LEARNING WHAT STRUCTURE TO PRESERVE FOR SOURCE-FAITHFUL GENERATIVE IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative super-resolution can synthesize plausible details that do not match the source. We introduce MetaSR, a source-assisted framework that learns how to use structural metadata and which structure to transmit under a budget. The receiver encodes the LR image and edge map with a shared VAE and fuses aligned tokens through native transformer attention. We define structural utility as the change in whole-image reconstruction quality produced by an assignment, relative to MetaSR w/o edge with the same trained receiver. A lightweight sender-side predictor learns reusable region-action scores from these counterfactual outcomes. Combining the scores with measured coding costs enables budgeted selection without online receiver queries, and changing the budget requires no retraining or recalibration. Across three $\times 4$ benchmarks, MetaSR improves PSNR by 0.47–1.60 dB over MetaSR w/o edge at matched aggregate transmission rates. On a held-out evaluation set with the receiver fixed, reusing the learned scores at five unseen budgets outperforms random allocation by 0.50 dB on average under shared metadata caps. These results show that receiver-dependent utility can guide which source structure to preserve for source-faithful reconstruction.

1 INTRODUCTION

Generative image super-resolution can recover sharp textures and coherent boundaries from severely degraded observations. Yet many high-resolution images can explain the same low-resolution input, and perceptually convincing detail need not correspond to the particular source. Perceptual plausibility and source fidelity are therefore distinct objectives (Blau & Michaeli, 2018; Wang et al., 2024a). The Rich Detail Range perspective similarly emphasizes retaining source-relevant detail in AI-enhanced visual systems (Wang, 2026).

We introduce MetaSR for a source-assisted setting: the sender observes the HR source and transmits compact structural metadata alongside a degraded observation. Our receiver integrates aligned image and structure tokens through shared-VAE encoding and native transformer attention to preserve thin structures, repeated contours and local geometry. After training the receiver to use source structure, we fix it and learn which evidence the sender should preserve under a budget.

Because the sender observes the HR source, direct HR coding is a valid system-level alternative. Appendix I presents an exploratory H.264 comparison that characterizes the rate-dependent trade-off for the evaluated configurations.

Visually prominent structure is not necessarily useful for reconstruction. In our allocation experiments, source-saliency heuristics can reduce fidelity, whereas learned selection improves it under the same metadata caps. This motivates learning structural utility from the receiver's reconstruction outcomes.

We therefore learn receiver-dependent counterfactual utility: the change in whole-image reconstruction quality when an assignment's decoded edge map replaces an all-zero edge map for the same source and degraded observation. With the MetaSR receiver fixed, expensive offline outcomes supervise an inexpensive deployment predictor. Because the budget enters selection rather than the predictor, one score table can serve new caps without retraining.

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.

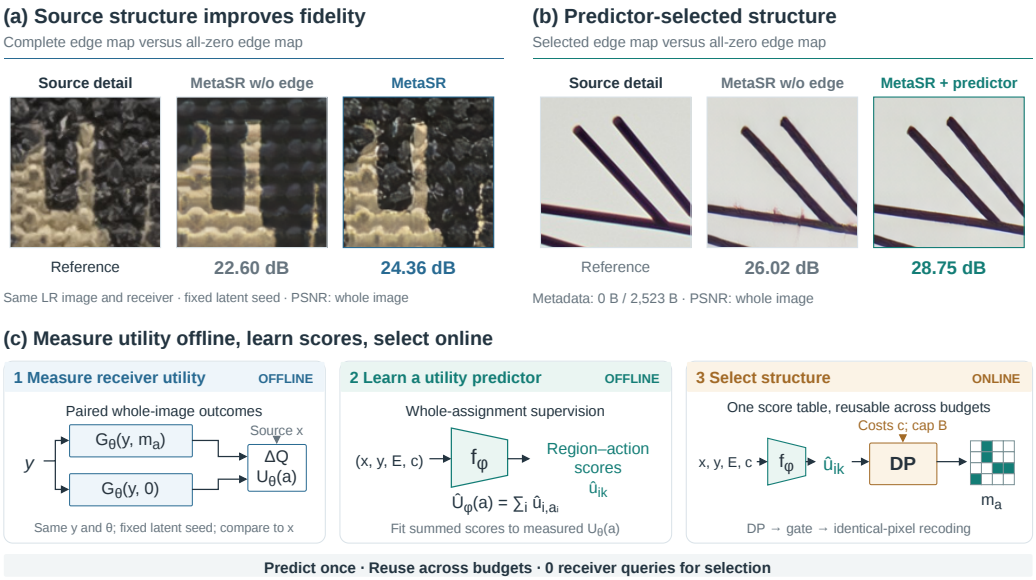

Figure 1: MetaSR learns how to use and preserve source structure. (a) MetaSR versus MetaSR w/o edge, with the same LR image, receiver and fixed latent seed. (b) An example of predictor-selected structure under an additional edge-density constraint, compared with MetaSR w/o edge using the same LR image, receiver and latent seed; metadata rates differ. See Figure 4 for the diagnostic setup. PSNR is whole-image; crops are recorded pixels. (c) Counterfactual gains supervise region-action scores; deployment prediction and budgeted DP selection require no receiver queries.

MetaSR connects structural conditioning with learned decisions about what to transmit. We evaluate source-faithful reconstruction, assignment ranking, allocation against source-only rules and finite receiver search, and reuse at five budgets absent from training. A standard discrete optimizer consumes predicted utilities and actual coded costs. Our contributions are:

- MetaSR, a source-assisted generative SR framework with coded structural metadata and conditioning through a shared VAE and native attention over aligned tokens, making source geometry available to the reconstruction model.
- Receiver-dependent counterfactual utility and an amortized region-action predictor learned from whole-image outcomes, enabling budgeted sender decisions without online receiver queries and reuse of one score table across budgets.
- Evidence for both reconstruction and selection: source-fidelity gains across three SR datasets, conditioning and predictor ablations, decision-useful rankings, a quantified finite-search gap, and reuse at five unseen budgets.

2 Related Work

Generative SR and source fidelity. Image SR has progressed from convolutional, residual and transformer-based reconstruction (Dong et al., 2014; Lim et al., 2017; Liang et al., 2021) to adversarial perceptual methods (Ledig et al., 2017; Wang et al., 2019; 2021) and diffusion restoration (Wang et al., 2024a; Lin et al., 2024; Wu et al., 2024b; Wang et al., 2024b; Wu et al., 2024a; Sun et al., 2025; Zhang et al., 2026). These methods primarily reconstruct from the degraded observation. Our setting additionally lets the sender retain compact evidence from the source, connecting generative restoration to the perception–distortion distinction (Blau & Michaeli, 2018) and source-detail preservation (Wang, 2026).

Structural conditioning. Edge-guided SR (Yang et al., 2017; Kim & Chun, 2018), semantic modulation (Wang et al., 2018), reference transfer (Zhang et al., 2019; Yang et al., 2020), ControlNet (Zhang et al., 2023), and T2I-Adapter (Mou et al., 2024) study ways to inject auxiliary information.

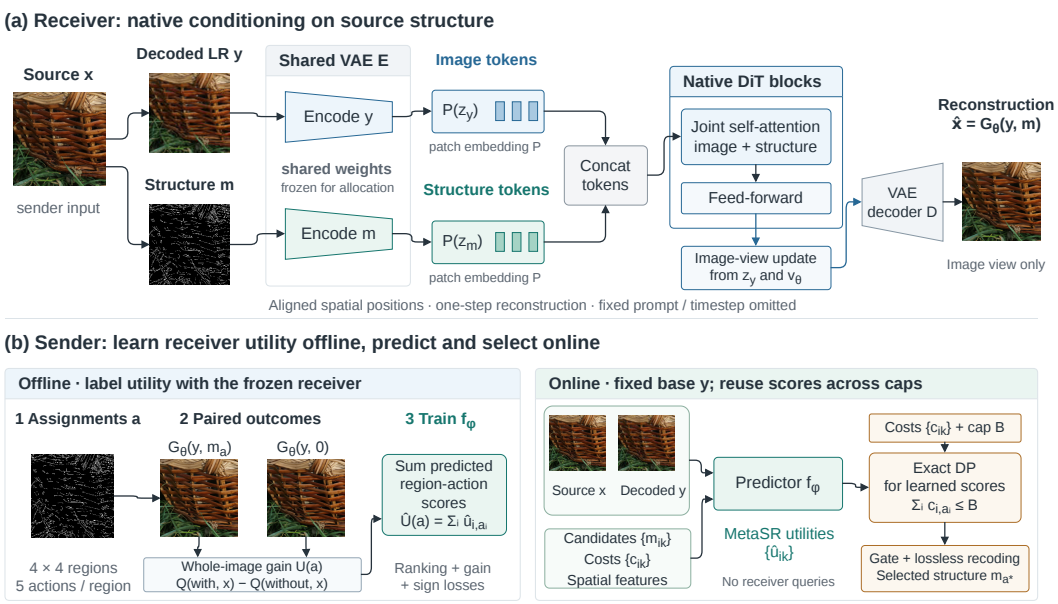

Figure 2: MetaSR receiver and sender. Top: shared VAE encoding, aligned image/structure tokens, native DiT attention and image-view decoding. Bottom: offline whole-assignment receiver gains supervise the predictor; online source observations and candidate costs yield utilities, followed by budgeted DP. Only the coded base, selected metadata and signaling cross the communication link.

Our MetaSR framework couples coded source structure with a receiver that integrates aligned image and structure tokens using a shared VAE and native attention. Its sender learns to select these conditions by their reconstruction value: an available condition may be easy to encode yet provide little benefit.

Learning utility under a budget. Generative compression (Mentzer et al., 2020) and learning for downstream optimization (Elmachtoub & Grigas, 2022) motivate evaluating information by the objective it serves. Here both the base observation and structural payload incur transmission cost. The learned target is the whole-image outcome of a fixed MetaSR receiver, and a standard multiple-choice knapsack solver uses its additive surrogate. With candidate extraction and optimization fixed, we test how learned receiver utility guides allocation compared with handcrafted saliency.

3 MetaSR: Structural Conditioning and Utility Learning

MetaSR couples a structure-conditioned generative receiver with a sender that learns which source evidence to preserve (Figure 2). We train the receiver to use structural metadata, then fix its weights while whole-image outcomes supervise a reusable utility surrogate. Online, predicted utilities and measured coding costs determine the transmitted structure.

3.1 Source-Assisted Setup and Rate Budget

MetaSR represents source structure using edge maps transmitted as metadata. Let x denote the HR source image, y the decoded LR image, and m_a the decoded edge map specified by assignment a. The receiver reconstructs $\hat{x}_a = G_\theta(y, m_a; \xi)$, where ξ is the fixed latent seed for that source. Paired allocation comparisons fix y, θ and ξ. The total transmitted cost, in bits, is

$$R_{\text{total}}(a) = R_{\text{base}}(y) + R_{\text{meta}}(a), \tag{1}$$

including image files, structural payload and signaling. Reported bpp divides this cost by the number of evaluated HR pixels. The metadata cap B is also in bits. Allocation holds the base fixed and constrains metadata cost; full-system comparisons also charge the base stream.

A 512×512 HR source image contains sixteen 128×128 regions. Each has five actions: zero edge, quarter-resolution sparse edges, half-resolution sparse edges, half-resolution dense edges, and full-resolution sparse edges. Canny extraction (Canny, 1986) uses a contextual halo before cropping; decoded edge tiles are expanded and assembled into m_a. Edge density is the fraction of nonzero pixels in a decoded edge map. We define the *full-edge coding cost* as the sum of tile payload costs and 40 signaling bits per region when all sixteen regions use the full-resolution sparse action (Eq. (6)). The candidate family is fixed, so learning changes how available evidence is valued and selected.

3.2 Native Image–Structure Conditioning

Our MetaSR receiver builds on CogVideoX-2B and DOVE-style one-step restoration (Yang et al., 2025; Chen et al., 2025). A shared VAE encodes the upsampled decoded LR image and RGB-replicated decoded edge map. Image encoding uses a posterior sample; structure uses the posterior mode. We concatenate the embedded latent views, assign corresponding image and structure tokens the same spatial positional indices, and process them jointly through native DiT attention. The image-view output alone is converted into a restored latent and decoded. This aligned interface lets the transformer learn how transmitted geometry informs appearance reconstruction (Figure 2).

The native-token receiver reported in Table 2, after the shared 500-update ISR adaptation (fine-tuning for ISR), is frozen and used for all utility labeling, predictor training, and budgeted-allocation experiments. Paired gains measure this receiver's response to different evidence. Appendix A gives the one-step reconstruction equation and training details.

3.3 Whole-Image Counterfactual Utility

For a quality measure Q with higher values indicating better reconstruction, define

$$U_\theta(a \mid x, y) = Q(G_\theta(y, m_a; \xi), x) - Q(G_\theta(y, 0; \xi), x)\,. \tag{2}$$

Here, source fidelity means agreement with the particular HR source instance rather than perceptual plausibility alone. We therefore use the change in whole-image PSNR on limited-range luminance as the allocation utility, while SSIM, LPIPS and DISTS provide complementary measures of reconstruction quality. Each outcome includes structural interactions across regions and depends on the source, decoded observation and fixed MetaSR receiver.

Maximizing receiver-evaluated utility under cap B would require

$$a^*(B) \in \underset{a:\, R_{\text{meta}}(a) \le B}{\arg\max}\; U_\theta(a \mid x, y). \tag{3}$$

Scoring assignments this way requires full receiver reconstructions. We amortize these offline measurements into a sender-side surrogate, retaining measured costs for online selection.

3.4 Amortized Region-Action Utility Prediction

Training proposals cover 25%, 50% and 75% of each source's full-edge coding cost. The fixed MetaSR receiver supplies their whole-assignment labels U_θ. A 76,665-parameter predictor observes the source, decoded LR, their RGB difference, edge-map candidates, spatial/action identities, edge density and measured cost descriptors. A shared local encoder represents each action; a global appearance encoder and mean action context supply source-level information. The head predicts all eighty region-action scores in one pass; the budget is not an input. Zero-edge anchoring defines

$$u_{ik} = f_\phi(x, y, \mathcal{E}, c)_{ik} - f_\phi(x, y, \mathcal{E}, c)_{i0}, \qquad \widehat{U}_\phi(a) = \sum_i u_{i,a_i}. \tag{4}$$

Here $\mathcal{E}$ is the bank of edge-map candidates and $u_{i0} = 0$. The region-action scores are learned jointly from whole-image gains to form an additive decision surrogate.

Within each source/budget group, training combines listwise ranking, regret-weighted pairwise ranking and gain regression, with sign consistency across groups:

$$\mathcal{L} = \sum_b \bar{w}_b(\mathcal{L}^b_{\text{list}} + 0.5\mathcal{L}^b_{\text{rank}} + 0.1\mathcal{L}^b_{\text{gain}}) + 0.2\mathcal{L}_{\text{sign}},$$
$$(\bar{w}_{25}, \bar{w}_{50}, \bar{w}_{75}) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}). \tag{5}$$

Ranking emphasizes decisions among assignments for the same source and observation. Model and gate selection use development data only; Appendix B gives the exact objectives and selection procedure.

3.5 Budgeted Selection with Measured Costs

A lossless JBIG2 payload s_{ik} (International Telecommunication Union, 2000) for region i and action k costs

$$c_{ik} = \begin{cases} 0, & k = 0, \\ 8|s_{ik}| + 40, & k > 0, \end{cases} \qquad R_{\text{meta}}(a) = \sum_{i=1}^{16} c_{i,a_i}. \tag{6}$$

The forty signaling bits identify the region, action and payload length. These are actual coded costs, not predicted rates. For metadata cap B, sparse dynamic programming solves the multiple-choice knapsack problem

$$a^{\text{DP}}(B) \in \underset{a \in \{0,\dots,4\}^{16}}{\arg\max} \sum_i u_{i,a_i} \quad \text{subject to} \quad \sum_i c_{i,a_i} \le B. \tag{7}$$

Dynamic programming exactly optimizes the supplied additive surrogate under the metadata cap. A development-calibrated gate selects the all-zero edge map when predicted gain is at most 0.05 dB. Selected actions can then be replaced by cheaper encodings with identical decoded pixels; saved bits are not reallocated. These steps preserve budget feasibility and the gated receiver input. The same utility table serves multiple caps for a fixed base observation; changing the base requires new predictions. The sender makes no receiver queries during deployment.

We distinguish *MetaSR w/o edge* (an all-zero edge map), *MetaSR* (an edge map encoded without a predictor), and *MetaSR + predictor* (a predictor-selected edge map). MetaSR transmits the complete sparse edge map in the main reconstruction benchmark, using a separate whole-map packet with its own measured header and payload costs (Appendix A).

4 Experiments

We first evaluate MetaSR's reconstruction system and structural conditioning, then test assignment ranking, budgeted allocation, predictor mechanisms and reuse at unseen budgets.

4.1 Evaluation Setup

The predictor is trained on 640 DIV2K sources and selected on 80 separate development sources. All allocation, utility and budget-reuse analyses use the same 80-image held-out evaluation set, with 40 LSDIR (Li et al., 2023) and 40 DIV8K (Gu et al., 2019) sources held out from predictor training and checkpoint selection. Each has a 512×512 HR source image and a fixed degraded 128×128 LR image. Ablation checkpoints use development-only selection; receiver weights, decision rules and selected assignments are fixed before evaluating additional held-out outcomes.

Paired decisions use the same LR image, MetaSR receiver and fixed latent seed within each source. Caps charge actual JBIG2 metadata plus 40 signaling bits per nonzero-mode region, excluding the identical base stream. PSNR uses rounded RGB8, BT.601 limited-range luminance, peak 255 and a four-pixel border crop. Correlated caps and proposals remain grouped by source; confidence intervals use 10,000 source-bootstrap replicates, stratified by dataset for ranking and unseen-budget analyses. Appendix A specifies source roles and metrics.

4.2 Source-Faithful Reconstruction and Conditioning

Table 1 evaluates MetaSR alongside seven generative baselines (Wang et al., 2024a; Lin et al., 2024; Wu et al., 2024b; Wang et al., 2024b; Wu et al., 2024a; Sun et al., 2025; Zhang et al., 2026) on 100 DIV2K validation (Agustsson & Timofte, 2017), 93 DRealSR (Wei et al., 2020) and 100 RealSR (Cai et al., 2019) images. Methods match exact aggregate image, structure and signaling bytes within each dataset, allowing different coded bases and per-image rates. MetaSR here transmits the complete sparse edge map without a predictor. At approximately 0.53 total bpp, it improves over MetaSR w/o edge by 0.47/1.60/1.34 dB, with consistent SSIM, LPIPS and DISTS improvements.

Table 1: Generative SR at $\times 4$ and 0.53 total bpp per dataset. Methods match exact dataset-level bytes, including image, structure, and signaling; per-image rates can differ. MetaSR transmits the complete losslessly repacked sparse edge map without a predictor. Bold/underline: best/second among the displayed methods.

	DIV2K				DRealSR				RealSR			
Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
StableSR [IJCV24]	23.94	0.6031	0.1780	0.1314	26.18	0.5844	0.2849	0.2718	23.55	0.5794	0.2732	0.2477
DiffBIR [ECCV24]	22.69	0.5419	0.2176	0.1550	24.32	0.5387	0.3596	0.2958	21.95	0.5551	0.2954	0.2608
SeeSR [CVPR24]	26.50	0.6869	0.1570	0.1100	28.32	0.7255	0.2178	0.2288	25.37	0.6850	0.2121	0.2213
SinSR [CVPR24]	27.31	0.7066	0.1374	0.1004	28.49	0.6665	0.2677	0.2509	26.23	0.6669	0.2466	0.2394
OSEDiff [NeurIPS24]	24.83	0.6581	0.1606	0.1171	28.01	0.7355	0.1916	0.2161	25.17	0.6883	0.1951	0.2125
PiSA-SR [CVPR25]	25.29	0.6586	0.1455	0.1129	28.43	0.7348	0.1913	0.2158	25.56	0.6996	0.1790	0.2040
TADSR [CVPR26]	24.92	0.6440	0.1678	0.1395	28.51	0.7289	0.2144	0.2408	25.22	0.6717	0.2165	0.2316
MetaSR w/o edge	27.46	0.7309	0.1354	0.0938	28.75	0.7611	0.1728	0.1924	26.29	0.7310	0.1672	0.1880
MetaSR	**27.93**	**0.7537**	**0.1120**	**0.0760**	**30.35**	**0.8060**	**0.1493**	**0.1762**	**27.64**	**0.7877**	**0.1338**	**0.1650**

Table 2: Controlled structural fusion at $\times 4$: a common starting checkpoint, 20,000 updates followed by the same 500-update ISR adaptation. Only the edge-fusion mechanism varies; decoded JPEG-Q40 LR images and complete edge maps are shared. Bold: best among the compared methods.

	DIV2K				DRealSR				RealSR			
Fusion	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Cross-attention	25.73	0.6428	0.1740	0.1093	28.43	0.7213	0.1907	0.1969	25.39	0.6735	0.1941	0.1982
T2I-Adapter-style [AAAI24]	25.82	0.6534	0.1643	0.1040	28.76	0.7430	0.1741	0.1887	25.57	0.6893	0.1774	0.1895
ControlNet-style [ICCV23]	25.85	0.6528	0.1639	0.1093	29.00	0.7470	0.1830	0.1952	25.77	0.6928	0.1839	0.1938
Native token fusion	**26.99**	**0.7083**	**0.1398**	**0.0936**	**30.00**	**0.7975**	**0.1592**	**0.1839**	**27.32**	**0.7678**	**0.1481**	**0.1731**

Table 1 compares complete systems at matched total rates. Fixed-LR comparisons hold the receiver and latent seed constant to assess transmitted structure; Table 2 compares conditioning mechanisms. Its four variants share initialization, data and inputs: decoded JPEG-Q40 LR images and complete edge maps. Training uses 20,000 optimizer updates, followed by 500 additional optimizer updates for ISR adaptation. An update denotes one optimizer step after gradient accumulation; evaluation uses the checkpoints after adaptation. Only the edge-fusion mechanism varies. Here, native token fusion outperforms the evaluated alternatives on all four metrics across the three benchmarks. Against ControlNet-style fusion, the highest-PSNR alternative, its paired PSNR gains are 1.138/0.997/1.550 dB, with 95% source-bootstrap intervals [1.052, 1.222]/[0.746, 1.232]/[1.392, 1.705]. Appendix C gives the protocol, parameter counts and LPIPS intervals.

4.3 Learning Decision-Useful Assignment Rankings

A frozen pool contains 5,016 physically deduplicated nonempty proposals. Within-source/cap statistics average valid caps, then equally weight sources: Spearman 0.708 (95% CI [0.668, 0.745]; 74/80 sources) and pairwise accuracy 78.3% [76.5%, 80.0%]. A fixed-rule pool gives correlation 0.700. Figure 3(a) shows an exploratory predicted-rank quintile trend, from -0.119 to $+0.725$ dB mean measured gain.

Calibration diagnostics give MAE 0.527 dB, bias -0.204 dB, and balanced sign accuracy 66.6%. Appendix F reports full ranking, sign, calibration and coverage diagnostics.

4.4 From Predicted Utility to Budgeted Decisions

Table 3 compares policies at the three training budgets, 25%/50%/75% of full-edge coding cost. Edge density scores the fraction of nonzero candidate edge pixels, and residual-weighted edge scores overlap with the absolute source–upsampled-LR residual. These controls and random utilities use

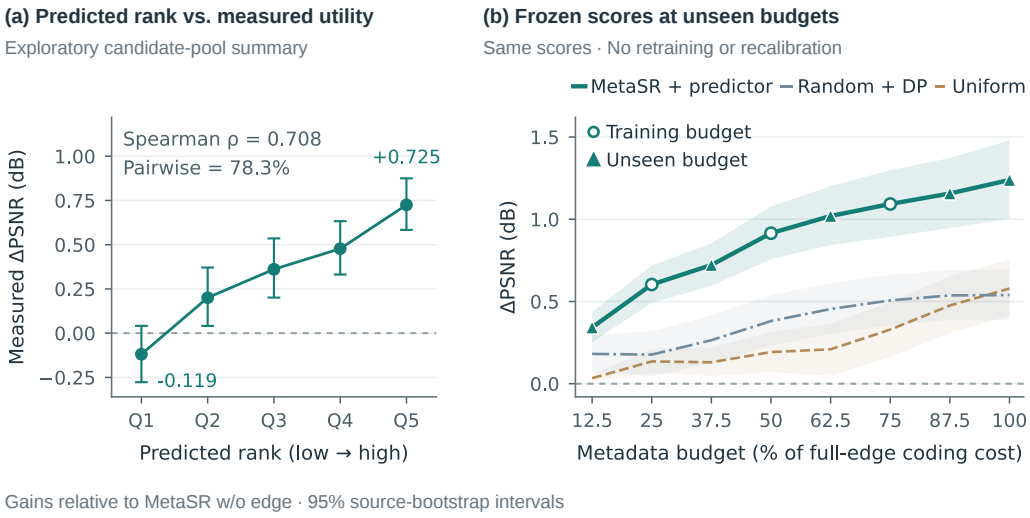

Figure 3: Predicted ranking and frozen-score reuse. (a) Source/cap rank quintiles summarize 5,013 rank-valid proposals (out of 5,016 total); grouped statistics are annotated. This exploratory pool partly contains predictor-generated proposals. (b) The same scores support eight caps without retraining or recalibration: open circles mark training budgets; filled triangles mark unseen budgets. Random averages ten fixed draws. Gains are over MetaSR w/o edge; error bars and bands are pointwise 95% source-bootstrap intervals. Shared caps permit different realized metadata rates.

Table 3: Allocation on the 80-image held-out evaluation set, averaged within source over 25%, 50%, and 75% caps of full-edge coding cost. Gain is measured whole-image PSNR relative to MetaSR w/o edge. Actual metadata bytes include signaling and exclude the fixed LR stream; caps are shared, realized rates differ.

Selection rule	ΔPSNR (dB)↑	Metadata bytes
Residual-weighted edge + DP	-0.175	3,256.5
Edge density + DP	-0.146	3,533.7
MetaSR w/o edge	0.000	0.0
Uniform regional budget	+0.219	2,829.9
Random utility + DP (10 seeds)	+0.355	3,798.9
MetaSR + predictor	**+0.870**	4,048.6
Measured-Pool Reference ($\leq$ 24)	+1.031	3,954.2

the same candidates, costs and DP. Random averages ten fixed seeds. Uniform assigns one-sixteenth of the cap to each region and chooses its first feasible mode in a fixed priority without redistributing leftover bits.

MetaSR + predictor gains 0.870 dB over MetaSR w/o edge, compared with 0.355 for random and 0.219 for uniform allocation. The saliency rules instead give -0.146 and -0.175 dB. This is a comparison of complete policies under shared caps, including their different realized bytes. Disabling the gain gate still gives 0.867 dB; Appendix E reports training-seed variability.

Why whole-image utility rather than structural saliency? Visible structure is not itself a reliable proxy for reconstruction value: under the same candidate family and metadata caps, the saliency rules reduce mean PSNR, while receiver-trained utility yields positive gains (Table 3). This is consistent with Appendix H: single-region insertions have a mean absolute effect of 0.252 dB, but their signed mean is -0.053 dB (95% CI $[-0.196, 0.074]$), with no established positive mean effect. These observations motivate learning utility from whole-image reconstruction outcomes rather than treating local edge strength or source saliency as reconstruction value.

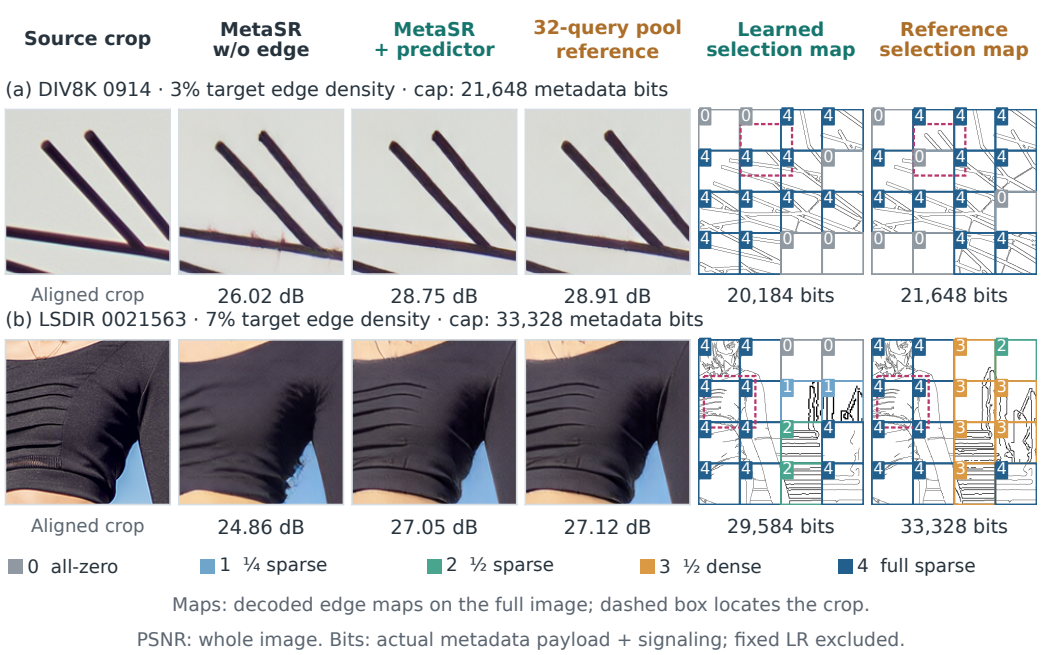

Figure 4: What the sender preserves. Post-hoc fixed-LR diagnostics with target edge densities of 3% (a) and 7% (b), each within ± 0.5 percentage points. Frozen predictor scores select the transmitted edge maps. The 32-query pool reference selects the best feasible reconstruction from a shared 32-query pool under the same edge-density band and metadata cap. Predictor selection uses no receiver queries. Comparisons use the same receiver and fixed latent seed. Crops retain recorded pixels; PSNR is whole-image. Maps show decoded edges/actions; dashed boxes locate crops. Metadata bits include signaling and exclude the shared base; common caps permit unequal spending.

Table 4: Predictor ablations: mean $\pm$ sample SD across three training seeds (first five rows), followed by a paired single-seed control (20260901). Checkpoints are development-selected. Ranking uses the common fixed pool; PSNR gain over MetaSR w/o edge averages all 80 sources and three training budgets. Bytes count metadata and signaling, excluding the fixed LR; shared caps permit unequal spending.

Predictor	Spearman↑	Pairwise (%)↑	Gain (dB)↑	Actual bytes
Full	0.701 ± 0.006	78.05 ± 0.22	0.834 ± 0.032	4187.0 ± 131.2
w/o explicit residual	0.672 ± 0.008	76.76 ± 0.19	0.775 ± 0.022	4320.4 ± 37.5
w/o cost descriptors	0.695 ± 0.030	77.80 ± 1.26	0.807 ± 0.061	4202.4 ± 130.4
w/o global context	0.686 ± 0.022	77.29 ± 0.88	0.764 ± 0.046	4230.5 ± 24.3
w/o sign loss	0.686 ± 0.009	77.21 ± 0.46	0.809 ± 0.011	4297.2 ± 51.1
Full (single seed)	0.708	78.29	0.870	4,048.6
Regression only (single seed)	0.614	73.84	0.599	4,239.2

Gap to finite receiver search. The Measured-Pool Reference selects by actual quality from at most 24 distinct decoded edge maps, including the learned proposal, and reaches 1.031 dB. Adaptive Finite Search allows at most 32 receiver queries and reaches 1.098 dB: its gap to deployment is 0.228 dB (95% source-bootstrap CI [0.156, 0.318]). The learned policy requires zero online receiver queries. Appendix D gives search coverage, other scorer controls and the measured sender workload.

4.5 Predictor Components

Across the three seeds in Table 4, removing explicit residual features or global context reduces allocation gain by 0.060/0.071 dB on average, with the same direction in every seed. The cost-

descriptor effect is smaller and varies across seeds. Sign loss improves average balanced sign accuracy, while its allocation benefit remains uncertain. The separate single-seed regression-only control gives 0.599 dB versus 0.870 dB for the corresponding Full model, a paired advantage of 0.271 dB [0.183, 0.375]. Appendix E reports sign diagnostics and paired source-bootstrap intervals.

4.6 Reusing Scores at Unseen Budgets

The predictor was trained at 25%, 50% and 75% of full-edge coding cost. We reuse its frozen score table, candidate costs, DP and gain gate at eight caps from 12.5% to 100% in 12.5-point steps on the same 80 sources. The five unseen caps include interpolation and settings outside the training-budget range. No retraining, calibration or budget-specific score fitting is performed.

At the five unseen caps, MetaSR + predictor gains 0.896 dB on average over MetaSR w/o edge and outperforms uniform and random allocation by 0.610 dB (95% CI [0.492, 0.735]) and 0.501 dB [0.386, 0.620], respectively. Both overall advantages remain positive at every unseen cap (Figure 3(b)). Because shared caps do not guarantee equal realized rates, we additionally report approximate matched-rate controls in Appendix G.1.

From learned utility to reusable decisions. Score reuse follows from separating utility prediction from budget enforcement. For a fixed source and decoded LR observation, the predictor computes one table of region-action scores. Changing the metadata cap changes which combinations the DP can afford, while preserving the learned scores. Measured coding costs connect this shared utility table to the transmission constraint. Different caps may favor different mixtures of zero-edge, coarse and full-resolution actions: the scores stay fixed while the transmitted structure changes. The unseen-budget sweep therefore tests whether the same learned objective continues to guide useful choices as the feasible set changes.

The ranking and ablation results help interpret this division of roles. Whole-image supervision teaches the predictor to compare assignments by their effect on the fixed MetaSR receiver. In the paired single-seed control, combining ranking, regression and sign objectives gives better assignment ranking and allocation than regression alone. Removing residual features or global context reduces allocation gains across all three seeds, supporting their role in valuing local actions. Together, these results connect the receiver's reconstruction outcomes to learned assignment preferences, and show how those preferences can be reused by budgeted selection.

What structure is selected? Figure 4 uses two sources from the held-out evaluation set to illustrate frozen-predictor decisions under an additional edge-density constraint. Both gain whole-image PSNR with learned metadata, although local artifacts remain. Tile maps expose resolutions, zero-edge regions and differing learned/reference choices. The selections satisfy both the metadata cap and the edge-density constraint.

5 Limitations and Conclusion

Limitations. Our utility predictor is trained for a fixed MetaSR receiver and a fixed family of structural candidates, so changes to either would require re-estimating receiver-dependent utility. The additive surrogate provides an efficient approximation for budgeted selection but does not model arbitrary interactions among regions; finite receiver search measures the remaining empirical gap within the tested candidate support. Finally, the allocation experiments optimize metadata under a fixed base stream, and shared metadata caps do not always produce identical realized rates. Learned structural representations, interaction-aware utilities and joint base–structure allocation are directions for future work.

Conclusion. MetaSR combines coded source structure, native structural conditioning and learned utility to address source-faithful generative SR. Its receiver uses transmitted geometry to improve reconstruction, while whole-image counterfactual outcomes teach the sender which evidence to preserve. The resulting scores support budgeted allocation, approach finite receiver-search performance, and remain useful at unseen budgets without online receiver queries. Compact coarse-to-fine edge coding also supports low-rate operation, with rate-dependent gains (Appendix I).

REPRODUCIBILITY STATEMENT

The appendix specifies receiver adaptation, candidates and transport, data splits, predictor training and selection, and metrics. Supporting records link configurations, decoded edge maps, payloads and reconstructions. We independently verified source identities, selected assignments, transmitted costs, and PSNR values from saved reconstructions. Ablations report three-seed variability and a separate single-seed regression-only control. Utility analyses document deduplication, degenerate groups and source-level uncertainty. Unseen-budget evaluation reuses frozen scores and policies without retraining or recalibration and reports actual bytes and offline receiver workload; deployment selection uses no receiver queries.

AI USE STATEMENT

AI tools assisted with manuscript editing, experimental code implementation and debugging, consistency checks against experimental records, and code-based figure and table preparation. The authors determined the research ideas, experimental design and model architecture. Generative image models were not used to directly synthesize architecture diagrams or edit experimental panels, which show recorded dataset images, decoded structural maps and saved reconstructions. The authors are responsible for the final content and its verification.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017_workshops/w12/html/Agustsson_NTIRE_2017_Challenge_CVPR_2017_paper.html.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Cai_Toward_Real-World_Single_Image_Super-Resolution_A_New_Benchmark_and_a_ICCV_2019_paper.html.

John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851. URL https://doi.org/10.1109/TPAMI.1986.4767851.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 85218–85237. Curran Associates, Inc., 2025. doi: 10.52202/085713-2855. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7b04ec5f2b89d7f601382c422dfe07af-Paper-Conference.pdf.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810. URL https://doi.org/10.1109/TPAMI.2020.3045810.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision – ECCV 2014*, volume 8692 of *Lecture Notes in Computer Science*, pp. 184–199. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10593-2_13. URL https://link.springer.com/chapter/10.1007/978-3-319-10593-2_13.

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

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. DIV8K: DIVerse 8K resolution image dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3512–3516, 2019. doi: 10.1109/ICCVW.2019.00435. URL https://doi.org/10.1109/ICCVW.2019.00435.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

International Telecommunication Union. Information technology—lossy/lossless coding of bi-level images. Technical Report Recommendation T.88 (02/2000), ITU-T, 2000. URL https://handle.itu.int/11.1002/1000/4845.

Kwanyoung Kim and Se Young Chun. SREdgeNet: Edge enhanced single image super resolution using dense edge detection network and feature merge network. *arXiv preprint arXiv:1812.07174*, 2018. doi: 10.48550/arXiv.1812.07174. URL https://arxiv.org/abs/1812.07174.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1775–1787, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/NTIRE/html/Li_LSDIR_A_Large_Scale_Dataset_for_Image_Restoration_CVPRW_2023_paper.html.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844, 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Liang_SwinIR_Image_Restoration_Using_Swin_Transformer_ICCVW_2021_paper.html.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017_workshops/w12/html/Lim_Enhanced_Deep_Residual_CVPR_2017_paper.html.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Computer Vision – ECCV 2024*, volume 15117 of *Lecture Notes in Computer Science*, pp. 430–448. Springer, 2024. doi: 10.1007/978-3-031-73202-7_25. URL https://link.springer.com/chapter/10.1007/978-3-031-73202-7_25.

Fabian Mentzer, George D. Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11913–11924, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8a50bae297807da9e97722a0b3fd8f27-Abstract.html.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226. URL https://ojs.aaai.org/index.php/AAAI/article/view/28226.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Parmar_On_Aliased_Resizing_and_Surprising_Subtleties_in_GAN_Evaluation_CVPR_2022_paper.html.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A Dual-LoRA approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_Pixel-level_and_Semantic-level_Adjustable_Super-resolution_A_Dual-LoRA_Approach_CVPR_2025_paper.html.

Haohong Wang. Rich detail range (RDR): Redefining perceptual picture quality in the era of AI displays. In *2026 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1–2. IEEE, February 2026. doi: 10.1109/ICNC68183.2026.11416982. URL https://doi.org/10.1109/ICNC68183.2026.11416982.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024a. doi: 10.1007/s11263-024-02168-7. URL https://link.springer.com/article/10.1007/s11263-024-02168-7.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 606–615, 2018. doi: 10.1109/CVPR.2018.00070. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Recovering_Realistic_Texture_CVPR_2018_paper.html.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Computer Vision—ECCV 2018 Workshops*, volume 11133 of *Lecture Notes in Computer Science*, pp. 63–79. Springer, 2019. doi: 10.1007/978-3-030-11021-5_5. URL https://link.springer.com/chapter/10.1007/978-3-030-11021-5_5.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Wang_Real-ESRGAN_Training_Real-World_Blind_Super-Resolution_With_Pure_Synthetic_Data_ICCVW_2021_paper.html.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_SinSR_Diffusion-Based_Image_Super-Resolution_in_a_Single_Step_CVPR_2024_paper.html.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13

(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ece.uwaterloo.ca/~z70wang/publications/ssim.html.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision – ECCV 2020*, volume 12353 of *Lecture Notes in Computer Science*, pp. 101–117. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58598-3_7. URL https://link.springer.com/chapter/10.1007/978-3-030-58598-3_7.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92529–92553, 2024a. doi: 10.52202/079017-2938. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8223b0ad64007423ffb308b0dd92298-Abstract-Conference.html.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25456–25467, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_SeeSR_Towards_Semantics-Aware_Real-World_Image_Super-Resolution_CVPR_2024_paper.html.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5791–5800, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yang_Learning_Texture_Transformer_Network_for_Image_Super-Resolution_CVPR_2020_paper.html.

Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep edge guided recurrent residual learning for image super-resolution. *IEEE Transactions on Image Processing*, 26(12):5895–5907, 2017. doi: 10.1109/TIP.2017.2750403. URL https://ieeexplore.ieee.org/document/8030140/.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

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. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Tianyi Zhang, Zheng-Peng Duan, Chun-Le Guo, Peng-Tao Jiang, Bo Li, Ming-Ming Cheng, and Chongyi Li. Time-aware one step diffusion network for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30543–30552, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Time-Aware_One_Step_Diffusion_Network_for_Real-World_Image_Super-Resolution_CVPR_2026_paper.html.

Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7982–7991, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Image_Super-Resolution_by_Neural_Texture_Transfer_CVPR_2019_paper.html.

A Implementation and Data Construction

Native image–structure conditioning. Our MetaSR receiver builds on CogVideoX-2B (Yang et al., 2025) and DOVE's one-step latent–pixel restoration formulation (Chen et al., 2025). A shared VAE encoder $\mathcal{E}_{\mathrm{V}}$ maps the upsampled decoded LR and RGB-replicated decoded edge map $\rho(m_a)$ into latent views. With patch embedding $\mathcal{P}$,

$$z_y = \mathcal{E}_{\mathrm{V}}(\mathrm{Up}(y); \xi), \qquad z_m = \mathcal{E}_{\mathrm{V}}^{\mathrm{mode}}(\rho(m_a)),$$
$$Z = [\mathcal{P}(z_y); \mathcal{P}(z_m)]. \tag{8}$$

Corresponding image and structure tokens share spatial positional indices and interact through native joint DiT attention. Image encoding samples its posterior; structure uses the posterior mode. With image-view velocity v_θ and coefficient $\bar{\alpha}_t$ from the pretrained diffusion schedule, reconstruction is

$$\hat{x} = \mathcal{D}_{\mathrm{V}}\left(\sqrt{\bar{\alpha}_t}\, z_y - \sqrt{1-\bar{\alpha}_t}\, v_\theta(Z; t)\right). \tag{9}$$

Only the restored image view enters the shared VAE decoder. This conditioning architecture provides the fixed MetaSR receiver for the utility-learning experiments.

Receiver architecture and inference. The MetaSR receiver uses CogVideoX-2B's causal VAE and a transformer with 30 layers, 30 attention heads of width 64, spatial patch size two, and 16 latent input/output channels. The transformer has 1,693,783,872 parameters; this count excludes the VAE and text encoder. A fixed empty-prompt embedding and restoration timestep $t = 399$ are used. The decoded LR is bilinearly upsampled to the output grid, normalized to $[-1, 1]$, and encoded with one fixed posterior sample per source. The binary decoded edge map is replicated across RGB channels and encoded with the shared VAE's posterior mode. Spatially corresponding image and structure tokens share positional indices. The implementation restores the batch/view order before selecting the image-view velocity and performing the one-step conversion in Eq. (9).

Inference uses FP16, disabled TF32 and a padded batch of four image crops. Large images use reflect-padded 512×512 image crops assembled and cropped back to the evaluated geometry. MetaSR w/o edge receives an all-zero edge map in the same two-view receiver. For fixed-base allocation, all assignments share the fixed latent seed and image latent. The MetaSR w/o edge RGB8 reconstruction is replayed exactly before evaluating new assignments. When rate matching changes the coded base, its decoded observation is re-encoded with the same fixed latent seed.

Receiver learning. Starting from pretrained CogVideoX-2B, we train the MetaSR receiver with DOVE-style latent–pixel adaptation. Stage 1 freezes the VAE and optimizes the DiT with normalized latent error,

$$\mathcal{L}_{\mathrm{s1}} = \frac{\|\hat{z} - z_x\|_2^2}{\|z_x\|_2^2}. \tag{10}$$

Stage 2 uses reconstruction and perceptual supervision. Image samples use $\|\hat{x}-x\|_2^2 + \lambda_1 \mathcal{L}_{\mathrm{per}}(\hat{x}, x)$. For a video sequence of $n > 1$ frames, an additional temporal term penalizes discrepancies in adjacent-frame changes:

$$\mathcal{L}_{\mathrm{frame}} = \frac{1}{n-1} \sum_{j=2}^{n} \|(\hat{x}^{(j)} - \hat{x}^{(j-1)}) - (x^{(j)} - x^{(j-1)})\|_1. \tag{11}$$

Receiver training uses DIV2K images and HQ-VSR videos, with 2,055 high-quality videos in the latter collection. All evaluations here concern individual images.

The evaluated receiver includes a final 500-update ISR adaptation (fine-tuning for ISR) with a fresh optimizer. Its recorded still-image adaptation uses 744 DIV2K and 2,590 Flickr2K images after exclusions, 512×512 crops, two devices, per-device batch one and two accumulation steps. AdamW uses learning rate 5×10^{-6}, 100 warmup updates followed by a constant schedule, maximum gradient norm 0.1 and seed 42. The transformer is fully fine-tuned; no validation-based selection occurs during these 500 updates. With one frame the temporal term vanishes, leaving mean squared pixel error plus DISTS. Recorded augmentation varies Canny thresholds, smoothing and a legacy patch mask to expose sparse structural conditions; it does not simulate exact JBIG2 budgets. The receiver is then frozen for all utility labeling, predictor training and allocation evaluation.

Structural candidates and transport. Each valid 128×128 region is extracted with a 32-pixel halo, giving a 192×192 context. Actual neighboring pixels are used inside the source image, with reflection at image boundaries. Grayscale context is area-resized before Canny extraction; the valid center is then cropped. Sparse actions use thresholds 100/200 and the dense action uses 50/100, all with L_2 gradients. The coded valid maps are 32×32, 64×64, 64×64, and 128×128 for quarter-sparse, half-sparse, half-dense and full-sparse actions. Nearest-neighbor expansion reconstructs each binary valid region. The zero-edge action has an all-zero map and no payload.

Independent tile payloads use lossless generic/embedded JBIG2. Each nonzero-mode record includes a five-byte header: a 16-bit region identifier, a 3-bit mode identifier and a 21-bit payload-byte length, followed by the payload. The application knows image geometry and stream boundaries. Candidate costs are actual byte counts, not inferred from edge density or PNG file sizes. Gating precedes replacement by the least costly action with identical receiver-visible pixels; the gate uses the original DP score and saved bits are not reassigned.

For the complete-edge benchmark, the full-resolution sparse raster, including padding, is losslessly represented by JBIG2 or a one-bit PNG. Both receive a 24-byte MEG1 header; the smaller complete packet is transmitted and ties favor JBIG2. This is a representation choice for the same complete raster, not predictor allocation. The different packet format is retained when charging the benchmark's total bytes.

Data roles and metrics. The predictor split assigns the original 800 DIV2K training identities to 640 training, 80 development and 80 historical test sources before deriving pairs or structural assignments. Each identity supplies one fixed $512/128$ source/LR pair using a seeded Real-ESRGAN-style two-stage degradation. The held-out evaluation set uses 40 LSDIR and 40 DIV8K sources selected by identity before outcome evaluation, disjoint from auditable local predictor-development and receiver-adaptation identities. Complete upstream pretraining inventories are unavailable, so this disjointness check covers local training records. All allocation, utility and budget-reuse analyses share this evaluation set, held out from predictor training and checkpoint selection.

PSNR uses rounded RGB8, BT.601 limited-range luminance and peak 255. Benchmark border crop equals the nominal scale; allocation and utility analysis use four pixels. SSIM (Wang et al., 2004) averages RGB channels with an 11×11 Gaussian window of standard deviation 1.5. LPIPS (Zhang et al., 2018) uses SqueezeNet v0.1; DISTS (Ding et al., 2022) uses full RGB images in $[0, 1]$ without resizing or border crop. FID (Heusel et al., 2017) uses the PyIQA clean Inception 2048-dimensional feature convention (Parmar et al., 2022) over each complete evaluation set. With 93 or 100 images per group, FID ranks are descriptive finite-sample comparisons (Chong & Forsyth, 2020). DISTS also serves as a receiver-adaptation objective. No-reference metrics retain their original pretrained evaluation configurations.

Scope of uncertainty estimates. Source-bootstrap intervals quantify source-sampling uncertainty for the fixed evaluated models and policies, with repeated caps and proposals grouped by source. The intervals are pointwise. Variation across predictor-training seeds is reported separately; receiver-training variability is outside these estimates.

B Predictor Architecture, Training and Selection

Network. The local action encoder receives ten channels: source RGB, bicubic-upsampled decoded LR RGB, their RGB difference and one binary candidate map. Appearance channels are bilinearly resized to 48×48, while edge maps are resized using nearest-neighbor interpolation. A separate global source encoder receives the nine appearance channels at 64×64. Each encoder has three width-32 3×3 convolutions, strides one/two/two, SiLU activations and GroupNorm in its first two blocks. Average pooling and projections yield 48-dimensional action features and 64-dimensional source features. Action weights are shared across all eighty region-action pairs; global and local encoders have separate weights.

The head combines global content, mean action context, eight-dimensional mode/row/column embeddings and three scalar features: edge density, cost divided by full-edge coding cost, and log-cost divided by log full-edge coding cost. Two width-96 hidden layers produce each score. Zero-edge

subtraction makes every region's zero action exactly zero. The network has 76,665 parameters and receives neither a proposed assignment nor a requested budget as an online input.

Proposals and targets. Each training source has twelve proposals at each of 25%, 50% and 75% of its full-edge coding cost, for 23,040 source–cap–proposal records. Development has 2,880 records. Proposals begin with the all-zero edge map and feasible uniform modes, then add randomized budget-feasible allocations. The offline training proposal mixture includes edge density, overlap with source–LR residual, overlap with the residual between the source and the MetaSR w/o edge reconstruction and an activity indicator. The receiver output in the third component is used only for constructing offline training proposals. Development proposals omit that component. Neither online predictor inputs nor deployment decisions contain a receiver reconstruction or measured quality label. Repeated decoded edge maps can reuse measured outcomes across caps.

Objective. For source s, cap group b and proposal j, let g_{sbj} be measured whole-image gain and p_{sbj} the sum of raw predicted unary values. With $T = 0.1$, let $q = \mathrm{softmax}(g/T)$ and $r = \mathrm{softmax}(p/T)$. The listwise and regression losses are

$$\mathcal{L}^b_{\mathrm{list}} = -\left\langle \sum_j q_{sbj} \log r_{sbj} \right\rangle_s, \qquad \mathcal{L}^b_{\mathrm{gain}} = \langle h_{0.1}(p_{sbj} - g_{sbj}) \rangle_{s,j}, \tag{12}$$

where $h_\beta(e) = e^2/(2\beta)$ for $|e| < \beta$ and $|e| - \beta/2$ otherwise. Pairwise ranking includes unordered proposal pairs with $|\Delta g| \geq \delta = 0.02$ dB, weighting hinge losses by measured regret:

$$\mathcal{L}^b_{\mathrm{rank}} = \left\langle \frac{\sum_{j<\ell:\,|\Delta g_{j\ell}|\geq\delta} |\Delta g_{j\ell}| [\delta - \mathrm{sgn}(\Delta g_{j\ell}) \Delta p_{j\ell}]_+}{\sum_{j<\ell:\,|\Delta g_{j\ell}|\geq\delta} |\Delta g_{j\ell}|} \right\rangle_s . \tag{13}$$

Empty pair groups do not contribute. Sign consistency uses $\log(1 + \exp[-\mathrm{sgn}(g)p/T])$, excludes $|g| < 0.02$ dB and doubles the weight of harmful targets. It is normalized by the sum of included weights. Equation (5) gives the combined weighting. The model is trained for 40 epochs with batch eight sources, AdamW learning rate 3×10^{-4}, weight decay 10^{-4} and gradient-norm clipping one.

Development selection and deployment. Checkpoint selection minimizes mean normalized top-choice regret on the original development proposal pool; zero-spread groups contribute zero and ties favor earlier epochs. Seed 20260901 selects epoch 34; additional seeds 20260902/20260903 select epochs 38/37. Development folds calibrate the decision policy and zero-edge gate before held-out evaluation. The selected threshold is 0.05 dB at each primary cap.

Sparse DP enumerates the feasible cost–utility frontier after each region and prunes dominated states. Integer costs are not quantized, and supplied finite floating-point utilities are accumulated as exact dyadic values. Backtracking selects the maximum-utility feasible assignment; ties prefer lower cost and then lexicographic modes. This guarantee applies to the supplied surrogate. Gating can override that optimizer with the zero-edge assignment, and identical-pixel replacement can lower the transmitted cost without changing its gated reconstruction.

C Structural-Fusion Comparison

Table 2 evaluates MetaSR's structural-conditioning design against three alternative fusion mechanisms under a controlled training protocol, using the same decoded JPEG-Q40 observations and complete edge maps for 100 DIV2K, 100 RealSR and 93 DRealSR images. Native token fusion leads all three datasets in PSNR, SSIM, LPIPS and DISTS. This comparison evaluates the receiver's fusion mechanism; the matched-total-rate benchmark measures the complete transmission system, and the utility experiments evaluate sender selection.

All four fusion variants start from the same pretrained no-edge DOVE checkpoint, built on CogVideoX-2B, and use the same training dataset for 20,000 optimizer updates, followed by 500 additional optimizer updates for ISR adaptation. Each update is one optimizer step after gradient accumulation. Evaluation uses each variant's checkpoint after ISR adaptation. Training data and split, preprocessing and augmentation, optimizer, learning-rate schedule, loss, effective batch size,

Table 5: Shared protocol for the controlled fusion comparison. Only the edge-fusion mechanism varies. Trainable counts include the backbone and fusion modules; frozen VAE and text components are excluded.

Setting	Shared across all four fusion variants
Starting weights	Same pretrained no-edge DOVE checkpoint (CogVideoX-2B backbone), before fusion-specific training
Fusion training	20,000 optimizer updates
ISR adaptation	500 additional optimizer updates; evaluation uses the checkpoint after adaptation
Data and sampling	Same training split, preprocessing, augmentation and effective batch size
Optimization	Same optimizer, learning-rate schedule, loss, weight decay and gradient clipping
Randomness and arithmetic	Same seed policy and training precision
Evaluation inputs	Same decoded JPEG-Q40 LR images and complete edge maps
Inference	Fixed latent seed within each source, FP16, timestep 399 and padded batches of four
Metrics	Same PSNR, SSIM, LPIPS and DISTS implementations
Trainable configuration	**Parameters, excluding VAE/text encoder**
Native token fusion	1,693,783,872 (1693.8M)
Cross-attention	1,693,783,902 (1693.8M)
T2I-Adapter-style [AAAI24]	1,701,177,472 (1701.2M)
ControlNet-style [ICCV23]	2,181,302,848 (2181.3M)

Table 6: Source-paired differences between native token fusion and ControlNet-style fusion, the highest-PSNR alternative in each dataset. Positive PSNR gain and LPIPS reduction favor native fusion. Pointwise 95% intervals use 10,000 within-dataset source bootstrap replicates, conditional on the evaluated models.

Dataset	Sources	PSNR gain (dB) [95% CI]	LPIPS reduction [95% CI]
DIV2K	100	1.138 [1.052, 1.222]	0.0241 [0.0209, 0.0275]
DRealSR	93	0.997 [0.746, 1.232]	0.0238 [0.0190, 0.0287]
RealSR	100	1.550 [1.392, 1.705]	0.0358 [0.0300, 0.0420]

seed policy and training arithmetic are identical across variants; only the edge-fusion mechanism differs. Evaluation fixes the decoded JPEG-Q40 LR observations, complete edge maps, fixed latent seeds and metric implementations. Transformer inference uses FP16, restoration timestep 399 and padded batches of four. Table 5 summarizes the shared protocol and trainable parameter counts.

The adapted native-token checkpoint in Table 2 is the same frozen receiver used for utility labeling, predictor training, allocation and unseen-budget evaluation. The no-edge initialization precedes fusion-specific training; MetaSR w/o edge denotes this adapted native receiver evaluated with an all-zero edge map. Table 6 resamples paired source identities within each dataset 10,000 times, retaining each source's two method outcomes together. ControlNet-style is selected as the comparator because it has the highest reported alternative mean PSNR on all three datasets. These post-hoc intervals are unadjusted for comparator selection. LPIPS reduction is alternative minus native, so positive values favor native fusion.

D Allocation References, Variability and Cost

The common expanded reference pool contains at most 24 distinct decoded edge maps per source/cap. It includes the learned DP proposal, fixed heuristics and randomized proposals. Multiple raw mode assignments can decode to the same edge map; physical deduplication avoids repeated receiver reconstructions while retaining each assignment's neural score. A Measured-Pool Reference uses actual receiver quality to choose from this finite support. Adaptive Finite Search instead allows up to 32 nonzero physical queries per source/cap, starting from feasible choices and expand-

Table 7: All three predictor-training seeds: mean ± sample standard deviation. Ranking uses the common physical nonempty support; balanced sign accuracy includes both beneficial and harmful assignments. Allocation retains all 80 sources and three training budgets. Actual bytes include signaling and exclude the fixed LR; shared caps permit different spending.

Predictor	Spearman	Pairwise (%)	Bal. sign (%)	Gain (dB)	Bytes
Full	0.701 ± 0.006	78.05 ± 0.22	64.49 ± 6.37	0.834 ± 0.032	4187.0 ± 131.2
w/o explicit residual	0.672 ± 0.008	76.76 ± 0.19	60.70 ± 6.05	0.775 ± 0.022	4320.4 ± 37.5
w/o cost descriptors	0.695 ± 0.030	77.80 ± 1.26	65.52 ± 2.07	0.807 ± 0.061	4202.4 ± 130.4
w/o global context	0.686 ± 0.022	77.29 ± 0.88	63.11 ± 2.72	0.764 ± 0.046	4230.5 ± 24.3
w/o sign loss	0.686 ± 0.009	77.21 ± 0.46	59.49 ± 4.15	0.809 ± 0.011	4297.2 ± 51.1

ing measured neighborhoods; cached queries count toward its limit and MetaSR w/o edge is a free baseline. Both references measure attainable quality within their finite evaluated support.

Among 240 searches on the held-out evaluation set, 216 use all 32 queries and 24 exhaust a smaller feasible physical space; fifteen contain no nonzero assignment. Mean gains are 1.031 dB for the measured pool and 1.098 dB for adaptive search, versus 0.870 dB for deployment. Each gap is evaluated against its respective reference under shared metadata caps, which allow different actual transmitted bytes and exclude the common fixed base.

The three independently trained predictors yield 0.870282, 0.809670 and 0.823399 dB on the held-out evaluation set, giving mean 0.834450 and sample standard deviation 0.031782 dB with the receiver fixed. A prior development refit with targeted labels gave 1.0132 dB compared with 1.0289 dB for the initial model and 1.0264 dB for count-matched random labels; the initial model was retained. The fixed-pool comparisons with a whole-assignment scorer do not establish uniform superiority of the unary representation.

Uncertainty and sender workload. Allocation confidence intervals resample source identities 10,000 times, retaining each source's caps together. The CPU benchmark uses sixteen development sources, three repeats per source, four threads and one benchmark warmup at a cap of 50% of the full-edge coding cost. It runs on Intel Xeon Platinum 8468 hardware (two sockets, 96 physical cores total, 192 logical cores, approximately 1007 GiB RAM), Ubuntu 22.04.5, Python 3.11.16 and PyTorch 2.9.1. Model loading is excluded, application tensors and candidates are fresh, and the OS page cache is uncontrolled.

Mean predictor forward time is 21.43 ms, DP compilation/selection 1.71 ms, candidate extraction plus 64 JBIG2 encodes 414.30 ms, and the full file-based edge-allocation path 634.66 ms. The complete path also includes reading/resizing inputs, predictor preparation, equivalence checks and gate/canonicalization. These sender timings exclude base encoding, model loading, network transport and receiver reconstruction. All online receiver-query counts are zero.

E Predictor Mechanism Ablations

Table 4 reports means and sample standard deviations over seeds 20260901/02/03 for the five feature/loss variants, followed by a separate comparison of Full and regression-only at the original primary seed, 20260901. Table 7 additionally reports balanced sign accuracy. Each epoch was selected using the same development-set normalized top-choice regret, with ties favoring the earlier epoch; checkpoints and thresholds were fixed before held-out evaluation. The primary-seed selected epochs are 34/25/40/40/18 for Full/residual/cost/global/sign, respectively; the additional regression-only control selects epoch 37.

Controlled feature and loss removals. All arms retain the nominal 76,665-parameter architecture and head dimensions; feature masks apply during both training and evaluation. The residual ablation zeros explicit source-minus-LR channels in local and global inputs, retaining source and LR images. The cost ablation zeros only normalized and log coded-cost descriptors, retaining edge density, edge-map candidates and the actual external DP costs. The global ablation zeros both global source

context and across-action mean context before the prediction head. The sign ablation removes only the sign-loss term, retaining listwise, pairwise and gain regression terms and low-budget weighting.

Full reuses the three original models. The four feature/sign ablations each train for 40 epochs with the same three seeds, original 640/80 training/development split, whole-assignment labels, initialization and sample-order policies, batch size eight, AdamW learning rate 3×10^{-4}, weight decay 10^{-4} and gradient clipping at one. Evaluation uses the fixed MetaSR receiver and held-out evaluation set. Ranking rescoring retains the common physical nonempty proposal pool and the fixed-rule legacy pool. Allocation fixes the three primary caps, DP, 0.05 dB gate and identical-pixel recoding without re-spending saved bits. All choices are frozen before new receiver outcomes are opened.

Regression-only control. An additional control keeps every input, the 76,665-parameter architecture and the same training and development-selection settings, but retains only whole-assignment gain regression. Its loss is $0.1(2\mathcal{L}_{\mathrm{gain},25} + \mathcal{L}_{\mathrm{gain},50} + \mathcal{L}_{\mathrm{gain},75})/4$, using the original Smooth-L1 transition of 0.1 dB and removing listwise, pairwise and sign terms. This preserves the regression coefficient and low-budget weighting. The protocol fixes the original primary seed before training. Its 480 evaluation decisions pair regression-only and the unchanged primary Full model across 80 sources and three caps.

Regression-only gives common-pool Spearman 0.6143 and pairwise accuracy 73.84%. Full's paired differences are 0.0935 [0.0613, 0.1257] and 4.45 percentage points [3.02, 5.88]. On the fixed-rule pool, regression-only gives 0.5804 correlation and 73.80% pairwise accuracy, versus Full's 0.7001 and 79.17%. Allocation gain is 0.5992 dB versus 0.8703 dB; the paired Full advantage is 0.2710 dB [0.1830, 0.3748]. Mean metadata bytes are 4,239.2 versus 4,048.6, and harmful-assignment fractions are 12.50% versus 5.42%. These are single-seed complete-policy comparisons under shared caps, including gate and spending behavior.

This additional evaluation has 417 distinct decoded edge maps: 226 cached outcomes and 191 new reconstructions. Its 136 padded batch-of-four calls use 544 slots, including 272 MetaSR w/o edge replay slots across 68 sources and 81 padding slots; measured receiver time is 576.6 seconds.

Variation and interpretation. The three-seed Full allocation mean is 0.834 ± 0.032 dB, distinct from the original primary model's 0.870 dB. Averaged across seeds, Full minus residual/global ablation is 0.060/0.071 dB, with paired source-bootstrap 95% intervals [0.018, 0.109]/[0.037, 0.113]; both differences have the same sign in all three seeds. Global context's main-pool correlation difference remains uncertain. The cost difference is 0.027 dB [0.004, 0.056], but reverses in the second seed. Sign loss improves average balanced sign accuracy by 4.99 percentage points [2.68, 7.58]; its allocation difference is 0.026 dB [-0.001, 0.058], and Full and no-sign have the same mean harmful-assignment fraction, 6.81%. The allocation benefit of sign loss therefore remains uncertain. Intervals use 10,000 paired source-bootstrap replicates stratified by dataset.

The original five-arm offline evaluation contains 3,600 decisions and 2,297 distinct decoded edge maps, including 693 cached outcomes and 1,604 new reconstructions. The 494 padded batch-of-four calls use 1,976 slots, including 96 padding slots and 276 MetaSR w/o edge replay slots across 69 sources.

F Direct Utility-Prediction Analysis

The frozen predictor is evaluated on the same 80-image held-out evaluation set used for allocation, without training, calibration, new receiver inference or outcome-dependent proposal generation. The common pool contains 6,136 raw references and 5,256 physical references. Selecting the cheapest original raw representation for each edge map and excluding all-zero edge maps leaves 5,016 primary proposal observations. Scores sum the original raw unary values; canonical modes locate measured labels but do not replace those prediction scores.

Correlations and pairwise accuracy are computed within each source/cap group, averaged over valid caps within a source, then equally across sources. There are 222 valid ranking groups out of 240, covering 74 sources (40 LSDIR and 34 DIV8K) and 5,013 of the 5,016 primary observations. Fifteen groups contain only the all-zero edge map; three singleton groups also have undefined rankings. All remain in coverage. Magnitude and sign measures have 75 contributing sources. Exact measured

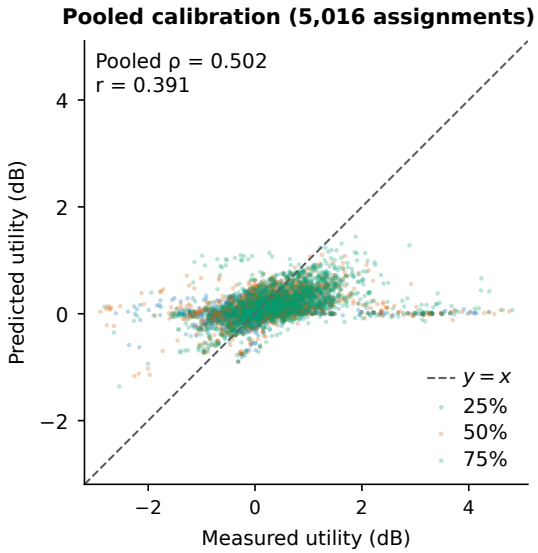

Figure 5: All 5,016 pooled prediction/gain observations for the fixed MetaSR receiver. This calibration diagnostic mixes sources and caps; the grouped ranking and budget-reuse evidence appears in main Figure 3.

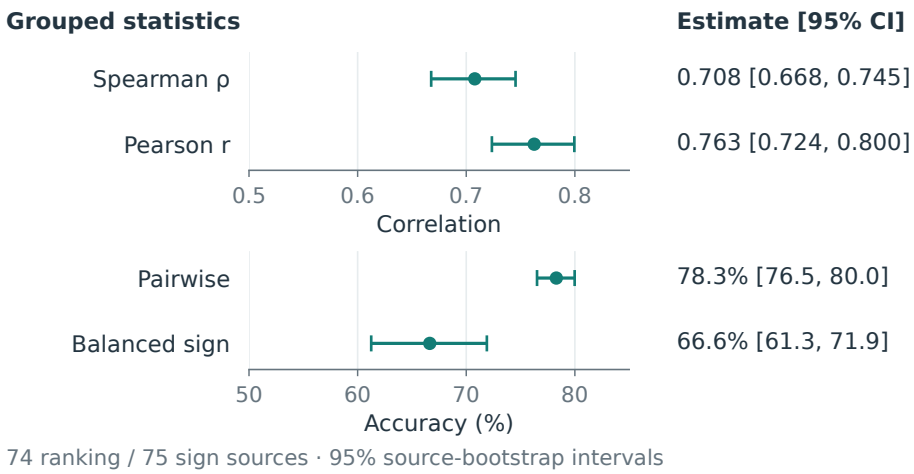

Figure 6: Grouped ranking and sign statistics for the primary predictor (seed 20260901) on the frozen, partly predictor-generated common proposal pool. Bars show 95% percentile intervals from 10,000 dataset-stratified source-bootstrap replicates. Correlations and pairwise accuracy have 74 contributing sources; balanced sign accuracy uses source-balanced confusion cells from 75 sources.

ties are excluded from pairwise comparisons and prediction ties receive half credit. The 54,795 strict proposal pairs remain grouped by source for uncertainty estimation.

Source-stratified percentile bootstrap uses 10,000 replicates, preserving the 40/40 dataset composition and all caps for a source. Mean Spearman is 0.707859 with 95% interval [0.667723, 0.745390], Pearson 0.762556 [0.723611, 0.799516], and pairwise accuracy 0.782922 [0.765171, 0.799799]. MAE is 0.527482 dB, RMSE 0.795828 dB and bias -0.203942 dB. Source-balanced sign accuracy is 0.714928 and balanced sign accuracy 0.666405 [0.612508, 0.718974] (95% source-bootstrap interval; 75 sources). Figure 6 summarizes these grouped statistics and intervals. Harmful proposals are harder to detect: negative recall is approximately 51.8%, compared with positive recall 81.5%.

The rank-binned display in Figure 3(a) is an exploratory re-expression of this frozen, partly predictor-generated common pool, chosen after inspecting existing results. Ascending raw predicted scores within each source/cap define five equal-mass bins; boundary proposals contribute fractional weight, and exact prediction ties share their mean measured gain before splitting (none occur here). Bin means average valid caps within each source, then weight valid sources equally. The ordinate denotes mean whole-image PSNR gain over MetaSR w/o edge. Its pointwise percentile intervals use 10,000 stratified bootstrap replicates (seed 20260922), drawing 40 source IDs per dataset with replacement from the full 80, including undefined curves. All caps and bins travel together; undefined curves are ignored within each replicate mean. The Q1 interval $[-0.277, 0.041]$ dB includes zero. The display summarizes the average within-pool ranking trend; its pointwise intervals do not test adjacent-bin differences or per-image monotonicity. Calibration is assessed separately, and deployment gains are measured by the allocation experiments.

The auxiliary pooled scatter in Figure 5 has Spearman 0.502 and Pearson 0.391, aggregating across sources and caps. A fixed-rule legacy proposal pool gives grouped Spearman 0.7001 for the primary seed; the other two training seeds give 0.6964/0.6993 on the common physical pool. Sign sensitivity excluding true gains inside a prespecified ± 0.01 dB band gives balanced accuracy 64.1%; small predicted magnitudes count as abstentions/errors. Restricting pairwise comparisons to measured gaps above 0.02 dB gives 79.4% accuracy. Together, these diagnostics characterize assignment ranking, gain calibration and sign detection.

G Utility Reuse Across Unseen Budgets

Frozen protocol. The budget sweep uses the 80-image held-out evaluation set (40 LSDIR and 40 DIV8K), with no additional checkpoint selection or calibration. For each source, the edge-map candidates, actual JBIG2 costs, raw unary score table, selected predictor checkpoint, receiver checkpoint and fixed latent seed are frozen. The budget is not a predictor input. Only the DP constraint changes across 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5% and 100% of full-edge coding cost, rounded down to integer bits. The training budgets are 25%, 50% and 75%; unseen 37.5%/62.5% interpolate between them, while 12.5%/87.5%/100% extend outside that range. The original gate and identical-pixel canonicalization remain fixed, with no reallocation of saved bits. Figure 3(b) retains the complete sweep.

Uniform uses a fixed regional share and mode priority without redistributing unused budget. Random assigns ten fixed PCG utility draws per source, reusing each draw across all caps with the same DP and measured costs. Its result is the average of all ten draws, not the best draw. MetaSR w/o edge is the shared zero-gain baseline. These controls share the candidate family and cap with learned allocation, but can choose different bytes and edge maps. Table 8 reports gain and actual transmitted metadata bits; all rates exclude the identical LR stream and include 40 signaling bits per nonzero-mode region.

Aggregation and uncertainty. Per-cap results first average random draws within each source, then weight sources equally. The unseen-budget aggregate first averages the five unseen caps within each source. Percentile intervals use 10,000 source-level bootstrap replicates stratified by dataset, preserving all caps and methods for each sampled source. The unseen-cap gains of learned allocation are 0.342, 0.722, 1.020, 1.157 and 1.239 dB in increasing cap order, averaging 0.895815 dB. Its mean paired advantages are 0.610337 dB over uniform [0.492494, 0.734840] and 0.500630 dB over random [0.386167, 0.620220]. Overall per-cap intervals favor learned allocation over both controls at all five unseen caps. At 12.5% on DIV8K, the domain-specific paired advantage over random is 0.107 dB with interval $[-0.125, 0.326]$, leaving this comparison inconclusive.

Realized rates and failure cases. At 100%, mean metadata cost is 55,079.8 bits for learned allocation, 44,009.8 for uniform and 40,264.5 for random. Figure 7 plots gains against these realized rates. Over unseen caps, learned, uniform and random have mean budget utilization of 76.63%, 42.31% and 75.81%. Their harmful-assignment fractions are 6.25%, 20.50% and 25.27%, while their zero-edge fractions are 11.75%, 30.25% and 6.25%. These fractions average within source and across unseen caps (and random draws where applicable), characterizing both scoring and realized allocation behavior.

Table 8: Frozen MetaSR utility across eight metadata caps on the 80-image held-out evaluation set. Gains are mean whole-image ΔPSNR over MetaSR w/o edge; actual metadata bits include JBIG2 payload and signaling but exclude the fixed LR stream. Random averages ten decisions per source and cap. Training denotes budgets used to generate predictor-supervision proposals. Shared caps do not imply equal realized rates; MetaSR w/o edge has zero gain and zero metadata bits.

		ΔPSNR (dB)↑			Actual metadata bits		
Cap	Budget type	Uniform	Random	MetaSR + predictor	Uniform	Random	MetaSR + predictor
12.5%	Unseen	+0.034	+0.181	**+0.342**	1,604.5	8,988.1	8,411.6
25%	Training	+0.136	+0.177	**+0.603**	8,688.7	18,089.6	16,949.2
37.5%	Unseen	+0.130	+0.264	**+0.722**	16,690.3	26,702.4	25,402.5
50%	Training	+0.192	+0.381	**+0.915**	24,669.1	33,683.0	33,186.4
62.5%	Unseen	+0.209	+0.454	**+1.020**	30,316.1	37,737.4	40,270.9
75%	Training	+0.329	+0.507	**+1.093**	34,560.0	39,400.3	47,030.0
87.5%	Unseen	+0.476	+0.537	**+1.157**	38,850.8	40,028.1	51,773.9
100%	Unseen	+0.579	+0.539	**+1.239**	44,009.8	40,264.5	55,079.8

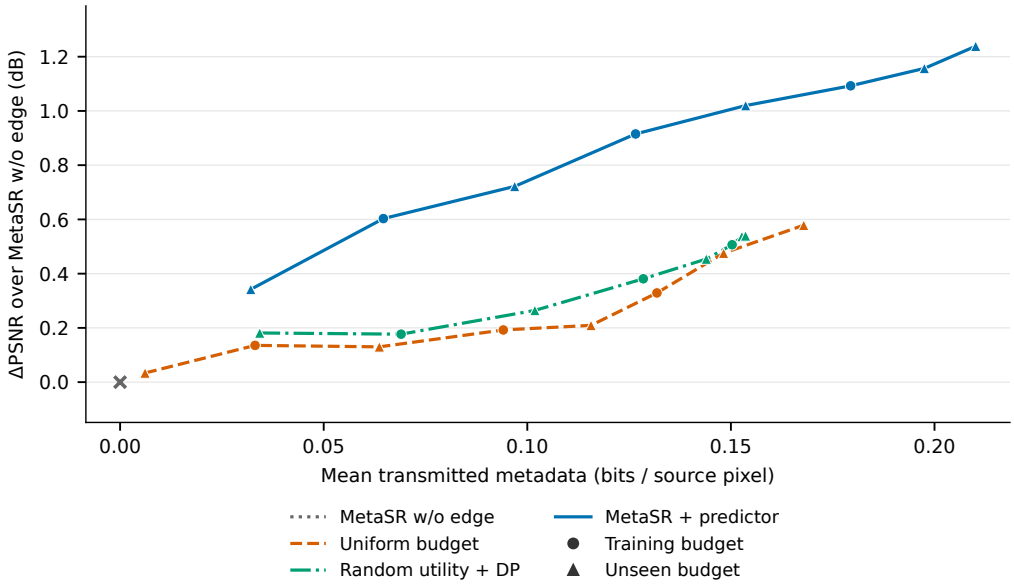

Figure 7: Reconstruction gain versus actual metadata rate for all eight caps. The rate denominator is the evaluated 512×512 HR image, and costs include coded structure plus signaling. The fixed LR stream is excluded. Lines connect measured policy points, without matching their realized rates.

G.1 Post-Hoc Controls at Common Realized Rates

Across five unseen caps, mean actual metadata cost is 36,187.74 bits for learned allocation, 30,744.09 for random and 26,294.30 for uniform. These differing realized rates motivate controls using existing source-level outcomes and measured metadata-plus-signaling costs. The common LR stream is fixed and excluded.

Nearest measured random draw. For each source and unseen cap, select the draw minimizing the absolute bit difference from learned allocation; ties select the smallest fixed random seed. No quality enters selection. Accept a pair when its difference is at most 1% of learned bits; when learned sends zero bits, require an exactly zero-bit match. The matching and aggregation rules were fixed before computing these paired effects, after access to the existing aggregate results and rate

coverage. We report the primary 1% tolerance and both 2%/5% sensitivities without selecting a tolerance by quality.

Selecting the nearest of ten random draws by actual metadata rate alone, with a fixed seed tie-break, matches 192/400 unseen-budget source–cap pairs within 1% (70/80 sources). Averaging eligible caps within source gives a learned advantage of 0.432 dB (95% CI [0.340, 0.537]). Excluding the 25 zero-rate pairs leaves 167 pairs from 65 sources and 0.465 dB [0.372, 0.572]. This post-hoc control uses saved reconstructions and supports approximate rate matching on the covered subset. The 1% subset contains 40 LSDIR and 30 DIV8K sources. The 2% and 5% tolerances cover 216 and 246 of 400 pairs from the same 70 sources, giving 0.454 dB [0.360, 0.560] and 0.460 dB [0.363, 0.569]. At 1%, the mean absolute bit mismatch is 8.89 bytes; positive-rate matches have mean absolute relative mismatch 0.247% and maximum 0.924%. Removing zero-rate pairs leaves 40 LSDIR and 25 DIV8K sources. Each source contributes the mean of its eligible caps; 10,000 paired source-bootstrap replicates preserve each eligible dataset count. This comparator is a rate-selected random draw rather than the ten-draw mean policy. Missing source/cap pairs remain in the coverage denominator, and none are filled by interpolation.

Complementary interpolation control. All eight caps contribute to the following analysis, which is not restricted to the five unseen operating points.

Estimator and support. For each source s and each policy, let the eight observed pairs (r_{sj}, g_{sj}) contain actual bits and whole-image gain. Random first averages cost and gain over all ten fixed draws at a cap, so its curve represents expected cost and expected gain. Repeated identical-rate points have the same gain and are deduplicated. Piecewise-linear interpolation defines $\widetilde{g}_{s,L}(r)$ and $\widetilde{g}_{s,R}(r)$. The common observed interval is

$$\ell_s = \max\{\min_j r_{sj,L}, \min_j r_{sj,R}\}, \qquad h_s = \min\{\max_j r_{sj,L}, \max_j r_{sj,R}\}. \tag{14}$$

For positive-width intervals we compute

$$d_s = \frac{1}{h_s - \ell_s} \int_{\ell_s}^{h_s} \big[\widetilde{g}_{s,L}(r) - \widetilde{g}_{s,R}(r)\big], dr, \qquad \overline{d} = \frac{1}{|\mathcal{S}_+|} \sum_{s \in \mathcal{S}_+} d_s. \tag{15}$$

This weights eligible sources equally regardless of interval width. Interval selection depends only on observed rates; there is no quality-selected envelope or extrapolation. The primary analysis uses the eight observed policy points without adding a zero-rate anchor. Of 80 sources, 72 have positive-width overlap (40 LSDIR, 32 DIV8K), five share only the zero-rate point, and three have disjoint observed intervals. All remain in the coverage report; the latter eight do not enter the integral mean.

Results and sensitivity. The main estimate is $+0.577$ dB with 95% interval [0.477, 0.685]. Sixty-eight eligible sources have positive integrated differences and four have negative differences. Within LSDIR and DIV8K the means are 0.513 and 0.657 dB, respectively. Intervals use 10,000 percentile source-bootstrap replicates (seed 20260923), stratified by dataset while retaining the eligible 40/32 counts. All caps, draws and paired curves for a sampled source remain together.

As a sensitivity analysis, each learned unseen-cap rate is compared with the interpolated random curve only when that rate lies in common support. There are 199 supported source/cap points out of 400, spanning 77 sources. Averaging supported caps within source and then weighting sources equally gives $+0.508$ dB [0.407, 0.616]. The supported-source counts at the five unseen caps, in increasing order, are 42, 74, 41, 23 and 19.

Adding the already measured MetaSR w/o edge point $(0, 0)$ to each curve is a separate sensitivity. The integral estimate becomes $+0.491$ dB [0.404, 0.585] on the same 72 positive-width sources, with eight zero-width cases. Supported unseen-point coverage becomes 254/400 across all 80 sources, giving $+0.449$ dB [0.360, 0.545] under the same within-source-first averaging. Each interval preserves its own eligible dataset counts.

Scope of the controls. These post-hoc estimates compare complete learned and random policies within the reported actual-rate support, using the frozen models and existing draws. Interpolated points are analytical estimates from saved reconstructions, and the primary interval average combines training and unseen budgets. Uniform serves as a policy-efficiency control in the shared-cap experiments.

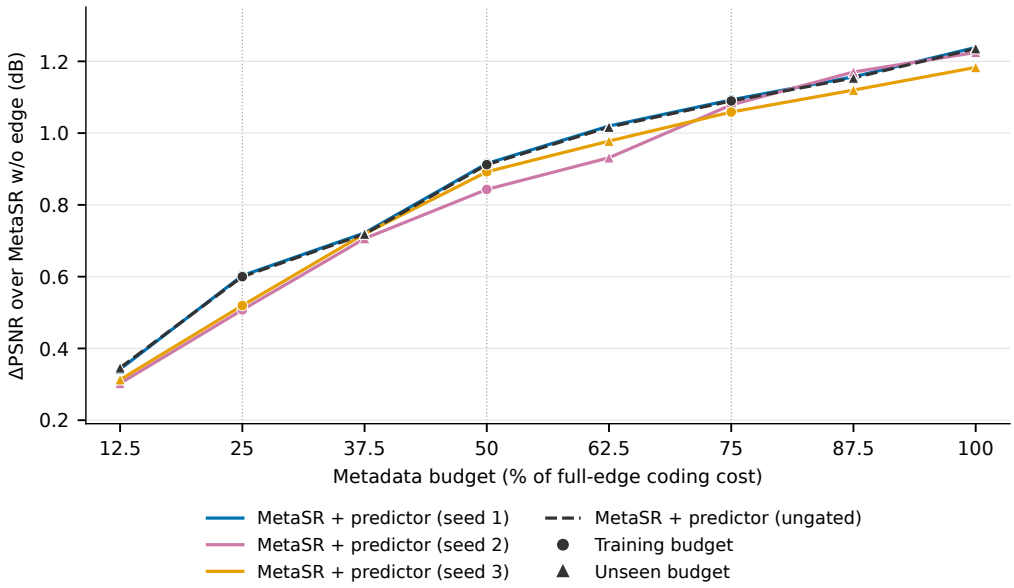

Figure 8: Unseen-budget sensitivity to the three previously trained predictor seeds and to removal of the primary predictor's gate. No seed or checkpoint is reselected using this budget sweep. Points retain all eight caps and use the same fixed MetaSR receiver.

Training-seed and gate sensitivity. Without selecting new checkpoints, the two additional predictor-training seeds give unseen-cap mean gains of 0.867012 and 0.862564 dB, compared with 0.895815 for the originally selected seed. Disabling its gate gives 0.894110 dB. Figure 8 shows the complete curves for these predictor and gate controls with the MetaSR receiver fixed.

Offline evaluation workload. The completed sweep records 10,240 decisions across 80 sources and 640 source/cap groups, with 5,804 distinct source/edge-map pairs, of which 3,587 reuse cached reconstructions and 2,217 require new receiver evaluation. Its 652 padded batch-of-four calls include 2,608 output slots: 2,217 new outputs, 288 MetaSR w/o edge replay slots and 103 padding slots. Recorded evaluation work totals 2,602.154 seconds, including surrounding source processing. These offline measurements evaluate the policies selected from reused predictor score tables.

H Single-Region Structural Interventions in MetaSR

On sixteen sources from the held-out evaluation set selected by identity hash (eight LSDIR and eight DIV8K), 256 assignments each retain one full-resolution sparse region in a 4×4 grid and leave the others with zero edge pixels, using actual lossless JBIG2 payloads. The LR input, MetaSR receiver, fixed latent seed, FP16 and padded-batch protocol remain fixed. Utility is whole-image PSNR gain over MetaSR w/o edge on rounded RGB8 with a four-pixel crop. Mean absolute gain is 0.252 dB (95% source-stratified bootstrap interval [0.159, 0.356]); signed mean gain is -0.053 dB [-0.196, 0.074], with no established positive mean effect. All sixteen sources remain in the 10,000-replicate bootstrap and descriptive aggregates, including three sources with entirely zero-valued candidate edge maps; 73 assignments have empty maps and exactly zero gain. The 29.7% within ± 0.01 dB uses a descriptive threshold. These measurements characterize the receiver's response to single-region interventions at varying payload sizes; predictor training uses separate whole-assignment labels.

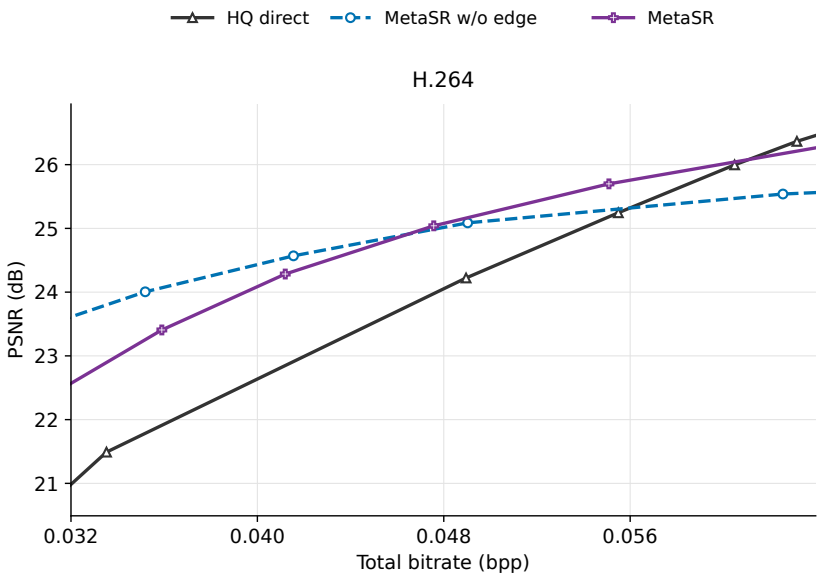

Figure 9: Exploratory H.264 control on 100 RealSR native source–LR pairs. The displayed upper hulls summarize measured configurations aggregated over the evaluation set; segments interpolate between measurements. All curves charge their actual image and structural payloads. MetaSR uses lossy coarse-to-fine edge refinement without a predictor.

I Scope Relative to Direct Image Coding

The comparison uses the same 100 source/native-LR pairs as the RealSR study, a fixed MetaSR receiver and fixed latent seeds. The measured grid contains 13 complete MetaSR configurations, eight MetaSR w/o edge configurations and 33 direct-HR configurations. Each point aggregates all 100 images. This exploratory grid was extended after preliminary outcomes to resolve the low-rate regime.

Applying the same upper-hull rule to all three curves gives an interpolated joint advantage interval of approximately 0.0465–0.0611 bpp. Retaining all Pareto vertices instead gives two intervals, approximately 0.0465–0.0488 and 0.0538–0.0590 bpp. A direct measured comparison is MetaSR at 0.05509 bpp/25.698 dB versus direct HR at 0.05549 bpp/25.245 dB. Low-rate losses to MetaSR w/o edge and eventual direct-HR superiority are retained.

The structural profile here is a lossy coarse-to-fine guide with its actual transmitted refinement cost, distinct from the complete lossless edge raster in the JPEG benchmark. The measured operating regime reflects anchor coding, structural representation and configuration allocation jointly.