.24	0.11	0.15	0.13	0.15	0.16
	Coher.	0.12	0.12	0.12	0.11	0.11	0.16	0.16	0.31	0.11	0.27	0.12	0.12	0.14	0.12	0.15
	Plot	0.15	0.15	0.18	0.16	0.13	0.17	0.18	0.29	0.11	0.23	0.13	0.15	0.14	0.14	0.16
	Engage	0.14	0.15	0.17	0.13	0.15	0.17	0.20	0.26	0.13	0.20	0.12	0.17	0.16	0.17	0.17
	Prompt	0.14	0.13	0.15	0.12	0.10	0.13	0.18	0.26	0.11	0.22	0.11	0.15	0.14	0.12	0.15
	Plan	0.15	0.15	0.19	0.18	0.17	0.17	0.15	0.24	0.17	0.22	0.14	0.18	0.15	0.19	0.17
	Poetry	0.14	0.13	0.17	0.12	0.09	0.13	0.15	0.19	0.09	0.20	0.12	0.13	0.19	0.12	0.14
	Nuance	0.13	0.16	0.16	0.13	0.10	0.14	0.17	0.25	0.11	0.24	0.12	0.15	0.12	0.13	0.15
	Purple	0.16	0.11	0.20	0.16	0.18	0.16	0.15	0.23	0.15	0.26	0.14	0.14	0.16	0.17	0.17
	Tell	0.12	0.17	0.20	0.20	0.19	0.24	0.18	0.25	0.14	0.23	0.13	0.12	0.12	0.18	0.18
	Uncreat.	0.11	0.13	0.10	0.14	0.11	0.17	0.18	0.21	0.11	0.17	0.11	0.15	0.10	0.13	0.14
	Dialogue	0.10	0.11	0.11	0.12	0.10	0.15	0.14	0.21	0.09	0.18	0.11	0.14	0.13	0.12	0.13
	Tone	0.16	0.14	0.25	0.15	0.11	0.18	0.16	0.19	0.13	0.22	0.17	0.23	0.18	0.12	0.17
AnchorPoints	Amat.	0.16	0.17	0.20	0.19	0.17	0.20	0.20	0.23	0.16	0.22	0.15	0.20	0.19	0.20	0.19
	Profile	0.18	0.17	0.21	0.22	0.18	0.20	0.21	0.22	0.18	0.22	0.19	0.19	0.20	0.20	0.20
	Coher.	0.13	0.13	0.14	0.13	0.10	0.14	0.18	0.24	0.13	0.25	0.14	0.16	0.15	0.12	0.15
	Plot	0.14	0.13	0.15	0.15	0.13	0.16	0.14	0.20	0.12	0.19	0.13	0.16	0.13	0.12	0.15
	Engage	0.17	0.15	0.17	0.17	0.16	0.21	0.19	0.24	0.14	0.23	0.17	0.20	0.16	0.18	0.18
	Prompt	0.13	0.13	0.15	0.14	0.12	0.16	0.15	0.23	0.11	0.23	0.12	0.15	0.13	0.14	0.15
	Plan	0.13	0.15	0.18	0.17	0.15	0.18	0.17	0.23	0.14	0.22	0.12	0.14	0.12	0.15	0.16
	Poetry	0.18	0.17	0.19	0.16	0.16	0.19	0.21	0.21	0.18	0.21	0.18	0.21	0.18	0.19	0.19
	Nuance	0.19	0.17	0.18	0.20	0.16	0.21	0.20	0.24	0.15	0.23	0.20	0.23	0.18	0.20	0.20
	Purple	0.17	0.17	0.19	0.18	0.14	0.20	0.18	0.22	0.14	0.20	0.16	0.18	0.17	0.18	0.18
	Tell	0.17	0.16	0.20	0.18	0.15	0.18	0.21	0.25	0.16	0.25	0.17	0.19	0.17	0.18	0.19
	Uncreat.	0.15	0.16	0.15	0.14	0.13	0.15	0.17	0.21	0.16	0.18	0.17	0.15	0.19	0.13	0.16
	Dialogue	0.17	0.17	0.20	0.18	0.14	0.19	0.21	0.24	0.15	0.22	0.16	0.19	0.17	0.17	0.18
	Tone	0.15	0.15	0.14	0.14	0.13	0.15	0.14	0.18	0.14	0.18	0.16	0.17	0.14	0.13	0.15
EssenceBench	Amat.	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Profile	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Coher.	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Plot	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Engage	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Prompt	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Plan	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Poetry	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Nuance	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Purple	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Tell	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Uncreat.	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Dialogue	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32
	Tone	0.36	0.30	0.39	0.31	0.28	0.28	0.37	0.25	0.30	0.27	0.35	0.32	0.34	0.31	0.32

Table 37: **Cross-metric SP on LongForm at 20% retention.** Underlining marks the selection metric; pale blue-gray marks each single-metric baseline's selection target with the lowest mean NMAE.

Method	Selection metric	Amat.	Profile	Coher.	Plot	Engage	Prompt	Plan	Poetry	Nuance	Purple	Tell	Uncreat.	Dialogue	Tone
Ours	All metrics	0.99	0.99	0.97	0.98	0.99	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.99	0.99
Random	None	0.99	0.98	0.97	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.99	0.98	0.98	0.98
tinyBenchmarks	Amat.	0.98	0.97	0.94	0.96	0.97	0.97	0.97	0.93	0.97	0.96	0.98	0.97	0.97	0.97
	Profile	0.98	0.96	0.95	0.97	0.97	0.97	0.97	0.91	0.98	0.95	0.98	0.97	0.97	0.97
	Coher.	0.95	0.97	0.93	0.95	0.97	0.95	0.97	0.95	0.97	0.95	0.97	0.96	0.96	0.97
	Plot	0.98	0.97	0.97	0.96	0.98	0.97	0.97	0.95	0.98	0.96	0.97	0.98	0.97	0.97
	Engage	0.97	0.96	0.96	0.96	0.98	0.97	0.98	0.94	0.97	0.95	0.97	0.98	0.97	0.97
	Prompt	0.94	0.97	0.93	0.94	0.97	0.95	0.96	0.90	0.97	0.93	0.96	0.94	0.94	0.96
	Plan	0.95	0.97	0.96	0.96	0.97	0.97	0.96	0.95	0.96	0.95	0.97	0.95	0.96	0.96
	Poetry	0.97	0.98	0.96	0.97	0.99	0.98	0.98	0.92	0.98	0.94	0.97	0.97	0.96	0.98
	Nuance	0.97	0.98	0.95	0.97	0.98	0.97	0.98	0.96	0.97	0.97	0.97	0.98	0.96	0.97
	Purple	0.97	0.96	0.94	0.96	0.97	0.95	0.95	0.93	0.96	0.96	0.96	0.96	0.97	0.96
	Tell	0.97	0.98	0.96	0.96	0.97	0.97	0.96	0.95	0.98	0.96	0.97	0.97	0.96	0.98
	Uncreat.	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.98	0.96	0.98	0.97	0.97	0.96
	Dialogue	0.97	0.97	0.96	0.95	0.97	0.96	0.97	0.94	0.97	0.96	0.98	0.98	0.97	0.97
	Tone	0.97	0.98	0.95	0.97	0.97	0.97	0.99	0.94	0.97	0.96	0.98	0.96	0.97	0.97
ZipBench	Amat.	0.96	0.96	0.97	0.96	0.97	0.97	0.96	0.91	0.97	0.95	0.96	0.96	0.95	0.96
	Profile	0.96	0.97	0.95	0.96	0.98	0.96	0.97	0.90	0.97	0.92	0.96	0.95	0.95	0.98
	Coher.	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.97	0.96	0.98	0.97	0.98	0.97
	Plot	0.95	0.97	0.93	0.94	0.97	0.96	0.95	0.90	0.96	0.91	0.96	0.95	0.95	0.97
	Engage	0.98	0.97	0.96	0.97	0.98	0.96	0.98	0.95	0.97	0.97	0.97	0.97	0.97	0.97
	Prompt	0.95	0.96	0.94	0.96	0.97	0.96	0.97	0.94	0.96	0.95	0.97	0.95	0.95	0.95
	Plan	0.98	0.96	0.97	0.96	0.98	0.97	0.97	0.95	0.97	0.97	0.98	0.98	0.96	0.96
	Poetry	0.97	0.97	0.96	0.96	0.98	0.98	0.96	0.93	0.97	0.95	0.97	0.97	0.96	0.98
	Nuance	0.97	0.97	0.96	0.96	0.98	0.97	0.97	0.93	0.98	0.97	0.97	0.97	0.97	0.97
	Purple	0.97	0.97	0.96	0.97	0.97	0.96	0.95	0.93	0.98	0.94	0.98	0.97	0.96	0.97
	Tell	0.97	0.98	0.96	0.97	0.99	0.97	0.97	0.94	0.98	0.96	0.97	0.97	0.97	0.98
	Uncreat.	0.94	0.97	0.93	0.95	0.97	0.95	0.97	0.93	0.96	0.94	0.97	0.96	0.95	0.95
	Dialogue	0.94	0.96	0.94	0.95	0.97	0.96	0.97	0.90	0.96	0.95	0.97	0.94	0.95	0.95
	Tone	0.97	0.96	0.95	0.98	0.98	0.97	0.97	0.92	0.97	0.95	0.98	0.97	0.97	0.97
mRMR	Amat.	0.99	0.98	0.98	0.99	0.99	0.96	0.98	0.95	0.99	0.99	0.98	0.98	0.98	0.98
	Profile	0.99	0.97	0.95	0.98	0.99	0.98	0.98	0.95	0.98	0.97	0.98	0.99	0.97	0.97
	Coher.	0.99	0.98	0.97	0.98	0.98	0.97	0.98	0.96	0.98	0.98	0.98	0.98	0.97	0.98
	Plot	0.99	0.98	0.97	0.97	0.98	0.97	0.98	0.97	0.99	0.98	0.99	0.99	0.99	0.99
	Engage	0.97	0.98	0.94	0.97	0.98	0.97	0.98	0.93	0.97	0.98	0.98	0.98	0.97	0.98
	Prompt	0.97	0.98	0.96	0.98	0.99	0.98	0.98	0.93	0.99	0.97	0.99	0.97	0.97	0.98
	Plan	0.98	0.98	0.93	0.97	0.98	0.99	0.99	0.93	0.96	0.97	0.98	0.98	0.97	0.98
	Poetry	0.98	0.99	0.97	0.98	0.99	0.99	0.98	0.96	0.98	0.97	0.99	0.99	0.98	0.99
	Nuance	0.98	0.98	0.95	0.97	0.99	0.98	0.98	0.93	0.98	0.96	0.99	0.98	0.98	0.99
	Purple	0.98	0.98	0.96	0.99	0.98	0.98	0.98	0.93	0.98	0.96	0.98	0.98	0.97	0.98
	Tell	0.98	0.97	0.94	0.96	0.95	0.95	0.96	0.92	0.97	0.96	0.98	0.98	0.98	0.95
	Uncreat.	0.98	0.97	0.97	0.98	0.99	0.98	0.98	0.94	0.97	0.99	0.99	0.97	0.99	0.98
	Dialogue	0.99	0.98	0.98	0.99	0.99	0.97	0.97	0.94	0.99	0.97	0.99	0.99	0.97	0.99
	Tone	0.99	0.98	0.97	0.97	0.98	0.97	0.98	0.98	0.98	0.97	0.98	0.99	0.98	0.99
AnchorPoints	Amat.	0.98	0.98	0.97	0.96	0.98	0.98	0.97	0.97	0.98	0.97	0.98	0.97	0.98	0.98
	Profile	0.96	0.98	0.93	0.93	0.98	0.95	0.97	0.94	0.97	0.95	0.97	0.96	0.96	0.97
	Coher.	0.98	0.99	0.95	0.98	0.99	0.99	0.96	0.95	0.97	0.97	0.98	0.98	0.96	0.98
	Plot	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.97	0.99	0.97	0.99	0.99	0.98	0.98
	Engage	0.98	0.97	0.96	0.96	0.99	0.97	0.97	0.95	0.97	0.97	0.98	0.98	0.98	0.97
	Prompt	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.97	0.98	0.96	0.98	0.98	0.98	0.98
	Plan	0.98	0.98	0.96	0.97	0.98	0.97	0.98	0.96	0.98	0.97	0.99	0.98	0.98	0.98
	Poetry	0.94	0.97	0.94	0.95	0.98	0.96	0.98	0.96	0.96	0.98	0.97	0.96	0.96	0.95
	Nuance	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.98	0.97	0.97
	Purple	0.96	0.97	0.95	0.98	0.98	0.97	0.99	0.95	0.98	0.97	0.97	0.97	0.96	0.97
	Tell	0.96	0.98	0.97	0.96	0.98	0.97	0.97	0.95	0.98	0.94	0.97	0.96	0.95	0.97
	Uncreat.	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.95	0.97	0.97	0.96	0.96	0.96	0.98
	Dialogue	0.97	0.98	0.96	0.97	0.98	0.98	0.98	0.96	0.98	0.97	0.97	0.97	0.96	0.98
	Tone	0.97	0.97	0.97	0.99	0.98	0.98	0.98	0.97	0.97	0.97	0.98	0.97	0.97	0.98
EssenceBench	Amat.	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Profile	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Coher.	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Plot	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Engage	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Prompt	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Plan	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Poetry	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Nuance	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Purple	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Tell	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Uncreat.	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Dialogue	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97
	Tone	0.95	0.97	0.96	0.95	0.98	0.97	0.96	0.93	0.97	0.94	0.95	0.95	0.95	0.97