SemanticSRJudge

Spatially-Grounded VLM Evaluation for Super-Resolution Quality Assessment

Vishwajeet Shukla | Ankit Dhankhar | Ajay Bedi · Adobe Inc., India · dhankhar@adobe.com

Adobe

ICML 2026 · Workshop

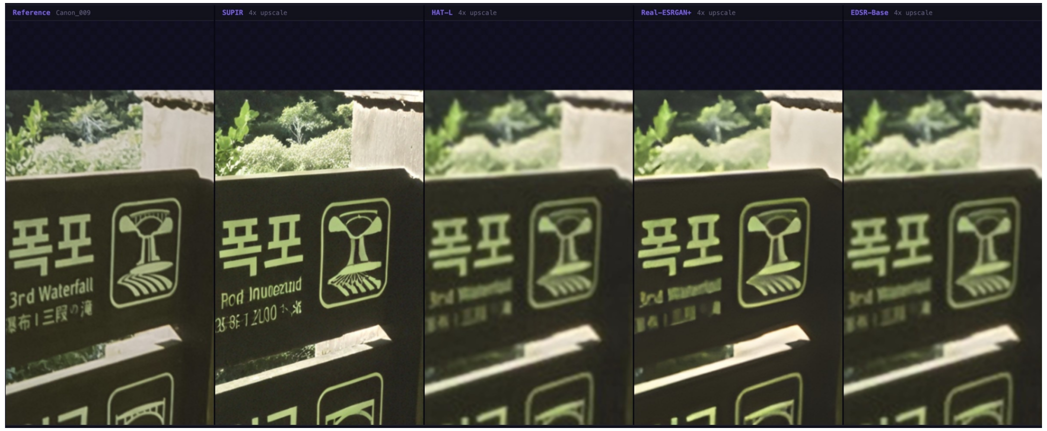
THE PROBLEM

Why one number is not enough

Super-resolution is judged almost entirely by **single numbers**. PSNR / SSIM reward pixel fidelity; LPIPS / DISTS / TOPIQ-FR add learned perceptual distance.

None attributes a score to a ***location*** or a ***failure type*** — and the perception-distortion trade-off forces any scalar to pick a side.

> **The practical question:** *where on this image did the model fail, and which class of failure was it?*



4× SR on a camera-noise sign. SUPIR sharpens foliage yet hallucinates the text — "Pod Inuoezud" vs the reference "3rd Waterfall." A global score never localizes that.

Two blind spots

1 Scalar metrics pick a side

The perception-distortion trade-off means fidelity and perceptual metrics disagree by construction — neither localizes the failure, and both reduce a rich image to one ranking.

2 Full-image VLM judges dilute

Shown the whole frame, a VLM's attention spreads out. Ringing, halos and hallucinated micro-texture occupy a few pixels, wash out at full-frame scale, and inflate the scores users care about.

The deciding evidence lives in **small, local departures from the reference**.

OUR APPROACH

Route the judge to the failure

Find **where the SR output stops looking like the reference** — then judge those regions at native resolution.

1

Detect

A frozen DINOv2 encoder compares reference and SR patch-by-patch in semantic feature space and builds a drift map.

2

Crop

Top-quartile drift regions become K = 3 native-resolution Reference | SR crops with numbered boxes.

3

Judge

GPT-5.4 scores the full image plus the crops on seven axes; a matched vanilla call isolates the routing.

METHOD · DETECT

A semantic drift map, not an edge map

A frozen **DINOv2 ViT-B/14** (self-supervised on 142M images) encodes both images at three depths; cosine dissimilarity gives a per-patch drift map:

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

ℓ 4 edges & ringing

ℓ 8 material & grain

ℓ 12 structure & hallucination

Reference · SR output · drift heatmap. Bright = largest semantic drift; these regions become the crops shown to the judge.

Why not Canny / Sobel? Gradient detectors ***invert*** — they fire hardest at over-sharpened, badly reconstructed edges. DINOv2 stays stable under photometric shifts and captures hallucinated grain and deformed structure. Deterministic, label-free, < 2 s on CPU.

A diagnosis, not a verdict

SemanticSRJudge returns a **vector, not a scalar** — seven complementary scores. A model's failure signature is the ***shape*** of its profile, which is what makes the calibration and content reversals legible.

UQ Upsampling Quality

TP Texture Preservation

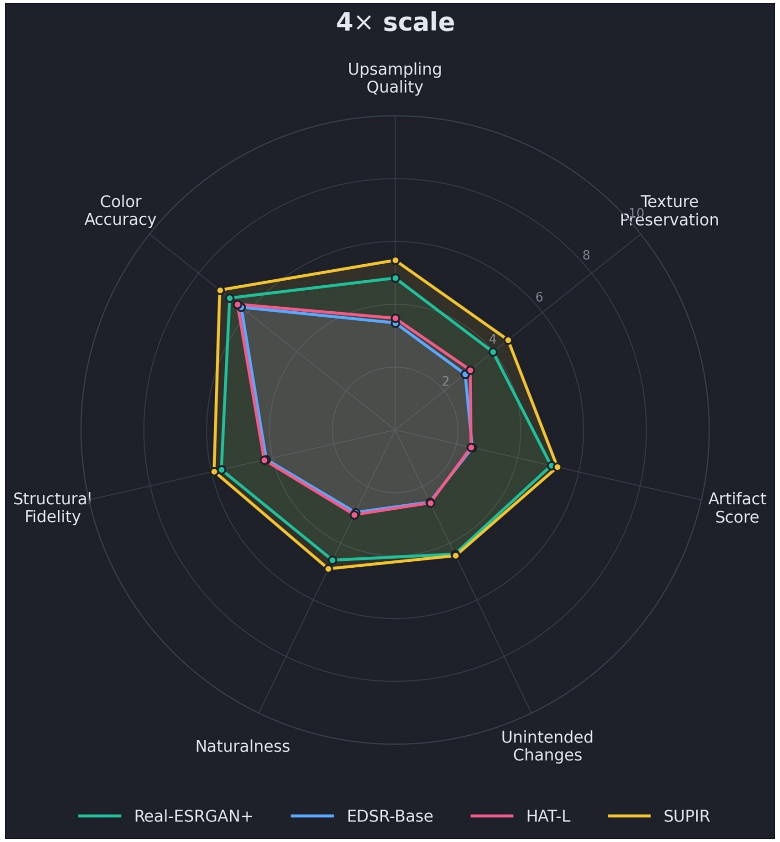
AS Artifact Score

UC Unintended Changes

Nat Naturalness

SF Structural Fidelity

CA Color Accuracy



Each polygon is one model: shape = failure signature, area = overall quality.

Four architectures, four datasets, two scales

317	**4**	**2**	**5,072**
source stems	SR models	scales (2×, 4×)	matched judge calls

Models span the design space

EDSR CNN, 1.4M

Real-ESRGAN+ GAN, 16.7M

HAT-L attention, 40.8M

SUPIR diffusion, multi-billion

Datasets cover synthetic & real

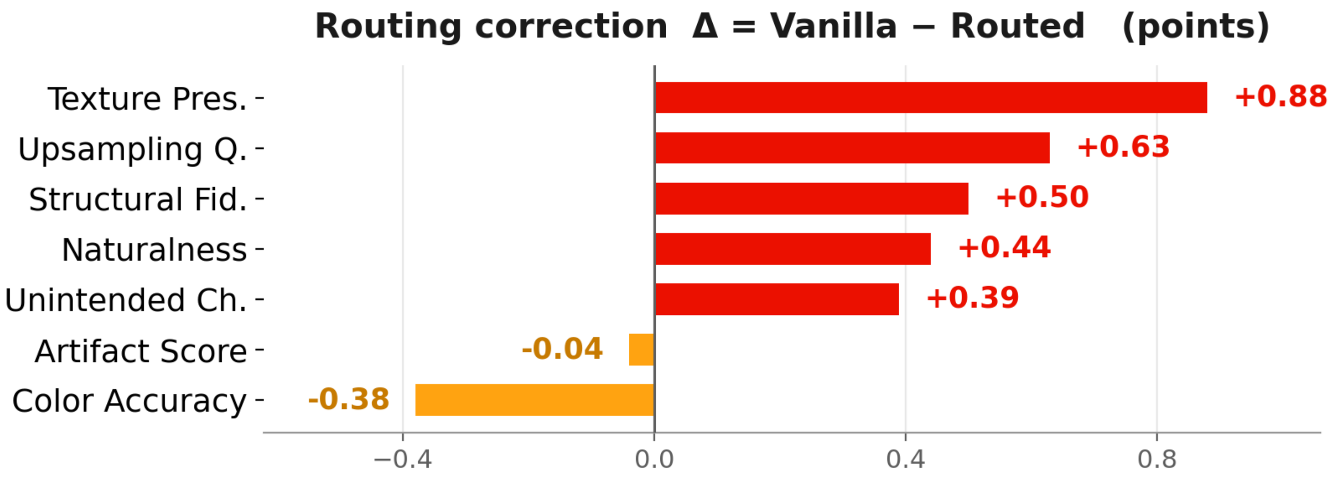
DIV2K 100 · synthetic

RealSR Canon / Nikon 50 + 50 · camera noise

SemanticSR-Bench 117 · 7 content categories

FINDING 1

Region routing corrects optimistic scoring



Pairing every routed call with its vanilla twin exposes a **systematic, dimension-dependent optimism**.

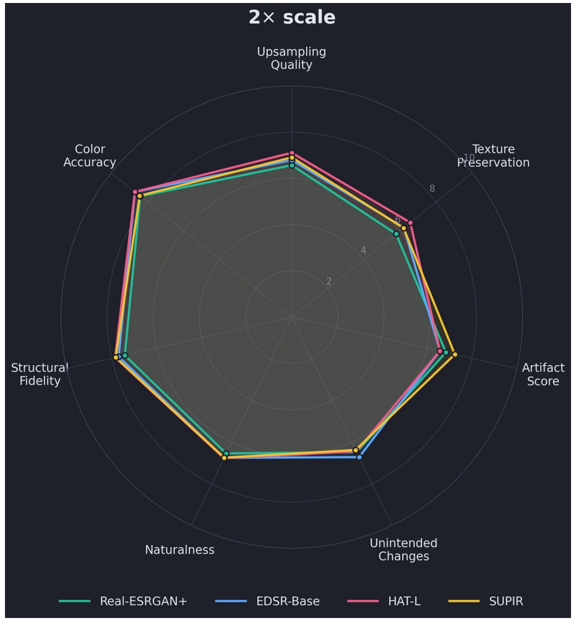
The correction is largest on the **locally-supported** axes (texture, upsampling, structure), near-zero on artifacts, and even reverses on color — its ordering tracks how spatially concentrated each failure mode is.

FINDING 2

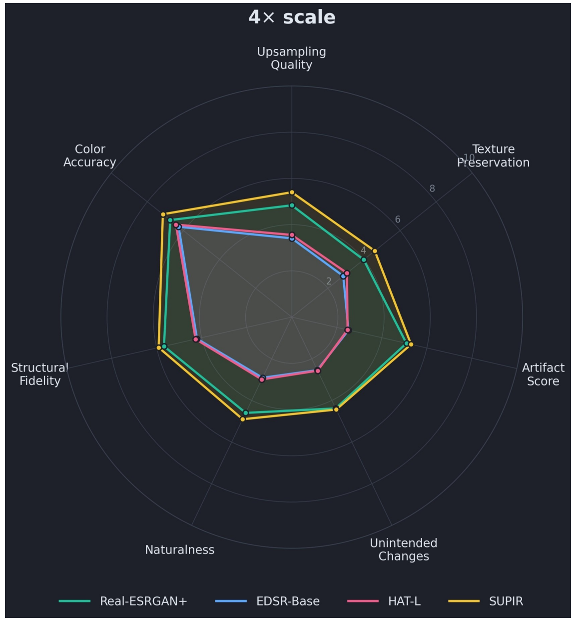
A scale-dependent prior reversal

The reversal is **architectural**. At 2×, attention (HAT-L) wins on synthetic data (DIV2K, SemanticSR-Bench) while diffusion (SUPIR) already wins the camera-noise sets; at 4× SUPIR wins Upsampling Quality on **all four datasets**.

2× scale



4× scale



Adobe

FINDING 3

Model preference flips with content

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**

On SemanticSR-Bench the 4× winner splits along **natural vs. structured** content.

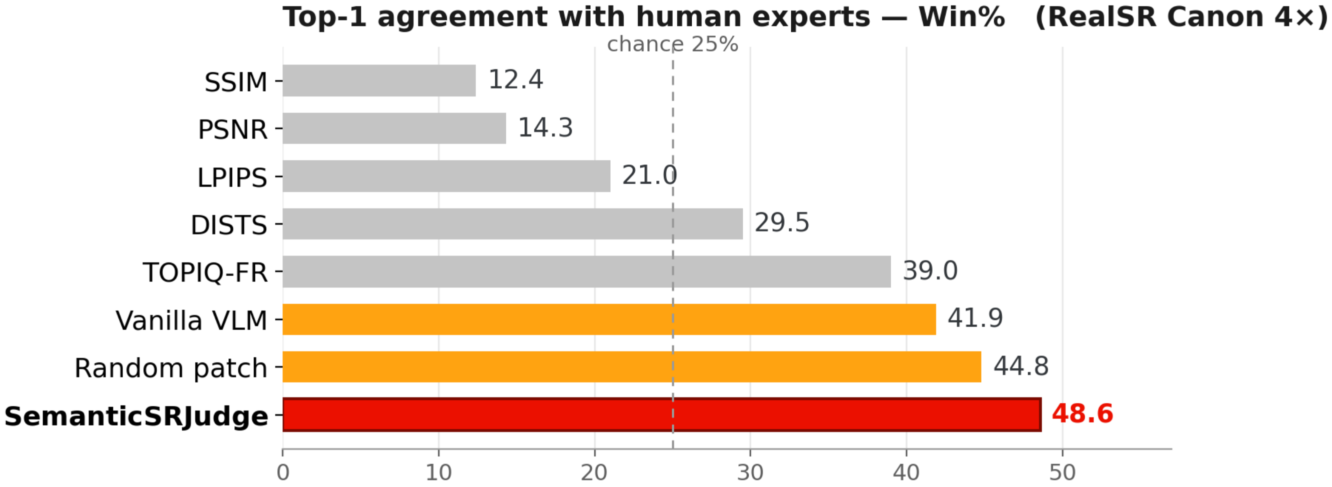
Diffusion wins faces, animals, nature, indoor — plausible micro-texture matches human preference.

The GAN prior wins text & patterns, where diffusion invents wrong character strokes and tile geometry.

Averaged over a dataset, this reversal is invisible.

FINDING 4

Validated by people



Against 3 experts on RealSR Canon 4×, SemanticSRJudge reaches **48.6% Win%** — and the vanilla VLM already beats every classical metric, so the routing is what makes it lead on ***both*** axes.

Only judge to clear ρ > 0.30 on rank correlation. Inter-annotator agreement 79% top-1.

Synchronized five-panel reviewer dashboard

Informed routing — and it transfers

Informed routing, not just more crops

Random crops at the same budget reach only **44.8**; drift-routed crops reach **48.6**. The gain is the routing — it lands on the axes the calibration analysis flagged as over-scored.

Robust across judges

Replacing GPT-5.4 with open-source **Qwen3-VL-8B** preserves the generative > classical ordering, with **>90% pairwise agreement** and $\rho = 0.67$ on Upsampling Quality.

Two controls: routing beats a same-budget random baseline, and the conclusions hold on a different, open VLM.

TAKEAWAY

Route the VLM to where SR fails, at the resolution it fails.

✓ Spatial

scores attached to regions, not the whole frame

✓ Dimension-specific

seven axes expose how a model fails, not just how much

✓ Content-aware

catches scale and category reversals scalars can’t see

A diagnostic that feeds both model **selection** and targeted **training**.

Thank you

Questions welcome

Vishwajeet Shukla | Ankit Dhankhar | Ajay Bedi · dhankhar@adobe.com

Adobe