GLOBAL BEFORE LOCAL: SEARCH AND SELECTION IN FLOW-BASED INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

When more inference computation is available, should an inverse solver refine one reconstruction longer or explore alternatives? We study this through global-before-local (GBL) inference, which advances multiple candidates, selects a subset, and completes the retained candidates with the same solver. Broader search can find adequate candidates yet still lose when selection discards them. We separate best-candidate advantage from selection loss, quantify the recovery needed to beat equal-work refinement, and use a Bayes benchmark to distinguish information limits from selector error. GBL with measurement-based selection gains 3.88 dB in mean peak signal-to-noise ratio (PSNR) over refinement in CelebA phase retrieval. Predictive selection improves reconstruction quality in held-out inpainting and exceeds refinement in AFHQ irregular masking. On CelebA 4× point-decimation super-resolution, development data show that even the best searched candidate trails the same-budget single trajectory and forecast a 1.30 dB mean-PSNR deficit for predictive selection; held-out evaluation confirms the ordering with a 1.15 dB deficit. These results show that the value of broader search depends not only on the candidates it finds, but also on how much of that advantage selection preserves.

1 INTRODUCTION

When more inference computation is available, should a generative inverse solver refine one reconstruction longer or explore alternatives? Inverse problems reconstruct an unknown image from incomplete or noisy measurements. Diffusion priors support this reconstruction (Chung et al., 2023; Zhang et al., 2025), while flow-based solvers adapt pretrained flow models to the same goal (Kim et al., 2025; Pourya et al., 2026; Martin et al., 2025). Exploring and pruning multiple trajectories offers another way to allocate inference computation (Makris et al., 2026; Guimarães & Perona, 2026). We study this choice through **global-before-local (GBL)** inference (Figure 1).

Our finding: broader search can produce adequate candidates more often yet still lose to deeper refinement when selection fails to preserve their quality.

What is global-before-local inference? The global stage advances several candidate trajectories to explore alternative reconstructions. Each trajectory follows the solver's updates from an initialization to a reconstructed image. GBL then selects a subset, and the local stage refines the retained candidates with the same solver. This creates two decisions: how broadly to search, and which candidates to keep. For evaluation, we complete every candidate with the same remaining solver steps. This separates failures to generate an adequate reconstruction from failures to select one, and measures how much selection must improve to beat equal-work refinement. A Bayes analysis distinguishes limits due to candidate quality, selection information, and the selection rule (Table 1).

- **A controlled diagnosis of search and selection.** We separate candidate advantage from selection loss, quantify the recovery needed to beat equal-work refinement, and use independent development data to forecast whether a fixed alternative selector will outperform refinement on held-out images (Sections 4.2 and 6.5). A Bayes benchmark distinguishes information limits from selector error.
- **Evidence for distinct allocation regimes.** Across tasks and solvers, candidate advantage can survive selection, be erased by it, or be absent. In CelebA phase retrieval, measurement-based selection (**GBL-M**) gains 3.88 dB over refinement; in the prospective super-resolution test, even the best of eight candidates falls 0.99 dB below refinement on average.

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.

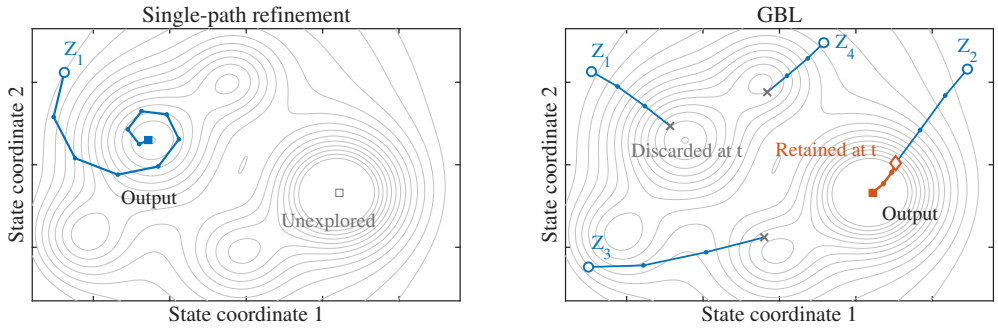

Figure 1: **GBL separates exploration across trajectories from selection and local refinement.** A single trajectory is refined to an output (left). GBL advances several candidate trajectories, selects one, and completes it with the same solver (right). Open circles denote random initial states Z_i, crosses discarded candidates, the diamond the selected candidate, and squares outputs. Coordinates and contours schematically illustrate the two inference paths.

- **Predictive selection as a test of recoverable selection loss.** **GBL-P** predicts completed quality from partial trajectories (Li et al., 2025; Zhao et al., 2026). Neither selector uses the target image at inference. With candidates and remaining solver steps fixed, it improves central-inpainting PSNR on both datasets. The recovery falls short of the amount needed to beat refinement in CelebA central inpainting, but exceeds that amount in AFHQ irregular masking.

2 Related Work

Generative inverse solvers. Diffusion and flow-based solvers combine learned priors with measurement constraints (Chung et al., 2023; Zhang et al., 2025; Kim et al., 2025; Pourya et al., 2026; Martin et al., 2025). D-Flow differentiates through a flow for controlled generation (Ben-Hamu et al., 2024). GBL allocates computation across trajectories with the base solver's updates unchanged.

Exploration and pruning. In language-model reasoning, the effectiveness of test-time search and revision depends on problem difficulty (Snell et al., 2025). Filtering posterior sampling uses sequential Monte Carlo to approximate inverse-problem posteriors (Dou & Song, 2024). Particle-Filtering-based Latent Diffusion (PFLD) adds measurement-based reweighting, resampling, and pruning around an existing diffusion solver (Nazemi et al., 2024). Accelerated Tempered Guided Diffusion (A-TGD) explores candidates and prunes once by measurement likelihood (Makris et al., 2026). Progressive Seed Pruning narrows the candidate set using intermediate denoised estimates under a model-evaluation budget (Guimarães & Perona, 2026). Our likelihood-pruning control varies candidate count and selection time while keeping the inverse solver fixed, using **GBL-M**'s current measurement negative log-likelihood (NLL) rule.

Predicting candidate quality. Process supervision trains verifiers from intermediate reasoning steps (Lightman et al., 2024). For generative sampling, TreeDiff and Soft Value-Based Decoding (SVDD) predict future rewards from partial generative states (Zhao et al., 2026; Li et al., 2025). **GBL-P** trains on the quality of completed candidates and selects once before their remaining solver steps. Our diagnostic analysis uses decision theory (Blackwell, 1953) to separate candidate quality from losses due to limited information and the choice of selector. Building on established multi-trajectory exploration and likelihood pruning, our contribution is to separate candidate advantage from selection loss in controlled comparisons and quantify the recovery needed to beat equal-work refinement.

3 GLOBAL-BEFORE-LOCAL INFERENCE

Inference and computation. Let $Y = h(X_\star) + \varepsilon$, where $X_\star$ is the unknown image, h the measurement operator, and ε observation noise. We first consider retaining one candidate ($k = 1$), as in the main comparisons. GBL draws $N \geq 1$ initial states Z_i from the solver's initialization distribution p_0, with $i \in [N] = \{1, \ldots, N\}$. Each candidate takes m of T uniformly spaced outer steps over $[0, 1]$, before selection at $t = m/T$. The prefix $P_{0:m}$ runs these steps and returns state and history $H_{i,m}$. The remaining solver steps, called the continuation $C_{m:T}$, complete a candidate if it is selected:

$$Z_i \sim p_0, \qquad H_{i,m} = P_{0:m}(Y, Z_i), \qquad \widehat{X}_i = C_{m:T}(H_{i,m}, Y). \tag{1}$$

These maps fix the prior, solver settings, and prescribed randomness. GBL advances all candidates, selects one using computed prefix information, and completes that candidate with the unchanged solver. Outer-update work is $W = Nm + (T - m)$. Single-path refinement instead divides the same $[0, 1]$ time interval into W updates on one trajectory. Equal outer-update work counts total completed updates across candidates. Boundary evaluations and feature computation are counted separately (Appendix B.4). Equal outer-update work need not mean equal FLOPs or wall-clock time.

Offline evaluation. For analysis, we also complete the discarded candidates under the same continuation. Candidate quality is $V_i = u(\widehat{X}_i, X_\star)$, where u is a utility such as PSNR (higher is better). These target-dependent values evaluate selection and supply development labels. They are unavailable to the selector at inference. Reference quality V_1 comes from Z_1 after T total outer updates.

Best-candidate quality M_N is the highest of $V_1, \ldots, V_N$. **Best-candidate gain G_N** is the mean improvement of M_N over reference quality V_1.

$$M_N = \max_{1 \leq i \leq N} V_i, \qquad G_N = \mathbb{E}[M_N - V_1]. \tag{2}$$

Expectations average over images, observation noise, and solver randomness. Assume $\mathbb{E}|V_i| < \infty$. Changing local steps changes candidate qualities. Changing only selection leaves them fixed. Section 4 relates best-candidate gain to selected-output quality and equal-work single-path refinement. Appendix A.1 extends the analysis to retaining several candidates.

4 DIAGNOSING SEARCH AND SELECTION

4.1 DID SEARCH FAIL, OR DID SELECTION FAIL?

We first ask whether search found an adequate candidate and whether selection kept it. Let V_{sel} denote the final quality of the candidate selected by selection rule π, and fix a quality threshold τ. A missing-candidate failure means no completed candidate reaches τ. A discard failure means a candidate reaches τ, but the selected output does not:

$$\underbrace{\mathbb{P}(V_{\mathrm{sel}} < \tau)}_{\text{total failure}} = \underbrace{\mathbb{P}(M_N < \tau)}_{\text{missing-candidate failure}} + \underbrace{\mathbb{P}(M_N \geq \tau,\ V_{\mathrm{sel}} < \tau)}_{\text{discard failure}}. \tag{3}$$

Adequacy depends on the continuation, so allocation comparisons use each configuration's own completed candidates. With one generated candidate ($N = 1$), there is no discard failure.

The partition counts inadequate outputs. Selection loss is how much quality is lost because the selector does not keep the best candidate. Its mean is $L^\pi = \mathbb{E}[M_N - V_{\mathrm{sel}}]$. Unlike failure counts, G_N and L^π measure mean quality differences. **Finding an adequate candidate does not establish that the selector could recognize it from the information available at selection time.**

4.2 WHEN IS BROADER SEARCH WORTH THE WORK?

Broader search is worthwhile when its best-candidate advantage exceeds the quality lost in selection. Let D_W be the mean quality of equal-work single-path refinement. The *search headroom*, $\mathbb{E}[M_N] - D_W$, is the advantage available before selection. Subtracting selection loss gives the advantage that reaches the output:

$$\mathbb{E}[V_{\mathrm{sel}}] - D_W = (\mathbb{E}[M_N] - D_W) - L^\pi. \tag{4}$$

Table 1: **Search and selection limits imply different next steps.** Conditions use population means at fixed candidates, continuation, and selection inputs.

Condition	Interpretation and next step
$\mathbb{E}[M_N] \leq D_W$	**No candidate advantage.** Selection alone cannot win. Change candidate generation or use refinement.
$\mathbb{E}[M_N] > D_W$, $B_{t,1} \leq D_W$	**Insufficient selection information.** A better rule using the same inputs cannot win. Reconsider the selection inputs or allocation.
$B_{t,1} > D_W$, $\mathbb{E}[V_{\text{sel}}] \leq D_W$	**Selector gap.** A better selector can win using the same inputs. Test whether its recovery closes the gap.
$\mathbb{E}[V_{\text{sel}}] > D_W$	**Breadth wins.** The current selector preserves enough candidate quality to beat refinement.

If headroom is nonpositive, selection alone cannot beat refinement. With positive headroom, breadth wins only if selection preserves enough of it. This separates a need for better candidates from a need for better selection.

When the current selector falls below refinement and $L^\pi > 0$, the fraction of selection loss that must be recovered to tie refinement is

$$r_{\text{req}} = \frac{D_W - \mathbb{E}[V_{\text{sel}}]}{L^\pi}. \tag{5}$$

Recovering this fraction of selection loss ties refinement, and exceeding it wins. A value of one permits at most a tie. A value above one rules out even a tie through selection alone. Section 6.3 compares required and achieved recovery.

To forecast a new selector's advantage before final testing, add its recovery to the current selector's difference from refinement, estimating the two terms on independent development images (Section 6.5).

4.3 What Part of Selection Loss Is Recoverable?

Selection loss reflects both what the available information can reveal and how well the selector uses it. Let $\mathcal{F}_t$ denote the available inputs, such as measurement residuals and changes in predicted clean images. Selectors restricted to these inputs and independent development data are *feasible policies*. Expectations condition on the development data used to fit the selector (Appendix A.2).

The *Bayes retained quality*, $B_{t,1}$, is the best mean quality achievable with these inputs. It is attained by choosing the candidate with the largest expected final quality, $\mu_{i,t} = \mathbb{E}[V_i \mid \mathcal{F}_t]$. The gap from mean best-candidate quality is the *information-limited loss*, $I_{t,1}$:

$$B_{t,1} = \mathbb{E}[\max_i \mu_{i,t}], \qquad I_{t,1} = \mathbb{E}[M_N] - B_{t,1}. \tag{6}$$

No selector using $\mathcal{F}_t$ can recover this first gap (Appendix A.1; illustrated in Appendix A.3). The remaining *selector gap*, $S^\pi = B_{t,1} - \mathbb{E}[V_{\text{sel}}]$, is quality lost by the particular selection rule.

Decomposition 1 (Selection loss and feasible-policy improvement)**.** *Suppose the compared policies use the same integrable candidate qualities, utility, and available information. Every feasible policy π satisfies*

$$L^\pi = I_{t,1} + S^\pi. \tag{7}$$

Both terms are nonnegative. If another feasible policy has higher true mean quality than π, the improvement is a lower bound on π's selector gap S^π.

Proof sketch. No candidate exceeds M_N, and no feasible policy exceeds the Bayes mean. Hence $\mathbb{E}[M_N] \geq B_{t,1} \geq \mathbb{E}[V_{\text{sel}}]$. Adding and subtracting $B_{t,1}$ splits L^π into the nonnegative gaps $\mathbb{E}[M_N] - B_{t,1} = I_{t,1}$ and $B_{t,1} - \mathbb{E}[V_{\text{sel}}] = S^\pi$. Any feasible policy's true mean gain over π is at most $B_{t,1} - \mathbb{E}[V_{\text{sel}}] = S^\pi$. Appendix A.1 proves the general case for up to k candidates.

This benchmark also refines the comparison with a single path:

$$\mathbb{E}[V_{\text{sel}}] - D_W = (B_{t,1} - D_W) - S^\pi.$$

Algorithm 1: **GBL-P** inference with optional calibrated override ($k = 1$)

Input: Observation Y, candidate count N, selection step m, total steps T, initialization distribution p_0, fitted predictor f, and fixed threshold q for override.
Mode: direct predictive selection (default) or calibrated override.
Solver maps: Equation (1). Tie rules: Appendix B.3.
Output: Final reconstruction from one selected candidate.

Advance candidates. $t \leftarrow m/T$, $Z_i \sim p_0$, $i \in [N]$
$H_{i,m} \leftarrow P_{0:m}(Y, Z_i)$, $i \in [N]$
Predict final quality. $\phi_{i,t} \leftarrow$ features from $Y, H_{1,m}, \ldots, H_{N,m}$, $i \in [N]$
$\widehat{\mu}_{i,t} \leftarrow f(\phi_{i,t})$, $i \in [N]$
Choose predicted best. $j_P \leftarrow \arg\max_{i \in [N]} \widehat{\mu}_{i,t}$, $j_{\text{sel}} \leftarrow j_P$
if *calibrated override is enabled* **then**
    $j_M \leftarrow$ **GBL-M** choice (lowest measurement NLL)
    $\widehat{\Delta} \leftarrow \widehat{\mu}_{j_P,t} - \widehat{\mu}_{j_M,t}$
    if $\widehat{\Delta} \leq q$ **then Keep measurement choice.** $j_{\text{sel}} \leftarrow j_M$
Complete selected candidate. return $C_{m:T}(H_{j_{\text{sel}},m}, Y)$

If $B_{t,1} \leq D_W$, the available selection inputs cannot support a win. If $B_{t,1} > D_W$, some selector using those inputs can beat refinement. Table 1 translates these cases into next steps.

Completed candidates measure headroom, total selection loss, and failures. Held-out comparisons then test how much loss a different selector recovers. These measurements locate the allocation bottleneck; separating information-limited loss from the remaining selector gap requires the Bayes benchmark (Appendix A.1).

5 Candidate Selection: GBL-M and GBL-P

We compare two GBL selectors with candidates and continuation fixed (Equation (1)). Both retain one candidate. Appendix C.5 extends selection to several candidates.

GBL-M: Measurement-based selection. GBL-M selects the candidate whose current clean estimate has the lowest measurement NLL. It uses already computed scores, requiring no fitted predictor. This ranks agreement with the observation now, before final quality is known.

GBL-P: Predictive selection. GBL-P predicts final quality from partial trajectories. We train f on development candidates, pairing selection-time features $\phi_{i,t}$ with final quality V_i from completed trajectories. Features summarize measurement fit, trajectory changes, and comparisons among candidates. Squared-error regression targets mean quality given these features (Appendix B.2).

At inference, the fixed predictor outputs $\widehat{\mu}_{i,t} = f(\phi_{i,t})$ from selection-time features, without the target image or future states. Its default **direct predictive selection** keeps the highest prediction and completes that candidate with the unchanged solver. This approximates the Bayes ranking using the supplied features.

Predictions can be wrong, so we can add a threshold before changing the GBL-M choice. This optional addition to GBL-P is **calibrated override**. It switches only when the predicted gain over GBL-M exceeds q. Otherwise it retains GBL-M. We calibrate q from predicted minus actual gains where the selectors disagree (Appendix B.3). The predictor and threshold are fixed before testing. Changing continuation requires a transfer check or refitting.

6 Experiments and Discussion

Our experiments ask when broader search creates an advantage, how much selection preserves, and whether this diagnosis can guide allocation before final evaluation (Table 1).

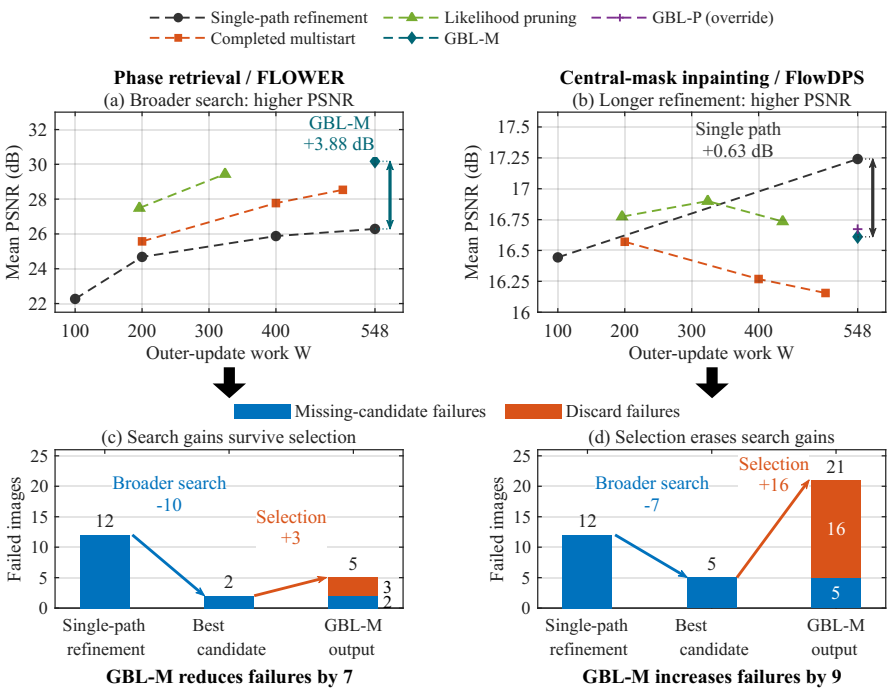

Figure 2: **Broader search reduces candidate absence in both tasks, but selection reverses the benefit in central inpainting.** (a,b) Mean PSNR on CelebA against outer-update work. Arrows compare equal work ($W = 548$). (c,d) At that work, bars show single-path failures, cases with no adequate completed GBL candidate, and failures after **GBL-M** selection. Development thresholds are 6.24 dB for phase retrieval and 14.15 dB for central inpainting, with 128 test images per task. The purple **GBL-P** marker uses a transferred predictor with calibrated override.

6.1 Common Setup and Comparisons

We use CelebA (Liu et al., 2015) at 128×128 and AFHQ-Cat (Choi et al., 2020) at 256×256, each with a pretrained optimal-transport flow prior from Martin et al. (2025). Phase retrieval uses FLOWER (Pourya et al., 2026). Central inpainting uses FlowDPS (Kim et al., 2025) with square masks of side 96 and 192 pixels, respectively. Both masks remove 56.25% of the image. Observation noise has standard deviation 0.05. We diagnose search and selection using full-image peak signal-to-noise ratio (PSNR, dB). Learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018) complements it with perceptual quality (lower is better; implementation in Appendix B.1).

Each allocation evaluation uses 128 images per setting. GBL advances eight candidates for 64 updates, retains one, and completes 36 more ($W = 548$). Controls use one 548-step path or five completed 100-step paths selected by final measurement NLL ($W = 500$, under the same cap). CelebA likelihood pruning tunes candidate count and selection time on development images. Work W counts solver updates. Appendices B.4 and E report model calls and latency. Phase-retrieval and inpainting failure thresholds use development PSNR percentiles of the native 100-step solver. Effects are paired by image. Appendices B.2 and B.5 specify splits and uncertainty.

6.2 When Does Broader Search Improve Reconstruction?

Broader search improves PSNR in CelebA phase retrieval and AFHQ central inpainting; refinement wins in CelebA central inpainting. Table 2 compares both datasets under the same outer-update budget. In CelebA phase retrieval, **GBL-M** gains 3.88 dB in mean PSNR, raises the PSNR tenth percentile from 6.70 to 16.79 dB, and exceeds completed multistart. In an eight-image GH200 run, **GBL-M** takes 30.32 versus refinement's 41.01 s/image (Appendix B.4).

Table 2: **Does broader search improve quality under the same work cap?** Mean full-image PSNR and LPIPS on 128 images per dataset and task, with $W \leq 548$ outer updates. Completed multistart uses 500 updates. CelebA phase likelihood pruning coincides with **GBL-M** at this cap. Bold marks the best mean for each metric within each setting. Appendix Tables 9 and 20 give additional metrics and controls.

	Phase / FLOWER			Central / FlowDPS		
Method	Work W	PSNR ↑	LPIPS ↓	Work W	PSNR ↑	LPIPS ↓
CelebA (128×128)						
Single-path refinement	548	26.29	0.1549	548	**17.24**	0.2668
Completed multistart	500	28.53	0.1050	500	16.16	0.2382
Likelihood pruning	548	**30.17**	**0.0794**	436	16.73	**0.2322**
GBL-M	548	**30.17**	**0.0794**	548	16.61	0.2356
AFHQ-Cat (256×256)						
Single-path refinement	548	19.63	0.3586	548	13.87	0.5407
Completed multistart	500	20.14	0.3127	500	**14.61**	**0.4896**
GBL-M	548	**21.31**	**0.2777**	548	14.46	0.4979

At equal work in central inpainting, refinement outperforms **GBL-M** on CelebA, whereas **GBL-M** outperforms refinement on AFHQ. Both PSNR intervals exclude zero (Table 8).

The PSNR and perceptual rankings need not agree: in CelebA central inpainting, refinement has the highest PSNR, whereas the search-based allocations yield lower LPIPS. In AFHQ central inpainting, completed multistart achieves higher mean PSNR than **GBL-M** and the lowest LPIPS. Its native solver also beats its long path (Table 20). Appendix C.1 compares published solvers at their native settings on CelebA.

A same-task solver control shows that allocation depends on the solver as well as the inverse problem (Appendix C.6).

6.3 Why Can Good Candidates Still Yield Poor Outputs?

Selection erases the candidate advantage in CelebA central inpainting. CelebA's best-candidate advantage of 1.145 dB is smaller than **GBL-M**'s 1.775 dB selection loss, leaving a 0.630 dB deficit to refinement (Equation (4)). In AFHQ central inpainting, a 2.361 dB candidate advantage exceeds the 1.770 dB selection loss, leaving a 0.590 dB output advantage (Appendix E). These means do not distinguish an information limit from a selector gap (Table 1).

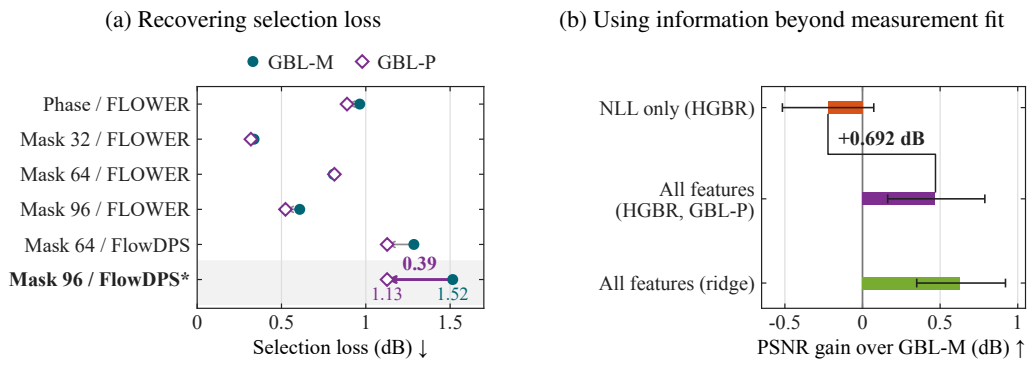

Figure 3: **Predictive selection can recover lost quality, and trajectory features improve on measurement-only prediction.** (a) Five-fold cross-fitting on 64 CelebA images per setting; the starred recovery has a paired 95% interval excluding zero. (b) Central-mask feature comparison on 128 held-out CelebA images (192 training images), with paired 95% intervals. The bracket compares the two HGBR predictors.

Failure counts show the same pattern: broader search finds adequate candidates more often in both CelebA tasks, but selection reverses the central-mask benefit (Figure 2; threshold sensitivity in Table 10). In AFHQ central inpainting, direct **GBL-P** reduces discard failures from 17 to 8 while the single missing-candidate failure remains.

The measurement-based ranking problem persists after candidate completion. On the development candidates used in Figure 3(a), terminal measurement fit and PSNR have near-zero average rank correlation for the largest central mask under both solvers (Appendix D.1). Delaying selection with the same NLL rule also fails to consistently improve quality (Appendix C.5).

With FLOWER fixed, the smallest Fourier grid has the largest selection loss, while search headroom shows no clear trend across grids (Appendix C.4).

Better selection does not yet close the CelebA central gap. Tying refinement requires recovering 35.5% of selection loss. On the allocation test set, the transferred ensemble recovers 21.1%, leaving a 0.254 dB shortfall; calibrated override also remains below refinement (Appendix C.3, Table 17).

6.4 How Much Does Predictive Selection Recover?

Predicting final quality recovers selection loss in both central-inpainting datasets. We compare **GBL-P** with **GBL-M** on held-out images with candidates and continuation fixed (Table 3). The closest cross-dataset comparison uses one predictor trained on 64 images and evaluated on 128: central-inpainting gains are 0.356 dB on CelebA and 0.403 dB on AFHQ. AFHQ phase retrieval shows no clear additional gain. Appendices B and E give the training and evaluation details.

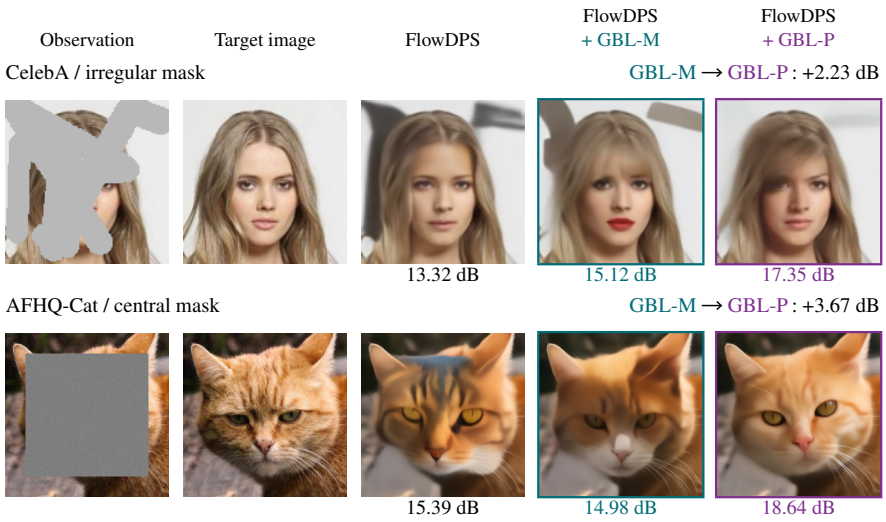

Figure 4: **Two selected improvements under predictive selection.** CelebA irregular-mask and AFHQ-Cat central-mask examples show **GBL-M** (teal) followed by direct **GBL-P** (purple), using identical candidates and remaining solver steps within each row. Native FlowDPS uses 100 updates. Labels give full-image PSNR and its change from **GBL-M** to **GBL-P**. CelebA observations suppress noise for display only; AFHQ observations retain measurement noise.

In AFHQ irregular masking, selection recovery is enough to beat refinement. Changing only the selector gains 0.686 dB over **GBL-M** and 0.582 dB over refinement. Completed multistart has the lowest LPIPS (Appendix E.1).

Mean gains coexist with individual losses in both datasets (Appendix D.5 and Table 19). Figure 4 illustrates selected improvements, while Figure 9 shows adverse AFHQ examples.

Using the same histogram-based gradient boosting regressor (HGBR), all features outperform NLL-only prediction by 0.692 dB (95% interval $[0.390, 1.003]$; Figure 3(b), Appendix D.1). Calibrated override reduces the worst observed losses at the cost of mean recovery: on CelebA irregular masks, direct selection gains 0.528 dB versus 0.144 dB for override (Table 17).

Table 3: **How much does direct predictive selection recover with candidates fixed?** Effects are direct **GBL-P** minus **GBL-M** on 128 images per evaluation, with identical candidates and continuation. CelebA intervals are pointwise 95%. AFHQ intervals (†) are 98.75%, Bonferroni-adjusted across four primary PSNR contrasts for a nominal 95% familywise level. Irregular masking uses the same adjustment; its paired geometry protocol is in Appendix E.1.

Evaluation	Mean PSNR (dB)		Paired effect (dB)	
	GBL-M	**GBL-P**	Mean [interval]	10th pct.
CelebA central / single predictor	16.72	17.08	**+0.356** [0.105, 0.617]	−1.444
CelebA central / predictor ensemble	16.63	17.10	**+0.470** [0.162, 0.788]	−1.529
AFHQ phase / single predictor†	21.31	21.12	−0.192 [−0.868, 0.450]	−0.033
AFHQ central / single predictor†	14.46	14.86	**+0.403** [0.063, 0.744]	−1.356
AFHQ irregular / single predictor†	15.11	15.80	**+0.686** [0.330, 1.043]	−1.355

6.5 Can the Diagnosis Predict a New Allocation Outcome?

When even the best candidate falls short, refine deeper. We test this prospectively on CelebA $4\times$ point-decimation super-resolution with FlowDPS. Independent development cohorts estimate **GBL-M**'s difference from refinement and a fixed **GBL-P** predictor's recovery. Their sum forecasts the predictive selector's advantage on 128 held-out images. Both predictor and forecast are fixed before final evaluation (Table 4; Appendix F).

Table 4: **The diagnosis predicts that refinement will remain preferable.** Mean full-image PSNR differences (dB) from independent development cohorts and 128 held-out images; refinement uses 548 updates.

Quantity	Development / forecast	Held-out evaluation
Best-candidate advantage	−1.055	−0.986
GBL-M − refinement	−1.202	−1.100
GBL-P − **GBL-M**	−0.098	−0.053
GBL-P − refinement	−1.300 (forecast)	−1.152

Even the best candidate trails refinement by 1.055 dB on average, with its interval below zero, while selection loses only 0.147 dB. This favors equal-work refinement over improving selection (Table 1). Held-out evaluation confirms the forecast: -1.300 dB (95% interval $[-1.640, -0.952]$) predicted versus -1.152 dB (adjusted 97.5% interval $[-1.424, -0.884]$) observed. Refinement also has the lowest LPIPS. The allocation forecast is supported, while **GBL-P** shows no clear gain over **GBL-M** (Appendix F.3).

7 Conclusion and Future Work

Broader search pays off when its candidate advantage survives selection. Our experiments reveal three regimes: that advantage survives, selection erases it, or search creates none. Separating candidate advantage from selection loss makes the allocation choice actionable: it shows how much a better selector must recover and when selection alone cannot help. Predictive selection recovers quality on fixed candidates, enough to exceed refinement in AFHQ irregular masking but short of the required recovery in CelebA central inpainting.

The prospective super-resolution study shows how the diagnosis can identify, before final testing, a setting where deeper refinement is preferable to improving selection among the tested candidates. Development data indicate insufficient candidate quality, and held-out evaluation confirms the forecast in favor of refinement.

Future work could adapt candidate counts and selection times per image, retain alternatives to reduce individual losses, and test perceptual objectives and zero-shot predictor transfer across domains and solvers. We will extend the **GBL-M** timing measurements to **GBL-P** and compare allocations at matched wall-clock budgets, including prediction overhead.

Reproducibility Statement

Appendix A states the assumptions and provides complete proofs of the theoretical results. Algorithm 1 specifies the inference procedure. Section 6.1 and Appendices B.1–B.3 describe the datasets, observation models, solver settings, data splits, and predictor training and selection procedures. Appendices B.4 and B.5 detail computation accounting and statistical procedures, including uncertainty estimation and multiple-comparison correction. Appendix C.4 describes the observation protocols underlying the two phase-retrieval comparisons. Appendix E provides the AFHQ-Cat evaluation protocol, including image reuse, evaluation splits, and task-specific predictors. Appendix F specifies the disjoint development and evaluation cohorts, the forecast fixed before final evaluation, and the held-out testing procedure.

Ethics Statement

The work uses an existing face-image research dataset and inherits its consent, representation, and distributional limitations. Reconstructions are hypotheses consistent with limited observations, not verified identities or recovered facts. The reported population effects do not justify safety-critical or identity-sensitive deployment. The AFHQ-Cat evaluation additionally inherits the collection and distributional limitations of its animal-face dataset. Generated details need not match the target.

AI Use Statement

Generative AI tools were used to support several aspects of the research and writing process. AI tools were used primarily to assist with implementation, analysis, and presentation. Specifically, AI tools assisted with polishing and revising manuscript text, drafting and restructuring portions of the paper, and improving notation and the presentation of mathematical arguments. They were also used for literature discovery, including identifying potentially relevant prior work, and for research execution, including discussing experimental designs and implementation details. For mathematical work, AI tools assisted with checking and improving proofs and intermediate calculations; all final mathematical claims and derivations were independently reviewed and verified by the authors. AI tools also supported code development and review, figure and table preparation, and editorial consistency checks across the manuscript. The authors made the final scientific and methodological decisions. The authors also verified the cited literature and reported results and take full responsibility for the final text, mathematical claims, code, experiments, figures, tables, and conclusions, including all AI-assisted content.

References

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-Flow: Differentiating through flows for controlled generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3462–3483. PMLR, 2024.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *International Conference on Learning Representations*, 2024.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Rogério Guimarães and Pietro Perona. Inference-time scaling of diffusion models via progressive seed pruning. In *European Conference on Computer Vision*, 2026.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. FlowDPS: Flow-driven posterior sampling for inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12328–12337, 2025.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Andreas Makris, Paul Fearnhead, and Chris Nemeth. Tempered guided diffusion. *arXiv preprint arXiv:2605.03712*, 2026. URL https://arxiv.org/abs/2605.03712.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025.

Amir Nazemi, Mohammad Hadi Sepanj, Nicholas Pellegrino, Chris Czarnecki, and Paul Fieguth. Particle-filtering-based latent diffusion for inverse problems. *arXiv preprint arXiv:2408.13868*, 2024. URL https://arxiv.org/abs/2408.13868.

Mehrsa Pourya, Bassam El Rawas, and Michael Unser. FLOWER: A flow-matching solver for inverse problems. In *International Conference on Learning Representations*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20895–20905, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Jiachi Zhao, Zehong Wang, Yamei Liao, Chuxu Zhang, and Yanfang Ye. Controllable graph generation with diffusion models via inference-time tree search guidance. In *Proceedings of the ACM Web Conference 2026*, pp. 1195–1205. ACM, 2026.

APPENDIX

A CORE ANALYSIS AND PROOFS

A.1 SELECTION-LOSS DECOMPOSITION AND FEASIBLE POLICIES

For these proofs, $\mathcal{F}_t$ is the sigma-algebra generated by the compared policies' computed summary inputs and independent development data. For fixed $1 \le k \le N$, the allowed retained sets are $\mathcal{A}_k = \{A \subseteq [N] : 1 \le |A| \le k\}$, where $[N] = \{1, \dots, N\}$. Retained-set quality is $V(A) = \max_{i \in A} V_i$. The decision now chooses a set to retain. Its quality is the best completed quality preserved within that set, so the Bayes objective below uses the conditional expected set maximum. Bayes retained quality, information-limited loss, and outer-update work generalize to

$$B_{t,k} = \mathbb{E}\left[\max_{A \in \mathcal{A}_k} \mathbb{E}[V(A) \mid \mathcal{F}_t]\right], \qquad I_{t,k} = \mathbb{E}[M_N] - B_{t,k}, \qquad W = Nm + |A|(T - m).$$

Compare policies using the same integrable candidate qualities, utility, and computed information. For fitted policies, expectations condition on independent development data. A feasible policy chooses an $\mathcal{F}_t$-measurable $A \in \mathcal{A}_k$. Define $V^\pi = V(A)$, $L^\pi = \mathbb{E}[M_N - V^\pi]$, and $S^\pi = B_{t,k} - \mathbb{E}[V^\pi]$. Here S^π is the selector gap. Retained-set quality equals selected-output quality for $k = 1$ (Decomposition 1). For $k > 1$, final selection can lose further quality (Appendix C.5).

The family $\mathcal{A}_k$ is finite and nonempty. Both $|V(A)|$ and $|M_N|$ are bounded by $\sum_i |V_i|$, ensuring integrability. Fix an ordering of $\mathcal{A}_k$, and let $A^\star$ be the first set maximizing $\mathbb{E}[V(A) \mid \mathcal{F}_t]$. These conditional means are $\mathcal{F}_t$-measurable, so $A^\star$ is feasible. Every feasible policy's chosen set is determined by $\mathcal{F}_t$, and its conditional mean quality cannot exceed this maximum. Taking expectations and using $V(A^\star) \le M_N$ gives

$$\mathbb{E}[V^\pi] \le B_{t,k} = \mathbb{E}[V(A^\star)] \le \mathbb{E}[M_N].$$

Thus $B_{t,k}$ is attained among all feasible policies, and $S^\pi \ge 0$ and $I_{t,k} \ge 0$. Adding and subtracting $B_{t,k}$ gives

$$L^\pi = \mathbb{E}[M_N] - \mathbb{E}[V^\pi] = \underbrace{\mathbb{E}[M_N] - B_{t,k}}_{I_{t,k}} + \underbrace{B_{t,k} - \mathbb{E}[V^\pi]}_{S^\pi}.$$

For another feasible policy ρ, its true mean gain over π satisfies

$$\mathbb{E}[V^\rho - V^\pi] = S^\pi - S^\rho \le S^\pi.$$

A positive true mean gain therefore bounds the current policy's excess loss from below. Also, $S^\pi \ge 0$ gives $I_{t,k} \le L^\pi$, and hence $I_{t,k} \le \min_\pi L^\pi$ for finitely many fixed policies. We treat the minimum estimated loss descriptively because selecting it introduces downward bias. Confirmatory intervals use fixed-policy contrasts.

For the single-candidate comparisons, completed candidates reveal the realized best-candidate quality and the quality selected by each evaluated policy. Averaging their differences estimates total selection loss. Computing $B_{t,k}$ requires finding the highest conditional expected set quality for each possible selection input and averaging those values. The full joint distribution of the selection information and candidate qualities determines this quantity. The finite comparisons of saved trajectories reported here do not directly measure that optimum. Thus the remaining observed loss is not an estimate of $I_{t,k}$. Discard failures can reflect insufficient summary information or a policy that uses it poorly, so these counts do not determine S^π.

A.2 COMPUTED INFORMATION AND MONOTONICITY

For the trajectory-selection comparisons, $\mathcal{F}_t$ specifies the computed summary inputs supplied to the compared selectors, together with the independent development data. The solver state and history $H_{i,m}$ support continuation, but selection is modeled using these summary inputs. This defines an information restriction. The Bayes optimum ranges over all $\mathcal{F}_t$-measurable policies and imposes no runtime restriction on the rule. Policies may ignore some inputs, and the optimum need not coincide with the best value in a particular fitted predictor class or under full solver-state information. A

full solver state may determine a future reconstruction without revealing its quality relative to the unknown target. Thus information-limited loss is relative to the declared selection inputs.

Hold candidate qualities fixed and suppose $\mathcal{F}_t \subseteq \mathcal{F}_{t'}$. The finite maximum is convex, so conditional Jensen gives the inequality below. The tower property gives the equality:

$$\begin{aligned}\mathbb{E}\left[\max_{A\in\mathcal{A}_k} \mathbb{E}[V(A) \mid \mathcal{F}_{t'}] \mid \mathcal{F}_t\right] &\geq \max_{A\in\mathcal{A}_k} \mathbb{E}[\mathbb{E}[V(A) \mid \mathcal{F}_{t'}] \mid \mathcal{F}_t] \\ &= \max_{A\in\mathcal{A}_k} \mathbb{E}[V(A) \mid \mathcal{F}_t].\end{aligned}$$

Taking expectations gives $B_{t',k} \geq B_{t,k}$. For $k' \geq k$, inclusion $\mathcal{A}_k \subseteq \mathcal{A}_{k'}$ also gives $B_{t',k'} \geq B_{t',k} \geq B_{t,k}$. Since $\mathbb{E}[M_N]$ is unchanged, $I_{t',k'} \leq I_{t,k}$. These bounds optimize over feasible policies, so a fixed scoring rule need not improve. These inequalities require unchanged candidate qualities. They need not hold if continuation changes.

A.3 The binary observation example

Let $X_\star$ be uniform on distinct images $\{x_a, x_b\}$ and $Y = h(X_\star) + \varepsilon$, with independent noise $\varepsilon \sim \mathcal{N}(0, \Sigma)$ and positive definite Σ. The Z_i are independent uniform draws from $\{x_a, x_b\}$, independent of the target and observation. Continuation returns $\widehat{X}_i = Z_i$. At time t, selection uses $\mathcal{F}_t = \sigma(Y, Z_1, \ldots, Z_N)$ and returns one sampled candidate. For $d \geq 0$, define the squared observation-space separation $d^2 = (h(x_a) - h(x_b))^\top \Sigma^{-1} (h(x_a) - h(x_b))$. With standard normal cdf Φ, the minimum probability of returning the wrong target image is

$$p_{\text{fail}}(N, d) = 2^{-N} + (1 - 2^{1-N})\Phi(-d/2). \tag{8}$$

Conditional on either target, all N independent draws miss it with probability $(1/2)^N = 2^{-N}$. All draws equal x_a with probability 2^{-N}, as they do for x_b. These events are disjoint, so both types appear with probability $1 - 2 \cdot 2^{-N} = 1 - 2^{1-N}$. If every draw equals the target, selection cannot fail.

When both types appear, independence of the draws from $(X_\star, Y)$ leaves the equal-prior Gaussian testing problem unchanged. For $d > 0$, its log-likelihood ratio is

$$\Lambda(Y) = (h(x_a) - h(x_b))^\top \Sigma^{-1} \left(Y - \frac{h(x_a) + h(x_b)}{2}\right).$$

Substituting $Y = h(X_\star) + \varepsilon$ shows that $\Lambda(Y)$ is Gaussian with variance d^2 and mean $d^2/2$ under x_a or $-d^2/2$ under x_b. The Bayes rule chooses x_a when $\Lambda(Y) \geq 0$, giving error $\Phi((0 - d^2/2)/d) = \Phi(-d/2)$ under either target by symmetry. The information $\mathcal{F}_t$ reveals which types are available and Y, so this rule is feasible and optimal. Adding the disjoint missing-target and both-types testing failures gives Equation (8).

With utility 0 for a correct output and -1 otherwise, mean quality is the negative error probability. The best candidate is wrong only when every draw misses the target, while the Bayes selector also incurs the testing errors derived above. Hence $\mathbb{E}[V_1] = -1/2$, $\mathbb{E}[M_N] = -2^{-N}$, and Bayes retained quality under the specified sufficient information is $B_{t,1} = -p_{\text{fail}}(N, d)$. Thus $G_N = \frac{1}{2} - 2^{-N}$ and $I_{t,1} = (1 - 2^{1-N})\Phi(-d/2)$. Equivalently, $B_{t,1} - \mathbb{E}[V_1] = G_N[1 - 2\Phi(-d/2)]$. For $N \geq 2$, candidate count controls best-candidate gain, while observation separation d controls the fraction an optimal selector can preserve. For $N = 1$, the output is the sole independent draw, giving failure $1/2$. For $d = 0$, positive definite Σ implies $h(x_a) = h(x_b)$, so the observations cannot distinguish the equally likely targets. Using $\Phi(0) = 1/2$ in Equation (8) gives failure $1/2$ for every N. As N grows, missing-candidate failures vanish while testing error remains. Independent binary draws and exact continuation are specific to this example.

B Implementation and Evaluation

B.1 Observations, continuations, and allocation choices

We use the pretrained optimal-transport (OT) flow priors from Martin et al. (2025), distributed with FLOWER (Pourya et al., 2026): CelebA at 128×128 and AFHQ-Cat at 256×256. The prior

weights remain fixed throughout our experiments. Appendix E gives the AFHQ-Cat protocol. Cross-dataset comparisons also change the prior and resolution, and the phase–central comparison changes both task and solver. The main CelebA phase allocation uses FLOWER with the native Fourier-magnitude operator and 256×256 measurements. The Fourier-grid diagnostic instead uses the DAPS operator (Zhang et al., 2025). Appendix C.4 details both observation protocols. For CelebA, the Allocation evaluation, Single-model evaluation, Ensemble evaluation, and Irregular-mask evaluation use FlowDPS for inpainting with a binary observation mask. Noise has standard deviation 0.05 on the $[-1, 1]$ scale. Their central and three-polyline irregular masks each remove $96^2 = 9{,}216$ pixels. The Selection-loss evaluation and Candidate-retention evaluation additionally examine central masks with side lengths $32, 64, 96$ using FLOWER and $64, 96$ using FlowDPS (Figures 3(a) and 7). Additional tables report normalized mean squared error (NMSE), where lower is better. For perceptual utility we use $-$LPIPS, while reporting LPIPS itself. We use pretrained AlexNet LPIPS v0.1 (package 0.1.4) with model-scale $[-1, 1]$ inputs and no additional input rescaling, evaluating full images without cropping or reconstruction clipping. Tenth percentiles refer to PSNR or paired PSNR effects, as labeled.

Native solver runs use $T = 100$ outer steps over $[0, 1]$. Other step counts span the same interval. The reference GBL allocation is $N = 8, m = 64, k = 1$. Eight candidates are fixed, without claiming optimality. For CelebA, likelihood pruning selects $N \in \{2, 4, 8\}$ and $t \in \{.32, .48, .64, .80\}$ by development mean PSNR under work caps $100, 200, 400, 548$. These choices are fixed before test evaluation. Central single-path refinement uses only 100 and 548 steps. Completed multistart selects by final measurement NLL. At the largest cap, completed multistart uses five candidates, and central likelihood pruning uses eight at development-selected $t = 0.48$. In the CelebA Allocation evaluation, central-mask **GBL-P** uses calibrated override with its ensemble and threshold fixed from separate images. At the largest phase-retrieval budget, likelihood pruning coincides with the reference **GBL-M** allocation.

We leave the original solver's initializations, coupled random draws, update steps, and final image conversion unchanged. **GBL-P** uses only prefix features to select an output from saved completions, adding no projections or clamps. Completing development candidates supplies training labels at an offline cost. OT-prior metrics use solver endpoints mapped to image scale. Clipping affects display only.

B.2 Data splits and predictor fitting

The CelebA Allocation evaluation shares 32 images for choosing allocations and failure thresholds and 128 test images across tasks. The Irregular-mask evaluation uses 192 separate predictor-training images. The Selection-loss evaluation uses 64 images per task–solver setting. Fitting and resampling keep each image's candidates together. All test outcomes are included. In the central splits (Table 5), Single-model evaluation's test images enter Ensemble evaluation's training set, but each evaluation excludes its own training images from testing. The same image-level splits apply across initialization runs.

Table 5: CelebA central-mask training, calibration, and evaluation images. The Single-model evaluation and Ensemble evaluation results appear in Table 3.

Use	Images	Split relationship
Single-model evaluation training	64	Initial predictor-training set
Single-model evaluation test	128	Separate from Single-model evaluation training
Ensemble evaluation training	192	Single-model evaluation training and test images
Override calibration	128	Separate calibration images
Initialization evaluation, calibration initializations	128	Separate from calibration, same set of initializations
Initialization evaluation, new initializations	128	Same evaluation images, new initializations
Ensemble evaluation test	128	Separate from all preceding labeled sets, new initializations

Single-model evaluation and Ensemble evaluation fit candidate-level squared-error regression to full-image PSNR at completion. Histogram-based gradient boosting regressors (Friedman, 2001) use learning rate 0.05, 200 iterations, at most 15 leaves, at least 20 samples per leaf, ℓ_2 regularization 1, and no depth limit. For a candidate drawn from the training population, squared-error regression targets mean completed quality conditional on its feature vector. For square-integrable targets, expected loss is the conditional variance plus the squared error from this mean. The features can contain less information than $\mathcal{F}_t$, so this target need not equal $\mu_{i,t} = \mathbb{E}[V_i \mid \mathcal{F}_t]$. **GBL-P** uses these feature-based predictions to approximate the Bayes ranking. Single-model evaluation fits 512 equally weighted candidate examples from 64 images. Ensemble evaluation partitions 192 images (1,536 candidates, weight $1/8$ each) into five folds, keeping candidates from the same image together. Each fit excludes one fold, and we average the five models' PSNR predictions. Predictors are fixed before testing.

Task-specific irregular HGBR uses the same hyperparameters, 192 development images, and five image-level folds. The irregular-mask ridge control (Hoerl & Kennard, 1970) uses all 13 raw features, regularization 1, and no external standardization. Both average five predictions from fits excluding one fold each. These fits differ from the transferred central ensemble.

B.3 Features and selection rules

Table 6 defines the 13 inputs to **GBL-P**'s predictor for reproducibility. Its feature numbers are used to specify the input subsets compared in Figure 8. Current and preceding checkpoints are steps 64 and 48 of the 100-step solver. HGBR inputs are unstandardized and exclude targets, future states, and image identifiers. Clean estimates are the solver's predictions of clean images from the current flow states, before the checkpoint's local correction. To compute measurement NLL, we apply the measurement operator to the clean estimate and subtract the observation. The residual norm is the Euclidean norm of this vector. The recorded NLL score is half the sum of its squared components, weighted by the solver's recorded diagonal inverse variance and likelihood scale. These quantities use the solver's image and observation scales.

Table 6: Trajectory features used to predict candidate quality.

Feature	Definition
1. Measurement NLL	Current measurement negative log-likelihood
2. Residual norm	Current measurement residual norm
3. Measurement rank	Rank by NLL, with ties resolved by lower candidate index
4. Top-two margin	Second-lowest minus lowest NLL
5. NLL change	Current minus preceding NLL
6. NLL slope	NLL change divided by the checkpoint time increment
7. Rank change	Current minus preceding rank
8. Rank stable	Indicator of unchanged rank
9. Clean estimate change	ℓ_2 distance between current and preceding clean estimates
10. Local correction	Recorded local-correction ℓ_2 norm
11. Measurement gradient	Recorded measurement-gradient ℓ_2 norm
12. Clean estimate dispersion	Mean ℓ_2 distance to the other candidates' clean estimates
13. Selection time	Numerical checkpoint time

Ranks and margins compare candidates within each image. The top-two margin has one value shared by that image's candidates, whereas ranks and clean estimate dispersion are candidate-specific. Dispersion averages distances to the other $N - 1$ clean estimates. For $N = 1$, margin and dispersion are zero. All predictor comparisons here select at $t = 0.64$, so time is constant and retained as part of the common feature schema. Slope denominators are floored at 10^{-12}. Single-model evaluation and the Allocation evaluation use recorded clean-estimate change norms. Ensemble evaluation recomputes them in double precision from stored clean estimates. Local-correction features require the additional corrections counted below.

Direct predictive selection maximizes predicted candidate quality. NLL and prediction ties use the lower candidate index, except Single-model evaluation prediction ties use the higher index. Neither held-out evaluation has ties at the highest prediction.

On calibration images where $j_P \neq j_M$, define overestimation $r = \widehat{\Delta} - (V_{j_P} - V_{j_M})$. Agreement needs no calibration because both rules return the same candidate. Whenever $r \leq q$, the actual gain obeys

$$V_{j_P} - V_{j_M} = \widehat{\Delta} - r \geq \widehat{\Delta} - q.$$

Calibrated override therefore requires this adjusted predicted gain to be positive before changing the measurement choice.

Sort the n disagreement residuals as $r_{(1)} \leq \cdots \leq r_{(n)}$ and set the corrected 90th percentile $q = r_{(\lceil 0.9(n+1) \rceil)}$. The $n+1$ rank correction comes from conformal prediction (Vovk et al., 2005). Standard split-calibration validity uses a fixed scoring rule and exchangeable calibration and test residuals. If the required rank exceeds n, we set $q = +\infty$, which disables overrides when the required order statistic is unavailable.

For irregular calibration, each image's predictions come from the single fit that excluded its fold. HGBR has 164 disagreements and $q = 2.6988$ (rank 149), while the ridge model has 162 and $q = 2.1604$ (rank 147). On separate test images, each irregular predictor instead averages all five fits. These irregular thresholds serve as empirical guards because calibration and testing use different predictors. For the transferred central ensemble, the fixed ensemble scores separate calibration images. Its 105 disagreements give $q = 3.4728$ (rank 96). Thresholds are rounded. The central Allocation evaluation and transfer in the Irregular-mask evaluation do not refit or recalibrate. Calibrated override changes the **GBL-M** choice only when $\widehat{\Delta} > q$, retaining it at equality without extra tolerance. Direct predictive selection can differ from zero-threshold calibrated override when maximal predictions tie. Calibration does not guarantee that PSNR is preserved for every image under new mask geometry or distributions.

B.4 Work accounting and measured latency

The following boundary-evaluation counts apply to phase retrieval and inpainting; Appendix F.4 specifies the prospective super-resolution implementation.

For one retained candidate, $(N, m) = (4, 32), (8, 32), (8, 48), (8, 64)$ use $196, 324, 436, 548$ updates, respectively. Five complete native trajectories use 500, below the 548 cap.

For reference **GBL-M**, pruning occurs after the next flow-model evaluation but before its local correction. Seven discarded candidates execute 65 flow-model evaluations and 64 updates each, and the retained candidate executes 100 of each: 555 flow-model evaluations and 548 completed updates. In the implemented FlowDPS solver, each outer update uses three local correction iterations, with one operator vector-Jacobian product (VJP) per iteration. Thus **GBL-M** uses $3 \times 548 = 1{,}644$ FlowDPS operator VJPs. FLOWER uses one per outer update, giving 548.

GBL-P uses the checkpoint's local-correction norm as a feature. It therefore needs that correction even for the seven candidates that **GBL-M** could discard beforehand. These seven additional corrections require $7 \times 3 = 21$ FlowDPS operator VJPs, giving 1,665 with the same 555 flow-model evaluations. For FLOWER, the seven extra corrections give 555 operator VJPs with the same 555 model evaluations. These **GBL-P** counts follow from the additional local corrections. Predictor scoring adds no flow-model evaluation or operator VJP. Equal outer-update work is not equal total FLOPs.

Runtime evaluation. On eight fixed images using NVIDIA GH200 hardware, **GBL-M** stops discarded trajectories during inference and takes 30.32 s/image for phase retrieval and 29.65 for central masking. Single-path refinement at 548 updates takes 41.01 and 37.11 s/image. Image coverage differs from the 128-image quality evaluation. These end-to-end timings cover **GBL-M** and single-path refinement. Irregular single-path refinement was timed while other candidate reconstructions ran concurrently. Single-model evaluation's median CPU aggregation and eight-candidate scoring times are 6.1 and 0.87 ms/image. For the five-model ensemble, aggregation and scoring take 101.9 and 4.66 ms/image. These component timings exclude data loading and saving.

B.5 Statistical procedures

The Allocation evaluation and Irregular-mask evaluation use 20,000 paired image-bootstrap resamples, conditional on fitted policies, initializations, and development choices. Each resample draws

images with replacement, keeping all compared methods' outcomes for an image together, and recomputes the mean paired difference. The 2.5th and 97.5th percentiles of these means give the pointwise 95% interval.

For each comparison, the one-sided bootstrap-tail probability is the number of nonpositive bootstrap means plus one, divided by the number of resamples plus one. We apply Holm correction to these probabilities for three primary comparisons: phase **GBL-M** versus single-path refinement and versus completed multistart, and central **GBL-P** with calibrated override versus **GBL-M**. The adjusted bootstrap-tail probabilities are 0.000150, 0.000400, and 0.018249. The displayed confidence intervals remain pointwise. Irregular task-specific **GBL-P** versus **GBL-M** is a separate primary contrast, with task-specific versus transfer secondary.

With the fitted selection policies fixed, intervals use 10,000 image resamples in Single-model evaluation and 20,000 in Ensemble evaluation. A separate 1,000-replicate Single-model evaluation retraining bootstrap resamples training and evaluation images and refits the predictor, giving percentile $[0.100, 0.747]$ dB and basic $[-0.035, 0.612]$ dB intervals. The basic interval reflects the bootstrap percentile interval about the original mean paired effect: its lower endpoint is twice that effect minus the upper bootstrap percentile, and its upper endpoint is twice that effect minus the lower bootstrap percentile. Neither procedure includes uncertainty from choosing the method or feature set.

Supplementary allocation and direct predictive selection versus calibrated override contrasts were specified after outcomes and reuse the 20,000 paired resamples. Each interval is computed directly from paired differences. Tenth percentiles use linear interpolation between adjacent order statistics. Failure thresholds are native single-path refinement's development fifth, tenth, and twentieth percentiles, fixed before testing. Counts use these fixed thresholds and omit their estimation uncertainty.

C Additional Allocation Evidence

C.1 Published solvers and qualitative examples

Native-solver evaluation. Table 7 uses a separate 128-image phase-retrieval test set with paired observations. We adapt D-Flow to the same prior and operator. Its development-selected configuration uses L-BFGS with 30 outer calls and at most 20 inner iterations, three explicit-midpoint integration steps (six flow-model evaluations per solve), and the same Gaussian measurement objective. It averages 3,933 flow-model evaluations, 3,921 flow-model backward VJPs, and about 347 s/image. Each solver uses its own configuration and computational budget. Figure 5 shows the example nearest to the 75th percentile of paired PSNR effects for FLOWER with **GBL-M** versus native single-path FLOWER in Table 7.

Table 7: **Phase-retrieval reconstruction with published inverse solvers.** The Native-solver evaluation uses native configurations on a separate 128-image test set. D-Flow is adapted to the same prior and operator. The Allocation evaluation uses its own test set and work accounting.

Method	PSNR (dB) ↑	NMSE ↓
D-Flow adaptation (Ben-Hamu et al., 2024)	19.95	0.345
FlowDPS (Kim et al., 2025)	18.45	0.488
FLOWER (Pourya et al., 2026)	24.90	0.217
Global-before-local configurations (ours)		
FlowDPS + **GBL-M**	24.73	0.159
FLOWER + **GBL-M**	**27.39**	**0.096**

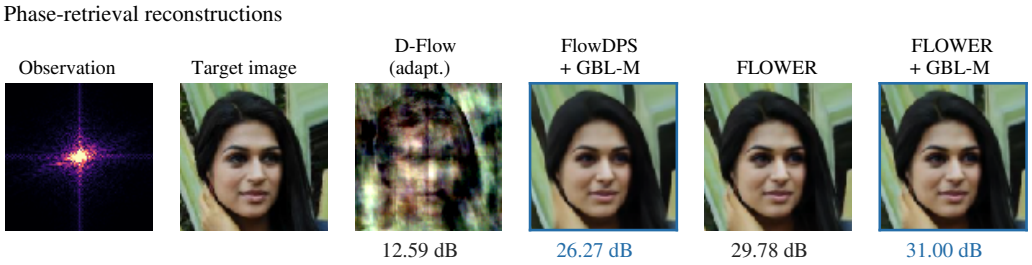

Figure 5: **GBL-M improves this phase-retrieval reconstruction.** This CelebA example from the Native-solver evaluation compares native FLOWER with FLOWER plus **GBL-M**. D-Flow (adapted) and FlowDPS plus **GBL-M** provide additional solver comparisons. Labels give full-image PSNR, and blue frames identify GBL outputs. The observation view suppresses noise for display.

C.2 Paired allocation effects and failure thresholds

Table 8 collects the paired effects of **GBL-M** relative to equal-work refinement in both datasets.

Table 8: **Allocation gains and losses relative to equal-work refinement.** Each row reports **GBL-M** minus single-path refinement in full-image PSNR (dB), with 128 paired images and $W = 548$ for both methods. CelebA intervals are pointwise 95%. AFHQ intervals are 98.75%, Bonferroni-adjusted over the four primary AFHQ phase/central PSNR contrasts (Appendix E).

Setting	Mean effect	Paired interval	Interval convention
CelebA phase	$+3.884$	$[2.272, 5.641]$	95%, pointwise
CelebA central	-0.630	$[-0.888, -0.370]$	95%, pointwise
AFHQ phase	$+1.677$	$[-0.535, 3.960]$	98.75%, adjusted
AFHQ central	$+0.590$	$[0.143, 1.020]$	98.75%, adjusted

Table 9: **Lower-tail reconstruction quality in the CelebA allocation evaluation.** The 128-image comparisons and work budgets are given in Table 2. The tenth percentile is full-image PSNR (dB). Failure thresholds are 6.242 dB for phase retrieval and 14.147 dB for central masking. Central **GBL-P** uses a transferred ensemble with calibrated override at $W = 548$ and has mean PSNR 16.67 dB.

Configuration	10th percentile	Failure (%)
Phase retrieval / FLOWER		
Single-path refinement	6.70	9.38
Completed multistart	8.87	5.47
GBL-M	16.79	3.91
Central-mask inpainting / FlowDPS		
Single-path refinement	14.24	9.38
Completed multistart	12.75	23.44
Likelihood pruning	13.29	17.19
GBL-M	13.32	16.41
GBL-P, calibrated override	13.32	16.41

On the 128 test images in the Allocation evaluation, phase-retrieval **GBL-M** exceeds completed multistart by 1.637 dB, with paired 95% interval $[0.648, 2.795]$. This uses 548 versus 500 outer updates under a common cap. For **GBL-M** versus single-path refinement in phase retrieval, the median and tenth percentile of the paired PSNR effect are -0.419 and -1.784 dB.

Under irregular masking, likelihood pruning exceeds task-specific **GBL-P** with calibrated override by 0.278 dB $[-0.076, 0.644]$, using $W = 324$ versus 548. These supplementary contrasts were specified after the original outcomes. The likelihood-pruning contrast changes allocation and selection together, and its paired interval spans zero.

Table 10 gives the threshold grid. Primary tenth-percentile thresholds are 6.2417 dB (phase), 14.1467 (central), and 13.9215 (irregular). Thresholds stay fixed across continuations, and each count uses that configuration's completed candidates.

Table 10: **The allocation contrast persists at alternative failure thresholds.** These fifth- and twentieth-percentile thresholds supplement the tenth-percentile analysis in Figure 2(c,d). The Allocation evaluation uses 128 images, each policy's completed candidates, and Equation (3). Single-path refinement uses 548 updates.

Dev. percentile	τ (dB)	Policy	Missing-candidate failures	Discard failures	Total (%)
Phase retrieval / FLOWER					
5	4.322	Single-path refinement	8	0	8 (6.25)
5	4.322	**GBL-M**	0	2	2 (1.56)
20	13.458	Single-path refinement	33	0	33 (25.78)
20	13.458	**GBL-M**	9	1	10 (7.81)
Central-mask inpainting / FlowDPS					
5	13.262	Single-path refinement	7	0	7 (5.47)
5	13.262	**GBL-M**	2	10	12 (9.38)
20	14.472	Single-path refinement	15	0	15 (11.72)
20	14.472	**GBL-M**	7	17	24 (18.75)

C.3 Selection improvement needed to match single-path refinement

With one retained candidate ($k = 1$), fix the candidate qualities, utility, and selection information. Here V^π is the selected-output quality of the current policy, as defined in Appendix A.1. For a feasible comparator ρ, the common mean best-candidate quality cancels when we subtract the two losses:

$$\begin{aligned} L^\pi - L^\rho &= (\mathbb{E}[M_N] - \mathbb{E}[V^\pi]) - (\mathbb{E}[M_N] - \mathbb{E}[V^\rho]) \\ &= \mathbb{E}[V^\rho - V^\pi]. \end{aligned}$$

Thus changing selection reduces total selection loss by exactly the amount that it raises mean selected-output quality.

Suppose the current mean is below D_W and $L^\pi > 0$. Reaching or exceeding single-path refinement requires a mean increase of at least $D_W - \mathbb{E}[V^\pi]$. Dividing by the current loss gives

$$\frac{L^\pi - L^\rho}{L^\pi} \geq \frac{D_W - \mathbb{E}[V^\pi]}{L^\pi}.$$

Equality gives a tie in mean quality, and strict inequality gives an advantage. A required fraction above one is impossible because total selection loss cannot be negative. A required fraction of one means $D_W = \mathbb{E}[M_N]$, so matching requires zero total selection loss and outperforming is impossible.

The information restriction can make even a smaller fraction unattainable. Appendix A.1 shows that the Bayes selector attains $B_{t,1}$. Its mean improvement over the current policy is $B_{t,1} - \mathbb{E}[V^\pi] = S^\pi$, the maximum loss that a feasible selector can recover with these inputs. Consequently, some feasible policy can reach or exceed D_W exactly when $B_{t,1} \geq D_W$. A strict advantage is possible exactly when $B_{t,1} > D_W$. The mean best-candidate advantage over equal-work single-path refinement is $\mathbb{E}[M_N] - D_W = G_N - (D_W - \mathbb{E}[V_1])$. The recovery condition in Section 4.2 concerns this advantage, not G_N alone.

For the 128 test images in the Allocation evaluation, central **GBL-M** has a mean deficit of 0.62954 dB to single-path refinement and total selection loss of 1.77458 dB. A replacement selector must recover the entire deficit to tie the baseline, giving $0.62954/1.77458 = 35.5\%$. Direct **GBL-P** raises mean quality by 0.37529 dB on the same candidates (Table 17). It therefore recovers $0.37529/1.77458 = 21.1\%$ of total selection loss, leaving $0.62954 - 0.37529 = 0.25425$ dB, reported as a 0.254 dB shortfall in observed mean. We evaluate the already fitted policies against this target on the allocation cohort, which is separate from the Single-model evaluation and Ensemble evaluation test cohorts. For phase, mean best-candidate quality $\mathbb{E}[M_N]$ exceeds single-path refinement by 4.591 dB and

GBL-M loses 0.708 dB, leaving its observed 3.884 dB advantage. These components are rounded separately.

C.4 Changing Fourier observations with the solver fixed

The Fourier-grid evaluation (Figure 6) fixes 64 separate CelebA images, FLOWER, the prior, the 100-step continuation rule, and eight initializations. DAPS (Zhang et al., 2025) oversampling $0, 2, 4$ gives 128, 192, and 256 measurement pixels per side through zero-padding. Noise standard deviation is 0.05. Levels were fixed before outcomes. Each grid defines a different observation problem and produces its own completed candidate qualities under the same continuation rule. **GBL-M** selects at step 64 from these saved completed candidates. This comparison measures how **GBL-M** responds to the Fourier grid, using the original 10,000 paired image-bootstrap resamples. Figure 6 connects best-candidate gain G_N to selected-output gain $G_N - L^\pi$ for each grid, with arrow lengths equal to selection loss. The smallest-minus-largest selection-loss increase is 1.156 dB $[0.608, 1.757]$. Best-candidate gain G_N for the smallest grid minus that for the largest is 0.018 dB $[-1.537, 1.595]$, spanning both positive and negative differences. The Fourier grids are not nested measurement sets, and their candidate qualities differ. Appendix A.2 instead compares nested information with candidate qualities fixed. This experiment measures how total selection loss changes across observation problems, without identifying information-limited loss $I_{t,k}$.

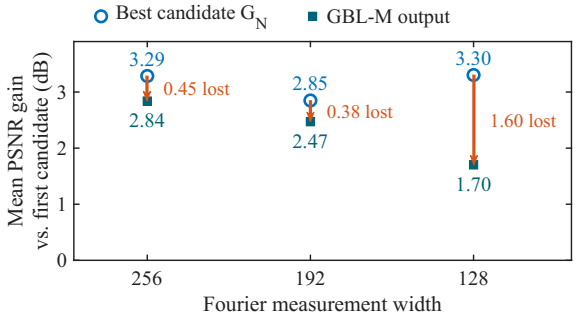

Figure 6: **With FLOWER fixed, the smallest Fourier grid has the largest selection loss.** Best-candidate gain G_N and selected-output gain $G_N - L^\pi$ are measured relative to the first candidate's completed quality V_1. Arrows show selection loss on the same 64 CelebA images across three measurement grids. The equal-work comparison below instead uses 548-step refinement as reference.

Equal-work refinement. We compare the same candidates with one 548-step path per image and grid (Table 11). The 100-step outputs were reproduced exactly; each long path shares the first candidate's initialization, measurement, and first 100 trajectory draws. Intervals use 20,000 paired image-bootstrap resamples. Table 12 applies Equation (4) with the long path as reference. The clearest difference across grids is selection loss: width 128 exceeds width 256 by 1.156 dB (pointwise 95% interval $[0.617, 1.747]$). The corresponding headroom difference is -0.115 dB $[-1.673, 1.349]$, and the output-advantage difference is -1.270 dB $[-3.094, 0.404]$; these two contrasts use Bonferroni-adjusted 97.5% intervals. Width 192 has mean headroom -0.388 dB with an interval spanning zero. Thus selection loss changes with the observation grid, while its relationship to equal-work headroom remains unresolved. Figure 6 uses the original 10,000-resample analysis relative to V_1.

Table 11: **Allocation across Fourier measurement grids.** Each width uses the same 64 images and FLOWER. PSNR is in dB; lower LPIPS is better. W counts outer updates. Best-of-eight selects by target PSNR for diagnosis; LPIPS is measured on the same candidate.

		Width 128		Width 192		Width 256	
Method	W	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
Native solver (100 steps)	100	15.153	0.2826	24.809	0.1386	26.233	0.1209
Single-path refinement	548	18.068	0.2353	28.047	0.1041	29.015	0.1032
Completed multistart	500	15.547	0.2675	27.210	0.0979	29.164	0.0821
GBL-M	548	16.855	0.2470	27.280	0.0970	29.072	0.0826
Best of eight (diagnostic)	800	18.457	0.2122	27.659	0.0939	29.518	0.0762

Table 12: **Candidate advantage minus selection loss determines the output advantage.** Means and pointwise 95% paired image-bootstrap intervals (dB). The first two columns give the terms in Equation (4); their difference gives the third.

Width	Headroom $\mathbb{E}[M_N] - D_W$	Loss L^π	Advantage $\mathbb{E}[V_{\text{sel}}] - D_W$
128	0.389 [-0.778, 1.524]	1.602 [1.143, 2.138]	-1.213 [-2.389, -0.062]
192	-0.388 [-1.251, 0.525]	0.378 [0.225, 0.558]	-0.766 [-1.624, 0.142]
256	0.504 [-0.553, 1.828]	0.447 [0.232, 0.704]	0.057 [-1.023, 1.387]

Observation protocols in the two phase studies. The $+3.884$ dB main allocation effect and the -0.766 dB 192-pixel Fourier-grid effect come from different observation protocols and nonoverlapping cohorts of 128 and 64 images. The main evaluation uses the native Fourier-magnitude operator with 256×256 measurements; the 192-pixel Fourier-grid diagnostic uses DAPS with 192×192 measurements. The native operator pads 64 pixels per side, keeps images on the $[-1, 1]$ scale, and uses an uncentered orthonormal FFT magnitude with a numerical offset. DAPS pads 32 pixels per side, rescales images to $[0, 1]$, and uses centered FFT magnitudes. The noise standard deviation agrees, but the measurement dimensions and trajectory randomness differ. Main trajectories use separate random draws for each branch and total step count; Fourier long paths share the first 100 draws with the 100-step baseline, and candidate branches share continuation draws after step 64. The prior, FLOWER settings, time grids, selection rule, and PSNR convention agree. These are therefore not direct replications, and the comparison does not isolate the effect of any single protocol difference.

These comparisons use saved per-image reconstruction scores and experimental settings.

C.5 Selection time, retained-set quality, and output quality

Candidate-retention evaluation. Figure 7 compares nested retained sets under **GBL-M** on 64 development images per setting. At the selection time, it retains the k candidates with lowest current measurement NLL. Each retained candidate completes the same continuation, after which the output is chosen by lowest final measurement NLL. For each image,

$$M_N - V_{\text{sel}} = (M_N - V(A)) + (V(A) - V_{\text{sel}}).$$

The first term is quality lost when forming the retained set A. The second is quality lost in the final choice within A. The gap between the dashed and solid curves is the mean of this second term. Retaining $k = 1, 2, 4, 8$ candidates uses $548, 584, 656, 800$ outer updates, respectively, excluding boundary and feature work.

With completed candidate qualities fixed, increasing k keeps all previously retained candidates in A, so its maximum quality $V(A)$ cannot decrease. A newly retained candidate can have lower final measurement NLL but worse PSNR, causing the selected output to deteriorate. For FlowDPS with the 96×96 central mask, mean selected-output quality falls from 17.070 to 16.451 dB between $k = 1$

and $k = 8$. With every candidate retained, $V(A) = M_N$ and the first term vanishes, but the final choice can still lose quality. This explains why the fixed measurement rule can behave differently from the Bayes optimum in Appendix A.2.

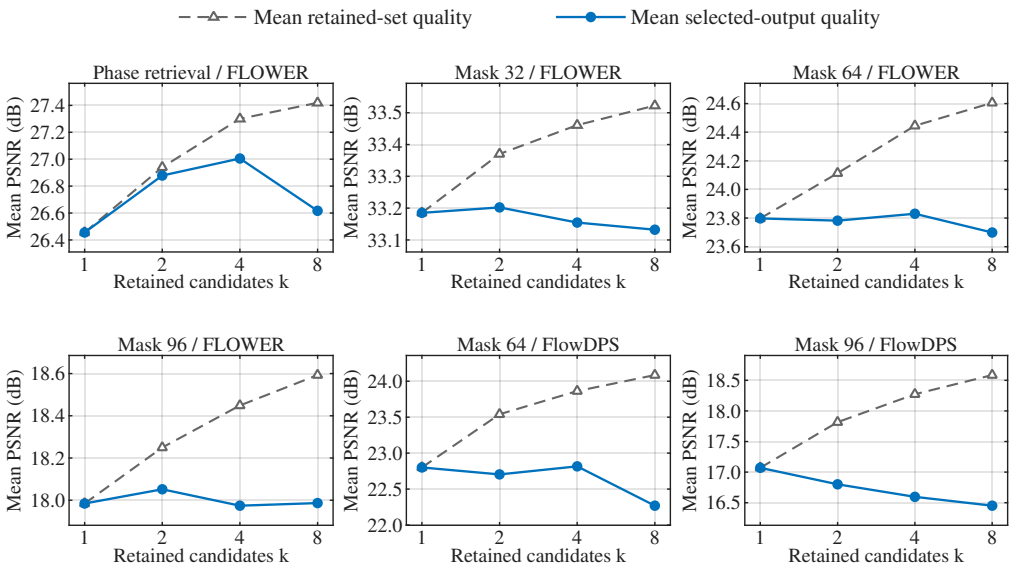

Figure 7: **Retaining more candidates can preserve better reconstructions but yield a worse final selection.** Dashed curves show the mean quality of the best completed candidate in each retained set, $\mathbb{E}[V(A)]$. Solid curves show mean selected-output quality $\mathbb{E}[V_{\text{sel}}]$ after final measurement NLL selection. Their gap is loss at this final selection. The Candidate-retention evaluation uses 64 CelebA images per panel and retains k of $N = 8$ candidates after 64 of 100 steps ($t = .64$). More retained candidates require more continuation work. Mask numbers are square side lengths in pixels.

Selecting at completion lowers PSNR in the largest-mask FlowDPS setting. The same saved trajectories also allow selection at $t = 0.16, 0.32, 0.48, 0.64, 0.80, 1.00$, retaining one candidate with lowest current NLL and completing it with the fixed continuation. Table 13 reports all six settings. With $N = 8$ and $T = 100$, work is $W = 100(1 + 7t)$, so later decisions use more updates. At $t = 1$, all candidates have already completed and selection uses their final NLL. For FlowDPS with the largest central mask, the mean falls from 17.070 dB at $t = .64$ to 16.451 dB at completion. This endpoint equals the $k = 8$ result above because both rules choose among all completed candidates by final NLL. The fixed rule uses the current score at each time. The optimum in Appendix A.2 can instead use all accumulated information, including earlier scores. Delaying selection and improving the selection rule therefore address different parts of the problem.

Table 13: **Later measurement-based selection does not consistently improve final PSNR.** Mean PSNR on the same 64 development images per setting, with eight fixed candidate trajectories and one selected output. The times use $212, 324, 436, 548, 660, 800$ outer updates, respectively.

Task	Solver	Selection time t					
		0.16	0.32	0.48	0.64	0.80	1.00
Phase	FLOWER	24.761	26.588	26.722	26.455	27.044	26.617
Central 32	FLOWER	33.177	33.178	33.163	33.185	33.325	33.132
Central 64	FLOWER	23.860	23.900	23.814	23.797	24.098	23.699
Central 64	FlowDPS	22.305	22.763	23.159	22.801	23.054	22.269
Central 96	FLOWER	18.080	18.032	17.987	17.984	18.150	17.987
Central 96	FlowDPS	17.126	17.104	17.026	17.070	16.843	16.451

Central-mask numbers denote side lengths in pixels. Values summarize the saved development sweep, with no new predictor fitting.

C.6 Same-task solver control at equal outer-update work

We compare FLOWER and FlowDPS on the same 64 CelebA images with 96×96 central masks. This reuses the supplementary cohort. The original 100-step outputs were reproduced exactly; each long path shares its baseline's initialization, measurement, and first 100 trajectory draws. Both step grids span $[0, 1]$, with evidence weight one per update. Intervals use 20,000 paired image-bootstrap resamples, conditional on the chosen protocol. Work W matches outer updates within each solver; model calls and latency are accounted for separately (Appendix B.4).

Table 14: **Allocation depends on the solver even with the task fixed.** Mean PSNR (dB) and LPIPS on 64 CelebA images with 96×96 central masks. Native refinement and multistart help identify degradation under longer refinement. Best-of-eight selects by target PSNR; LPIPS is measured on the same candidate.

Method	W	FLOWER		FlowDPS	
		PSNR	LPIPS	PSNR	LPIPS
Native solver (100 steps)	100	17.931	0.2358	16.567	0.2464
Single-path refinement	548	16.870	0.2767	17.261	0.2693
Completed multistart	500	18.036	0.2331	16.372	0.2404
GBL-M	548	17.984	0.2337	17.070	0.2269
Best of eight (diagnostic)	800	18.593	0.2259	18.586	0.2143

Table 15: **Paired contrasts for the same-task solver control.** Positive PSNR and negative LPIPS differences favor the first method. †: 97.5% individual intervals adjusted for the two primary supplementary PSNR comparisons; all other intervals are pointwise 95%.

Solver	Contrast	Metric	Mean	Interval
FLOWER	**GBL-M** − long path	PSNR	1.114	[0.840, 1.392]†
FLOWER	**GBL-M** − long path	LPIPS	-0.04299	[-0.05116, -0.03508]
FLOWER	**GBL-M** − multistart	PSNR	-0.052	[-0.182, 0.082]
FLOWER	**GBL-M** − multistart	LPIPS	0.00055	[-0.00311, 0.00408]
FLOWER	multistart − long path	PSNR	1.166	[0.935, 1.399]
FLOWER	multistart − long path	LPIPS	-0.04354	[-0.05190, -0.03563]
FlowDPS	**GBL-M** − long path	PSNR	-0.191	[-0.627, 0.229]†
FlowDPS	**GBL-M** − long path	LPIPS	-0.04248	[-0.05331, -0.03185]
FlowDPS	**GBL-M** − multistart	PSNR	0.698	[0.250, 1.152]
FlowDPS	**GBL-M** − multistart	LPIPS	-0.01354	[-0.02307, -0.00443]
FlowDPS	multistart − long path	PSNR	-0.889	[-1.306, -0.470]
FlowDPS	multistart − long path	LPIPS	-0.02894	[-0.03915, -0.01808]

The **GBL-M**-minus-refinement PSNR effect is 1.114 dB for FLOWER and -0.191 dB for FlowDPS, whose interval spans zero. Their paired difference is 1.305 dB (pointwise 95% interval $[0.921, 1.698]$), showing that allocation behavior depends on the solver even on the same task. The LPIPS difference is inconclusive: -0.00051 $[-0.01244, 0.01203]$. FLOWER's long-path PSNR falls from the native baseline's 17.931 to 16.870 dB, and completed multistart slightly exceeds **GBL-M**. The breadth advantage therefore partly reflects degradation under longer refinement.

D Additional Selection Evidence

D.1 Selection loss and feature controls

The Selection-loss evaluation (Figure 3(a)) uses five-fold image cross-fitting with Appendix B.2's hyperparameters. Each development image is scored by the model fitted without its fold, so different images can be scored by different fitted models. The training sets of these fits overlap. In Ensemble evaluation, every test image is scored by the same frozen ensemble averaging five fitted models, and none of those test images is used for fitting.

For FlowDPS with 96×96 central masks, the paired 95% interval for **GBL-P** versus **GBL-M** starts at 0.025 dB. Figure 3(a) reports the recovery and remaining selection loss. The loss reduction equals the signed mean paired PSNR effect, with negative values indicating increased loss. Intervals use 10,000 paired image resamples.

The largest central mask has low terminal ranking agreement under both solvers. Table 16 adds best-candidate gain and terminal score agreement to the same six development settings. For each image, we rank its eight completed candidates by PSNR and by negative final measurement NLL, then compute their Spearman correlation. The table averages these within-image correlations over 64 images. Positive correlation means better measurement fit tends to accompany better reconstruction. The average is near zero for the largest central mask under both solvers, even after all candidates have completed. These correlations describe the NLL ranking. Policies using additional trajectory information are assessed separately below.

Table 16: **Terminal measurement fit has near-zero mean rank correlation with PSNR under the largest central mask.** The same development candidates support Figure 3(a) and Table 13. Best-candidate gain uses $N = 8$ and the same 100-step continuation. Rank correlation compares negative final NLL with final PSNR.

Task	Solver	Best-candidate gain G_N (dB)	Mean rank correlation
Phase	FLOWER	3.486	0.660
Central 32	FLOWER	0.438	0.109
Central 64	FLOWER	0.929	−0.049
Central 64	FlowDPS	1.889	0.206
Central 96	FLOWER	0.661	0.022
Central 96	FlowDPS	2.019	−0.057

G_N is relative to the first candidate's completed quality V_1. Rank correlation uses negative NLL so that higher scores mean better fit. Central-mask numbers denote side lengths in pixels.

Comparison with equal-work refinement uses D_W rather than the first-candidate reference V_1 (Section 4.2).

The rank comparison motivates testing information beyond NLL. The model comparison in Figure 3(b) and the additional feature-subset controls in Figure 8 use Ensemble evaluation's 192 training images, 128 test images, and five image-level folds. Measurement NLL uses feature 1, current measurement uses 1–4, history / rank uses 5–8, and clean estimate / correction uses 9–12. All-feature models additionally include time (feature 13). In this central-mask feature comparison, ridge regression standardizes features within each training split and uses regularization 1. This differs from the raw-feature irregular-mask control (Appendix B.2). HGBR uses raw features. We average five models' predictions. The NLL-only HGBR uses the full model's settings and folds, with equal weight

per candidate and no feature standardization. It learns to predict PSNR from NLL, so its ranking can differ from **GBL-M**'s direct lowest-NLL rule. Figure 3(b) compares the HGBR predictors and the all-feature ridge model. Figure 8 reports the remaining ridge feature-subset controls relative to **GBL-M**. These feature comparisons were specified after the primary evaluation. Intervals reuse Ensemble evaluation's 20,000 paired resamples.

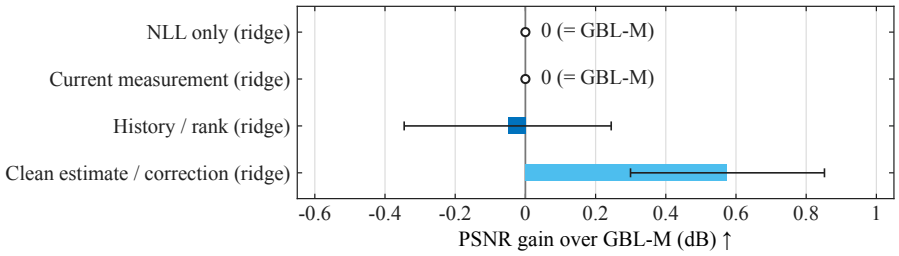

Figure 8: **Clean-estimate and correction features improve ridge selection over measurement selection.** These feature-subset controls complement the model comparison in Figure 3(b). Bars show mean paired PSNR effects relative to **GBL-M** on 128 CelebA central-mask test images in the Ensemble evaluation, with pointwise paired 95% intervals. Open circles labeled 0 denote identical choices to **GBL-M**.

D.2 Held-out failure partition

After the primary evaluation, we apply the failure partition to the central-mask Single-model evaluation and Ensemble evaluation tests in Table 3. It uses $\tau = 14.1467$ dB from 32 central allocation-development images, separate from both tests. For **GBL-M** and direct **GBL-P**, selected-output quality V_{sel} is compared with best-candidate quality M_N from the same eight completed candidates.

Because both selectors use the same candidates, missing-candidate failures are fixed at one in Single-model evaluation and five in Ensemble evaluation. Switching from **GBL-M** to direct **GBL-P**, discard failures fall from 20 to 16 in Single-model evaluation (total 21 to 17) and from 17 to 12 in Ensemble evaluation (total 22 to 17). In each evaluation, nine images cross from below to at or above the threshold, while five in Single-model evaluation and four in Ensemble evaluation cross in the opposite direction. These counts are separate from the sixteen discard failures in the central Allocation evaluation.

D.3 Direct predictive selection and calibrated override

Table 17 compares selection rules within the central-mask Allocation evaluation (Table 9) and the Irregular-mask evaluation (Table 18). Both use the Allocation evaluation's 128 test images. Within each setting, the predictor, observations, eight candidates, features, $t = .64$, and continuation are fixed. Central masking uses transferred HGBR, irregular masking task-specific HGBR, with fixed thresholds for calibrated override. On central masks, mean LPIPS is 0.23560 for **GBL-M**, 0.23101 for direct predictive selection, and 0.23429 for calibrated override.

Table 17: **Calibrated override reduces the worst observed PSNR loss and the mean gain.** The central-mask and irregular-mask settings correspond to Tables 9 and 18, respectively.

Selection rule	Paired full-image PSNR effect (dB)				Changed selections
	Mean	95% interval	10th percentile	Worst	
Central mask, transferred HGBR					
Direct predictive selection	0.375	$[0.107, 0.647]$	-1.415	-3.215	111
Calibrated override	0.062	$[0.008, 0.133]$	0	-0.162	5
Irregular mask, task-specific HGBR					
Direct predictive selection	0.528	$[0.176, 0.869]$	-1.470	-10.347	107
Calibrated override	0.144	$[0.026, 0.304]$	0	-0.875	7

Effects are relative to **GBL-M**, with pointwise paired image-bootstrap intervals. Changed selections: number of images whose choice differs from **GBL-M**, out of 128.

In this posthoc comparison, calibrated override rejects useful changes as well as harmful ones. Averaged over all 128 images, rejected changes forgo 0.723 dB of benefit and avoid 0.409 dB of harm in central masking. The corresponding values are 0.808 and 0.423 dB for irregular masking. Avoided harm minus forgone benefit gives the mean override-minus-direct effects of -0.313 and -0.385 dB, respectively. Thus the reductions in the worst observed loss come with a loss of mean gain. Of the 128 choices, 123 remain unchanged in central masking and 121 in irregular masking. These unchanged choices contribute zero paired effects, placing both tenth percentiles at zero even though some changed choices reduce PSNR.

D.4 Predictive selection under irregular masking

The Irregular-mask evaluation (Table 18) uses the Allocation evaluation's 128 test images, FlowDPS, and the same number of missing pixels as central masking. Allocation is chosen on 32 separate development images under the 548-update cap. Task-specific HGBR and ridge regression are fitted on 192 separate development images, while the transferred predictor uses the central ensemble.

Table 18: **GBL-P with calibrated override improves mean PSNR over GBL-M under irregular masking.** The Irregular-mask evaluation uses calibrated override for **GBL-P**. Figure 4 illustrates direct predictive selection on the same task. Likelihood pruning uses less work and has higher observed mean PSNR than task-specific override.

Policy	Work W	Full-image PSNR (dB)		Failure (%)
		Mean	10th percentile	
GBL-M	548	17.62	14.41	8.59
GBL-P (override), task-specific	548	17.77	14.85	6.25
GBL-P (override), transferred	548	17.72	14.60	7.81
GBL-P (override), ridge control	548	17.78	14.85	6.25
Native solver (100 steps)	100	17.11	13.39	12.50
Likelihood pruning	324	18.05	14.86	5.47

Failure: PSNR below 13.921 dB. Likelihood pruning: $N = 8$, $t = 0.32$ (development-selected).

The task-specific predictor exceeds the transferred predictor by 0.045 dB (paired 95% interval $[-0.026, 0.130]$), leaving their ordering unresolved. Discard failures number eleven for **GBL-M**, eight for task-specific **GBL-P**, ten for its transferred version, eight for ridge regression, and seven for likelihood pruning. None has a missing-candidate failure in this sample. The native 100-step solver has sixteen failures. Counts use the fixed development threshold on these evaluation images.

The CelebA example in Figure 4 was chosen after evaluation.

D.5 Utilities, individual losses, and initialization

In the Selection-loss evaluation for FlowDPS with 96×96 central masks, **GBL-P** has mean paired PSNR effects versus **GBL-M** near 0.39 dB for both full-image and missing-pixel PSNR. The full-image-PSNR-trained predictor has mean LPIPS 0.22930, compared with 0.22687 for **GBL-M**. In the separate held-out Ensemble evaluation, **GBL-M** and direct **GBL-P** have mean LPIPS 0.22712 and 0.21981, respectively; their missing-pixel PSNR means remain 14.18 and 14.65 dB. These descriptive comparisons hold candidates and continuation fixed.

For **GBL-P** versus **GBL-M**, worst paired PSNR effects are -3.531 dB in Single-model evaluation and -4.352 dB in Ensemble evaluation. Single-model evaluation changes 86 of 128 selections, averaging 0.530 dB among changes. The other 42 contribute zeros.

Initialization evaluation (splits in Table 5). The predictor from the Ensemble evaluation gains 0.402 dB $[0.056, 0.749]$ with new initializations on 64 of its test images. On a separate 128-image evaluation set with the same initializations as calibration, it gains 0.732 dB $[0.454, 1.024]$. Changing initializations on these same images gives 0.382 dB $[0.133, 0.637]$. All three gains are mean paired PSNR effects.

E AFHQ-Cat Details and Additional Results

The AFHQ experiments test whether the search–selection diagnosis and predictive-selection effects recur across datasets.

Protocol. Section 6.1 specifies datasets and allocations. AFHQ phase retrieval uses the native Fourier-magnitude operator: each 256×256 RGB channel, scaled to $[-1, 1]$, is zero-padded by 128 pixels on each side, yielding a 512×512 measurement grid (oversampling factor 2). The operator returns $\sqrt{|\,\mathrm{FFT}_{\mathrm{ortho}}(x_{\mathrm{pad}})|^2 + 10^{-8}}$; independent Gaussian noise with standard deviation 0.05 is added on this scale, without rescaling the image to $[0, 1]$ or normalizing the magnitudes.

Both AFHQ tasks use the same 64 development and 128 evaluation images, separated at image level. These image sets reuse images from earlier AFHQ experiments. Each task fits a new single HGBR on 512 candidate examples, using the features and hyperparameters in Appendices B.2 and B.3. Selection uses the fitted task-specific predictor directly.

The AFHQ allocation means and predictive-selection effects appear in Tables 2 and 3. As an additional phase-retrieval reference, native FLOWER (100 steps) has mean PSNR 14.749 dB and LPIPS 0.4742. Direct **GBL-P** has phase-retrieval LPIPS 0.2829. Table 20 gives the central and irregular inpainting controls.

For central inpainting, the mean best-candidate PSNR is 16.231 dB. Together with Table 20, this gives the headroom and selection-loss decomposition in Section 6.3. Differences use unrounded per-image scores. Best-candidate quality is a target-dependent diagnostic.

Computation. The 548-step single path is a separate run sharing GBL's first initialization. Model-call and correction counts follow Appendix B.4. Full candidate banks were completed for counterfactual analysis. On one development image per setting, stopping discarded trajectories during inference produced exactly the same **GBL-M** output as selection from fully completed candidates.

Uncertainty and failure definitions. In phase retrieval, **GBL-M** has a positive observed mean advantage over equal-work refinement on AFHQ, but its adjusted interval spans zero (Table 8). Both CelebA and AFHQ phase comparisons have negative paired medians despite positive mean effects. The four primary PSNR contrasts compare **GBL-M** with single-path refinement and direct **GBL-P** with **GBL-M** in each task. We use 20,000 paired image-bootstrap resamples. Individual 98.75% intervals are Bonferroni-adjusted for a nominal 95% familywise level, conditional on the fitted predictors, cohort, and random streams. Resampling varies evaluation images while holding predictor fitting and the evaluation protocol fixed. For secondary LPIPS effects of **GBL-P** minus **GBL-M**, pointwise 95% intervals are $[-0.00679, 0.01724]$ (phase) and $[-0.00621, 0.00653]$ (central). Both intervals span zero.

Failure thresholds are development tenth percentiles of native 100-step PSNR: 6.711579 dB (phase) and 12.336281 dB (central). Both selectors use the same candidates, fixing missing-candidate failures at one per task. Discard failures change from seven to six in phase and from 17 to eight in central under predictive selection. Table 19 reports individual PSNR effects over the full evaluation sets.

Table 19: **Mean predictive-selection gains coexist with individual losses on AFHQ-Cat.** Effects are full-image PSNR of **GBL-P** minus **GBL-M**. Counts use the sign of the unrounded paired effect over all 128 images per task. Medians include zero effects. Worst denotes the smallest observed effect. Paired-effect tenth percentiles appear in Table 3.

	Image counts			Paired PSNR effect (dB)	
Setting	Improved	Degraded	Unchanged	Median	Worst
Phase / FLOWER	16	14	98	0.000	−19.276
Central / FlowDPS	70	38	20	+0.226	−3.936
Irregular / FlowDPS	76	33	19	0.458	−4.063

Illustrations. After evaluation, we chose the central-inpainting example in Figure 4 to illustrate increased PSNR under predictive selection. The phase-retrieval and central-inpainting examples in Figure 9 were each selected nearest the tenth percentile among negative selection effects, with deterministic tie-breaking. The full-cohort tenth percentiles appear in Table 3. Observations retain measurement noise. Central observations use range mapping and clipping for display, while Fourier observations use $\log(1 + |y|)$ and min–max normalization. These transforms do not enter metrics or selection.

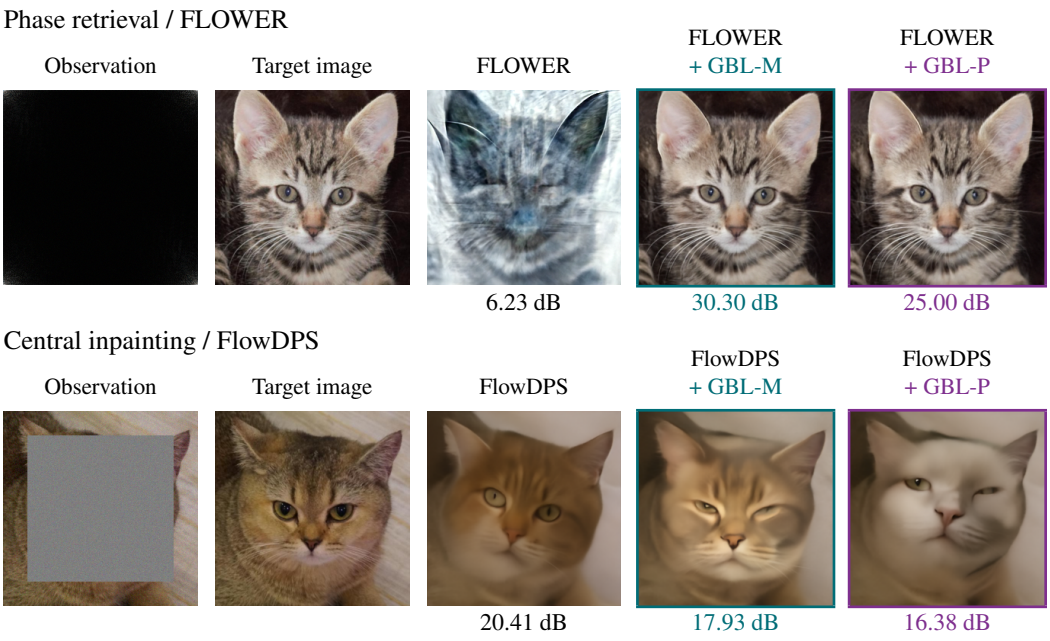

Figure 9: **GBL-P replaces higher-PSNR measurement choices in both AFHQ examples.** With candidates and continuation fixed, direct **GBL-P** loses 5.30 dB relative to **GBL-M** in phase retrieval (top) and 1.55 dB in central inpainting (bottom). For each task, the example is nearest the tenth percentile among images with negative **GBL-P** minus **GBL-M** PSNR effects. Labels give full-image PSNR. Native solvers use 100 steps and GBL uses 548 outer updates.

E.1 Matched-area central and irregular masking

This paired geometry comparison uses the same 64 development and 128 evaluation images, with disjoint training and test sets. The irregular study was designed after the central results were known,

using these previously evaluated images. Both geometries share the prior, resolution, initialization and trajectory random draws, and FlowDPS settings. Each mask removes exactly $36{,}864/65{,}536$ pixels; the irregular mask uses three content-independent polylines with distance-ranked dilation. Each geometry uses direct selection from its own single HGBR, fitted on development images with the settings in Appendix B.2, then fixed before evaluation.

Table 20: **Matched-area AFHQ-Cat central and irregular masking.** FlowDPS on the same 128 evaluation images; both masks remove 56.25%. PSNR is full-image dB. The PSNR tenth percentile describes reconstructions, not paired effects. Failure counts use geometry-specific development thresholds. Completed multistart uses five completed paths ($W = 500$); GBL uses $W = 548$.

Method	W	PSNR	LPIPS	10th pct.	Missing / discard / total
Central					
Native solver (100 steps)	100	14.356	0.5018	11.573	26 / 0 / 26
Single-path refinement	548	13.870	0.5407	11.305	30 / 0 / 30
Completed multistart	500	14.610	0.4896	11.745	1 / 17 / 18
GBL-M	548	14.461	0.4979	11.870	1 / 17 / 18
GBL-P, direct	548	14.864	0.4981	12.608	1 / 8 / 9
Irregular					
Native solver (100 steps)	100	14.968	0.4763	12.112	18 / 0 / 18
Single-path refinement	548	15.217	0.4935	11.996	17 / 0 / 17
Completed multistart	500	15.473	0.4601	12.713	0 / 8 / 8
GBL-M	548	15.112	0.4674	12.511	0 / 11 / 11
GBL-P, direct	548	15.798	0.4639	13.014	0 / 6 / 6

Predictive selection preserves enough candidate advantage to beat refinement under irregular masking. Best-candidate headroom is 1.792 dB, while selection loss falls from 1.897 to 1.211 dB. Discard failures fall from 11 to 6, with no missing-candidate failures (Table 20). Failure thresholds are development tenth percentiles of native 100-step PSNR: 12.351467 dB for irregular masks and 12.336281 dB for central masks.

Table 21: **Predictive selection exceeds refinement in AFHQ irregular inpainting.** These contrasts supplement the predictive-minus-measurement PSNR result in Table 3. †: a 98.75% interval using the same four-comparison adjustment as the central-mask study. Other intervals are pointwise 95%. Positive PSNR and negative LPIPS differences favor the first method.

Contrast	Metric	Mean	Interval
GBL-M − long path	PSNR	-0.105	[-0.527, 0.320]†
GBL-M − long path	LPIPS	-0.02601	[-0.03397, -0.01784]
GBL-P − **GBL-M**	LPIPS	-0.00350	[-0.00994, 0.00287]
GBL-M − multistart	PSNR	-0.361	[-0.693, -0.027]
GBL-M − multistart	LPIPS	0.00734	[0.00049, 0.01425]
GBL-P − long path	PSNR	0.582	[0.255, 0.907]
GBL-P − long path	LPIPS	-0.02951	[-0.03725, -0.02163]

Paired intervals use 20,000 image-bootstrap resamples, holding predictors and random streams fixed; fitting and protocol-selection uncertainty are excluded. Measurement selection has unresolved PSNR ordering against refinement and trails completed multistart. The LPIPS interval in Table 21 leaves perceptual recovery uncertain. Multistart has the lowest mean LPIPS. Table 19 reports individual losses. The irregular-minus-central difference in predictive PSNR gain is also unresolved: 0.283 dB $[-0.078, 0.642]$. For measurement selection versus refinement, the difference is -0.695 dB $[-1.054, -0.338]$. The corresponding LPIPS differences are -0.00365 $[-0.01222, 0.00492]$ and 0.01674 $[0.00721, 0.02625]$, respectively. These secondary comparisons use pointwise 95% intervals.

Model-call and local-correction counts follow Appendix B.4. Completing discarded candidates adds offline evaluation work.

F PROSPECTIVE EVALUATION ON POINT-DECIMATION SUPER-RESOLUTION

This appendix details the prospective evaluation in Section 6.5, including the separated cohorts, forecast calculation, and additional outcomes.

F.1 DATA SEPARATION AND FIXED PROTOCOL

All images come from the official CelebA test split and were absent from the earlier experimental image sets checked for this evaluation. After removing duplicate images, a partition fixed without inspecting image content or reconstruction outcomes assigns disjoint cohorts for diagnosis (64 images, A), predictor training (128), recovery estimation (64, B), and final evaluation (128, C). Images used for preliminary checks are excluded. Separation is by image content, not person identity. Use of these images during development of the pretrained prior has not been independently verified.

The task reconstructs 128×128 RGB images from 32×32 observations by $4\times$ point decimation: $h(x)_{c,i,j} = x_{c,4i,4j}$, using zero-based spatial indices. There is no blur or interpolation kernel. Gaussian noise has standard deviation 0.05 on the model's $[-1, 1]$ scale. Preprocessing uses a central 178-pixel crop resized to 128×128 and the existing CelebA optimal-transport prior. FlowDPS keeps its fixed update grid, three correction iterations, and correction step size 15. The candidate protocol is $T = 100, N = 8, m = 64, k = 1$, hence $W = 548$; selection occurs after exactly 64 completed outer updates. Trajectory features use that state and the state after 48 updates, with no later information. The time feature records the start of the last completed update, $0.63 = 63/100$. Selection follows that update at $t = 64/100 = 0.64$. The long path uses 548 updates over the same time interval and shares the first candidate's initialization; completed multistart selects among five 100-step paths.

Predictive selection uses the 13 features in Appendix B.3 and five HGBRs with the hyperparameters in Appendix B.2. Each model excludes one image-level fold of the 128 training images. The same five fitted models are averaged for direct selection on both the recovery and final cohorts. PSNR uses the full image after affine conversion to $[0, 1]$, without clipping. LPIPS follows the full-image protocol in Appendix B.1.

F.2 DIAGNOSIS AND FORECAST BEFORE FINAL EVALUATION

Image assignment and the protocol were fixed first, followed in order by diagnosis, predictor fitting, independent recovery estimation, and the forecast. The forecast was fixed before any final-cohort reconstruction began. Final evaluation used the same predictor ensemble as recovery estimation. The final cohort was evaluated once, without outcome-dependent changes.

Let $\widehat{H}_A$ be the mean best-candidate advantage over refinement on the diagnosis cohort, and $\widehat{L}_{M,A}$ its measurement-selection loss. The forecast applies Equation (4) as $\widehat{\Delta}_P = (\widehat{H}_A - \widehat{L}_{M,A}) + \widehat{R}_B$, where $\widehat{R}_B$ is the fixed predictive selector's independent mean recovery. Table 4 reports the component means and forecast. Table 22 gives their uncertainty.

The forecast interval uses 20,000 bootstrap draws, independently resampling diagnosis and recovery images and adding their mean effects. It quantifies uncertainty in the development estimate conditional on the fitted ensemble and protocol, rather than serving as a prediction interval for a future cohort. Held-out evaluation tests whether the forecast transfers.

Table 22: **The prospective headroom deficit persists on held-out images.** Pointwise 95% intervals (dB) for the means in Table 4. Development intervals use the diagnosis cohort, the independent recovery cohort for **GBL-P** minus **GBL-M**, and both cohorts for the forecast.

Quantity	Development / forecast	Held-out evaluation
Best-candidate advantage	[-1.358, -0.732]	[-1.231, -0.739]
GBL-M − refinement	[-1.538, -0.854]	[-1.353, -0.846]
GBL-P − **GBL-M**	[-0.159, -0.042]	[-0.112, 0.001]
GBL-P − refinement	[-1.640, -0.952]	[-1.391, -0.917]

For the two primary held-out contrasts, Bonferroni-adjusted 97.5% intervals are $[-0.122, 0.008]$ for **GBL-P** minus **GBL-M** and $[-1.424, -0.884]$ for **GBL-P** minus refinement, giving nominal 95% familywise coverage.

F.3 HELD-OUT ALLOCATION AND INDIVIDUAL EFFECTS

Table 23 gives the absolute reconstruction metrics. The forecast has 0.148 dB absolute error. The measurement-selection gap changes by 0.102 dB from development to test, and predictive recovery by 0.046 dB.

Table 23: **Longer refinement remains best in the prospective setting.** Means on 128 held-out CelebA images under $4\times$ point decimation. W counts outer updates. Best of eight uses target PSNR; its LPIPS is measured on that same candidate.

Method	W	PSNR ↑	LPIPS ↓
Native solver (100 steps)	100	19.556	0.2631
Single-path refinement	548	**21.874**	**0.2200**
Completed multistart	500	20.559	0.2340
GBL-M	548	20.774	0.2305
GBL-P, direct	548	20.722	0.2323
Best of eight (diagnostic)	800	20.888	0.2288

Final intervals use 20,000 paired image-bootstrap resamples. Measurement-selection loss is 0.147 dB $[0.058, 0.263]$ on development images and 0.114 dB $[0.072, 0.161]$ on held-out images. Held-out predictive-selection loss is 0.166 dB $[0.112, 0.228]$. Completed multistart minus refinement is -1.315 dB with 95% interval $[-1.566, -1.064]$. LPIPS effects relative to refinement are 0.01050 $[0.00057, 0.02082]$ for measurement selection, 0.01235 $[0.00296, 0.02214]$ for predictive selection, and 0.01398 $[0.00450, 0.02396]$ for completed multistart; positive values are worse. Predictive minus measurement LPIPS is 0.00185 $[-0.00091, 0.00473]$. These are pointwise 95% intervals.

The failure partition here compares each image with its own equal-work refinement output. On 101 of 128 images, even the best candidate does not exceed that reference. Among the remaining 27, measurement selection fails to preserve an advantage on 4 and predictive selection on 6, leaving 23 and 21 improved outputs. This image-specific reference differs from the development quality thresholds used in the earlier failure analyses.

Predictive selection changes 32 choices: 11 improve PSNR and 21 worsen it; 96 are unchanged. Across all 128 images, the predictive-minus-measurement median is 0, the tenth percentile is -0.279 dB, and the worst observed effect is -1.998 dB. Relative to refinement, the predictive median, tenth percentile, and worst effect are -1.202, -2.689, and -5.236 dB.

F.4 COMPUTATION

Each GBL reconstruction uses $8 \times 64 + 36 = 548$ outer updates. Offline evaluation completes all eight candidates (800 updates) and the long-path reference (548 updates). Per image, this requires 1,348 model forward evaluations and 4,044 observation-operator vector–Jacobian products (VJPs), with no model VJPs. Computing trajectory features adds 45 operator forward evaluations and no corrections. CPU prediction averages about 7.1 ms/image. These costs include completed discarded candidates and the reference path. Prospective diagnosis also requires completed development candidates and predictor training.