SemanticSRJudge: Spatially-Grounded VLM Evaluation for Super-Resolution Quality Assessment

Vishwajeet Shukla [1] **Ankit Dhankhar** [1] **Ajay Bedi** [1]

Abstract

Single-image super-resolution (SR) has advanced rapidly, but its evaluation still relies heavily on scalar metrics such as PSNR, SSIM, and LPIPS. These metrics give useful aggregate signals, but they do not explain *why* one model is better than another, nor do they reveal the localized failure modes that distinguish modern SR architectures. Different SR architectures also fail in qualitatively different ways: some sharpen aggressively at the cost of edge ringing, others preserve content faithfully but flatten fine texture, and diffusion-based models can hallucinate detail that has no support in the reference. Full-image vision-language model (VLM) judges provide richer feedback, but their attention is diluted over the entire image, while many SR errors are small, spatially concentrated, and content-dependent.

We introduce **SemanticSRJudge**, a training-free framework that makes these tradeoffs visible. A frozen DINOv2 backbone identifies regions where an SR output semantically drifts from its reference, and a VLM judge evaluates those regions alongside the full image. This turns a single global judgment into a structured diagnostic across seven perceptual dimensions, revealing where each architecture succeeds, where it fails, and what kind of failure it commits. We also introduce **SemanticSR-Bench**, a content-stratified benchmark covering seven semantic categories, designed to expose model preferences that are hidden at the dataset level.

Across $5{,}072$ matched judge calls spanning four SR architectures, four datasets, and both $2\times$ and $4\times$ scales, SemanticSRJudge consistently corrects the optimistic bias of full-image VLM scoring and recovers content-specific model tradeoffs. In a controlled human study on RealSR Canon $4\times$, DINOv2-guided routing raises mean Win% from 41.9% to 48.6% ($+6.7$pp) and improves mean Spearman correlation with human ratings from $+0.21$ to $+0.31$ ($+0.10$).

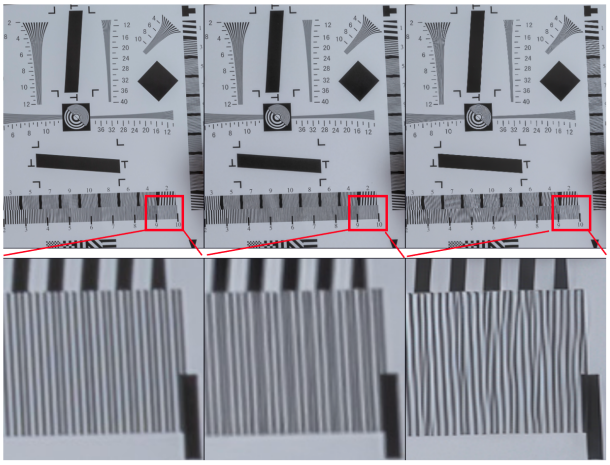

Figure 1. **SemanticSRJudge overview.** A frozen DINOv2 ViT-B/14 drift detector localises regions of maximal SR-induced semantic change between the reference (left) and an SR output (centre/right). The VLM judge (GPT-5.4) scores the full image together with top-K native-resolution crop pairs across seven perceptual dimensions, surfacing local failures (ringing, edge halos, texture hallucination) that are diluted at full-frame scale. We use $K{=}3$ as the production default, selected by a controlled ablation against a human anchor on RealSR Canon $4\times$.

1. Introduction

Super-resolution (SR) is evaluated almost exclusively through single-number metrics. PSNR and SSIM [3] reward pixel fidelity; LPIPS, DISTS, and TOPIQ-FR [4, 5, 6] add learned perceptual distance. None of them attribute a score to a location or to a failure type, and the perception-distortion trade-off [2] forces any scalar to choose between fidelity and naturalness. For a developer the more practical question is different: *where on this image did the model fail, and which class of failure was it?*

Vision-language model (VLM) judges partly close that

[1]Adobe Inc, India. {abedi,dhankhar}@adobe.com, vishwajeet993511@gmail.com. Correspondence to: Ajay Bedi <abedi@adobe.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

gap. Q-Bench [14], DepictQA [15], ExIQA [16], and self-evolving VLM judges [17, 19] elicit per-dimension scores from prompted multimodal models, and Su et al. [18] argue for decomposition along multiple perceptual axes. A vanilla VLM, however, sees the entire image and is dominated by the intact context: ringing halos and hallucinated micro-textures occupy a small fraction of pixels, dilute at full-frame scale, and produce optimistic scores on the texture and artifact dimensions where users actually notice problems [20]. This makes SR a particularly hard setting for generic VLM-based IQA: the global scene is usually plausible, while the deciding evidence lies in small local departures from the reference.

We introduce **SemanticSRJudge**, a training-free framework that routes a VLM to the regions of an SR output where failures concentrate. A frozen DINOv2 [7] ViT-B/14 encodes the reference and the SR image at three depths ($\ell \in \{4, 8, 12\}$), and cosine dissimilarity across those feature spaces yields a per-patch drift map that fuses fine-edge ringing, material hallucination, and structural deformation. The top-K connected components of this map become *native-resolution* crop pairs, which the VLM judges alongside the full image; the output is a structured seven-axis diagnostic rather than a single numeric verdict. Every region-conditioned call has a matched vanilla call on identical inputs minus the routing, so any score difference is attributable to the routing itself. Alongside the framework we introduce **SemanticSR-Bench**, a content-stratified benchmark covering seven semantic categories (faces, animals, scenes, text, patterns, art, indoor) designed so that content-conditional model preferences become visible.

Our main contributions with this paper include:

1. **A region-conditioned VLM judge for SR.** SemanticSRJudge uses a frozen self-supervised backbone (DINOv2 [7] ViT-B/14) as a drift detector that routes a VLM to the regions where SR failures concentrate. Classical spatial detectors are confounded by SR sharpening: over-sharpened edges produce stronger gradients precisely at the worst-reconstructed regions, inverting the detection signal. DINOv2's depth-stratified features [1] instead remain stable under photometric shifts and capture hallucinated grain, deformed structure, and texture substitution in a single drift map. The framework is training-free, requires no SR-specific labels or fine-tuning, and runs on any reference/SR pair with a VLM endpoint.

2. **A seven-axis perceptual schema for SR evaluation.** In place of a single scalar verdict, the judge returns scores on seven complementary dimensions (Upsampling Quality, Texture Preservation, Artifact Score, Unintended Changes, Naturalness, Structural Fidelity, Color Accuracy). Each axis isolates a distinct class of evidence, so an architecture's failure signature shows up as the *shape* of its radar profile, not just an aggregate number; this is what makes per-dimension calibration analysis and content-conditional reversals legible.

3. **Benchmark, protocol, and empirical validation at scale.** We introduce SemanticSR-Bench (seven content categories) together with a matched-pair vanilla/region-conditioned protocol that isolates routing's effect from prompt drift, dataset shift, and sampling noise. Across four SR architectures (Real-ESRGAN+, EDSR, HAT-L, SUPIR), four datasets (DIV2K, RealSR Canon, RealSR Nikon, SemanticSR-Bench), and both $2\times$ and $4\times$ scales (n=5,072 matched judge calls), SemanticSRJudge corrects the optimistic bias of full-image VLM scoring, reveals scale-dependent model preferences, and exposes content-specific reversals that standard scalar metrics miss. A controlled human study, random-patch ablation, and cross-backbone replication validate that the gains come from informed semantic routing rather than crop conditioning alone.

2. The SemanticSRJudge Pipeline

The pipeline has three stages: a DINOv2-based detector that localises the regions of greatest semantic drift, a crop extractor that surfaces those regions at native resolution, and a structured GPT-5.4 judge that scores seven perceptual dimensions on the full image together with the extracted crops. Each stage is described below, with emphasis on the design choices that turn out to matter most.

2.1. Semantic Drift Localisation via DINOv2

Classical spatial detectors such as Canny or Sobel are confounded by SR sharpening: over-sharpened edges produce *stronger* gradient responses at the worst-reconstructed regions, inverting the detection signal. DINOv2 [7] ViT-B/14 (frozen, self-supervised on 142M images) encodes material identity, surface texture, and structural role per patch. Cosine dissimilarity in this feature space picks up hallucinated grain, deformed structure, and texture substitution while remaining stable under photometric shifts. The representation is also depth-stratified [1]: early layers carry fine-grained spatial structure while later layers encode higher-level semantics. Concretely, Layer 4 carries fine-grained edge and high-frequency texture (ringing, fringing). Layer 8 carries material identity (hallucinated grain, incorrect weave). Layer 12 carries semantic context (structural deformations, object hallucinations). Averaging across $\ell \in \{4, 8, 12\}$ keeps the drift map responsive to all three failure classes.

Given $I_r, I_s \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}} \times 3}$, patch feature tensors $\mathbf{f}^\ell \in$

$\mathbb{R}^{P_h \times P_w \times 768}$ ($P_h = \lfloor H_{\text{img}}/14 \rfloor$, $P_w = \lfloor W_{\text{img}}/14 \rfloor$) yield

$$D_\ell(i,j) = 1 - \frac{\mathbf{f}_r^\ell(i,j) \cdot \mathbf{f}_s^\ell(i,j)}{\|\mathbf{f}_r^\ell(i,j)\| \, \|\mathbf{f}_s^\ell(i,j)\|}, \quad \mathcal{H} = \frac{1}{3} \sum_{\ell \in \{4,8,12\}} D_\ell. \tag{1}$$

Eight-connected component analysis on the top quartile of $\mathcal{H}$ selects the top-K regions by mean drift magnitude. A controlled ablation against the RealSR Canon $4\times$ human anchor (§S5) selects $K=3$ as the production default. The procedure is deterministic, training-free, and completes in under 2 s on CPU.

2.2. Region Crop Extraction and VLM Judging

For each top-K drift region we extract a native-resolution crop (at least 128×128, expanded to enclose the full connected component when the drift region spans a larger area) from I_r and I_s (downsampled to match resolution with bicubic downsampler) and stitch the two into a single (Reference | SR) panel. The full SR image is annotated with numbered bounding boxes at the drift-region centres. GPT-5.4 (temperature 0.2) receives, in order, the full reference, the annotated SR image, and the K panels. The system prompt fixes the seven perceptual dimensions and a 1–10 integer scale with explicit anchors. The user prompt instructs the judge to ground its scores in the full-frame impression and to treat the panels as diagnostic evidence for local failures. Full prompt text and image pre-/post-processing are in the supplement (§S1).

2.3. Vanilla Baseline

To isolate the contribution of region conditioning, we run a *vanilla* baseline that differs from SemanticSRJudge in exactly two respects: it presents only the full reference and the full SR output, and it omits the annotated bounding boxes. The DINOv2 step is skipped entirely. The VLM, system prompt, dimension list, scoring scale, decoding parameters, and image preprocessing are otherwise identical. Each vanilla call is paired with a SemanticSRJudge call on the same ⟨stem, model, scale⟩, which enables matched-pair analysis.

3. Experimental Setup

Models. **EDSR-Baseline** [8] (1.4 M parameters, ℓ_1 loss; pixel-objective CNN); **Real-ESRGAN+** [10] (16.7 M; GAN-trained); **HAT-L** [9] (40.8 M; hybrid attention transformer, both $2\times$ and $4\times$ public weights available); and **SUPIR** [11] (a multi-billion-parameter diffusion-based SR model with strong reported $4\times$ performance). All four operate at $2\times$ and $4\times$.

Datasets. (i) **DIV2K** validation (n=100), bicubic synthetic degradation; (ii) **RealSR Canon** (n=50), optically captured camera LR/HR pairs; (iii) **RealSR Nikon** (n=50), the same protocol on a Nikon DSLR; and (iv) **SemanticSR Bench** (n=117), a content-categorised collection covering portraits, landscapes, animals, text, patterns, textures, and indoor scenes. The full matrix is 317 stems $\times$ 4 models $\times$ 2 scales, giving 2,536 image pairs per judge and 5,072 judge calls.

Judges. **SemanticSRJudge** (§2.1–2.2) and the matched **Vanilla** baseline (§2.3). Both call Azure GPT-5.4 at temperature 0.2 with `max_tokens`=2000, run with four-thread parallelism and exponential backoff on rate-limit errors. Across the 5,072 raw calls, 24 (0.47%) returned an API or parse error (8 vanilla, 16 region-conditioned) and are excluded. Matched-pair analyses use the 2,516 pairs that returned a valid response under both judges.

Human study. Three trained annotators independently scored 50 RealSR Canon stems at $4\times$ on all four SR models, rating each of the seven SemanticSRJudge dimensions on the same 1–10 scale via a synchronised five-panel viewer (§S3). We aggregate the three rating sheets by mean per ⟨stem, model, dimension⟩ cell and use that mean as the human anchor; the inter-annotator agreement (mean per-dimension top-1 winner agreement of 79.0% vs. 25% chance, mean Kendall $\tau=+0.211$, ranging from $\tau=+0.696$ on Upsampling Quality to $\tau=-0.214$ on Color Accuracy) is reported in §S3. The aggregated subset (200 paired scores per dimension) serves two roles: it is the anchor for selecting $K=3$ as the production default (§S5), and it is the test set for the matched-pair Win%/correlation comparison in §4.5.

4. Results

4.1. Region Conditioning Calibrates Optimistic Full-Image Scoring

Pairing every vanilla call with the matched SemanticSRJudge call on the same ⟨stem, model, scale, dataset⟩ exposes a systematic, dimension-dependent bias in the full-image judge. Table 2 reports the offset $\Delta = \overline{Vanilla} - \overline{Semantic}$ over all 2,536 matched pairs on the cross-dataset corpus. The offset is dimension-specific (Table 2): it is largest on the locally-supported dimensions (Texture Preservation, Upsampling Quality, Structural Fidelity, all $|\Delta|>0.5$ pts), shrinks to zero on Artifact Score, and reverses sign on Color Accuracy. The ordering of $|\Delta|$ tracks the locality of each failure mode: textures live in localised patches, while precise colour-balance evaluation remains a known limitation of multimodal judges.

Table 1. **SemanticSRJudge scores (↑, 0–10) across four datasets, four models, and two scales.** Each cell is the mean over n stems on the seven SemanticSRJudge dimensions: UQ=Upsampling Quality, TP=Texture Preservation, AS=Artifact Score, UC=Unintended Changes, Nat=Naturalness, SF=Structural Fidelity, CA=Color Accuracy. Bold marks the best model per dataset×scale on each dimension. Per-stem standard errors and the matched vanilla counterparts are in the supplement (§S2).

			SemanticSRJudge (2×)							**SemanticSRJudge** (4×)						
Dataset	**Model**	n	UQ	TP	AS	UC	Nat	SF	CA	UQ	TP	AS	UC	Nat	SF	CA
DIV2K	RealESRGAN+	100	6.73	5.93	7.13	6.66	6.69	7.81	8.27	6.12	5.07	6.72	5.75	5.95	7.12	7.49
	EDSR	100	7.06	6.33	8.03	**8.20**	7.70	8.13	8.83	5.63	4.67	7.58	**7.45**	6.56	7.22	8.26
	HAT-L	100	**7.65**	**7.22**	**8.15**	8.04	**7.81**	**8.60**	**8.94**	6.51	5.66	**7.77**	7.42	**6.96**	**7.74**	**8.54**
	SUPIR	100	6.96	6.18	7.39	6.64	6.77	7.79	8.31	**6.73**	**5.92**	7.21	6.33	6.52	7.65	8.15
Canon	RealESRGAN+	50	**6.12**	**5.26**	5.98	5.46	5.68	6.80	7.30	4.20	3.28	4.56	3.36	3.82	4.66	6.02
	EDSR	50	5.12	4.28	**7.16**	**7.42**	**6.28**	6.72	**7.92**	3.54	2.72	**6.60**	**6.66**	**5.00**	5.16	**6.80**
	HAT-L	50	5.44	4.56	7.12	7.06	6.28	6.78	7.90	3.38	2.60	6.56	6.24	4.66	5.02	6.80
	SUPIR	50	6.04	5.04	5.72	4.78	5.20	**6.86**	7.26	**4.98**	**3.98**	5.36	3.68	4.38	**5.74**	6.08
Nikon	RealESRGAN+	50	6.04	5.02	6.48	5.84	5.86	7.16	8.06	4.50	3.40	5.38	4.24	4.42	5.66	6.60
	EDSR	50	5.42	4.56	7.46	**7.50**	6.54	7.08	8.28	3.82	2.92	**6.76**	**6.72**	**5.28**	5.66	7.28
	HAT-L	50	5.74	4.96	**7.56**	7.28	**6.74**	**7.26**	**8.30**	3.90	2.96	6.74	6.32	5.14	5.64	**7.36**
	SUPIR	50	**6.16**	**5.18**	6.40	5.18	5.68	7.04	7.88	**5.20**	**4.14**	5.67	4.24	4.71	**6.12**	6.82
SemanticSR-Bench	RealESRGAN+	117	6.57	5.78	6.85	6.48	6.55	7.42	8.39	4.84	3.98	5.09	4.38	4.60	5.68	6.74
	EDSR	117	6.79	6.21	6.59	**6.72**	6.76	7.69	8.69	3.41	2.85	2.52	2.55	2.91	4.21	6.27
	HAT-L	117	**7.11**	**6.56**	6.58	6.44	**6.77**	**7.85**	**8.69**	3.56	3.06	2.48	2.57	2.99	4.27	6.42
	SUPIR	117	6.91	6.19	**7.24**	6.38	6.75	7.82	8.42	**5.40**	**4.60**	**5.30**	**4.44**	**4.90**	**5.91**	**7.14**

Table 2. **Vanilla vs. SemanticSRJudge calibration offset** ($\Delta =$ *Vanilla* $-$ *Semantic*, points on the 0–10 scale), over 2,536 matched pairs across four datasets (n=100 DIV2K, 50 each Canon/Nikon, 117 SemanticSR-Bench), four models, and two scales. Region conditioning lowers scores most on the dimensions whose failures concentrate in small regions and barely moves the global ones.

Dimension	Vanilla	SemanticSRJ	Δ
Texture Preservation	5.81	4.94	+0.88
Upsampling Quality	6.36	5.73	+0.63
Structural Fidelity	7.25	6.75	+0.50
Naturalness	6.24	5.80	+0.44
Unintended Changes	6.27	5.89	+0.39
Artifact Score	6.29	6.33	−0.04
Color Accuracy	7.35	7.74	−0.38

4.2. Classical Metric Concordance and Divergence

Table 3 reports the standard classical IQA suite at 4×. The headline pattern is the classical perception-distortion tradeoff [2]: fidelity metrics (PSNR, SSIM, TOPIQ-FR) and perceptual metrics (DISTS, LPIPS) disagree on the model ranking on every dataset, and the fidelity ranking further compresses on the camera-noise regimes (Canon, Nikon) where attention models lose their high-frequency edge. Crucially, the SemanticSRJudge UQ ordering at 4× (SUPIR>RealESRGAN+>EDSR≈HAT) matches the *perceptual* ranking on every dataset, not the fidelity ranking — and it additionally returns six other dimensional scores that no scalar IQA metric provides.

4.3. Diffusion vs. Attention: Scale-Dependent Reversal Across Four Datasets

Table 1 reports the full 4 *models*×2 *scales*×4 *datasets* matrix. The model-preference structure is striking.

At 2×, the Upsampling Quality (UQ) winner varies by content source: HAT-L's hybrid-attention prior leads on synthetic and content-categorised data (DIV2K, SemanticSR-Bench) where a clean LR signal favours attention, while SUPIR's diffusion prior leads on the camera-noise regimes (Canon, Nikon).

At 4×, SUPIR wins UQ on all four datasets, by margins of up to 1.4 points on Canon. HAT-L and EDSR retain narrow leads on Artifact Score and Unintended Changes (their reference-faithful outputs hallucinate less), but their UQ collapses as the pixel ratio expands 16×. The reversal is visible in Figure 2, where HAT-L's UQ-dominant polygon at 2× becomes diffusion-shaped at 4×. The mechanism is architectural: attention and CNN priors remain reference-faithful at scales where the LR signal is recoverable, while diffusion priors synthesise plausible high-scale texture where other models struggle.

4.4. Content Composition Reverses Model Preference within SemanticSR-Bench

SemanticSR-Bench is partitioned into seven content categories (faces/portraits, nature/architecture, animals, text/documents, patterns/textures, art/illustration, indoor) precisely so that content-conditional model preferences can be detected. Recomputing SemanticSRJudge per category

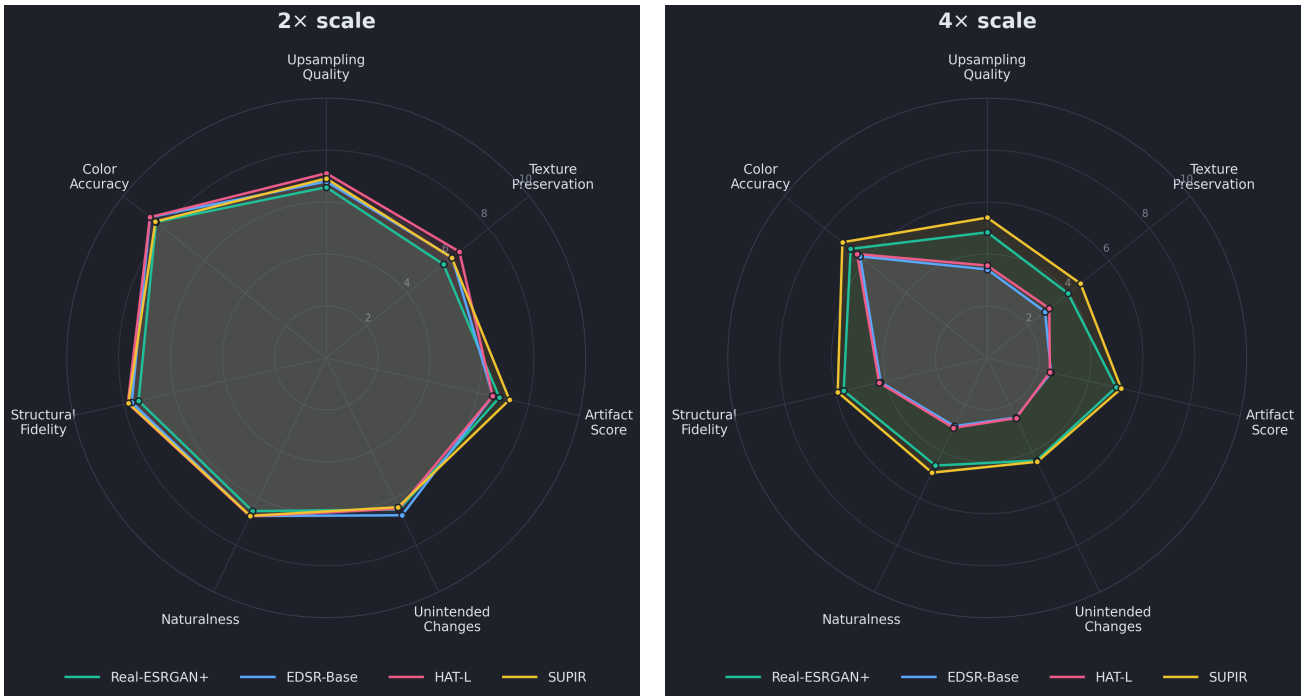

Figure 2. **Per-dimension radar profiles, all four models.** Axes (0–10, outer = 10): UQ, TP, AS, UC, Nat, SF, CA. Each polygon's *shape* encodes the failure-mode signature; *area* reflects overall quality. Left: $2\times$. Right: $4\times$. Across the four datasets, HAT-L's polygon is largest at $2\times$ on DIV2K and SemanticSR-Bench (synthetic and content-categorised data) while SUPIR wins UQ on Canon and Nikon (camera-noise data, where the diffusion prior has the edge); SUPIR's polygon is largest at $4\times$ on *all four* datasets, a prior-reversal not visible in any single classical metric.

at $4\times$ (Table 4) shows that the aggregate "SUPIR wins SemanticSR $4\times$ everywhere" breaks down once the category structure is restored.

The per-category structure splits cleanly along *natural-vs-structured* content. The diffusion prior (SUPIR) wins on Faces, Animals, Nature, and Indoor scenes — categories where plausible micro-texture synthesis matches human preference. The GAN prior (Real-ESRGAN+) wins on Text/Documents and Patterns, where the diffusion prior invents plausible-but-wrong character strokes and tile geometry. Per-category winners and dimension-counts are in Table 4; absolute scores are uniformly lowest on Text/Documents, the regime where every model in the comparison struggles at $4\times$. The practical takeaway is direct: for $4\times$ text and pattern restoration prefer a GAN-based prior, for all other categories prefer SUPIR. Without the per-category breakdown the "Real-ESRGAN+ wins on structured content" finding is invisible at the dataset level.

4.5. Human Alignment: Region Conditioning Lifts Win% and Spearman

We compare three VLM-judge conditions and five classical IQA baselines against the same human anchor on RealSR Canon $4\times$. The anchor is the mean rating of three independent expert annotators on 50 stems and four models. Per-dimension agreement statistics are in §S3. Canon is the harder of our regimes: the LR/HR pairs are optically captured and carry sensor noise, and the four SR models score within a much narrower Win% band than on DIV2K. That makes the subset a discriminating benchmark for any SR judge.

DINOv2-routed conditioning is the only condition that lifts both Win% and rank correlation simultaneously. On the eight-method comparison in Table 5, SemanticSRJudge (K=3) leads on Win%, Pearson r, *and* Spearman ρ, beating the vanilla VLM by $+6.7$pp Win% and a same-budget random-patch control by $+3.8$pp. The per-dimension breakdown in Table 5 confirms the gain concentrates on the locally-supported dimensions (Upsampling Quality, Texture Preservation, Structural Fidelity); Color Accuracy remains the ceiling, where human raters themselves disagree most.

The classical baselines split as expected, and none combine high Win% with positive rank correlation. PSNR and SSIM are weakly anti-correlated with the panel on this regime; LPIPS and DISTS recover a perceptual edge but neither clears ρ=$+0.20$. TOPIQ-FR is the strongest classical metric, yet still trails SemanticSRJudge by 9.6pp Win%. Notably, the vanilla VLM judge already matches or exceeds every classical scalar on the combined Win%/rank-correlation

Table 3. **Classical IQA metrics at $4\times$ (full coverage).** PSNR/SSIM/TOPIQ-FR ↑; DISTS/LPIPS ↓. **Bold**: best per metric×dataset. Sample sizes: DIV2K n=100, Canon n=50, Nikon n=50, SemanticSR-Bench n=117. LR inputs are bicubic-downsampled HR (synthetic-degradation protocol, applied uniformly across datasets). On DIV2K, fidelity (PSNR/SSIM/TOPIQ) ranks HAT first and perceptual (DISTS/LPIPS) ranks SUPIR first — the textbook perception-distortion tradeoff [2]. On RealSR Canon/Nikon and SemanticSR-Bench this ordering compresses or reverses: EDSR≈HAT on PSNR/SSIM (camera images have less recoverable high-frequency content), and Real-ESRGAN+ leads on perceptual metrics.

Dataset	Model	PSNR	SSIM	TOPIQ	DISTS	LPIPS
DIV2K	RealESRGAN+	24.84	0.72	0.59	0.121	0.228
	EDSR	29.29	0.84	0.66	0.133	0.257
	HAT	**29.75**	**0.85**	**0.68**	0.129	0.244
	SUPIR	25.05	0.70	0.61	**0.100**	**0.196**
Canon	RealESRGAN+	24.46	0.74	**0.49**	**0.161**	**0.261**
	EDSR	**26.06**	**0.74**	0.39	0.229	0.437
	HAT	26.01	0.73	0.32	0.226	0.439
	SUPIR	24.47	0.69	0.48	0.188	0.322
Nikon	RealESRGAN+	24.13	0.72	0.50	**0.177**	**0.285**
	EDSR	**25.91**	0.73	0.42	0.225	0.448
	HAT	25.88	**0.75**	0.41	0.227	0.452
	SUPIR	24.56	0.67	**0.53**	0.180	0.308
Sementic SR	RealESRGAN+	**24.03**	**0.75**	**0.51**	**0.146**	0.207
	EDSR	22.03	0.73	0.43	0.194	0.265
	HAT	21.91	0.73	0.42	0.192	0.278
	SUPIR	23.71	0.72	0.49	0.155	**0.203**

criterion, and the additional lift from DINOv2 routing is what turns it into the leading method on both axes. The lift comes from spatially routing the VLM to where the SR failures are, at the resolution at which they exist.

5. Discussion

Why classical metrics are not enough. The classical IQA suite on the Canon human anchor splits along the predicted lines (Table 5): pixel-fidelity metrics (PSNR, SSIM) are weakly anti-correlated with the panel on this regime, while perceptual metrics (LPIPS, DISTS, TOPIQ-FR) recover positive rank correlation but trail SemanticSRJudge (K=3) on the combined Win% / Spearman criterion. Even the vanilla VLM matches or exceeds every classical scalar before any region routing is added; the additional gain SemanticSRJudge contributes on top is what turns it into the leading method on both axes, while uniquely returning a per-dimension diagnostic rather than a single number.

Informed patch routing, not just local context. The random-patch ablation (§S6, Table S9) replaces the DINOv2-guided crops with K=3 uniformly random 128×128 patches at the same patch budget. Random patches yield only a marginal lift over the vanilla full-image baseline

Table 4. **Per-category $4\times$ winners on SemanticSR.** "Winner" is the model that wins the most of the seven SemanticSRJudge dimensions in the category (ties broken by UQ). Cell values are the winner's per-dimension scores (↑) for the first five dimensions (UQ, TP, AS, UC, Nat); SF and CA are omitted to fit the column width and follow the same pattern. Bold marks dimensions where the winner is the best of the four models. Diffusion (SUPIR) wins on natural content; Real-ESRGAN+ wins on structured content (text, fine patterns), where diffusion priors tend to hallucinate.

Category	Winner	UQ	TP	AS	UC	Nat
Faces	SUPIR (7/7)	**6.95**	**6.00**	**7.27**	**6.32**	**6.32**
Nature	SUPIR (6/7)	**5.75**	**4.95**	**5.75**	4.60	**5.20**
Animals	SUPIR (7/7)	**5.70**	**4.90**	**6.50**	**5.30**	**5.40**
Text/Docs	RealESRGAN+ (5/7)	3.50	3.00	3.25	**2.62**	**3.38**
Patterns	RealESRGAN+ (6/7)	**5.18**	**4.45**	**5.18**	**4.55**	**5.00**
Art/Illust.	SUPIR (4/7)	**4.81**	**4.06**	4.69	3.69	**4.44**
Indoor	SUPIR (7/7)	**5.82**	**5.00**	**5.68**	**4.77**	**5.23**

(within the GPT-5.4 sampling noise floor on the matched human anchor), whereas DINOv2-routed crops more than double the Spearman lift and are the only condition that clears the $\rho>0.30$ rank-correlation bar. The contrast establishes that the SemanticSRJudge gain comes from *semantic* routing to regions of meaningful SR failure, not merely from showing the VLM additional local context.

Why does region routing help most where it does? The Win% gain pattern across the seven SemanticSRJudge dimensions (Table 5) tracks the calibration offset hierarchy in Table 2: dimensions with the largest vanilla over-score (Texture Preservation, Upsampling Quality, Structural Fidelity) are precisely the ones on which DINOv2 routing yields the largest Win% and rank-correlation gains. The exceptions are Naturalness and Color Accuracy, both of which integrate global photometric impressions that three local crops cannot reproduce. The pattern is consistent with a single mechanism: locally-supported failure modes benefit from local evidence; globally-supported ones do not.

What SUPIR adds to the picture. The CNN, GAN, and attention-based models in this study fail in qualitatively different ways than SUPIR does. Real-ESRGAN+ oversharpens and hallucinates high-frequency texture; HAT-L and EDSR are reference-faithful but collapse on UQ at $4\times$. SUPIR synthesises plausible $4\times$ texture and also preserves Naturalness and Structural Fidelity better than Real-ESRGAN+ on every dataset. The resulting $4\times$ UQ ranking, SUPIR>Real-ESRGAN+>EDSR≈HAT-L, is more granular than any single classical metric resolves. The SemanticSR per-category breakdown then shows that this aggregate hides a content-conditional reversal: diffusion loses to a GAN prior on text and patterns. A practitioner using PSNR or LPIPS alone would miss this entirely.

Table 5. **Human alignment on RealSR Canon** $4\times$ **(**n=50 **stems, four models,** 200 **paired scores per dimension).** The human anchor is the mean rating across three independent expert annotators (mean per-dimension top-1 winner agreement 79.0%, mean Kendall τ=+0.211; full agreement table in §S3). Win% is the fraction of (stem, dim) cells on which the metric/judge's top-1 model matches the panel's (chance =25%); averaged over the seven SemanticSRJudge dimensions and matching the convention in Tables S8 and S9. Pearson r and tie-corrected Spearman ρ are computed per dimension across 200 pairs and averaged over the seven dimensions. We compare SemanticSRJudge (K=3, DINOv2-routed crops) against Vanilla VLM (K=0, full image) and a Random Patch control (K=3, uniform random crops) to isolate the contribution of informed region selection. Bold marks the best value in each column among the three VLM judges.

Method	Win%	Pearson r	Spearman ρ
PSNR	14.3	−0.04	−0.02
SSIM	12.4	−0.02	−0.02
TOPIQ-FR	39.0	+0.19	+0.23
LPIPS	21.0	+0.11	+0.10
DISTS	29.5	+0.22	+0.19
Vanilla VLM (K=0)	41.9	+0.25	+0.21
Random Patch (K=3)	44.8	+0.26	+0.24
SemanticSRJudge (K=3)	**48.6**	**+0.34**	**+0.31**

Cross-VLM Validation with Qwen3-VL-8B. To assess whether SemanticSRJudge's rankings depend on GPT-5.4's particular scoring disposition, we re-ran the full K=3, $4\times$ protocol on all 100 RealSR images (50 Canon, 50 Nikon) using the open-source QWEN3-VL-8B-INSTRUCT [13], producing 400 independent judge calls. Despite absolute-score differences of ~ 1.7 points on average, the two judges agree on the pairwise winner in over 90% of stems when comparing classical models (EDSR, HAT-L) against generative models (REAL-ESRGAN+, SUPIR). Per-dimension rank concordance is strongest on the quality-sensitive dimensions (ρ=0.67 on *upsampling_quality*, ρ=0.64 on *texture_preservation*, both $p<10^{-24}$, N=200), and weakest on the faithfulness dimensions (*unintended_changes*, *artifact_score*), where the two judges differ in how aggressively they penalise hallucinated GAN/diffusion textures. The net effect is an absolute calibration offset on those dimensions but a preserved cross-family ordering (generative>classical), confirming that SemanticSRJudge's ordinal rankings are robust to the underlying VLM while absolute calibration reflects each judge's training-induced perceptual prior.

From diagnosis to selection and training. The contribution gives users a tool that flags the dimensions and content categories on which each SR model succeeds or fails: where it preserves texture, hallucinates, collapses on text or patterns, or holds up on faces, animals, and natural scenes. This breakdown lets practitioners select models for their use case rather than rely on a single scalar that averages across regimes. The same per-dimension, per-category signal is actionable for SR developers: persistent weak spots identify concrete failure modes to target with curated data, dimension-weighted losses, or category-conditioned fine-tuning. In this sense the diagnostic is a direct input to both model selection and model improvement.

Limitations. The human study uses three annotators on 50 stems; the panel is small, and broadening to a larger multi-rater pool would tighten the absolute Win% estimates. We anchor on a single dataset and scale (RealSR Canon $4\times$), deliberately the harder regime; extending to DIV2K, Nikon, and to $2\times$ would test whether the Win% gain pattern transfers across degradation regimes. SUPIR is the only diffusion model in the comparison; SeeSR [12] and DiffBIR are natural next entries. Region routing underperforms on Color Accuracy and Artifact Score, where failures are spatially diffuse rather than locally concentrated; this is also where annotators themselves disagree most (τ=0.214 and τ=0.013 respectively), so the framework inherits a ceiling that no amount of local crop evidence can lift.

6. Conclusion

In this paper, we introduced SemanticSRJudge, a framework for mitigating the limits of scalar metrics (PSNR, SSIM, and LPIPS) in super-resolution evaluation. By routing a structured VLM judge to semantically divergent regions, it replaces a single global verdict with a spatial, seven-axis diagnosis of where failures reside.

Across four SR architectures, four datasets, and both $2\times$ and $4\times$ scales, the framework exposes tradeoffs hidden by dataset-level metrics. SemanticSRJudge corrects the optimistic bias of full-image VLM judging, while SemanticSR-Bench shows that model preference is not universal: diffusion-based SR can lead at high magnification, but GAN-based reconstruction remains competitive on text and pattern images where hallucination is especially damaging. The human study supports this result, with DINOv2-guided routing raising mean Win% from 41.9% to 48.6% and mean Spearman ρ from $+0.21$ to $+0.31$. Random-patch and cross-backbone checks further show that the gain comes from informed semantic routing and that rankings remain stable across VLM judges.

Overall, SemanticSRJudge provides the localized diagnostic signal, and SemanticSR-Bench provides the content-stratified setting needed to interpret it. Together, they support a shift from global SR scoring toward evaluation that is spatially explicit, dimension-specific, and sensitive to image category. Future work should expand the human study across more datasets, scales, SR models, and VLM families.

References

[1] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep ViT Features as Dense Visual Descriptors. *arXiv*:2112.05814, 2021.

[2] Y. Blau and T. Michaeli. The Perception-Distortion Tradeoff. *CVPR*, 2018.

[3] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TIP*, 2004.

[4] R. Zhang et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR*, 2018.

[5] K. Ding et al. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI*, 2020.

[6] C. Chen et al. TOPIQ: A Top-down Perspective for Image Quality Assessment. *IEEE TIP*, 2024.

[7] M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2023.

[8] B. Lim et al. Enhanced Deep Residual Networks for Single Image Super-Resolution. *CVPRW*, 2017.

[9] X. Chen et al. Activating More Pixels in Image Super-Resolution Transformer. *CVPR*, 2023.

[10] X. Wang et al. Real-ESRGAN: Training Real-World Blind SR with Pure Synthetic Data. *ICCVW*, 2021.

[11] F. Yu et al. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration in the Wild. *CVPR*, 2024.

[12] R. Wu et al. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. *CVPR*, 2024.

[13] Qwen Team. Qwen3-VL Technical Report. *arXiv*:2505.09875, 2025.

[14] H. Wu et al. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. *ICLR*, 2024.

[15] Z. Chen et al. Enhancing Descriptive Image Quality Assessment with a Large-scale Multi-modal Dataset. *IEEE TIP*, 2024.

[16] ExIQA: Explainable Image Quality Assessment Using Distortion Attributes. *arXiv*:2409.06853, 2024.

[17] Self-Evolving VLMs for IQA via Voting and Ranking. *arXiv*:2509.25787, 2024.

[18] S. Su et al. Rethinking Image Evaluation in Super-Resolution. *arXiv*:2503.13074, 2025.

[19] V. Shukla and A. Bedi. Generative Image Quality Assessment Agent. *Proc. 17th Int. Conf. Quality of Multimedia Experience (QoMEX)*, pp. 1–4. IEEE, 2025.

[20] V. Shukla, A. Bedi, and A. Jain. FaceQTransformer: Perceptual Face Quality Assessment and Enhancement for Generative Models. *Proc. 3rd Int. Workshop on Multimedia Content Generation and Evaluation (MM '25)*, pp. 115–120. ACM, 2025.

Supplementary Material

SemanticSRJudge: Spatially-Grounded VLM Evaluation for Super-Resolution Quality Assessment

S1. Prompt Design and VLM Configuration

Both judges use Azure OpenAI `gpt-5.4`[1] (v: 2026-03-05 Deployed 09/03/2026, 20:49:33), temperature 0.2, max tokens 2,000, and four-thread parallelism with exponential backoff (initial 5 s, max 60 s, up to five retries per call). The system prompt is identical across the two judges and fixes the seven dimensions, the 1–10 scale, and the JSON response schema. The vanilla judge receives only the reference and SR image; SemanticSRJudge additionally receives an annotated SR image and side-by-side reference/SR crops at the selected drift regions. Unless otherwise stated, SemanticSRJudge uses K=3 crops; the sensitivity analysis over K is reported in §S5.

System prompt.

You are an expert image quality assessor specialising in super-resolution (SR) evaluation. You will be given:

- *Image 1: The original high-quality REFERENCE image (ground truth, full resolution)*
- *Image 2: The SR model output (full image) annotated with numbered red boxes on the top-5 regions of highest semantic feature drift, detected by DINOv2 patch-level comparison*
- *Images 3–7: 128×128 native-resolution crops — LEFT = same patch from REFERENCE, RIGHT = SR model*

Your primary goal: compare SR output against the REFERENCE holistically across the full image. The numbered red-box regions highlight areas of highest semantic drift detected by DINOv2 — use them as diagnostic hints to guide your attention, but do NOT let localised crop-level observations dominate your image-level scores. Always base image-level scores on your overall impression of the entire frame first, then use crops to inform or confirm — never to override a strong global signal.

Scoring Criteria — score every dimension on a 0–10 integer scale where: 10 = indistinguishable from reference / no issue at all; 7 = minor deviation noticeable only on close inspection; 5 = moderate, clearly visible but not severe; 3 = severe issue clearly hurting quality; 1 = extreme distortion / major structural failure.

Per-region scores (assess each of the 5 numbered crops):

- `texture_match` *— Does SR texture (hair, fabric, bark, skin pores) match the reference? Penalise: smoothed-out detail, wrong pattern, hallucinated texture.*
- `sharpness` *— Is sharpness appropriate and matching the reference? Penalise: over-blurred, OR over-sharpened with ringing/halos.*
- `artifact_free` *— Freedom from artifacts: ringing, double edges, fringing, mosquito noise, checkerboard patterns, compression-like banding, false contours.*
- `unintended_change` *— Is the SR patch free of any content NOT in the reference? Penalise: hallucinated details, shape deformation, extra texture elements invented by the model, minute spurious additions.*

Image-level scores (holistic, across entire image including non-highlighted regions):

- `upsampling_quality` *— The primary perceptual quality score for this SR output. This is not a weighted average of the other dimensions — it is an independent, holistic judgement of the perceived visual quality of the entire image as a trained observer would experience it at normal viewing distance. Consider the cumulative impression: does the SR image feel resolved, coherent, and visually satisfying relative to the reference? This encompasses sharpness, tonal balance, micro-contrast, overall crispness, and the degree to which the image reads as a faithful, high-quality rendering of the scene — not merely whether individual crops are sharp. A high score requires the full image to hold up perceptually, not just selected regions. Penalise: global softness, loss of mid-frequency detail, muddy micro-textures, and any processing that reduces the perceived resolution or tonal richness of the full scene.*
- `texture_preservation` *— Fidelity of fine textures (hair, fur, fabric, foliage, skin) across image.*
- `artifact_score` *— Global freedom from ringing, halos, double-edge sharpening, fringing near high-contrast edges, banding, false textures.*
- `unintended_changes` *— Model introduced elements, deformations, or minute unwanted details not present in reference (hallucinations, shape warping, extra strokes).*
- `naturalness` *— Does the image look perceptually natural, without the "AI-generated" over-processed look? Penalise: uncanny textures, plastic skin, hyper-sharpened micro-detail that looks synthetic.*
- `structural_fidelity` *— Edges, boundaries, shapes, and large structures match reference. Penalise: deformed contours, bent lines, shifted boundaries.*
- `color_accuracy` *— Colour and luminance match reference. Penalise: hue shifts, saturation changes, tone mapping differences.*

Respond ONLY with a single valid JSON object — no markdown, no text outside JSON:

```
{
"region_scores": [
{
"region": <1--5>,
"texture_match": <0--10>,
"sharpness": <0--10>,
```

[1] OpenAI system cards: https://openai.com/research/index/?tags=system-cards

```
"artifact_free": <0--10>,
"unintended_changes": <0--10>,
"observations": "<1--2 sentences>"
}
],
"image_scores": {
"upsampling_quality": <0--10>,
"texture_preservation": <0--10>,
"artifact_score": <0--10>,
"unintended_changes": <0--10>,
"naturalness": <0--10>,
"structural_fidelity": <0--10>,
"color_accuracy": <0--10>
},
"strengths": "<1--2 sentences>",
"weaknesses": "<1--2 sentences>",
"overall_summary": "<3--4
sentences>"
}
```

Vanilla user prompt. "Image 1 = ground-truth REFERENCE. Image 2 = SR model output. Score all 7 dimensions and provide an overall summary."

SemanticSRJudge user prompt. "Image 1 = ground-truth REFERENCE. Image 2 = SR model output (annotated with K numbered bounding boxes at the regions of greatest DINOv2-detected semantic drift). Images 3 to $2{+}K$ = native-resolution $128{\times}128$ Reference | SR side-by-side panels at the K drift regions. Score all 7 dimensions, treating the annotated full-image impression as primary and the crops as diagnostic evidence for local failure modes. Also provide per-region observations." The placeholders $\{n_regions\}$ and $\{last_img\}$ in the prompt template are filled at runtime with K and $2{+}K$, so the text adapts to the chosen K.

The SemanticSRJudge response includes a `region_scores` array with four per-region dimensions (texture match, sharpness, artifact-free, unintended change). These per-region fields are not used for the Δ and Win% comparisons in the main paper, which use the seven shared global dimensions only.

S2. Full Vanilla-Judge Tables and Δ per Cell

Table S6 is the matched-pair counterpart to Table 1 (main). Per-cell Δ values ($\Delta = \overline{\textit{Vanilla}} - \overline{\textit{Semantic}}$ on each $\langle$*dataset, scale, model, dimension*$\rangle$ cell) are uniformly positive on UQ, TP, and SF and uniformly negative or near zero on AS and CA, the same hierarchy reported in Table 2 of the main paper, but reproducing that the offset is structural rather than driven by any single dataset or scale.

Figure S3. **SemanticSRJudge human-annotation interface.** A multi-panel synchronised viewer presents the reference and model outputs side by side. Left: reference HR image. Centre/right: SR outputs from the four models under comparison. Synchronised zoom and pan: any zoom or scroll action on one panel is immediately reflected on all panels simultaneously, enabling pixel-level side-by-side comparison without losing spatial context. Score drop-downs at the base of each model panel allow independent 1–10 rating on all seven SemanticSRJudge dimensions before submission, and the interface supports session continuity across multiple sittings.

S3. Human Annotation Methodology and Interface

Interface. Figure S3 shows the synchronised multi-panel viewer: the reference HR image alongside the four SR outputs, with any zoom or pan applied simultaneously to all panels so a rater cannot score a model crop without verifying the reference at the same location and scale.

Annotation procedure. Three trained annotators independently rated 50 RealSR Canon validation stems at $4\times$ across all four SR models (Real-ESRGAN+, EDSR, HAT-L, SUPIR) in the viewer of Figure S3. Stems were presented in randomised order with models randomly assigned to side-panel slots, removing left-right position bias. The interface enforced scoring on all seven dimensions before advancing.

Anchors and scale. Annotators used the same 1–10 scale fixed in the VLM system prompt (§S1), with extremes anchored verbatim against the VLM text ("10: indistinguishable from the reference at this viewing distance; 1: severe distortion, image is unusable").

Annotation set and aggregation. The subset is 50 stems $\times$ 4 SR outputs $=$ 200 cells per dimension per annotator (600 raw scores per dimension across the panel). We aggregate annotators by mean per $\langle$stem, model, dimension$\rangle$ cell to form the human anchor used in Table 5.

Inter-annotator agreement. Table S7 reports per-dimension agreement using three complementary statistics.

Table S6. **Vanilla VLM judge scores (↑, 0–10).** Same protocol as Table 1 but with no DINOv2-derived crops or annotations. **Bold**: best model per dataset×scale on that dimension.

			Vanilla (2×)							**Vanilla** (4×)						
Dataset	**Model**	n	UQ	TP	AS	UC	Nat	SF	CA	UQ	TP	AS	UC	Nat	SF	CA
DIV2K	RealESRGAN+	100	7.71	7.30	7.26	7.72	7.57	8.72	8.29	6.74	5.87	6.23	5.93	6.29	7.66	6.94
	EDSR	100	8.07	7.98	**8.96**	9.04	8.39	9.05	9.10	7.03	6.45	8.01	8.20	7.37	8.27	8.11
	HAT-L	100	**8.59**	**8.54**	8.83	**9.20**	**8.60**	**9.53**	**9.53**	**7.70**	**7.37**	**8.15**	**8.37**	**7.72**	**8.73**	**8.41**
	SUPIR	100	7.67	7.33	7.67	7.51	7.61	8.58	8.10	7.47	7.00	7.22	6.80	7.20	8.20	7.71
Canon	RealESRGAN+	50	6.38	5.50	5.46	5.44	5.82	7.12	6.94	4.10	3.18	3.46	2.70	3.60	4.44	5.30
	EDSR	50	5.92	5.04	7.36	**7.60**	6.56	7.08	7.40	3.88	3.26	**6.08**	**6.56**	**5.16**	**5.40**	**6.18**
	HAT-L	50	6.32	5.52	**7.38**	7.56	**6.76**	**7.48**	**7.64**	3.84	3.50	5.96	6.14	5.12	5.34	6.18
	SUPIR	50	**6.48**	**5.74**	5.76	5.14	5.80	6.92	6.74	**4.62**	**3.72**	4.20	3.04	4.06	5.06	5.06
Nikon	RealESRGAN+	50	6.60	5.88	6.12	6.16	6.32	7.64	7.54	4.80	3.92	4.16	3.82	4.40	5.80	5.66
	EDSR	50	6.90	6.48	**7.94**	**8.24**	7.40	8.08	8.14	5.04	4.42	**6.80**	**7.04**	**6.06**	6.58	6.70
	HAT-L	50	**7.38**	**7.08**	7.90	8.24	**7.62**	**8.48**	**8.30**	5.26	**4.72**	6.58	6.74	6.00	**6.66**	**6.78**
	SUPIR	50	6.84	6.42	6.58	6.44	6.60	7.70	7.48	**5.43**	4.69	5.24	4.45	5.08	6.31	6.14
SemanticSR	RealESRGAN+	117	6.99	6.38	6.63	6.52	6.80	7.82	7.81	4.95	4.17	3.89	3.55	4.35	5.41	5.71
	EDSR	117	7.43	7.16	**7.51**	**7.87**	7.47	8.40	8.50	3.78	3.01	2.22	2.62	2.85	4.59	5.71
	HAT-L	117	**7.80**	**7.57**	7.38	7.77	**7.62**	**8.62**	**8.67**	3.97	3.25	2.23	2.74	3.01	4.74	5.95
	SUPIR	117	7.04	6.52	7.04	6.55	6.99	7.71	7.85	**5.33**	**4.58**	**4.47**	**3.75**	**4.64**	**5.60**	**6.35**

Table S7. **Inter-annotator agreement across three independent annotators on the RealSR Canon 4× human study (n=50 stems, four SR models).** Top-1: fraction of stems on which the annotators picked the same best model on the dimension (chance 25%). Pairwise: mean agreement across the $\binom{4}{2}$ model pairs per stem (chance 50%). Kendall τ: rank agreement on the four-model ranking per stem (chance 0).

Dimension	Top-1 %	Pairwise %	Kendall τ
Upsampling Quality	93.3	94.1	+0.696
Texture Preservation	80.0	76.6	+0.416
Unintended Changes	80.0	68.1	+0.236
Naturalness	73.3	70.4	+0.318
Color Accuracy	80.0	34.8	−0.214
Artifact Score	73.3	46.9	−0.013
Structural Fidelity	73.3	44.2	−0.059
Mean (7 dims)	79.0	65.0	+0.211
Chance baseline	25.0	50.0	0.000

Top-1: fraction of stems where the annotators picked the same best model (chance 25%). *Pairwise*: mean agreement over the $\binom{4}{2}$ model pairs per stem (chance 50%). *Kendall* τ: rank agreement on the four-model ranking per stem (chance 0).

Agreement is strongest for visible local quality failures, especially Upsampling Quality and Texture Preservation. Color Accuracy, Artifact Score, and Structural Fidelity capture subtler global or model-specific effects, so we treat them as complementary diagnostic axes rather than as the sole basis for model ranking.

Top-K sensitivity. The full Top-K ablation is reported in Table S8 (§S5). The main effect is the transition from no crops to at least one drift-routed crop; differences among $K \in \{1, 2, 3\}$ are small and should be read as sensitivity results.

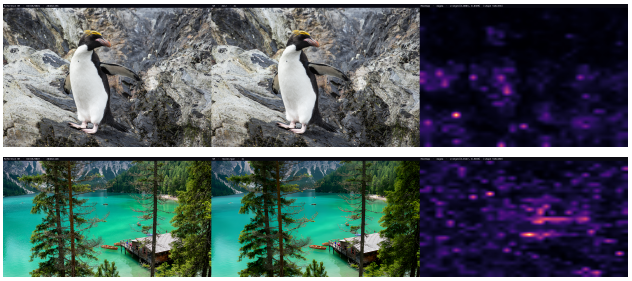

Figure S4. **Semantic drift visualisation for 4× super-resolution on DIV2K.** Each row shows the reference, SR output, and DINOv2 drift heatmap. Brighter regions indicate larger feature disagreement and determine the local regions shown to the VLM.

S4. Implementation Details

SemanticSR-Bench category composition. SemanticSR comprises 117 stems partitioned into seven content categories: Faces/Portraits (n=24), Indoor (n=22), Nature/Architecture (n=20), Art/Illustration (n=16), Text/Documents (n=14), Patterns/Textures (n=11), and Animals (n=10). Per-category scores are described in Table 4.

SR inference. EDSR-Baseline (ℓ_1, 1.4M params) and HAT-L (40.8M, hybrid attention) use their official PyTorch releases and publicly released pretrained weights. Real-ESRGAN+ uses the `RealESRGAN_x2plus` and

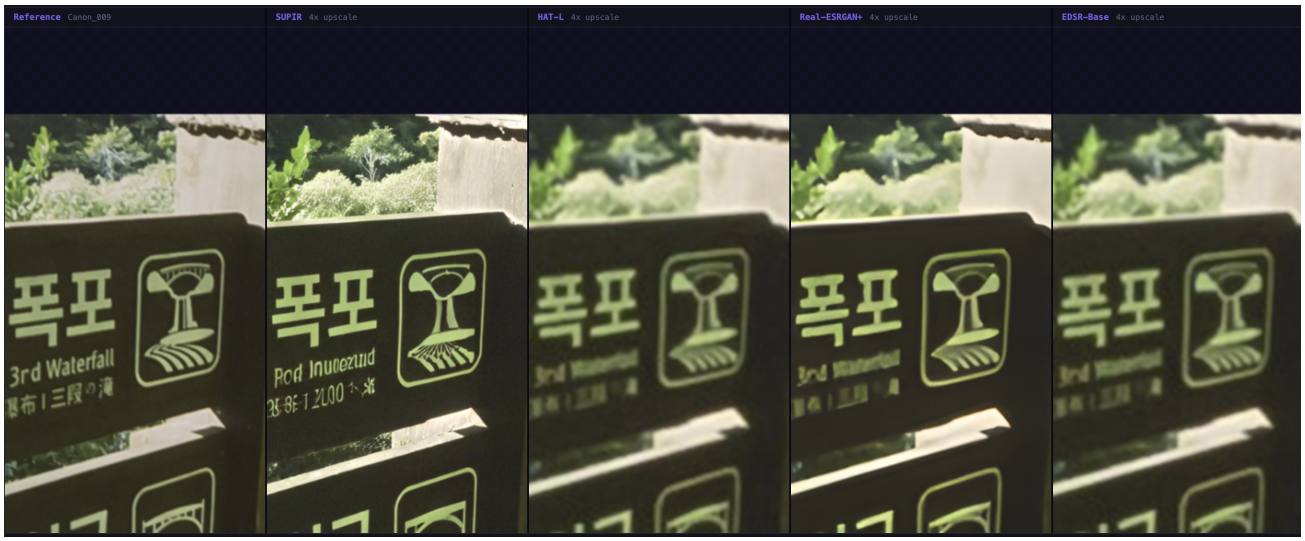

Figure S5. **Qualitative $4\times$ SR comparison on a low-light camera-noise stem.** Columns left to right: reference HR crop, SUPIR, HAT-L, Real-ESRGAN+, EDSR. SUPIR recovers signage stroke geometry and foliage micro-texture; HAT-L and Real-ESRGAN+ smooth or oversharpen the same regions. Consistent with the $4\times$ UQ ordering in Table 3 and Figure 2.

`RealESRGAN_x4plus` pretrained models. SUPIR uses the public release SUPIR-v0Q checkpoint with the default sampling configuration (45 steps, classifier-free guidance scale 7.5). All inference is performed at the native input resolution; outputs are resized to match the reference dimensions where necessary using Lanczos interpolation. Total SR inference time across the 317 stems $\times 4$ models $\times 2$ scales is approximately 14 hours on a single A100.

DINOv2 feature extraction. We use `vit_base_patch14_dinov2.lvd142m` (timm). Inputs are resized so $\max(H_{\text{img}}, W_{\text{img}}) \leq 518$ with both dimensions multiples of the 14-pixel patch size, preserving aspect ratio. Per-layer features are extracted via forward hooks on blocks 3, 7, and 11 (0-indexed). The drift map $\mathcal{H}$ averages the three per-layer $1 - \cos(\mathbf{f}_r, \mathbf{f}_s)$ tensors at full patch resolution; eight-connected component analysis on its top quartile yields the top-K regions ranked by mean drift magnitude. Reference features are computed once per stem and cached across the four SR models. Total DINOv2 cost across the full study is approximately 2.5 hours on CPU.

VLM judge. Judge results are written to JSON line-by-line so interrupted runs can resume. Total judge-call wall time across 5,072 calls is approximately 9 hours.

Classical metrics (cross-checks). PSNR and SSIM are computed via `scikit-image` (`compare_psnr`, `compare_ssim`; data range 255; multichannel). LPIPS, DISTS, and TOPIQ-FR are computed via `pyiqa` ($\geq$ v0.1.12) with default full-reference settings. All metrics are computed on native-resolution image pairs after Lanczos resampling to align stride-rounding mismatches in model inference.

Table S8. **Patch-count ablation on RealSR Canon $4\times$.** All rows are evaluated against the three-annotator panel mean. Lower is better for MAE.

Condition	Win%	Spearman ρ	MAE
K=0 (vanilla)	41.9	+0.211	2.131
Random patches (K=3)	44.8	+0.241	2.126
K=1	48.6	+0.294	2.083
K=2	50.5	+0.327	1.988
K=3 (default)	48.6	+0.310	1.960

S5. Patch-Count Sensitivity Analysis

Setup. Each additional crop adds $\sim$ 30 KB to the API payload and $\sim$ 1 s of latency, so K is both a behavioural and cost parameter. We ran the $K \in \{0, 1, 2, 3\}$ ablation on the RealSR Canon $4\times$ human anchor (§S3) using 50 stems across four SR models (Real-ESRGAN+, EDSR, HAT-L, SUPIR). All semantic conditions share the same DINOv2 region selection, crop extraction, prompt, GPT call, and retry procedures; only K varies. The random-patch row uses the same K=3 crop budget with crop centres drawn uniformly at random.

Metrics. Win% measures per-(stem, dimension) top-1 model identification against the human panel. Spearman ρ is computed per dimension over the 200 stem-model cells and then averaged. MAE is the mean absolute deviation between judge and panel scores on the 1–10 scale.

Results. The main gain comes from adding the first drift-routed crop. Relative to the vanilla judge, all region-conditioned settings improve human alignment, while differences among $K \in \{1, 2, 3\}$ are small. Rank metrics favor $K{=}2$, whereas MAE is lowest at $K{=}3$; we therefore use $K{=}3$ as the default setting but treat adjacent-K differences as exploratory rather than decisive. The random-patch row provides a same-budget control and shows that DINOv2-routed crops outperform uninformed local crops.

S6. Random-Patch Ablation: Does Informed Region Selection Matter?

Motivation. A natural question raised by the SemanticSRJudge design is whether the Win% and rank-correlation gains reported in §4.5 arise from the DINOv2 drift-detection routing specifically, or simply from providing *any* local context to the VLM. To isolate this, we ran a controlled third condition: the same GPT-5.4 judge, identical system prompt, and $K{=}3$ crops, but with the crop centres drawn *uniformly at random* within the image (seed fixed per stem for reproducibility) rather than at the top-K DINOv2 drift locations. All three conditions (DINOv2-routed, random, vanilla) were evaluated on the same 50 RealSR Canon $4\times$ stems and four open-source SR models against the panel-mean of the three human annotators.

Statistical reliability of the calibration claim. Under a paired stem-bootstrap on the full $n{=}50$ RealSR Canon $4\times$ human panel ($B{=}2000$, all four SR models and seven dimensions co-moved per stem to preserve within-stem dependence), Semantic K=3 reduces per-cell MAE relative to Random K=3 by -0.167 against the three-rater panel mean (95% CI $[-0.274, -0.071]$, $p{=}0.001$ by paired bootstrap and Wilcoxon signed-rank). Secondary endpoints — paired per-dimension Spearman ($+0.07\,[-0.04, +0.18]$) and per-stem Win% ($+20.0$ pp $[-2.1, +42.3]$) — are directionally consistent but individually less decisive than the primary MAE result. We pre-specify per-cell MAE as the primary calibration endpoint and flag broadening the human anchor (≥ 150 stems, ≥ 5 raters) as the highest-priority v2 extension.

Metrics. T1 and T2 measure per-(stem, dimension) top-1 and top-2 model identification accuracy. Pairwise accuracy measures agreement on strict human-preference pairs, and Spearman ρ is computed per dimension over 200 stem-model cells and averaged.

Table S9. **Random-patch control at the same crop budget.** Random crops improve modestly over the vanilla judge, while DINOv2-routed crops perform best across all reported alignment metrics.

Condition	T1	T2	Pair	Sp. ρ
Vanilla ($K{=}0$)	41.9%	82.9%	58.5%	+0.211
Random patches ($K{=}3$)	44.8%	83.8%	62.5%	+0.241
SemanticSRJudge ($K{=}3$)	**48.6%**	**89.5%**	**66.4%**	**+0.310**

Interpretation. Random crops modestly improve over the vanilla judge, suggesting that local context is useful. However, DINOv2-routed crops perform best across all metrics at the same crop budget. The result supports the claim that the gain comes from informed region selection rather than merely increasing the number of image views shown to the VLM.

S7. Limitations and Planned Extensions

1. **Human-panel size.** The human anchor uses three annotators on 50 RealSR Canon $4\times$ stems. This supports a controlled pilot comparison, but larger panels are needed for tight confidence intervals.
2. **Human-study scope.** The human study covers one dataset and one scale. Extending annotation to DIV2K, Nikon, and $2\times$ would test whether the alignment trends transfer across degradation regimes.
3. **Model coverage.** SUPIR is the only diffusion SR model in the comparison. Additional diffusion and restoration models would test whether the observed high-scale behavior is architecture-specific.
4. **Judge coverage.** The main human-alignment analysis (§4.5) uses GPT-5.4. The Qwen3-VL replication supports broad ordinal agreement, but a full open-weight replication would improve reproducibility.
5. **Crop-selection baselines.** The random-crop control suggests that informed region selection matters, but other crop policies remain untested, including uniform grids, saliency maps, edge-density regions, and full-reference error heatmaps.
6. **Fixed routing depth.** K is a fixed hyperparameter, not content-adaptive. Rank metrics slightly favour $K{=}2$ and MAE favours $K{=}3$ (§S5); we default to $K{=}3$. A per-image adaptive K is a natural extension.
7. **Absolute scores are judge-relative.** Absolute 1–10 values are VLM-dependent ($\sim$ 1.7-point mean GPT-5.4/Qwen3-VL offset) and should not be compared across judges. The validated output is the cross-model ranking, preserved across both judges (>90% agreement; §4.5).