# Assessing Face Super-Resolution with Semantic Guidance from Large Vision–Language Models

Tae-Seung Kim*, Hyeopgeon Lee, Young-Woon Kim
Dept. of Bigdata, Seoul Gangseo Campus of Korea Polytechnic College

***Abstract*— Face Super-Resolution (FSR) reconstructs high-resolution faces from low-resolution inputs. This paper assesses lightweight LVLM supervision for general (Real-ESRGAN), blind (BSRGAN), and prior-guided (GFPGAN) pipelines. We run paired pipelines—with and without LVLM—on identical degraded inputs. Evaluation uses PSNR, SSIM, LPIPS, NIQE, ArcFace, and side-by-side visuals. General FSR shows negligible metric shifts; blind FSR is sensitive to misclassified degradations under LVLM tagging; prior-guided FSR yields small auditing benefits with limited structural change. Overall, LVLMs work best as validators; measurable gains require deeper, controllable hooks.**

***Keywords—face super-resolution, blind super-resolution, prior-guided face restoration, vision–language models, semantic supervision, identity preservation.***

## 1 Introduction

Face Super-Resolution (FSR) [1], [2] reconstructs identity-preserving high-resolution faces from degraded inputs. FSR serves forensics, privacy-aware analytics, and video enhancement, where identity fidelity and artifact control are critical [1], [2]. Large Vision–Language Models (LVLMs) [3], [4] provide image-grounded reasoning, textual critiques, and lightweight quality checks that are model-agnostic and easy to insert.

FSR pipelines face coupled trade-offs among realism, distortion fidelity, and identity preservation [2]. General methods (e.g., Real-ESRGAN [1], [2]) improve textures but may overshoot under mismatched degradations. Prior-guided methods (e.g., GFPGAN [7]) stabilize identity but expose only coarse inference controls. Blind methods (e.g., BSRGAN [6]) aim for robustness yet remain sensitive to wrong degradation assumptions at test time. Prior work (LLV-FSR [5]) reports LVLM benefits mainly in prior-guided settings; effective scope and insertion points remain narrow and under-explored.

This paper assesses LVLM supervision across three deployed FSR families: general (Real-ESRGAN [1]), blind (BSRGAN [6]), and prior-guided (GFPGAN [7]). We run paired pipelines—FSR-only vs. FSR+LVLM—on identical degraded inputs to isolate LVLM effects. LVLM acts as a degradation tagger and a quality auditor with bounded, family-specific knobs; we do not redesign model internals.

We hypothesize that LVLM helps most as a validator and gatekeeper, while direct steering yields mixed outcomes when control surfaces are coarse.

We preview the findings. General SR shows little change, blind SR suffers from misclassified degradations, and prior-guided SR gains modest plausibility but flat metrics. Overall, LVLM works better as a validator than as a controller.

The rest of this paper is organized as follows. Section 2 reviews method families, metric and dataset rationales, and the LVLM design space. Section 3 details the experimental setup and paired procedures. Section 4 reports results, per-family analysis, and ablations. Section 5 concludes with practical guidance and limitations.

## 2 Related Work

### 2.1 FSR taxonomy and trade-offs

FSR methods are commonly grouped as general, prior-guided, reference-based, multi-task, and blind approaches. They balance three objectives: (i) distortion fidelity (PSNR/SSIM) [2], (ii) perceptual realism (LPIPS/NIQE) [2], and (iii) identity preservation (recognition consistency). Moving toward realism can hurt distortion metrics; rigid priors preserve identity but limit editability; blind pipelines improve robustness yet remain sensitive to degradation assumptions [2].

#### 2.1.1 General, Prior-Guided, Blind FSR

General FSR (Real-ESRGAN) [1], [2] introduces realistic degradations in training with adversarial/perceptual objectives to stabilize textures under noise and compression. Strengths include texture realism and deployment robustness; weaknesses include over/under-restoration when test degradations diverge from the synthetic mixture. Inference control is limited (scale, tiling, mild post-filters), making external guidance conservative by design [1]. The safest LVLM roles are artifact auditing (e.g., halos/color shift) and panel-level accept/reject prior to saving; direct parameter steering is intentionally minimal to avoid overshoot [1].

Prior-Guided FSR (GFPGAN) [7] leverages a facial prior (GAN latent/encoder) to stabilize identity and facial structure. Rigidity yields strong identity but limited editability at inference; exposed knobs (e.g., restoration weight, upscale, only_center_face) are coarse and mostly trade stability vs. detail [7]. Prior studies (LLV-FSR) [5] that inject LVLM cues into prior-guided regimes report structural gains, yet controllability remains bounded by the fixed prior and narrow interfaces. Effective LVLM roles are safe parameter hints (e.g.,

weight ranges, center-face on/off) and auditing (identity consistency, artifacts); deep semantic steering likely requires retraining or new hooks [5], [7].

Blind FSR (BSRGAN) [6] models unknown degradations via stochastic kernels, resizing, noise, and compression to learn resilient restoration. Its strength is wide-range robustness; its weakness is sensitivity to mis-specified degradation at test time. Because degradation is partially non-identifiable from LR alone, wrong assumptions can induce scale mismatch, ringing, over-smoothing, or color shifts [6]. LVLMs can add conservative degradation tags (e.g., "heavy JPEG," "motion blur likely") to route preprocessing or choose an upscale factor, but over-confident tags can harm performance; policies should favor low-risk defaults and limited retries [6].

### 2.2 LVLM CAPABILITIES AND SUPERVISION SIGNALS

Modern LVLMs [3], [4] (e.g., GPT-4V, Gemini) perform image-grounded reasoning, generate natural-language critiques, and emit lightweight tags (e.g., "ringing," "oversharpening," "compression blockiness"). They also support accept/reject QA workflows and can propose coarse parameter hints. These signals are attractive because they are model-agnostic and inexpensive to integrate, but they are probabilistic and prompt-sensitive, which motivates conservative policies [3], [4].

## 3 EXPERIMENTAL SETUP

### 3.1 ENVIRONMENT

Table 1 shows the operating system, hardware, and framework used in our experiments. Fixed seeds ensure reproducibility across retries, while FP32/FP16 precision settings balance numerical stability with efficiency.

<Table 1> Experimental environment configuration for reproducibility and consistency across runs.

| Item | Spec |
|---|---|
| OS | Windows (x64) |
| GPU | 1 x NVIDIA (≥ 8 GB VRAM) |
| Framework | Python 3.10, PyTorch 2.x, CUDA |
| Precision | FP32, optional FP16 |
| Reproducibility | Fixed seeds for data sampling and retries |

### 3.2 DATASET & METRICS

We use FFHQ [1], [2] for its identity diversity, pose/expression coverage, and high-quality HR images that reduce label noise for recognition analysis. LR–HR pairs are synthesized with a Real-ESRGAN-style degradation pipeline [1], [2] (blur → resize → noise → blur → JPEG), which better approximates real degradations than bicubic alone and stresses robustness to multi-stage distortions.

To generate LR inputs, we apply a five-stage degradation pipeline adapted from Real-ESRGAN [1] [2]:

- Kernel-1 (blur-A): random anisotropic kernel, size $k \in \{7,9,11\}$, $\sigma \in [0.2,3.0]$ [1].
- Resize: random up/down/keep, scale $s \in [0.5,2.0]$, then back to target [1].
- Noise: Gaussian $\sigma_n \in [0,15]$ or Poisson (optional grayscale channel) [1].
- Kernel-2 (blur-B): fresh kernel, same ranges as blur-A [1].
- Compression: JPEG quality $q \in [30,90]$ [1].
- I/O: read HR from inputs/high_res_imgs/, write LR to inputs/low_res_imgs/.

This procedure ensures realistic degradations while recording explicit parameter ranges for reproducibility.

For evaluation, we adopt PSNR/SSIM (distortion fidelity), LPIPS/NIQE (perceptual realism), and ArcFace cosine (identity preservation). This triad follows common practice: PSNR/SSIM may reward smoothness over detail, LPIPS/NIQE track perceived texture quality, and recognition features capture identity stability under perceptual changes [2]. Conflicts are expected (e.g., LPIPS↓/NIQE↓ may not raise PSNR/SSIM); reporting all three avoids metric myopia [1], [2].

Table 2 summarizes the dataset choice, the identical LR inputs for both FSR-only and FSR+LVLM runs, the metrics reported, and the artifacts saved for verification.

<Table 2> Dataset and evaluation metrics

| Aspect | Setting |
|---|---|
| Data | FFHQ HR; synthetic LR from the same HR set |
| Inputs | Identical LR for Pure and +LVLM runs |
| Outputs | Per-model folders: pure-fsr, fsr+lvlm, side-by-side panels |
| Metrics | PSNR↑, SSIM↑, LPIPS↓, NIQE↓, ArcFace Cosine↑ |

### 3.3 LVLM GUIDANCE DESIGN

LVLM supervision is lightweight and bounded.

- Inputs**:** downsampled LR image, current SR output, 3–5 cropped face patches, and a compact QA summary (ArcFace, LPIPS, NIQE, artifact flags).
- Outputs: (i) an audit score [0–100], (ii) diagnostic tags {blockiness, ringing, oversharpening, color shift}, and (iii) bounded

hints per family (see Table 4).

- Policy: one pass plus up to R retries. Only safe knobs are applied; if LVLM confidence < τ_conf, the baseline SR is accepted without modification.

This design constrains LVLM to act as an auditor and gatekeeper rather than an unrestricted generator.

### 3.4 Generation Procedure

We run paired pipelines with fixed seeds for reproducibility. The process has three steps:

1. Pure FSR: Run each family (general / blind / prior-guided) on the LR batch to obtain baseline SR outputs (PNG).
2. FSR+LVLM: Run the same LR batch with LVLM supervision, allowing only bounded knobs (see Table 4) per family, producing +LVLM SR outputs (PNG).
3. Artifact Saving: For each case, save side-by-side panels (LR | Pure | +LVLM) and a run-config CSV that records seeds and knob values for reproducibility.

<Table 3> Allowed LVLM knobs during generation

| Family | LVLM-allowed knobs | Notes |
|---|---|---|
| General (Real-ESRGAN) | mild color-balance, dehalo, gentle detail dampening | No structure edits |
| Blind (BSRGAN) | ×2 / ×4 scale switch, low denoise, JPEG deblock | One conservative change only |
| Prior-guided (GFPGAN) | restore-weight band, only_center_face toggle | Fixed prior; small nudges |

Table 3 bounds LVLM's control surface; differences in these limits help explain method-specific effects.

A unified QA schema (score 0–100, reason text) was applied with τ_conf = 80 and ≤1 retry. Prompts varied slightly by model—artifact scoring (Real-ESRGAN), degradation tagging (BSRGAN), or mild parameter hints (GFPGAN)—and all outputs were parsed by a common evaluation script (vlm_quality_score) to ensure consistent, bounded supervision.LVLM prompt templates Configuration

### 3.5 Evaluation Procedure

Evaluation is performed offline on saved images with identical preprocessing:

1. Preprocessing: Convert HR references and SR outputs (Pure, +LVLM) to a standardized color space and range. Face alignment is kept identical across variants.
2. Metric Computation: From standardized tensors, compute PSNR, SSIM, LPIPS, NIQE, and ArcFace Cosine similarity.
3. Aggregation: Report per-image metrics, compute family-wise means, and calculate Δ = (+LVLM − Pure). Tables and plots summarize the results.

ArcFace uses the same crops for Pure and +LVLM to ensure fairness. All outputs are stored as lossless PNG in sRGB.

## 4 Results & Discussion

### 4.1 Quantitative Results

Across families, +LVLM causes consistent drops in blind SR (↓PSNR/SSIM, ↑LPIPS/NIQE, ↓ArcFace), produces negligible shifts in general SR, and leaves prior-guided SR essentially flat on objective metrics.

<Table 4 > Distortion metrics (PSNR/SSIM)

*Higher is better, bold marks the better variant within each family*

| Method Group | Variant | PSNR ↑ | SSIM ↑ |
|---|---|---|---|
| Blind FSR | Pure FSR | **31.093** | **0.896** |
| | FSR+LVLM | 28.592 | 0.886 |
| General FSR | Pure FSR | **30.251** | **0.905** |
| | FSR+LVLM | 30.155 | 0.903 |
| Prior-Guided FSR | Pure FSR | **30.036** | **0.852** |
| | FSR+LVLM | **30.036** | **0.852** |

<Table 5> Perceptual & identity metrics (LPIPS/NIQE/ArcFace)

*LPIPS/NIQE: lower is better; ArcFace: higher is better. Bold marks the better variant within each family*

| Method Group | Variant | LPIPS ↓ | NIQE ↓ | ArcFace Cosine ↑ |
|---|---|---|---|---|
| Blind FSR | Pure FSR | **0.443** | **5.523** | **0.920** |
| | FSR+LVLM | 0.524 | 6.176 | 0.876 |
| General FSR | Pure FSR | **0.474** | **9.422** | **0.909** |
| | FSR+LVLM | 0.483 | 9.454 | 0.908 |

| | | | | |
|---|---|---|---|---|
| Prior-Guided FSR | Pure FSR | **0.259** | **5.756** | **0.884** |
| | FSR+LVLM | **0.259** | **5.737** | **0.883** |

### 4.2 Per-family analysis

For General FSR, we observed negligible changes between Pure FSR and FSR+LVLM. Metrics remained within measurement noise across all images. These results suggested limited advantage for LVLM advice over a strong upsampler.

For Blind FSR, we observed consistent degradations under LVLM guidance with BSRGAN. We attribute the drops to limited controllability at inference, not to a fully black-box model. The pipeline exposes only coarse knobs (scale and mild post-filters), so LVLM cues cannot steer internal priors. Degradation is non-identifiable in blind SR, and misclassification propagates to scale and filters, which reduces perceptual quality and identity scores.

For Prior-Guided FSR, we observed high LVLM quality scores despite unchanged objective metrics. GFPGAN's StyleGAN prior dominated restoration and constrained controllability. LVLM acted as a parameter tuner and auditor rather than a semantic controller.

### 4.3 Mitigations Tried

We increased retry counts and raised QA thresholds. We observed more compute and more artifacts without consistent gains. We reduced filters and used conservative color/dehalo settings. We observed fewer side effects but still no metric gains in BSRGAN. We tuned GFPGAN weight and center-face options. We observed minor perceptual shifts without PSNR/SSIM changes.

Our LVLM remained an external planner and auditor. Our SR models exposed only coarse knobs. This mismatch limited measurable improvements.

### 4.4 Ablation and QA Behavior

Raising thresholds increased retries with little benefit; lowering thresholds missed severe artifacts. Moderate correction strengths worked best; aggressive settings amplified artifacts. Conservative policies are recommended.

### 4.5 Limitation and Threats to Validity

This study limits scope to three FSR categories and a single dataset subset. This study relies on API-based LVLMs without access to internal embeddings. This study controls parameters via coarse interfaces that restrict semantic adjustments.

External validity may suffer under different degradations or domains. Construct validity may depend on the chosen QA prompts and scoring rules. Future replications should report prompts, seeds, and code for reproducibility.

The literature often reports LVLM gains under prior-guided setups. Those setups expose priors and structures that accept semantic cues. Our pipelines expose limited hooks, which constrains control.

General SR baselines already recover textures effectively. LVLM advice yields small or neutral changes when headroom is low. Blind SR depends on accurate degradation inference. Misclassification compounds errors through the whole chain.

GFPGAN preserves identity with a fixed generative prior. LVLM approval reflects perceived plausibility, not metric change. This gap explains high LVLM QA versus flat PSNR/SSIM.

Our system treats BSRGAN as black-box-like at the interface level. Internal degradation embeddings or priors are not exposed to LVLM signals. This interface mismatch limits corrective actions to scale and mild post-filters, so measurable gains are unlikely without deeper coupling.

### 4.6 Practical guidance and future directions

Use LVLM mainly as a gatekeeper: reference/consistency checks, artifact screening, and fail-case retries. Avoid over-steering generation when model knobs are coarse. Future work should expose controllable hooks; align degradation tags with internals, and couple QA signals with training or adaptation.

## 5 Conclusion

We compared FSR-only against FSR+LVLM across general, blind, and prior-guided pipelines. General SR shows near-neutral changes, suggesting little headroom for external advice. Blind SR degrades when guidance misclassifies degradations and steers coarse knobs. Prior-guided SR benefits from auditing but resists structural change under fixed priors. Across families, LVLM works best as a validator and gatekeeper, not a direct controller. Practical use should prioritize reference checks, artifact screening, and safe retries. Consistent metric gains will likely require deeper, controllable hooks inside SR models.

## 6 References

[1] Xintao Wang et al., "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," ICCV, 2021.

[2] (Review) Deep Learning for Face Super-Resolution: A Techniques Review, Springer, 2022.

[3] OpenAI, "GPT-4V(ision) Technical Report,"arXiv:2309.17421, 2023.

[4] Google DeepMind, "Gemini Pro Vision," Technical Report, 2024.

[5] LLV-FSR: Exploiting Large Language-Vision Prior for Face Super-Resolution, Proc. CVPR, 2024.

[6] Kai Zhang et al., "Designing a Practical Degradation Model for Deep Blind Image Super-Resolution," ICCV, 2021.

[7] Xintao Wang et al., "GFPGAN: Generative Facial Prior for Blind Face Restoration," arXiv:2101.04061, 2021.