FEW-QUERY QUALITY ASSESSMENT FOR UNSEEN MACHINE OBSERVERS: GAUGE-INVARIANT DECISION SUBSPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-oriented image quality is observer-dependent: visual models can prefer different processed versions of the same image. We study few-query calibration on a fixed content–candidate bank: source models provide complete responses, while a previously unseen observer reveals only a few candidate groups before its choices on other groups are predicted. A reconstruction-optimal low-rank basis can discard decisive candidate comparisons, and an empirical-Bayes posterior can depend on arbitrary factor coordinates unless its prior transforms with the basis. We introduce Gauge-Invariant Decision Subspaces (GIDS), which selects a decision-aligned subspace within a spectral dictionary and estimates its observer prior and residual noise in matching coordinates. Training simulates complete-group calibration; deployment uses a closed-form low-dimensional posterior. Theory establishes gauge-covariant prediction and a regret bound separating representation from posterior-identification error. Against a matched gauge-consistent SVD baseline, GIDS lowers observer-macro selection regret from 0.1282 to 0.1266 under CIFAR-10 architecture-family holdout, with mean improvements on prospectively frozen observers and a frozen CIFAR-100 domain. The results show the value of choosing a low-rank representation for decisions after calibration, rather than response reconstruction alone. The code is available at `https://anonymous.4open.science/r/ICLR2027-00EB`.

1 INTRODUCTION

Image quality assessment (IQA) traditionally predicts how people judge an image. In machine vision, the better image is the one that helps the deployed model. Two classifiers can rank the same processed candidates differently, so a quality decision for one model need not transfer to another. MPD documents differences between human and machine preference (Li et al., 2025a), and MIQD-2.5M shows variation across models and tasks (Wang et al., 2025). This observer dependence creates a deployment problem: the model receiving the images may be absent from the data used to build the evaluator. Running that model on every processed candidate would defeat the purpose of an inexpensive quality estimate. We study whether a few queries can locate a new machine observer within response patterns learned from other models. A query reveals all candidates for one content; the evaluator must then choose the new observer's preferred candidate on unqueried contents. Unlike a benchmark estimate, this requires a separate decision for every content.

A low-rank response model offers a compact account of how observers differ, but reconstructing responses is not the same as choosing candidates. Singular value decomposition (SVD) preserves high-energy directions across the response matrix; selection depends on differences within each content. A low-energy direction may still decide which candidate wins. A second issue arises during calibration: factor coordinates are arbitrary, yet a Bayesian posterior depends on the prior expressed in those coordinates. Changing the basis without changing the prior gives different predictions from equivalent factorizations.

We introduce Gauge-Invariant Decision Subspaces (GIDS) to address both issues. Within an orthogonal spectral dictionary, it learns a subspace for candidate selection after few-query calibration. Each candidate subspace is paired with an observer prior and a residual-noise estimate in its own

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

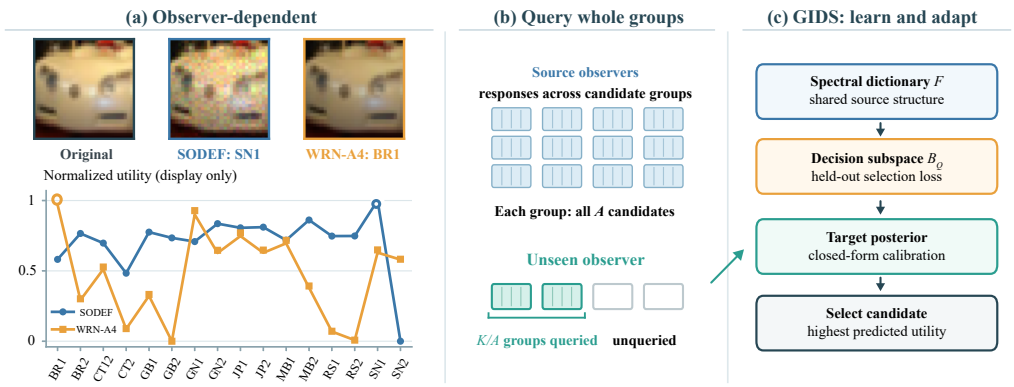


Figure 1: From observer-dependent quality to few-query deployment. (a) An illustrative evaluation image for which two prospectively frozen observers prefer different candidates; displayed utilities are observer-normalized. (b) Queries reveal complete candidate groups. (c) GIDS learns a decision subspace from source observers and calibrates a new observer in closed form.

coordinates, so the posterior does not depend on the chosen basis. We train our primary GIDS configuration with complete-group calibration episodes and also study direct decision fitting, labeled "GIDS (dense fit)" in figures and tables. Once learned, the subspace calibrates a new observer with a closed-form rank-three Gaussian posterior.

Our contributions are:

- We formalize few-query candidate selection on a fixed item bank with complete-group queries and architecture-family holdout, so that calibration is evaluated on observers excluded from representation learning;
- We learn a decision-aligned spectral subspace with coordinate-consistent empirical-Bayes calibration, preserving predictions under basis changes while keeping target inference closed form;
- We derive regret bounds separating representation and posterior-identification errors, then test GIDS under matched family-held-out and frozen evaluations, finding lower mean regret than SVD + EB with heterogeneous observer effects.

Our task and workflow are illustrated in Figure 1.

2 Related Work

Machine-oriented quality. Q-Bench evaluates low-level visual understanding and human-aligned quality (Wu et al., 2024a). Lens uses target-model confidence to guide camera settings, and Embodied-IQA studies image usability for robotic systems (Baek et al., 2025; Li et al., 2026). MPD and MIQD-2.5M document model-dependent machine utility across images and tasks (Li et al., 2025a; Wang et al., 2025). We ask whether source-response differences can guide the choices of a previously unseen observer after a few measurements.

Query-efficient evaluation. Active Testing and Active Surrogate Estimators estimate aggregate risk from selected labels or losses (Kossen et al., 2021; 2022). In language-model evaluation, a low-rank bandit approach seeks an overall best method (Zhou et al., 2025), while SparseEval and metabench select informative benchmark items (Zhang et al., 2026; Kipnis et al., 2025). tinyBenchmarks and Anchor Points use cross-model responses, while low-rank recovery and experimental design provide related tools (Maia Polo et al., 2024; Vivek et al., 2024; Cao et al., 2023; Jang et al., 2024). Such estimates depend on source–target correlation (Zhang et al., 2025). Our endpoint instead requires a per-content argmax, and each query reveals a complete candidate group.

Cold-start decision models. Matrix factorization and ordinal elicitation adapt to new subjects (Mnih & Salakhutdinov, 2007; Salakhutdinov & Mnih, 2008; Houlsby et al., 2014; Nguyen et al., 2024).

Collaborative ranking (Wu et al., 2018) and decision-focused learning (Elmachtoub & Grigas, 2022; Mandi et al., 2022; Shah et al., 2022; Gupta & Huang, 2024) show why reconstruction and downstream choices can favor different representations. Here source observers define the representation and a few complete groups identify the target observer within it.

Factor geometry. Factor models admit equivalent coordinates under rotations, scales, and permutations (Papastamoulis & Ntzoufras, 2022); Grassmann geometry represents their subspace without choosing one basis (Edelman et al., 1998). In cold-start calibration the prior must follow the coordinates, so GIDS estimates its prior and residual noise in each chosen basis.

3 Problem Formulation

Let $m \in \mathcal{M}$, $x \in \mathcal{X}$, and $a \in \mathcal{A}_x$ denote a machine observer, image content, and processed candidate. The response $r_{mxa} \in \mathbb{R}$ is task utility, with larger values preferred; in our experiments it is the within-group z-score of negative cross-entropy, computed separately for each observer and content. Source observers supply complete responses over a fixed item collection. The observer axis is inductive but the item axis is transductive: the target $m_\star$ is unseen, while source responses cover the content–candidate items later used for target evaluation.

One query chooses a content x and observes $\{r_{m_\star xa} : a \in \mathcal{A}_x\}$. If every group contains A candidates, a budget of K responses corresponds to K/A complete groups. Calibration and evaluation contents are disjoint. The prediction is

$$\hat{a}(x) = \arg\max_{a \in \mathcal{A}_x} \hat{r}_{m_\star xa}.$$

Writing $a^\star(x) = \arg\max_a r_{m_\star xa}$ and $g_{m_\star x} = \max_a r_{m_\star xa} - \min_a r_{m_\star xa}$, normalized selection regret is

$$\ell_{m_\star x} = \frac{r_{m_\star x a^\star} - r_{m_\star x \hat{a}}}{g_{m_\star x}}.$$

When all candidates have identical utility, we assign the group zero regret; the implementation floors the denominator at 10^{-8}. The target observer contributes no responses to the source representation, although the source observers cover the item collection. Throughout, "quality" means task-conditioned machine utility.

Only differences among candidates determine a group's winner. A content-wide or observer-wide offset carries no selection information. We center responses across observers and items before fitting the shared structure, and the listwise softmax removes any remaining groupwise offset. This leaves the model free to use its capacity for comparisons that can change a choice.

4 Gauge-Invariant Decision Subspaces

The method has three parts: a spectral dictionary from source responses, a subspace selected for candidate decisions, and a Bayesian posterior for the new observer.

4.1 SVD reconstruction with empirical-Bayes calibration

Let $\boldsymbol{\mu} \in \mathbb{R}^N$ be the itemwise source mean and $\mathbf{R} \in \mathbb{R}^{M \times N}$ the response matrix after observer and item centering. Truncated SVD gives an orthonormal item basis $\mathbf{B}_0 \in \mathbb{R}^{N \times d}$. We pair it with the following empirical-Bayes calibrator:

$$\mathbf{y}_\mathcal{S} = \boldsymbol{\mu}_\mathcal{S} + \mathbf{B}_{0,\mathcal{S}}\mathbf{z} + \boldsymbol{\epsilon}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_0), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}),$$

where $\mathcal{S}$ contains every candidate row in the queried groups. Source responses estimate both $\mathbf{C}_0$ and the residual variance σ_0^2 in the coordinates of $\mathbf{B}_0$. We use the same fixed likelihood-variance multiplier κ for this baseline and GIDS, with $\tilde{\sigma}_0^2 = \kappa\sigma_0^2$:

$$\boldsymbol{\Sigma}_{\mathbf{z}|\mathcal{S}} = (\mathbf{C}_0^{-1} + \tilde{\sigma}_0^{-2}\mathbf{B}_{0,\mathcal{S}}^\top \mathbf{B}_{0,\mathcal{S}})^{-1}, \tag{1}$$

$$\hat{\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{z}|\mathcal{S}} \tilde{\sigma}_0^{-2} \mathbf{B}_{0,\mathcal{S}}^\top (\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S}). \tag{2}$$

In figures and tables, "SVD + EB" describes this combination of SVD, covariance shrinkage (Ledoit & Wolf, 2004), and Gaussian empirical-Bayes calibration. We construct it as a matched control; it is not a separately published method.

4.2 GAUGE CONSISTENCY

For invertible $\mathbf{H}$, $\mathbf{B}' = \mathbf{BH}$ represents the same responses under $\mathbf{z}' = \mathbf{H}^{-1}\mathbf{z}$. The posterior prediction remains unchanged when $\mathbf{C}' = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-\top}$. Reusing the old prior can change posterior shrinkage despite identical source reconstructions. An orthogonal basis removes scale ambiguity but leaves rotations, so the prior still has to follow the basis.

4.3 COORDINATE-INVARIANT DECISION SUBSPACE

Let $\mathbf{F} \in \mathbb{R}^{N\times r}$ contain the leading orthonormal right-singular vectors of $\mathbf{R}$, and let $\mathbf{Q} \in \mathrm{St}(r, d)$, so $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_d$. GIDS defines

$$\mathbf{B}_\mathbf{Q} = \mathbf{FQ}, \quad \mathbf{A}_\mathbf{Q} = \mathbf{RB}_\mathbf{Q}, \quad \mathbf{C}_\mathbf{Q} = \mathrm{ShrinkCov}(\mathbf{A}_\mathbf{Q}), \quad \sigma^2_\mathbf{Q} = \frac{1}{MN}\|\mathbf{R} - \mathbf{A}_\mathbf{Q}\mathbf{B}_\mathbf{Q}^\top\|^2_\mathrm{F}.$$

We recompute the prior and noise for every learned subspace and use $\tilde{\sigma}^2_\mathbf{Q} = \kappa\sigma^2_\mathbf{Q}$ in its posterior. For any orthogonal $\mathbf{O}$, $\mathbf{Q}$ and $\mathbf{QO}$ span the same subspace. The statistical object is therefore the projector $\mathbf{QQ}^\top$ on the Grassmann manifold.

4.4 CALIBRATION-AWARE OBJECTIVE

Reconstruction chooses a subspace before considering how a target observer will be calibrated. We instead estimate $\hat{\mathbf{z}}_{m,\mathcal{S}}$ from queried rows of a source observer and evaluate its predictions on held-out groups:

$$\hat{\mathbf{r}}_{m,\mathcal{E}} = \boldsymbol{\mu}_\mathcal{E} + \mathbf{B}_{\mathbf{Q},\mathcal{E}}\hat{\mathbf{z}}_{m,\mathcal{S}}.$$

The loss therefore favors directions that can be identified from a small query set and still recover decisions on unseen groups.

Each source observer supplies such calibration episodes. Training cycles through complete-group prefixes of a fixed source-designed query order for $\mathcal{S}$, applies the deployment posterior, and samples $\mathcal{E}$ from a fixed, content-disjoint evaluation pool. Within a group, utilities define a temperature-controlled distribution $p_{m,x}$ and predictions define $\hat{p}_{m,x}$. We use

$$\mathcal{L}_\mathrm{sel} = -\frac{1}{|\mathcal{E}|}\sum_{x\in\mathcal{E}}\sum_{a\in\mathcal{A}_x} p_{m,x}(a)\log\hat{p}_{m,x}(a),$$

which is insensitive to groupwise offsets and focuses on selection comparisons. The objective is

$$\min_{\mathbf{Q}\in\mathrm{St}(r,d)} \mathbb{E}_{m,\mathcal{S},\mathcal{E}}\left[\mathcal{L}_\mathrm{sel}(m,\mathcal{S},\mathcal{E};\mathbf{Q},\mathbf{C}_\mathbf{Q},\sigma^2_\mathbf{Q})\right] + \frac{\lambda}{r^2}\|\mathbf{QQ}^\top - \mathbf{Q}_0\mathbf{Q}_0^\top\|^2_\mathrm{F}.$$

Here $\mathbf{Q}_0 = [\mathbf{I}_d; \mathbf{0}]$ selects the leading d dictionary directions. The projector anchor limits departure from this subspace without favoring an orientation. We optimize an unconstrained parameter and obtain $\mathbf{Q}$ by QR orthogonalization at each forward pass. Learning changes the representation while retaining the same deployment solver. At deployment, we freeze $\mathbf{B}_\mathbf{Q}, \mathbf{C}_\mathbf{Q}, \sigma^2_\mathbf{Q}$ and solve the low-dimensional target posterior.

4.5 TRAINING AND DEPLOYMENT

Figure 2 shows the controlled comparison. During training, gradients pass through the source-observer posterior to the QR-orthogonalized $\mathbf{Q}$, and each updated subspace receives a matching prior and noise estimate. At deployment, the basis and its statistical model are frozen. Each query returns all A responses for one content, after which a single rank-d Gaussian solve supplies predictions for the unqueried groups. The solve costs $\mathcal{O}(Kd^2 + d^3)$, small relative to obtaining the K target responses at $d = 3$.

5 THEORY

Two properties explain the method: predictions should survive a change of latent coordinates, and calibration error should be small where it can reverse a candidate choice. Proofs appear in Appendix A.

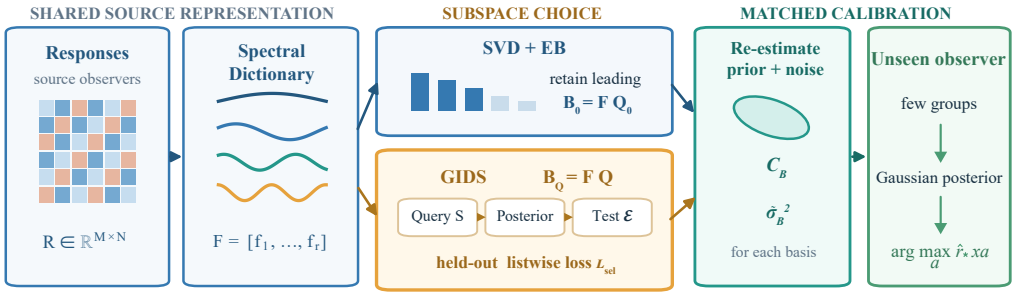


Figure 2: Comparison of GIDS with the SVD + EB control. SVD preserves response energy, whereas GIDS learns directions that preserve choices after calibration. Both share the dictionary, the rule for estimating coordinate-matched statistics, and the target solver. The key difference is subspace selection: SVD retains leading reconstruction directions, whereas GIDS differentiates held-out selection loss through simulated few-query calibration.

Proposition 1 (Gauge-covariant calibration)**.** Under $\mathbf{B}' = \mathbf{BH}$, $\mathbf{z}' = \mathbf{H}^{-1}\mathbf{z}$, and $\mathbf{C}' = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-\top}$, posterior coordinates satisfy $\hat{\mathbf{z}}' = \mathbf{H}^{-1}\hat{\mathbf{z}}$ and predictions are unchanged. For orthogonal $\mathbf{O}$, replacing $\mathbf{Q}$ by $\mathbf{QO}$ also leaves reconstruction, residual noise, episodic loss, and the projector anchor unchanged.

The first statement holds for any invertible coordinate change if the prior transforms with the basis. GIDS uses an orthogonally equivariant shrinkage estimator, leaving rotations as the residual gauge of its subspace search.

Proposition 2 (Calibration-margin regret)**.** Fix a true response group $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Bz} + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty \le \varepsilon_{\text{rep}}$. Over calibration noise, assume $\mathbf{z} - \hat{\mathbf{z}} = \mathbf{h}_\mathcal{S} + \boldsymbol{\eta}_\mathcal{S}$, where every projection $\mathbf{w}^\top\boldsymbol{\eta}_\mathcal{S}$ is zero-mean sub-Gaussian with variance proxy $\mathbf{w}^\top\mathbf{V}_\mathcal{S}\mathbf{w}$. For $a^* = \arg\max_a y_a$, let $\Delta_a = y_{a^*} - y_a$, $\mathbf{w}_a = \mathbf{b}_{a^*} - \mathbf{b}_a$, $\beta_a = |\mathbf{w}_a^\top\mathbf{h}_\mathcal{S}|$, and $v_a = \mathbf{w}_a^\top\mathbf{V}_\mathcal{S}\mathbf{w}_a$. Then

$$\mathbb{E}_{\text{cal}}[\text{Regret}_{\text{raw}} \mid \mathbf{y}, \mathcal{S}] \le \sum_{a \ne a^*} \Delta_a \min\left\{1, \exp\left[-\frac{(\Delta_a - 2\varepsilon_{\text{rep}} - \beta_a)_+^2}{2v_a}\right]\right\}.$$

For $v_a = 0$, interpret the probability factor as zero when $\Delta_a > 2\varepsilon_{\text{rep}} + \beta_a$, and one otherwise. The inequality bounds raw utility regret; dividing by $\max\{g_{m_\star x}, 10^{-8}\}$ gives the normalized version. The bound exposes three sources of error: the selected subspace must represent the response, the few queries must identify the observer coordinate, and query noise must not overwhelm a candidate margin. The bias β_a and uncertainty v_a are invariant under the gauge transformation in Proposition 1. For the Gaussian posterior, they arise from shrinkage, representation residual, and propagated query noise.

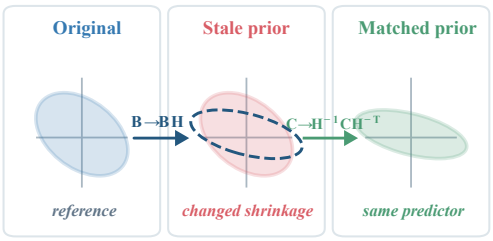
Proposition 3 (Dictionary and decision-subspace error)**.** For a centered target response $\mathbf{u}$ and $\mathbf{B_Q} = \mathbf{FQ}$,

$$(\mathbf{I} - \mathbf{B_Q}\mathbf{B_Q}^\top)\mathbf{u} = (\mathbf{I} - \mathbf{F}\mathbf{F}^\top)\mathbf{u} + \mathbf{F}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{F}^\top\mathbf{u}.$$

Thus $\varepsilon_{\text{rep}} \le \varepsilon_{\text{dict}} + \varepsilon_{\text{select}}(\mathbf{Q})$: dictionary width controls the first term, while subspace learning controls the second.

SVD minimizes global Frobenius energy, whereas Proposition 2 weights candidate-difference directions by margins and posterior variance. These objectives need not agree; Figure 3 gives a strict two-candidate counterexample.

The leading SVD subspace is feasible for GIDS, providing a reference point for the empirical objective.



(a) Gauge consistency. Co-transforming the latent prior preserves Bayesian predictions across equivalent factorizations.



(b) Objective mismatch. Lower response-reconstruction error can reverse the preferred candidate and incur larger regret.

Figure 3: The reasons for the insufficiency of reconstruction alone in few-query calibration and selection. (a) Transforming the latent prior with the basis preserves the calibrated estimator; a stale prior can change posterior shrinkage. (b) Bars show ground-truth and predicted candidate utilities. Labels report the sum of squared reconstruction errors (SSE), $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$, across the two candidates. The lower-SSE prediction selects the wrong candidate; the higher-SSE prediction preserves the winner.

6 Experiments

6.1 Experimental setup

Observers and splits. We evaluate 12 classification observers from six architecture families on CIFAR-10 (Krizhevsky, 2009). In each fold, we hold out one family while the others supply the dictionary, prior, run-specific validation, and query order. The 10,000 contents are partitioned into 6,000 for protocol development, 1,000 for source-only validation, 1,000 for target calibration, and 2,000 for evaluation. Source responses on calibration and evaluation items define the transductive representation; target responses enter only through queried calibration groups.

Queries and metric. Each content has eight operations at two severities. A cross-severity direction queries one eight-candidate severity group and evaluates the other on disjoint contents. Budgets $K \in \{8, 16, 32, 64, 128\}$ therefore provide 1, 2, 4, 8, or 16 complete groups. The calibrated methods in Table 1 share folds, normalization, and a greedy block D-optimal query order.

For complete-group negative cross-entropy utilities, normalization uses the sample standard deviation within each observer–content–severity group, floored at 10^{-8}. Thus a queried group can be normalized using only its observed responses. The itemwise mean $\boldsymbol{\mu}$ is computed across source observers; $\mathbf{R}$ then removes both observer and item means and adds back the grand mean.

Baselines. The source-mean control selects without target calibration. SVD + ridge estimates target coordinates in a truncated SVD basis with a source-selected penalty. PMF adapts probabilistic matrix factorization (Mnih & Salakhutdinov, 2007) to continuous source-observer responses using a maximum-a-posteriori source fit and regularized target-coordinate inference, with rank and regularization selected using source families only. SVD + EB uses the posterior of Section 4.1 with the leading reconstruction subspace, making it the closest comparison: GIDS changes the subspace and retains the target solver. The oracle observes every target response and gives a lower bound.

Selection and reporting. We select ridge penalties by source-only pseudo-family validation within each fold. We fixed the GIDS configuration after CIFAR-10 development and before prospective-observer and CIFAR-100 evaluation: dictionary rank capped at eight, decision rank three, anchor 0.1, listwise temperature 0.25, likelihood-variance multiplier 10, 200 steps, and three seeds. We average results over contents, then observers or families. Paired 10,000-sample bootstrap intervals describe variation across contents and, separately, across observers and families.

6.2 Main results

Architecture-family holdout. GIDS lowers observer-macro normalized regret from 0.1282 to 0.1266 across budgets (Table 1). The improvement appears in all ten direction–budget settings;

Table 1: CIFAR-10 observer-macro normalized selection regret (↓) over 12 held-out observers and two directions. Blocks separate baselines, GIDS variants, and the oracle. Gray marks the primary GIDS; bold and underline mark the lowest and second-lowest non-oracle point estimates. PMF is adapted from a published model; dense fit changes only the GIDS training objective.

Method	$K = 8 \downarrow$	$K = 16 \downarrow$	$K = 32 \downarrow$	$K = 64 \downarrow$	$K = 128 \downarrow$	Mean ↓
Source mean	0.1484	0.1484	0.1484	0.1484	0.1484	0.1484
SVD + ridge	0.1421	0.1402	0.1341	0.1312	0.1260	0.1347
PMF	0.1409	0.1307	0.1335	0.1265	<u>0.1210</u>	0.1305
SVD + EB	0.1338	0.1291	0.1284	0.1258	0.1241	0.1282
GIDS	<u>0.1333</u>	<u>0.1284</u>	<u>0.1264</u>	<u>0.1237</u>	0.1215	<u>0.1266</u>
GIDS (dense fit)	**0.1323**	**0.1276**	**0.1256**	**0.1231**	**0.1203**	**0.1258**
Oracle	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Target-observer regret over two directions and $K \in \{32, 64, 128\}$. "Pros." denotes prospectively frozen observers; the earlier mean covers ResNeSt-152, XCiT-M12, and CARD/LRR. Bold marks lower displayed regret; gray marks the prospective mean. Deltas use unrounded estimates.

Observer	Status	SVD + EB ↓	GIDS ↓	$\Delta \downarrow$	Rel. reduction ↑ (%)
SODEF	Pros.	0.1170	0.1170	−0.00004	0.04
WRN-A4	Pros.	0.0427	**0.0397**	−0.00296	6.94
Prospective mean	—	0.0799	**0.0784**	−0.00150	1.88
Earlier mean	—	0.0680	**0.0672**	−0.00081	1.18

averaging by architecture family also favors GIDS, by 0.00115. Figure 4 summarizes budget trends and frozen-observer effects.

At matched budgets, the paired difference is -0.00159, with a content-bootstrap 95% interval of $[-0.00205, -0.00113]$.

Prospectively frozen observers. We selected SODEF and WRN-A4 as prospective targets before observing their responses. At $K \geq 32$, their mean regret falls from 0.0799 to 0.0784, a 1.88% reduction (Table 2); the paired observer-bootstrap interval is $[-0.00296, -0.00004]$. The gain is driven mainly by WRN-A4, while SODEF changes little. Earlier-frozen targets favor GIDS on average as well.

6.3 Mechanism, ablations, and transfer

Gauge-consistency stress test. Equivalent factorizations should give the same calibrated prediction. Across 20 gauges with scales from 0.25 to 4, transforming the prior with the basis keeps the largest prediction difference near 10^{-13} and regret variance at zero. Reusing the old prior instead gives a mean prediction RMSE of 0.0945 and a mean regret range of 0.0191. Gauge consistency is therefore a property of the target estimator, not merely of source reconstruction.

Decision alignment. To locate the performance gain, we change the objective used to choose the subspace while keeping the rank-8 dictionary, rank-3 parameterization, optimizer, anchor, seeds, queries, and posterior fixed. Reconstruction training returns the leading SVD projector (maximum principal sine 5.98×10^{-4}) and reproduces the SVD + EB control's regret. Training for decisions lowers observer-macro regret from 0.1282 to about 0.1266 (Table 3); changing covariance or noise within that decision basis has a much smaller effect. Direct dense decision fitting reaches 0.1258 on the development domain; we retain the prespecified episodic configuration as the primary transfer method.

The learned subspace also changes candidate choices. GIDS raises selection accuracy by 1.14 percentage points; its pairwise-concordance gain reaches 0.30 points at $K = 128$. Across observers, greater concordance gain accompanies larger regret reduction ($\rho = -0.811$).

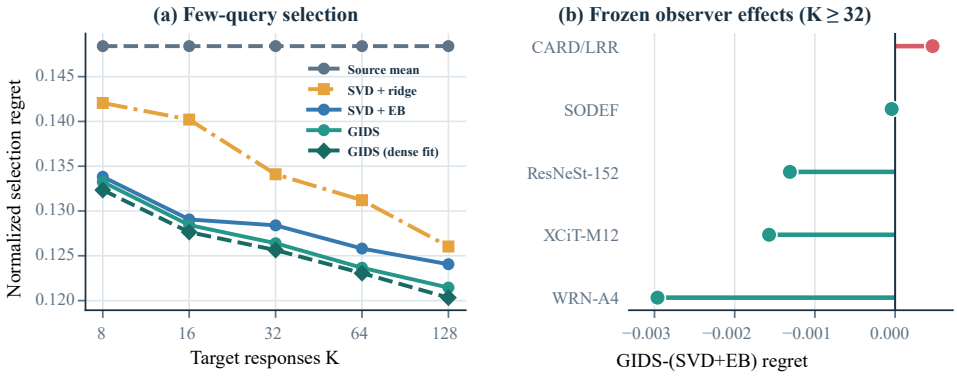


Figure 4: Few-query behavior and observer heterogeneity. (a) Normalized regret across query budgets, macro-averaged over 12 held-out observers and two directions. (b) Paired effects for the five frozen observers, averaged over two directions and $K \in \{32, 64, 128\}$; negative values favor GIDS.

Table 3: Decision-subspace and posterior ablations, macro-averaged over observers, directions, and budgets. Lower normalized regret is better (↓); gray marks the frozen GIDS configuration.

Variant	Normalized regret ↓
Decision basis + stale covariance/noise	0.12654
Decision basis + recomputed covariance	0.12659
GIDS	0.12664
GIDS without projector anchor	0.12665
SVD + EB	0.12823

CIFAR-100 transfer. On CIFAR-100, we retain the frozen configuration and evaluate ten observers from seven families. At $K \geq 32$, family-macro regret falls from 0.1107 to 0.1103; four of six observer-macro direction–budget settings improve (Table 4). The family-macro difference has a family-bootstrap 95% interval of $[-0.00157, 0.00057]$.

7 Discussion

Our experiments separate two requirements of a useful low-rank evaluator. Gauge consistency makes few-query calibration well defined: changing latent coordinates cannot change a prediction when the prior changes with them. Decision alignment determines which response directions deserve space in the low-rank model. The objective swap and posterior ablations attribute the measured regret reduction mainly to this second choice. A direction can account for little global response variance and still matter when it changes the winner of a content group.

Observer-level effects are not uniform. Some observers need little change from the reconstruction subspace; others benefit when the learned basis preserves their candidate comparisons. The same pattern explains why average regret and individual outcomes should be read together. GIDS is most useful when a source response matrix can be reused to calibrate many future models on a fixed candidate collection: the source cost is amortized, while each new observer needs only a few complete groups and a small posterior solve.

The SVD + EB comparison isolates representation choice: target query groups, rank, likelihood multiplier, and posterior solver remain fixed. The 1.14-percentage-point gain in selection accuracy confirms that GIDS changes discrete choices, not just predicted scores. Yet the observer- and family-bootstrap intervals for the primary regret difference cross zero. We interpret the narrower content-level interval as evidence for the evaluated observer collection, not a guarantee for every new architecture family.

Table 4: CIFAR-100 regret without retuning (↓). Direction 1 queries severity 1 and evaluates severity 2; direction 2 reverses them. Direction–budget rows are observer-macro; the gray $K \geq 32$ row is family-macro. Bold marks lower regret per row. Delta is GIDS minus SVD + EB.

Direction	K	SVD + EB ↓	GIDS ↓	Δ ↓
1	32	0.1052	**0.1047**	-0.0005
1	64	**0.0994**	0.0996	0.0002
1	128	**0.0979**	0.0981	0.0002
2	32	0.1200	**0.1197**	-0.0004
2	64	0.1119	**0.1116**	-0.0004
2	128	0.1101	**0.1098**	-0.0003
Macro	$K \geq 32$	0.1107	**0.1103**	-0.0004

For the matched query-policy comparison, SVD + EB and GIDS receive exactly the same complete groups. At $K \geq 32$, their mean regret deltas are -0.00225 under D-optimal queries, -0.00177 under five fixed random-query repeats, and -0.00249 under representative queries. The decision subspace therefore transfers across acquisition rules without policy-specific retraining.

8 Limitations

Our method is inductive over observers but transductive over items: source models must have responses on the candidate collection used at evaluation. An item encoder could extend decision-subspace learning to new images and operations. Each target query also requires task utility for every candidate in a group; for our classifiers, this requires labels. We evaluate classification observers, leaving detection, segmentation, and multimodal systems for future study.

Our query budget counts calls to the new observer, not the cost of constructing or updating the dense source-response bank. That cost can be amortized when many observers share a stable item collection, but changes to images or operations require new source responses. We have not measured this maintenance cost or tested an item-conditioned alternative.

9 Conclusion

Machine-oriented quality depends on the model that will use the image. We studied the cold-start case: an unseen observer must choose among processed candidates after only a few complete-group queries. A low-rank representation learned for response reconstruction can miss the comparisons that decide those choices, and its Bayesian calibrator must carry the prior along when latent coordinates change.

GIDS learns a decision subspace inside a spectral dictionary and calibrates the new observer in that subspace. It lowers mean selection regret in family-held-out evaluation, on prospectively frozen observers, and on a frozen second data domain, while keeping target inference closed form. The broader lesson is that a compact evaluator should be judged by the decisions it supports after calibration, not solely by the responses it reconstructs.

AI use statement

The authors used LLMs for basic proofreading of writing. We reviewed and approved all resulting text.

Reproducibility statement. The appendix documents proofs, the calibration algorithm, data, protocol, hyperparameters, and controls. Anonymous code runs the method and verifies principal paired comparisons from frozen outcomes.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.

Eunsu Baek, Sung-hwan Han, Taesik Gong, and Hyung-Sin Kim. Adaptive camera sensor for vision models. In *International Conference on Learning Representations*, pp. 94681–94706, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ec14daa5c50745f83fb27f685f8dfc22-Paper-Conference.pdf.

Stefano Blumberg, Paddy Slator, and Daniel Alexander. Experimental design for multi-channel imaging via task-driven feature selection. In *International Conference on Learning Representations*, pp. 39998–40024, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aec2dfc4f5e19acf05c15587c889dbc4-Paper-Conference.pdf.

Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/a1140a3d0df1c81e24ae954d935e8926-Abstract.html.

Steven Cao, Percy Liang, and Gregory Valiant. One-sided matrix completion from two observations per row. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3599–3624, 2023. URL https://proceedings.mlr.press/v202/cao23d.html.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024a. doi: 10.1109/TIP.2024.3378466.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752, 2025.

Zewen Chen, Haina Qin, Juan Wang, Chunfeng Yuan, Bing Li, Weiming Hu, and Leon Wang. PromptIQA: Boosting the performance and generalization for no-reference image quality assessment via prompts. In *Computer Vision – ECCV 2024*, pp. 247–264, 2024b. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/75_ECCV_2024_paper.php.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html.

Alan Edelman, Tomas A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Zilin Fang, Andrey Ignatov, Eduard Zamfir, and Radu Timofte. SQAD: Automatic smartphone camera quality assessment and benchmarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20532–20542, 2023.

Ruoyu Feng, Yixin Gao, Xin Jin, Runsen Feng, and Zhibo Chen. Semantically structured image compression via irregular group-based decoupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17237–17247, 2023.

Riccardo Fogliato, Pratik Patil, Nil-Jana Akpinar, and Mathew Monfort. Precise model benchmarking with only a few observations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9563–9575, 2024. doi: 10.18653/v1/2024.emnlp-main.536.

Hakan Emre Gedik, Shashank Gupta, and Alan Bovik. Learning where to look and how to judge: Resolution-agnostic image quality assessment with quality-aware saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37507–37517, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Vishal Gupta and Michael Huang. Decision-focused learning with directional gradients. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/907a9fb75a408f6c3a2ae1bf84c39e44-Abstract-Conference.html.

Neil Houlsby, Jose Miguel Hernandez-Lobato, and Zoubin Ghahramani. Cold-start active learning with robust ordinal matrix factorization. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 766–774, 2014. URL https://proceedings.mlr.press/v32/houlsby14.html.

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024. doi: 10.1017/S0962492924000023.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21329–21372, 2024. URL https://proceedings.mlr.press/v235/jang24e.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Minyoung Kim and Timothy Hospedales. A hierarchical bayesian model for few-shot meta learning. In *International Conference on Learning Representations*, pp. 40069–40096, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/af2df67c0d2f9c9a950278acf3965ffc-Paper-Conference.pdf.

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*, pp. 31734–31770, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4ebc26584810a189ef1e4f173aba4319-Paper-Conference.pdf.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24557–24570, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9b9cfd5428153ccfbd4ba34b7e007305-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Aobo Li, Jinjian Wu, Yongxu Liu, and Leida Li. Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28422–28431, 2024.

Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, Weisi Lin, and Guangtao Zhai. Image quality assessment: From human to machine preference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7570–7581, 2025a.

Chunyi Li, Jiahao Xiao, Jianbo Zhang, Farong Wen, Zicheng Zhang, Yuan Tian, Xiangyang Zhu, Xiaohong Liu, Zhengxue Cheng, Weisi Lin, and Guangtao Zhai. Image quality assessment for embodied AI. In *International Conference on Learning Representations*, pp. 39960–39982, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4310ae054ce265e56d8ea897971149b5-Paper-Conference.pdf.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Jun Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3479362c2eb7321ba0490d98eb032772-Abstract-Conference.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947, 2022. URL https://proceedings.mlr.press/v162/mandi22a.html.

Akshay Mehra, Yunbei Zhang, and Jihun Hamm. Understanding the transferability of representations via task-relatedness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d3602fc92fb8b9e0d55356c9e8815e2b-Abstract-Conference.html.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735, 2021. URL https://proceedings.mlr.press/v139/miller21b.html.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html.

Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/d7322ed717dedf1eb4e6e52a37ea7bcd-Abstract.html.

Hieu Trung Nguyen, Duy Nguyen, Khoa Doan, and Viet Anh Nguyen. Cold-start recommendation by personalized embedding region elicitation. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2766–2786, 2024. URL https://proceedings.mlr.press/v244/nguyen24a.html.

Panagiotis Papastamoulis and Ioannis Ntzoufras. On the identifiability of Bayesian factor analytic models. *Statistics and Computing*, 32(2):23, 2022. doi: 10.1007/s11222-022-10084-4.

Michael Poli, Stefano Massaroli, Stefano Ermon, Bryan Wilder, and Eric Horvitz. Ideal abstractions for decision-focused learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10223–10234, 2023. URL https://proceedings.mlr.press/v206/poli23a.html.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

Carlos Riquelme, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active learning for accurate estimation of linear models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2931–2939, 2017. URL https://proceedings.mlr.press/v70/riquelme17a.html.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pp. 880–887, 2008. doi: 10.1145/1390156.1390267.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/85690f81aadc1749175c187784afc9ee-Abstract.html.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0904c7edde20d7134a77fc7f9cd86ea2-Abstract-Conference.html.

Jinsong Shi, Pan Gao, Xiaojiang Peng, and Jie Qin. DSMix: Distortion-induced sensitivity map based pre-training for no-reference image quality assessment. In *Computer Vision – ECCV 2024*, pp. 1–17, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/08824.pdf.

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60238–60265, 2025. URL https://proceedings.mlr.press/v267/truong25c.html.

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*, pp. 1576–1601, 2024. doi: 10.18653/v1/2024.eacl-long.95.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Xiaoqi Wang, Yun Zhang, and Weisi Lin. Image quality assessment for machines: Paradigm, large-scale database, and models. *arXiv preprint arXiv:2508.19850*, 2025. URL https://arxiv.org/abs/2508.19850.

Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2763–2772, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/363d4c97bb411e4b07612915b76c06ae-Abstract-Conference.html.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024c. URL https://proceedings.mlr.press/v235/wu24ah.html.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison. In *Computer Vision – ECCV 2024*, pp. 360–377, 2024d. doi: 10.1007/978-3-031-72646-0_21.

Liwei Wu, Cho-Jui Hsieh, and James Sharpnack. SQL-Rank: A listwise approach to collaborative ranking. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5315–5324, 2018. URL https://proceedings.mlr.press/v80/wu18c.html.

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. In *Computer Vision – ECCV 2024*, pp. 143–160, 2024e. doi: 10.1007/978-3-031-72904-1_9.

Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, and Weisi Lin. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2662–2672, 2024.

Shengqi Xu, Xinpeng Zhou, Yabo Zhang, Ming Liu, Tao Liang, Tianyu Zhang, Yalong Bai, Zuxuan Wu, and Wangmeng Zuo. PreferThinker: Reasoning-based personalized image preference assessment. In *International Conference on Learning Representations*, pp. 107510–107539, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/af31604708f3e44b4de9fdfa6dcaa9d1-Paper-Conference.pdf.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *Computer Vision – ECCV 2024*, pp. 259–276, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6356_ECCV_2024_paper.php.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025.

Wei-Hsiang Yu, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Ranking-aware adapter for text-driven image ordering with CLIP. In *International Conference on Learning Representations*, pp. 80614–80633, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c86d3f43371270c70ed94e024852895d-Paper-Conference.pdf.

Kun Yuan, Hongbo Liu, Mading Li, Muyi Sun, Ming Sun, Jiachao Gong, Jinhua Hao, Chao Zhou, and Yansong Tang. PTM-VQA: Efficient video quality assessment leveraging diverse pretrained models from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2835–2845, 2024.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In *Advances in Neural Information Processing Systems*, volume 38, pp. 135638–135678, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c57bacb5ea621582235c54ce2a83136d-Abstract-Conference.html.

Peiyan Zhang, Haoyang Liu, Chaozhuo Li, Xing Xie, Sunghun Kim, and Haohan Wang. Foundation model-oriented robustness: Robust image model evaluation with pretrained models. In *International Conference on Learning Representations*, pp. 9050–9078, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25b0c393cb90a1a2f68b643de55407af-Paper-Conference.pdf.

Taolin Zhang, Hang Guo, Wang Lu, Tao Dai, Shu-Tao Xia, and Jindong Wang. SparseEval: Efficient evaluation of large language models by sparse optimization. In *International Conference on Learning Representations*, pp. 108282–108296, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b059de4629cccb411ef92d0d4157e87e-Paper-Conference.pdf.

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/137cb5dd61b2685bd2623967daee6860-Abstract-Conference.html.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023.

Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, and Yu Wang. FlashEval: Towards fast and accurate evaluation of text-to-image diffusion generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16122–16131, 2024.

Jin Zhou, Christian Belardi, Ruihan Wu, Travis Zhang, Carla Gomes, Wen Sun, and Kilian Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations*, pp. 65092–65111, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a38cfba4c717e5a37b200294615e546e-Paper-Conference.pdf.

A PROOFS

A.1 GAUGE-COVARIANT CALIBRATION

Let $\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S} = \mathbf{B}_\mathcal{S}\mathbf{z} + \boldsymbol{\epsilon}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Then

$$\boldsymbol{\Sigma}_{\mathbf{z}|\mathcal{S}} = (\mathbf{C}^{-1} + \sigma^{-2}\mathbf{B}_\mathcal{S}^\top\mathbf{B}_\mathcal{S})^{-1}, \quad \hat{\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{z}|\mathcal{S}}\sigma^{-2}\mathbf{B}_\mathcal{S}^\top(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S}).$$

Under $\mathbf{B}' = \mathbf{B}\mathbf{H}$ and $\mathbf{C}' = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-\top}$, the precision becomes $\mathbf{P}' = \mathbf{H}^\top\mathbf{P}\mathbf{H}$ and the information vector becomes $\mathbf{b}' = \mathbf{H}^\top\mathbf{b}$. Hence $\hat{\mathbf{z}}' = \mathbf{H}^{-1}\hat{\mathbf{z}}$ and $\mathbf{B}'_\mathcal{E}\hat{\mathbf{z}}' = \mathbf{B}_\mathcal{E}\hat{\mathbf{z}}$. If we keep $\mathbf{C}$ fixed, this equality can fail. With one observation, $\mathbf{B} = 1$, and $\mathbf{C} = c$, the predictive shrinkage coefficient changes from $c/(\sigma^2 + c)$ to $h^2c/(\sigma^2 + h^2c)$.

For $\mathbf{B}_\mathbf{Q} = \mathbf{F}\mathbf{Q}$ and $\mathbf{Q}' = \mathbf{Q}\mathbf{O}$, we have $\mathbf{B}_{\mathbf{Q}'} = \mathbf{B}_\mathbf{Q}\mathbf{O}$ and $\mathbf{A}_{\mathbf{Q}'} = \mathbf{A}_\mathbf{Q}\mathbf{O}$. An orthogonally equivariant covariance estimator gives $\mathbf{C}_{\mathbf{Q}'} = \mathbf{O}^\top\mathbf{C}_\mathbf{Q}\mathbf{O}$. The transformation leaves residual reconstruction and $\mathbf{Q}\mathbf{Q}^\top$ unchanged. Posterior predictions, episodic loss, noise, and the anchor are therefore invariant.

A.2 CALIBRATION-MARGIN REGRET BOUND

For one group, write $\mathbf{y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{z} + \boldsymbol{\delta}$ and $\hat{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{B}\hat{\mathbf{z}}$. We hold the realized true group $\mathbf{y}$ fixed and take probability only over calibration noise. Write $\mathbf{z} - \hat{\mathbf{z}} = \mathbf{h}_\mathcal{S} + \boldsymbol{\eta}_\mathcal{S}$, where $\mathbf{h}_\mathcal{S}$ is the conditional estimation bias and every $\mathbf{w}^\top\boldsymbol{\eta}_\mathcal{S}$ is zero-mean sub-Gaussian with variance proxy $\mathbf{w}^\top\mathbf{V}_\mathcal{S}\mathbf{w}$. Let $a^* = \arg\max_a y_a$, $\Delta_a = y_{a^*} - y_a$, and $\mathbf{w}_a = \mathbf{b}_{a^*} - \mathbf{b}_a$. Here $\mathbf{b}_a^\top$ is row a of the group basis $\mathbf{B}$. The predicted margin satisfies

$$\hat{y}_{a^*} - \hat{y}_a = \Delta_a - \left[(\delta_{a^*} - \delta_a) + \mathbf{w}_a^\top(\mathbf{z} - \hat{\mathbf{z}})\right].$$

If candidate a beats a^* in predicted utility, then

$$\mathbf{w}_a^\top\boldsymbol{\eta}_\mathcal{S} \geq \Delta_a - (\delta_{a^*} - \delta_a) - \mathbf{w}_a^\top\mathbf{h}_\mathcal{S} \geq \Delta_a - 2\varepsilon_{\text{rep}} - |\mathbf{w}_a^\top\mathbf{h}_\mathcal{S}|.$$

Let $\beta_a = |\mathbf{w}_a^\top\mathbf{h}_\mathcal{S}|$ and $v_a = \mathbf{w}_a^\top\mathbf{V}_\mathcal{S}\mathbf{w}_a$. The sub-Gaussian Chernoff bound gives, for $v_a > 0$,

$$\Pr_{\text{cal}}(\hat{y}_a \geq \hat{y}_{a^*} \mid \mathbf{y}, \mathcal{S}) \leq \min\left\{1, \exp\left[-\frac{(\Delta_a - 2\varepsilon_{\text{rep}} - \beta_a)_+^2}{2v_a}\right]\right\}.$$

When $v_a = 0$, the centered error projection is zero almost surely, yielding the convention in Proposition 2. Since selecting a implies $\hat{y}_a \geq \hat{y}_{a^*}$, summing $\Delta_a \Pr_{\text{cal}}(\hat{a} = a \mid \mathbf{y}, \mathcal{S})$ over $a \neq a^*$ proves Proposition 2. The fixed-response conditioning prevents the true winner and margins from changing inside the expectation.

Under a gauge transform, $\mathbf{w}'_a = \mathbf{H}^\top\mathbf{w}_a$, $\mathbf{h}'_\mathcal{S} = \mathbf{H}^{-1}\mathbf{h}_\mathcal{S}$, and $\mathbf{V}'_\mathcal{S} = \mathbf{H}^{-1}\mathbf{V}_\mathcal{S}\mathbf{H}^{-\top}$. Direct substitution gives $\beta'_a = \beta_a$ and $v'_a = v_a$.

For the posterior mean in Section 4.1, let

$$\mathbf{A}_\mathcal{S} = (\mathbf{C}^{-1} + \tilde{\sigma}^{-2}\mathbf{B}_\mathcal{S}^\top\mathbf{B}_\mathcal{S})^{-1}\tilde{\sigma}^{-2}\mathbf{B}_\mathcal{S}^\top.$$

Writing the queried centered response as $\mathbf{B}_\mathcal{S}\mathbf{z} + \boldsymbol{\delta}_\mathcal{S} + \boldsymbol{\epsilon}_\mathcal{S}$ gives

$$\mathbf{z} - \hat{\mathbf{z}} = (\mathbf{I} - \mathbf{A}_\mathcal{S}\mathbf{B}_\mathcal{S})\mathbf{z} - \mathbf{A}_\mathcal{S}\boldsymbol{\delta}_\mathcal{S} - \mathbf{A}_\mathcal{S}\boldsymbol{\epsilon}_\mathcal{S}, \tag{3}$$

$$\mathbf{h}_\mathcal{S} = (\mathbf{I} - \mathbf{A}_\mathcal{S}\mathbf{B}_\mathcal{S})\mathbf{z} - \mathbf{A}_\mathcal{S}\boldsymbol{\delta}_\mathcal{S}. \tag{4}$$

If $\boldsymbol{\epsilon}_\mathcal{S}$ is Gaussian with covariance $\boldsymbol{\Sigma}_\epsilon$, then $\mathbf{V}_\mathcal{S} = \mathbf{A}_\mathcal{S}\boldsymbol{\Sigma}_\epsilon\mathbf{A}_\mathcal{S}^\top$. This identifies the two calibration terms in Proposition 2 with the shrinkage and noise propagation of the implemented posterior.

A.3 REPRESENTATION DECOMPOSITION

Because $\mathbf{B}_\mathbf{Q} = \mathbf{F}\mathbf{Q}$ and both factors have orthonormal columns,

$$\mathbf{I} - \mathbf{B}_\mathbf{Q}\mathbf{B}_\mathbf{Q}^\top = \mathbf{I} - \mathbf{F}\mathbf{Q}\mathbf{Q}^\top\mathbf{F}^\top \tag{5}$$

$$= (\mathbf{I} - \mathbf{F}\mathbf{F}^\top) + \mathbf{F}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{F}^\top. \tag{6}$$

Multiplication by $\mathbf{u}$ proves Proposition 3; the ℓ_∞ bound follows from the triangle inequality. For the source matrix, the two residual components are orthogonal, so their squared Frobenius norms add. Explicitly, $\varepsilon_{\text{dict}} = \|(\mathbf{I} - \mathbf{F}\mathbf{F}^\top)\mathbf{u}\|_\infty$ and $\varepsilon_{\text{select}}(\mathbf{Q}) = \|\mathbf{F}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{F}^\top\mathbf{u}\|_\infty$; ε_{rep} may be taken as the infinity norm of their summed residual.

A.4 LISTWISE CALIBRATION AND EPISODIC SEARCH

Let $\mathbf{p} = \mathrm{softmax}(\mathbf{y}/\tau)$ and $\mathbf{q} = \mathrm{softmax}(\hat{\mathbf{y}}/\tau)$. Their maximizers match those of $\mathbf{y}$ and $\hat{\mathbf{y}}$. For $\hat{a} = \arg\max_a q_a$,

$$p_{a^*} - p_{\hat{a}} \le 2\,\mathrm{TV}(\mathbf{p}, \mathbf{q}) \le \sqrt{2\,\mathrm{KL}(\mathbf{p}\|\mathbf{q})}.$$

If $p_a \ge \rho > 0$ for all candidates, the logarithm is $1/\rho$-Lipschitz on the relevant interval, which yields

$$y_{a^*} - y_{\hat{a}} = \tau \log \frac{p_{a^*}}{p_{\hat{a}}} \le \frac{\tau}{\rho}\sqrt{2\,\mathrm{KL}(\mathbf{p}\|\mathbf{q})}.$$

Cross-entropy differs from this KL divergence only by the fixed target entropy, so the episodic loss is a calibrated selection surrogate under the probability-floor condition.

Let $D(\mathbf{Q}, \mathbf{Q}_0) = \|\mathbf{Q}\mathbf{Q}^\top - \mathbf{Q}_0\mathbf{Q}_0^\top\|_{\mathrm{F}}^2$. Since $\mathbf{Q}_0$ is feasible and $D(\mathbf{Q}_0, \mathbf{Q}_0) = 0$, a ξ-suboptimal solution obeys

$$\widehat{\mathcal{R}}_{\mathrm{epi}}(\hat{\mathbf{Q}}) + \frac{\lambda}{r^2} D(\hat{\mathbf{Q}}, \mathbf{Q}_0) \le \widehat{\mathcal{R}}_{\mathrm{epi}}(\mathbf{Q}_0) + \xi.$$

Fix the dictionary and the map from each subspace to its statistical model independently of the episodes used in this bound. Assume one independent, identically distributed episode per source observer, with loss in $[0, L]$ and L_Q-Lipschitz dependence on projector distance. An η-cover of $\mathrm{Gr}(r, d)$ has size at most $(c/\eta)^{d(r-d)}$. Applying Hoeffding's inequality on the cover and extending by Lipschitz continuity gives, with probability at least $1 - \delta$,

$$\sup_{\mathbf{Q}} |\mathcal{R}(\mathbf{Q}) - \widehat{\mathcal{R}}(\mathbf{Q})| \le L\sqrt{\frac{d(r-d)\log(c/\eta) + \log(2/\delta)}{2M}} + 2L_Q\eta =: \Gamma_M.$$

Combining uniform convergence with empirical optimality yields $\mathcal{R}(\hat{\mathbf{Q}}) \le \mathcal{R}(\mathbf{Q}_0) + 2\Gamma_M + \xi$. This conditional search bound describes the subspace class. Reusing the same source population to estimate the dictionary and prior, as in our experiments, does not by itself satisfy its independence assumptions; optimization steps are not additional independent observers.

A.5 RECONSTRUCTION–SELECTION COUNTEREXAMPLE

Consider a two-candidate group with $\mathbf{y} = (1, 0)$, $\hat{\mathbf{y}}_1 = (0.49, 0.51)$, and $\hat{\mathbf{y}}_2 = (2, 0)$. Then $\|\mathbf{y} - \hat{\mathbf{y}}_1\|_2^2 = 0.5202 < 1 = \|\mathbf{y} - \hat{\mathbf{y}}_2\|_2^2$, but $\hat{\mathbf{y}}_1$ selects the second candidate and incurs regret 1, whereas $\hat{\mathbf{y}}_2$ selects the first candidate and incurs regret 0.

B MULTIMODAL OBSERVER-PRIOR EXTENSION

The source-centered kernel-mixture prior can be used with a GIDS decision subspace. We evaluate it on frozen subspaces from GIDS with dense fitting, keeping each learned subspace and query policy fixed while changing only the observer prior. In this appendix, τ denotes kernel bandwidth, distinct from the listwise temperature. Let $\mathbf{u}_j$ be source observer j in the current coordinates and set

$$p(\mathbf{z}) = \frac{1}{M}\sum_{j=1}^{M} \mathcal{N}(\mathbf{z}; \mathbf{u}_j, \tau\mathbf{C}).$$

With $\tilde{\sigma}^2 = \kappa\sigma^2$ and $\mathbf{d}_{\mathcal{S}} = \mathbf{y}_{\mathcal{S}} - \boldsymbol{\mu}_{\mathcal{S}}$, all mixture components share posterior covariance

$$\boldsymbol{\Sigma}_\tau = \left[(\tau\mathbf{C})^{-1} + \tilde{\sigma}^{-2}\mathbf{B}_{\mathcal{S}}^\top\mathbf{B}_{\mathcal{S}}\right]^{-1}$$

and component j has posterior mean

$$\mathbf{m}_j = \boldsymbol{\Sigma}_\tau\left[(\tau\mathbf{C})^{-1}\mathbf{u}_j + \tilde{\sigma}^{-2}\mathbf{B}_{\mathcal{S}}^\top\mathbf{d}_{\mathcal{S}}\right].$$

Let $\mathbf{S}_\tau = \tilde{\sigma}^2\mathbf{I} + \mathbf{B}_{\mathcal{S}}(\tau\mathbf{C})\mathbf{B}_{\mathcal{S}}^\top$. The component weight is

$$\pi_j \propto \exp\left[-\frac{1}{2}(\mathbf{d}_{\mathcal{S}} - \mathbf{B}_{\mathcal{S}}\mathbf{u}_j)^\top\mathbf{S}_\tau^{-1}(\mathbf{d}_{\mathcal{S}} - \mathbf{B}_{\mathcal{S}}\mathbf{u}_j)\right], \qquad \sum_j \pi_j = 1,$$

where the common determinant cancels. The predictive mean remains closed form:

$$\hat{\mathbf{y}}_{\mathcal{E}} = \boldsymbol{\mu}_{\mathcal{E}} + \mathbf{B}_{\mathcal{E}} \sum_{j=1}^{M} \pi_j \mathbf{m}_j.$$

An atomic empirical prior, obtained as $\tau \to 0$, can only average observed source coordinates and deteriorates when its effective source count collapses. Positive τ identifies a source mode while allowing the target coordinate to move away from its center. We select $\tau = 0.1$ on the development corpus and freeze it for all confirmations.

For an orthogonal gauge $\mathbf{B}' = \mathbf{BO}$, $\mathbf{u}'_j = \mathbf{O}^\top \mathbf{u}_j$, and $\mathbf{C}' = \mathbf{O}^\top \mathbf{CO}$, we have $\boldsymbol{\Sigma}'_\tau = \mathbf{O}^\top \boldsymbol{\Sigma}_\tau \mathbf{O}$, $\mathbf{m}'_j = \mathbf{O}^\top \mathbf{m}_j$, and $\pi'_j = \pi_j$. Decoded predictions are therefore unchanged. A random-rotation stress test gives maximum absolute prediction error 1.11×10^{-16}. Because the components share a covariance, one rank-d inverse serves all of them; calibration remains low-dimensional and closed form.

Table 5: Kernel-mixture versus Gaussian calibration on frozen GIDS subspaces trained with dense fitting, averaged over both directions and all five budgets. CIFAR-10 uses observer-macro averaging; CIFAR-100 uses family-macro averaging. Bold marks lower regret per row; $\Delta = \text{Kernel} - \text{Gaussian}$ uses unrounded estimates. The bandwidth is selected on development data and then fixed.

Evaluation	Gaussian posterior	Kernel posterior	Δ
CIFAR-10 development	0.1258	**0.1237**	-0.0021
Frozen SODEF/WRN-A4	0.0807	**0.0789**	-0.0018
Frozen CIFAR-100	0.1140	**0.1078**	-0.0062

Table 6: Five confirmation observers with frozen GIDS subspaces trained with dense fitting, averaged over directions and budgets. Bold marks lower regret per row; gray marks the equal-observer mean. Deltas use unrounded estimates.

Observer	Gaussian posterior	Kernel posterior	Δ
SODEF	**0.1201**	0.1221	$+0.0020$
WRN-A4	0.0413	**0.0357**	-0.0057
ResNeSt-152	0.0372	**0.0335**	-0.0037
CARD/LRR	**0.1363**	0.1367	$+0.0004$
XCiT-M12	0.0463	**0.0387**	-0.0076
Equal-observer mean	0.0762	**0.0733**	-0.0029

Across SODEF and WRN-A4, 15 of 20 direction–budget cells improve. On frozen CIFAR-100, 59 of 70 family–direction–budget cells and six of seven family means improve. Across all five confirmation observers, 36 of 50 cells and three observer means improve; the equal-observer difference is -0.0029, with observer-bootstrap interval $[-0.00603, 0.00022]$. The gains vary across observers, reflecting the heterogeneous response modes captured by the mixture prior.

C Additional Controls

Budget comparison and sampling variation. At $K = 64$, GIDS obtains 0.1237 regret, compared with 0.1241 for SVD + EB at $K = 128$. The paired half-budget difference is -0.00040, with content-bootstrap 95% interval $[-0.00144, 0.00061]$. For the matched-budget primary comparison, resampling the 12 observers or six architecture families gives intervals $[-0.00441, 0.00059]$ and $[-0.00378, 0.00093]$, respectively. Content resampling measures variation for the fixed observer collection. The prospective observer-bootstrap interval in the main text resamples only SODEF and WRN-A4. On CIFAR-100, the content-bootstrap interval for the primary family-macro difference is $[-0.00074, 0.00005]$; both this interval and the family-bootstrap interval cross zero.

Tables 7–11 report rank and anchor sensitivity, query-policy controls, prior-structure controls, correctness, and matched selection-aware alternatives. Table 9 gives the direction–budget results for the published-method adaptations under the main continuous-utility protocol.

Table 7: Rank and projector-anchor sensitivity at $K \geq 32$. Regret and Δ from the frozen configuration are lower (↓); gray marks that configuration.

d	λ	Normalized regret ↓	Δ from frozen ↓
2	0.1	0.1287	0.0049
3	0.03	0.1238	-6.7×10^{-6}
3	0.1	0.1238	0
3	0.3	0.1238	7.8×10^{-6}
4	0.1	0.1236	-0.0003

Table 8: Query, prior, and correctness controls for GIDS. Selected correctness is the classification accuracy of the chosen candidate; the gap is oracle correctness minus selected correctness. Regret and gap are lower (↓); selected correctness is higher (↑).

Control	Variant	K	Normalized regret ↓	Selected correctness ↑	Correctness gap to best ↓
Prior structure	diagonal	8	0.1330	0.8779	0.0481
Prior structure	diagonal	16	0.1283	0.8801	0.0460
Prior structure	diagonal	32	0.1264	0.8810	0.0451
Prior structure	diagonal	64	0.1237	0.8808	0.0452
Prior structure	diagonal	128	0.1214	0.8816	0.0444
Prior structure	Full shrinkage	8	0.1333	0.8780	0.0481
Prior structure	Full shrinkage	16	0.1284	0.8800	0.0461
Prior structure	Full shrinkage	32	0.1264	0.8810	0.0451
Prior structure	Full shrinkage	64	0.1237	0.8808	0.0453
Prior structure	Full shrinkage	128	0.1215	0.8817	0.0444
Prior structure	isotropic	8	0.1345	0.8773	0.0488
Prior structure	isotropic	16	0.1293	0.8795	0.0466
Prior structure	isotropic	32	0.1267	0.8808	0.0453
Prior structure	isotropic	64	0.1240	0.8809	0.0452
Prior structure	isotropic	128	0.1217	0.8815	0.0446
Query policy	D-optimal	8	0.1333	0.8780	0.0481
Query policy	D-optimal	16	0.1284	0.8800	0.0461
Query policy	D-optimal	32	0.1264	0.8810	0.0451
Query policy	D-optimal	64	0.1237	0.8808	0.0453
Query policy	D-optimal	128	0.1215	0.8817	0.0444
Query policy	random	8	0.1359	0.8777	0.0484
Query policy	random	16	0.1310	0.8794	0.0467
Query policy	random	32	0.1282	0.8802	0.0459
Query policy	random	64	0.1252	0.8806	0.0455
Query policy	random	128	0.1234	0.8813	0.0448
Query policy	representative	8	0.1400	0.8770	0.0491
Query policy	representative	16	0.1313	0.8785	0.0475
Query policy	representative	32	0.1282	0.8797	0.0463
Query policy	representative	64	0.1246	0.8809	0.0452
Query policy	representative	128	0.1225	0.8815	0.0446

Matched published-method adaptations. PMF applies the Gaussian factor model of Mnih & Salakhutdinov (2007) to the same doubly centered source-observer response matrix used by the spectral baselines, with the source item mean as an intercept. We use maximum a posteriori (MAP) point estimation for the factors. For factor rank d, the dense-source objective is $\frac{1}{2}\|\mathbf{R} - \mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}^2 + \frac{\lambda}{2}(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2)$; its global low-rank solution soft-thresholds the source singular values by λ. We infer a held-out observer factor from its queried complete groups by regularized least squares with the

same λ. Rank $d \in \{1, 2, 3, 5\}$ and $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ are selected separately for each outer family, direction, and budget by equal-family regret in nested source-only pseudo-family validation. Our implementation adapts the published PMF model to this task; it does not reproduce the original authors' code. Because the source matrix is dense, its fitted directions remain spectral; PMF is therefore not an independent decision-focused factor learner.

The Anchor Point Predictor adapts the instance-level predictor of Vivek et al. (2024): for each evaluation item, it uses the source-fitted univariate regression with the smallest residual error among queried anchor responses. To keep the target information identical to GIDS, it receives the same complete-group query order rather than the original single-item medoid acquisition; continuous utilities replace bounded class probabilities. It is therefore a grouped-query predictor adaptation, not the full original Anchor Points pipeline. Both adaptations share the main folds, source-item access, target query budgets, and normalized selection regret. They were evaluated retrospectively on the existing CIFAR-10 response collection, after GIDS development.

Across all budgets, GIDS has regret 0.1266 versus PMF's 0.1305 and the adapted Anchor Point Predictor's 0.1627. The paired GIDS-minus-PMF difference is -0.00387; GIDS improves eight of ten direction–budget means, eight of twelve observer means, and four of six family means. PMF is better at direction 1, $K = 128$ and direction 2, $K = 64$ (Table 9). Paired observer- and family-bootstrap 95% intervals for the all-budget difference are $[-0.00882, 0.00006]$ and $[-0.00803, 0.00127]$, respectively; both cross zero. Thus the comparison supports a lower mean regret on this evaluated collection, not uniform superiority over new architecture families.

Table 9: Published-method adaptations and matched SVD control under the continuous-utility, complete-group protocol. Direction 1 queries severity 1 and evaluates severity 2; direction 2 reverses them. Entries are observer-macro normalized regret (↓); bold marks the lowest displayed method regret in each row. Negative GIDS−PMF favors GIDS. The Anchor Point Predictor uses the matched group-query order, not its original medoid acquisition.

Direction	K	SVD + EB ↓	PMF ↓	GIDS ↓	Anchor Point Predictor ↓	GIDS−PMF ↓
1	8	0.13014	0.13820	**0.12958**	0.15307	-0.00861
1	16	0.12180	0.12503	**0.12149**	0.15260	-0.00354
1	32	0.12061	0.12687	**0.11808**	0.16377	-0.00878
1	64	0.11951	0.12300	**0.11620**	0.16734	-0.00680
1	128	0.11752	**0.11366**	0.11469	0.16563	0.00103
2	8	0.13749	0.14352	**0.13692**	0.16133	-0.00659
2	16	0.13632	0.13638	**0.13537**	0.16144	-0.00101
2	32	0.13618	0.14012	**0.13472**	0.16614	-0.00540
2	64	0.13212	**0.13009**	0.13113	0.16730	0.00104
2	128	0.13060	0.12826	**0.12821**	0.16839	-0.00004

Decision alignment and deployment matching. We compare GIDS with two rank-three alternatives under the same family holdout, query order, budgets, seeds, coherent empirical-Bayes deployment, and normalized regret. The dense listwise alternative in Table 11 learns an unrestricted item basis; the unrestricted episodic alternative differentiates through posterior calibration but optimizes all item coordinates directly. GIDS improves 103 of 120 and 106 of 120 observer–direction–budget cells, respectively. Their observer-bootstrap intervals for GIDS-minus-control all-budget regret are $[-0.01878, -0.00444]$ and $[-0.01009, -0.00438]$.

Those controls change both the objective and the representation family, so we also swap the objective inside the same rank-eight spectral dictionary. We hold fixed the rank-three Stiefel parameterization, leading-SVD initialization, projector anchor, optimizer, 200 steps, seeds, query order, coherent covariance and noise, target posterior, and metric. Under a common observer-macro aggregation, dense within-content listwise loss lowers development regret from 0.1266 to 0.1258; the paired content-bootstrap 95% interval for dense-minus-episodic is $[-0.00128, -0.00040]$. The observer-bootstrap interval $[-0.00192, 0.00014]$ reflects architecture heterogeneity. Table 10 also reports the family-macro comparison.

With every target-calibration group available, dense and episodic subspaces reach 0.1191 and 0.1199 regret; least-squares target coordinates give nearly identical values. SVD + EB remains at about 0.1238. Dense fitting therefore improves the represented decision structure in this protocol. On CIFAR-100, the difference between GIDS (dense fit) and SVD + EB at the prespecified $K \geq 32$ endpoint is -9.2×10^{-5} (content-bootstrap 95% CI $[-4.26 \times 10^{-4}, 2.35 \times 10^{-4}]$), compared with -3.50×10^{-4} for episodic GIDS. Across five frozen confirmation observers—two prospective and three earlier-frozen—dense improves three observer means and episodic improves four. Dense fitting is stronger on the development data, while episodic simulation shows better transfer stability in these evaluations.

Table 10: Objective swap inside the same spectral dictionary: GIDS uses calibration episodes, whereas GIDS (dense fit) directly fits the decision loss. Regret uses matched aggregation; bold marks lower regret per row, and negative delta favors dense fitting.

Aggregation	GIDS ↓	GIDS (dense fit) ↓	Dense fit minus GIDS ↓	Dense fit wins ↑	Cells
Observer macro	0.1266	**0.1258**	-0.0008	69	120
Family macro	0.1210	**0.1203**	-0.0008	38	60

Table 11: Matched alternatives with unrestricted item-space representations, distinct from GIDS (dense fit) within the spectral dictionary. Bold marks lower regret per row. Delta is GIDS minus the alternative; negative delta favors GIDS.

Alternative	Alternative regret ↓	GIDS regret ↓	GIDS minus alternative ↓	Improved cells ↑	Total cells
Unrestricted episodic subspace	0.1338	**0.1266**	-0.0071	106	120
Unrestricted item-space listwise subspace	0.1379	**0.1266**	-0.0113	103	120

Decision-error composition. The mean-regret change corresponds to discrete candidate decisions rather than only small score perturbations. Table 12 compares the frozen GIDS seeds and SVD + EB predictions across 720,000 observer–direction–budget–seed–content comparisons. GIDS repairs 23,862 baseline selection errors while introducing 15,684 new errors, for 8,178 net repairs. Its mean regret change remains negative after removing any single observer (observer-macro range $[-0.00192, -0.00063]$). On contents in the top quartile of the true top-two margin, the mean GIDS-minus-baseline regret is -0.00372 across all budgets and -0.00517 for $K \geq 32$. Thus the aggregate gain includes fewer consequential argmax errors, not merely changes among near-tied candidates.

Table 12: Frozen top-choice transitions between SVD + EB and GIDS. Each GIDS seed is paired with the same baseline; cells repeat contents across seeds, budgets, directions, and observers. Δ is GIDS minus baseline, so negative cumulative regret favors GIDS.

Decision outcome	Cells	Fraction (%)	Cumulative Δ regret
Both correct	231,936	32.21	0.00
Baseline error repaired	23,862	3.31	-3282.87
New error introduced	15,684	2.18	$+2195.70$
Both wrong	448,518	62.29	-56.32

D Training and Deployment Procedure

Algorithm 1 separates source-only representation learning from target calibration. The new observer supplies only its queried complete groups; its responses never update the item subspace, prior, or noise model. The source observers supplying training episodes also contribute to the dictionary and

Algorithm 1 GIDS training and unseen-observer selection on a fixed item bank

Require: Source responses, $(\mathbf{R}, \boldsymbol{\mu})$, query order, ranks (r, d), training budgets, K.
Source training.
$\mathbf{F} \leftarrow$ leading r right-singular vectors of $\mathbf{R}$; $\mathbf{W} \leftarrow \mathbf{Q}_0$.
for each training step **do**
  QR-orthogonalize $\mathbf{W}$ to $\mathbf{Q}$; set $\mathbf{B}_\mathbf{Q} = \mathbf{F}\mathbf{Q}$.
  Recompute $(\mathbf{C}_\mathbf{Q}, \tilde{\sigma}^2_\mathbf{Q})$ from $\mathbf{R}$ in basis $\mathbf{B}_\mathbf{Q}$.
  for each source observer m **do**
    $\mathcal{S} \leftarrow$ scheduled query prefix; $\mathcal{E} \leftarrow$ disjoint held-out groups.
    Infer $\hat{\mathbf{z}}_{m,\mathcal{S}}$; evaluate $\mathcal{L}_{\mathrm{sel}}$ on $\mathcal{E}$.
  end for
  Update $\mathbf{W}$ via mean listwise loss and projector anchor.
end for
Recompute and freeze $(\mathbf{B}_\mathbf{Q}, \mathbf{C}_\mathbf{Q}, \tilde{\sigma}^2_\mathbf{Q})$.
Unseen-observer deployment.
Query K/A complete groups; observe $\mathbf{y}_\mathcal{S}$.
Infer $\hat{\mathbf{z}}_\mathcal{S}$ with the frozen posterior.
for each unqueried group x **do**
  $\hat{\mathbf{r}}_x \leftarrow \boldsymbol{\mu}_x + \mathbf{B}_{\mathbf{Q},x}\hat{\mathbf{z}}_\mathcal{S}$.
  $\hat{a}(x) \leftarrow \arg\max_a \hat{r}_{xa}$.
end for
Ensure: Selected candidate $\hat{a}(x)$ for each unqueried group.

prior, while the target architecture family remains excluded throughout learning.

D.1 Implementation details

We use dictionary width $r = \min(8, M-1)$ and decision dimension $d = 3$. Our CIFAR-10 folds have $r = 8$; some CIFAR-100 folds have fewer source observers and therefore a smaller dictionary. For $\mathbf{A}_\mathbf{Q} = \mathbf{R}\mathbf{B}_\mathbf{Q}$, let

$$\mathbf{S}_\mathbf{Q} = \frac{\mathbf{A}_\mathbf{Q}^\top \mathbf{A}_\mathbf{Q}}{M-1}, \qquad s_\mathbf{Q} = \frac{\mathrm{tr}(\mathbf{S}_\mathbf{Q})}{d}.$$

The implemented shrinkage covariance is

$$\mathbf{C}_\mathbf{Q} = \tfrac{1}{2}\mathbf{S}_\mathbf{Q} + \tfrac{1}{2}s_\mathbf{Q}\mathbf{I}_d + 10^{-6}\max\{s_\mathbf{Q}, 1\}\mathbf{I}_d,$$

and the residual variance is floored at 10^{-6} before multiplication by $\kappa = 10$. These isotropic regularizers preserve orthogonal equivariance.

The unconstrained parameter starts at $\mathbf{Q}_0$. Adam uses a learning rate of 0.02, runs for 200 steps, and applies gradient-norm clipping at 10. Each step includes all source observers and samples up to 96 complete evaluation groups without replacement. Budgets cycle through $8, 16, 32, 64, 128$. The target and predicted listwise distributions are softmaxes of the corresponding group utilities divided by temperature 0.25. The anchor coefficient is 0.1 on the mean squared difference between the $r \times r$ projectors. We report means over three fixed seeds, rather than an ensemble of their predictions.

The query order is computed once from the leading rank-three SVD basis using block D-optimal design with an identity ridge matrix, and is shared across the calibrated main-table methods. In matrix subscripts, $\mathcal{S}$ and $\mathcal{E}$ select all candidate rows belonging to their complete groups; the sum and normalization in $\mathcal{L}_{\mathrm{sel}}$ count evaluation groups.

D.2 Component map

Table 13 separates the roles of the method components. Gauge consistency defines the estimator across equivalent coordinates, while decision alignment chooses the subspace used by that estimator.

D.3 Notation and computational cost

Table 14 collects the symbols used in the method and theory. Constructing the rank-r dictionary requires one truncated SVD of the $M \times N$ source matrix. Posterior solves operate in dimension

Table 13: Method components, the quantities they control, and their empirical checks.

Component	Mathematical role	Empirical evidence
Spectral dictionary $\mathbf{F}$	Controls dictionary truncation error $\varepsilon_{\mathrm{dict}}$ while restricting the search space	Rank sensitivity and the reconstruction-matched control
Decision subspace $\mathbf{Q}$	Selects retained dictionary directions and determines $\varepsilon_{\mathrm{select}}(\mathbf{Q})$	Lower regret, higher selection accuracy, and improved pairwise concordance
Coherent $\mathbf{C}_{\mathbf{Q}}$ and $\sigma^2_{\mathbf{Q}}$	Defines posterior uncertainty in the current coordinates and preserves gauge covariance	Equivalent-gauge stress test and covariance/noise ablation
Projector parameterization	Identifies a subspace rather than an arbitrary oriented basis	Invariance under $\mathbf{Q} \mapsto \mathbf{QO}$ and the projector anchor
Complete-group episodes	Matches training information and loss to deployment-time argmax decisions	Query-policy controls and disjoint query/evaluation groups

d, while forming the basis and residual statistics during learning also depends on M and N. The deployment posterior requires $\mathcal{O}(Kd^2 + d^3)$ operations and stores a $d \times d$ covariance. Predicting all $N - K$ unqueried responses costs $\mathcal{O}((N - K)d)$. Target calibration uses the stored source quantities and a closed-form posterior.

Table 14: Core notation.

Symbol	Meaning
M, N	numbers of source observers and content–candidate items
r, d	spectral-dictionary width and decision-subspace dimension
$\mathbf{R}$	centered source response matrix in $\mathbb{R}^{M \times N}$
$\boldsymbol{\mu}$	itemwise mean response across source observers
$\mathbf{F}$	orthonormal rank-r spectral dictionary
$\mathbf{Q}$	rank-d Stiefel coordinate inside the dictionary
$\mathbf{Q}_0$	coordinates selecting the leading d dictionary directions
$\mathbf{B}_{\mathbf{Q}} = \mathbf{FQ}$	item basis of the learned decision subspace
$\mathbf{C}_{\mathbf{Q}}, \tilde{\sigma}^2_{\mathbf{Q}}$	latent prior covariance and effective likelihood variance
$\mathcal{S}, \mathcal{E}$	disjoint complete-group query and evaluation sets
K	number of queried responses, equal to A times queried groups
$\mathbf{h}_{\mathcal{S}}, \mathbf{V}_{\mathcal{S}}$	calibration-error bias and sub-Gaussian variance proxy

E Protocol Details

The CIFAR-10 item registry contains 10,000 contents. We reserve 6,000 as a development response field, 1,000 for source-only validation, 1,000 for target calibration, and 2,000 for evaluation. Source responses on the transductive calibration and evaluation items define our dictionary and prior; we used the development field to inform candidate generation, the protocol, and the fixed GIDS configuration. Each content stores 16 responses: Gaussian blur, motion blur, Gaussian noise, shot noise, JPEG, resize, contrast, and brightness at two severities. A cross-severity direction fixes one severity and therefore contains eight candidates per decision group. Twelve target observers span the ResNet, VGG, MobileNet, XCiT, ResNeXt, and WideResNet families. Two protocol-development observers serve only as source observers. For each target, we remove every observer from the same architecture family from the source population. We fix candidate generation, observer responses, content splits, and family membership before comparing methods.

The contents are images from the official CIFAR test sets. Table 15 specifies the eight operations; the clean image is displayed for reference in Figure 1 but is not a selectable candidate. Direction 1 queries severity 1 and evaluates severity 2 on disjoint contents, and direction 2 reverses the severities.

Table 15: Fixed candidate operations. Abbreviations are used in Figure 1; suffixes 1 and 2 denote severity.

Operation	Parameter	Severity 1	Severity 2
Gaussian blur (GB)	Gaussian radius	0.6	0.9
Motion blur (MB)	Horizontal kernel length	3	5
Gaussian noise (GN)	Standard deviation, 0–255 scale	8	16
Shot noise (SN)	Poisson peak	240	80
JPEG (JP)	Quality	60	30
Resize (RS)	Downsampling factor	0.75	0.625
Contrast (CT)	Enhancement factor	0.8	0.55
Brightness (BR)	Enhancement factor	0.8	0.6

Resize uses bilinear interpolation to downsample and restore the original size. JPEG uses 4:2:0 chroma subsampling. Noise realizations are deterministic for each content–candidate pair; noisy pixels are rounded and clipped to the 0–255 range.

Figure 1(a) displays all 16 candidates across both severities for the two prospectively frozen observers, with each observer's negative cross-entropy rescaled to $[0, 1]$ for visualization. Among evaluation contents on which their preferred candidates differ and both normalized top-two margins exceed 0.05, we display the content whose summed margin is closest to the eligible-set median (29 of 2,000 contents); stable content-ID order resolves ties. The figure asset records the selected ID, utilities, and response-file hashes.

Table 16 makes the transductive item protocol explicit. Source-observer responses on calibration and evaluation items define the shared item representation. Target-observer evaluation responses remain hidden until scoring; only selected complete groups from the target-calibration split are revealed. Thus "unqueried" denotes missing target-observer responses on items represented by the source observers.

Table 16: Response access by content split. "Source" excludes the held-out target architecture family.

Split	Contents	Source-observer use	Target-observer use
Response field	6,000	Candidate/response-field development	Never used for reported target scoring
Validation	1,000	Run-specific Ridge selection	Held-out target family excluded
Calibration	1,000	Enters the transductive dictionary and prior	Only queried complete groups revealed
Evaluation	2,000	Enters the transductive dictionary and prior	Hidden until final regret computation

E.1 Metric stability

Across 48,000 observer–content–direction groups, GIDS retains a favorable paired effect under unnormalized regret on the standardized utility scale. The result also holds when we floor denominators at the 1st, 5th, or 10th percentile of positive ranges, or remove the corresponding low-range contents. These checks vary the final range normalization after groupwise z-scoring.

F Within-Group Decision Preservation

Table 17 measures how well each representation preserves the comparisons that determine candidate selection. Selection accuracy checks whether the predicted and true maximizers agree. Pairwise concordance measures the fraction of the 28 within-content comparisons with matching signs. GIDS improves both measures as it receives more target-observer evidence. Across the 12 target observers, the concordance delta and regret delta have a Spearman correlation of -0.811 ($p = 0.0014$).

Table 17: Within-content decision preservation. Deltas are GIDS minus SVD + EB; positive selection accuracy and concordance favor GIDS, while negative regret favors GIDS. Selection accuracy means choosing the true utility maximizer. Accuracy and concordance deltas are in percentage points; the near-zero concordance delta at $K = 8$ is -0.0019 points before rounding.

Budget	Regret delta	Selection-accuracy delta	Pairwise-concordance delta
8	−0.00057	+0.42	≈ 0
16	−0.00063	+0.65	−0.02
32	−0.00200	+1.28	+0.18
64	−0.00215	+1.59	+0.28
128	−0.00261	+1.75	+0.30
All	−0.00159	+1.14	+0.15

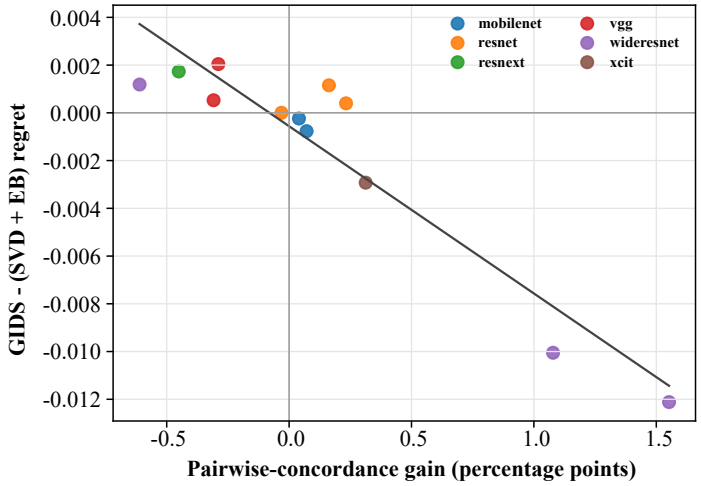


Figure 5: Observer-level mechanism diagnostic. Each point averages directions and budgets for one observer. Greater preservation of within-content comparisons accompanies lower selection regret; the fitted line summarizes this association.

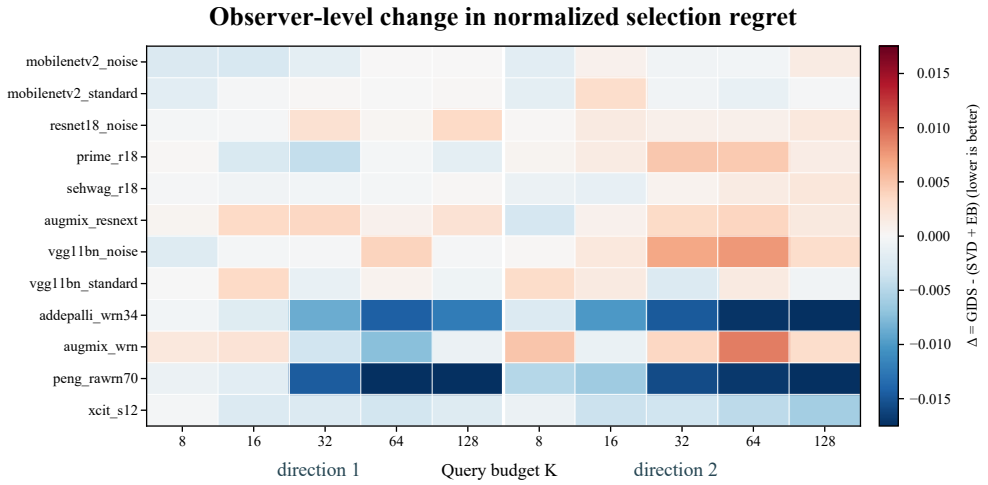


Figure 6: Observer–budget heterogeneity on the architecture-family-held-out evaluation. Each cell reports the paired change in normalized regret, GIDS minus the SVD + EB control, after averaging the three GIDS seeds. Blue cells favor GIDS. The method's mean gain arises from a structured but non-uniform response across observers, budgets, and transfer directions.

G Extended Related Work

From perceptual IQA to machine utility. Human-oriented IQA now spans multiscale and distortion-aware models, vision–language supervision, feature adaptation, domain adaptation, and multimodal scoring (Ke et al., 2021; Zhang et al., 2023; Agnolucci et al., 2024; Chen et al., 2024a; Wang et al., 2023; Xu et al., 2024; Shi et al., 2024; Li et al., 2024; Wu et al., 2024c). PromptIQA adapts a predictor to a new assessment requirement (Chen et al., 2024b). Q-Bench evaluates low-level visual understanding (Wu et al., 2024a), while Q-Instruct provides instruction data to improve it (Wu et al., 2024b). Text-conditioned image ordering also uses ranking rather than only scalar prediction (Yu et al., 2025). Q-Insight and ReLIQS develop complementary learned IQA systems (Li et al., 2025b; Gedik et al., 2026). Camera and video assessment add acquisition settings (Fang et al., 2023; Yuan et al., 2024; Wen et al., 2024). Perceptual attacks (Zhang et al., 2022) and semantic compression for downstream vision (Feng et al., 2023) further motivate studying machine utility separately from human opinion.

Quality assessment also extends beyond scalar scores: multimodal models compare and explain images (Wu et al., 2024d; You et al., 2024), though fine-grained and multi-image judgments remain difficult (Wu et al., 2024e). Distributional supervision (You et al., 2025) and generalized full-reference assessment (Chen et al., 2025) broaden how quality is judged. GIDS asks how the preferred candidate changes with a downstream observer absent from source training.

Few-shot priors and preference. Hierarchical Bayesian meta-learning shares global information across tasks for episode-specific posterior adaptation (Kim & Hospedales, 2024). PreferThinker infers an individual person's image preferences from a small reference set (Xu et al., 2026). Both concern adaptation from limited observations; our target is a machine observer choosing among processed candidates on a fixed source-response bank.

Evaluation, transferability, and observer shift. Efficient evaluation uses representative subsets, item-response models, empirical-Bayes estimates, and amortized or learned small evaluators (Fogliato et al., 2024; Truong et al., 2025; Zhao et al., 2024); broad frameworks also show the limits of aggregate scores (Liang et al., 2023). Transferability metrics help choose a model (You et al., 2021; Bolya et al., 2021; Mehra et al., 2024); we instead adapt decisions to one chosen observer. Performance under domain and corruption shift motivates family holdout (Koh et al., 2021; Miller et al., 2021; Mintun et al., 2021; Schneider et al., 2020; Croce et al., 2021). Foundation-model-based robustness evaluation uses a pretrained model zoo as a surrogate oracle over perturbed images (Zhang et al., 2024); our family holdout instead tests prediction for an unseen observer. Even related or weight-averaged models can differ off distribution (Wortsman et al., 2022), motivating target-observer calibration.

Query-efficient inference and decision alignment. Block D-optimal queries follow experimental design (Huan et al., 2024; Rainforth et al., 2024; Riquelme et al., 2017). Task-driven imaging design instead chooses measurement channels to support downstream analysis (Blumberg et al., 2024). GIDS learns a decision representation while retaining a fixed source-designed query policy, connecting its subspace objective to decision-focused abstractions (Poli et al., 2023). Confidence calibration concerns whether predicted probabilities reflect classification correctness (Guo et al., 2017); our observer calibration estimates latent response coordinates from a few utility measurements.