Enhancing VLM Perception Through Adaptive Super-Resolution: A Multi-Agent Feedback Approach

Rajib Chandra Ghosh[1,*], Egidia Cirillo[1], Alberto Moccardi[1] and Marcello Pelosi[1]

[1]*Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy*

Abstract

Vision-language models have recently demonstrated strong capabilities in multimodal understanding by jointly reasoning over visual and textual data. However, in real-world scenarios, their performance remains highly sensitive to input image quality, with low-resolution inputs often leading to incomplete or inaccurate semantic interpretations. In this work, we propose Agentic SR-VLM, a novel multi-agent framework that enhances VLM perception through adaptive super-resolution guided by semantic feedback. The proposed system consists of three interacting agents: Super-Resolution agent, Vision-Language Model agent, and an Evaluator agent, where all agents are coordinated by an Orchestrator within an iterative feedback loop. Unlike conventional super-resolution approaches that optimize for pixel-level fidelity, our method treats image enhancement as a semantic optimization problem, where improvements are measured based on the quality and richness of generated textual descriptions. The Evaluator agent computes a composite reward score, which guides the adaptive refinement of super-resolution parameters. Experimental results on the test dataset demonstrate that the proposed framework of our significantly improves semantic understanding, achieving an average reward of ~**0.80** in the final stage compared to ~**0.42** for baseline low-resolution inputs at the initial stage. Qualitative analysis further shows enhanced detail recognition, attribute extraction, and contextual grounding in generated descriptions. These findings highlight the effectiveness of integrating super-resolution with semantic feedback in a multi-agent setting, paving the way for more robust and perceptually aware multimodal systems.

Keywords

Super-Resolution, Vision-Language Models, Semantic Feedback, Agentic SR-VLM

1. Introduction

The rapid evolution of large language models (LLMs) has significantly advanced the field of artificial intelligence (AI), enabling systems to perform complex reasoning and generate human-like text across a wide range of applications. Building upon this success, recent research has extended LLM capabilities into the visual domain through vision-language models (VLMs), which jointly process images and text to perform tasks such as image captioning, visual question answering, and multimodal reasoning [1, 2]. Despite these advancements, a critical limitation persists: VLMs are highly sensitive to the quality of visual input, often producing incomplete or inaccurate interpretations when presented with low-resolution or degraded images.

Image super-resolution (SR) has long been studied as a means to reconstruct high-resolution images from low-quality inputs. Traditional approaches, including interpolation-based and reconstruction-based methods, have been largely superseded by deep learning techniques that leverage convolutional neural networks (CNNs) and, more recently, diffusion models to generate high-fidelity outputs [3, 4]. While these methods achieve impressive improvements in pixel-level quality, they typically optimize for low-level metrics such as peak signal-to-noise ratio and structural similarity index, which do not necessarily correlate with improvements in semantic understanding.

Recent efforts have begun to explore the integration of semantic information into the SR process.

WOA 2026: 27th Workshop From Objects to Agents, June 15–17, 2026, Salerno, Italy

*Corresponding author.

rajib.chandraghosh@unina.it (R. C. Ghosh); egidia.cirillo@unina.it (E. Cirillo); alberto.moccardi@unina.it (A. Moccardi); marcello.pelosi2@unina.it (M. Pelosi)

0009-0001-1137-3465 (R. C. Ghosh); 0009-0005-3227-6073 (E. Cirillo); 0009-0001-6136-9368 (A. Moccardi); 0009-0001-4718-4826 (M. Pelosi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-10-30

CEUR-WS.org/Vol-4307/paper32.pdf

For example, SeeSR [5] incorporates semantic priors extracted from pretrained models to guide image reconstruction, producing visually coherent and context-aware outputs. However, such approaches still treat SR as a one-shot enhancement process and lack mechanisms to evaluate whether the reconstructed images lead to better semantic interpretation by downstream models.

In parallel, advancements in multimodal such as Qwen2-VL [6], CLIP [1], and BLIP-2 [7] have demonstrated strong capabilities in aligning visual and textual representations. These models enable the extraction of rich semantic information from images, making them suitable candidates for evaluating the effectiveness of visual enhancement techniques. As presented in Amato et al. [8], a semantic framework for multimedia data processing, highlighting the importance of semantic enrichment in enhancing data interpretation. Nevertheless, existing pipelines typically separate image enhancement from semantic reasoning, preventing dynamic adaptation based on downstream performance.

To address these challenges, we propose Agentic SR-VLM, a novel multi-agent framework that unifies SR and semantic evaluation within an adaptive feedback loop. Our proposed system includes three particular agents: the SR agent, the VLM agent, and the Evaluator agent, which interact under the guidance of the Orchestrator that iteratively refines image quality based on semantic reward. Unlike conventional approaches, we treats SR as an optimization process guided by improvements in semantic understanding rather than purely visual metrics.

The main contributions of our research work are as follows:

- We introduce a multi-agent architecture that integrates SR and vision-language reasoning in a unified framework.
- We propose a semantic reward mechanism that evaluates improvements in image quality based on descriptive richness and semantic content.
- We develop an adaptive feedback loop that dynamically adjusts SR parameters to maximize semantic gain.
- We demonstrate that our approach significantly improves VLM-generated descriptions compared to baseline methods operating on low-resolution inputs.

2. Related Work

Due to the constraints of current methodologies, it becomes necessary to study the literature from two perspectives: SR and VLMs, as well as the new multi-agent approach. This is explained briefly in the next section along with the gap that our methodology aims to bridge.

2.1. Image Super-Resolution

Image SR has been a long-standing research problem in computer vision. Early approaches relied on interpolation techniques such as bicubic upsampling [9], which often produce blurry results. The introduction of deep learning revolutionized the field, with CNNs [10] achieving significant improvements in reconstruction quality. The NTIRE 2017 challenge [3] demonstrated the effectiveness of learning-based SR methods and established benchmarks for evaluating performance.

More recently, generative models, particularly generative adversarial networks [11] and diffusion models [12], have further advanced SR capabilities. SRGAN [4] introduced adversarial training to generate perceptually realistic textures, while diffusion-based approaches have enabled more stable and high-quality image generation. Reconstruction-based models have been used in other areas as well. For example, autoencoders help detect anomalies in messy and complex data [13], illustrating the broader utility of learning to reconstruct the expected structure from imperfect inputs, a principle that also underlies modern generative SR. SeeSR [5] represents a state-of-the-art method that incorporates semantic guidance into the SR process using pretrained models such as Stable Diffusion and ControlNet. However, these methods primarily focus on visual fidelity and do not explicitly optimize for downstream semantic tasks.

2.2. Vision-Language Models

VLMs have become a central component of modern multimodal AI systems. CLIP [1] introduced a contrastive learning framework that aligns image and text embeddings, enabling zero-shot classification and retrieval. BLIP-2 [7] further improved multimodal reasoning by leveraging pretrained language models and lightweight visual encoders. More recently, Qwen2-VL [6] has demonstrated strong performance in generating structured and detailed descriptions from images across diverse domains.

Despite their capabilities, VLMs are inherently dependent on the quality of visual input. Low-resolution or noisy images often result in degraded performance, as important visual details may be lost. Existing research has largely focused on improving model architectures, with limited attention given to enhancing input quality in a task-aware manner.

2.3. Semantic-Aware Image Enhancement

A growing body of work has explored the integration of semantic information into image enhancement tasks. Approaches such as SeeSR and the Recognize Anything Model (RAM) guided diffusion models incorporate object-level and contextual information to guide image reconstruction. These methods demonstrate that semantic priors can improve visual coherence and realism.

However, most existing approaches rely on static pipelines and do not incorporate feedback from downstream tasks. As a result, they cannot adaptively refine images based on whether semantic understanding has improved, limiting their effectiveness in multimodal applications. In recommender systems, for example, Ferraro et al. [14] move from explanation toward exploration by explicitly promoting diversity as a quality objective, demonstrating that optimizing for richer, more informative output can outperform single-metric approaches.

2.4. Multi-Agent and Feedback-Based AI Systems

Multi-agent systems have gained increasing attention as a framework for coordinating multiple specialized components to solve complex problems. In the context of LLMs, agent-based architectures enable iterative reasoning, tool usage, and self-improvement through feedback loops [15, 16]. These systems have been successfully applied to tasks such as planning, code generation, and autonomous decision-making. For instance, Amato et al. [17] proposed an agent-based framework for recommending cultural tours, demonstrating the effectiveness of multi-agent systems in handling complex and dynamic tasks. More recently, Ferraro et al. [18] proposed a generative agent-based modeling approach for insider threat detection, showing that LLM-driven agents can produce diverse and realistic behaviors within a coordinated multi-agent setting.

Despite their potential, the application of multi-agent frameworks to image processing and multimodal learning remains relatively unexplored. In particular, integrating SR with semantic evaluation in a feedback-driven loop represents a novel direction that has not been extensively studied.

From the literature, several key limitations emerge:

- Super-resolution methods optimize for pixel-level fidelity rather than semantic usefulness.
- Vision-language models depend on input quality but lack mechanisms for adaptive enhancement.
- Multi-agent systems provide iterative optimization, yet are rarely applied to visual enhancement tasks.

To address these gaps, we introduce an Agentic SR-VLM framework that jointly optimizes image quality and semantic understanding, bridging the gap between low-level vision and high-level reasoning.

3. Methodology

This paper presents an agent-based framework that applies an SR method to improve VLMs' perception by enhancing image quality and providing semantic feedback on how well the VLM generates object

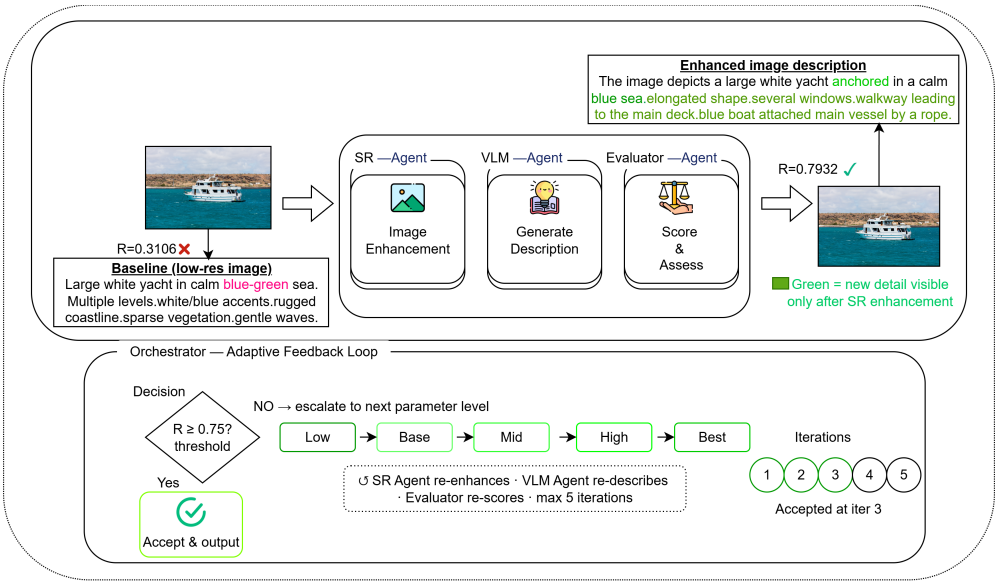

Figure 1: Overview of the proposed **Agentic SR-VLM** framework for enhancing VLM perception through adaptive SR. The pipeline iteratively enhances a low-resolution input image using an **SR Agent** with adaptively selected parameters, generates a natural language description via a **VLM Agent**, and evaluates semantic quality through an **Evaluator Agent** that computes reward **R**. The **Orchestrator** drives the feedback loop, escalating SR parameters until VLM description quality meets the acceptance threshold ($R \geq 0.75$) or the maximum iteration limit is reached. Green text indicates semantic details newly perceived by the VLM only after adaptive SR enhancement

descriptions. The SeeSR [5] model is primarily based on pixel-level fidelity; however, we provide an additional level of semantic feedback regarding the quality of the descriptors generated by the VLM. This proposed agent-based architecture contains three agents that interact through a central Orchestrator: the SR Agent, the VLM Agent, and the Evaluation Agent. Figure 1 illustrates the overall architecture.

3.1. Super-Resolution Agent

The SR Agent is built upon the **SeeSR** framework, a diffusion-based SR model that integrates Stable Diffusion with ControlNet conditioning and RAM. The RAM module extracts semantic tags from the input image, which are embedded and injected into the diffusion process to guide content-aware generation. To balance quality and efficiency, the SR Agent operates over a predefined parameter grid.

$$\mathcal{P} = \{(4.5, 30), (5.5, 50), (6.5, 50), (5.5, 70), (7.5, 50)\}$$

where each pair corresponds to (`cfg_scale`, `num_inference_steps`), and a fixed upscaling factor of 4× is used. The `cfg_scale` parameter controls classifier-free guidance strength. Higher values encourage stronger adherence to semantic priors but may introduce artifacts. The `num_inference_steps` parameter determines the number of denoising iterations, with higher values yielding better visual quality at increased computational cost. The parameter grid is arranged to support a progressive refinement strategy, in which different parameter combinations are explored iteratively to balance semantic consistency, image quality, and computational efficiency.

3.2. Vision-Language Model Agent

The VLM Agent is responsible for generating natural language descriptions from visual content, which are used for semantic evaluation. We employ **Qwen2-VL-2B-Instruct** [6], a multimodal that is capable of generating structured descriptions from images.

Given an input image, the VLM produces a comprehensive description capturing key aspects of the scene, including objects and entities, their visual attributes (e.g., color, shape, condition), spatial

relationships, and overall context. These descriptions provide a high-level semantic representation of the image content.

The VLM Agent performs two tasks sequentially in the pipeline. First, it creates the description based on the initial low-quality image. Next, it generates the enhanced description based on the SR image. Using the same model for both stages ensures consistency, so that differences between descriptions reflect changes in image quality rather than model variation. For the consistency of prompts across all images regardless of domain and content, we rely on a universal prompting scheme that promotes detailed, informative responses without any customization for the task. The VLM Agent serves as the semantic observer in the pipeline. Its role does not include any intervention in the SR process. Instead, it only generates the information that can be used by the Evaluator Agent to detect the progress in the visual domain, whether meaningful visual improvements have been achieved.

3.3. Evaluator Agent

The Evaluator Agent is responsible for quantifying the semantic improvement achieved by the SR process. Rather than relying on only traditional pixel-level metrics (e.g., PSNR, SSIM), it assesses whether the enhanced image enables richer and more informative descriptions compared to the original low-resolution input.

To this end, the Evaluator computes a composite reward score $R \in [0, 1]$ based on four complementary components. Each component captures a different aspect of descriptive quality and semantic gain.

$$R = 0.25W_1 + 0.25W_2 + 0.20W_3 + 0.30W_4 \tag{1}$$

where W_1 is the Word Score, W_2 is the Unique Word Score, W_3 is the Object Score, and W_4 is the Semantic Improvement Score.

Word Score. Measures the length of the enhanced description as a proxy for detail richness. Longer descriptions generally indicate that more visual information has been perceived.

Unique Word Score. Measures vocabulary diversity to penalize repetitive or low-information descriptions, ensuring that increased length corresponds to meaningful content rather than redundancy.

Object Score. Estimates the number of distinct objects present in the scene by counting articles and quantifiers (e.g., “a”, “the”, “two”, “several”). This serves as a lightweight proxy for object-level understanding.

Semantic Improvement Score. Measures the degree of semantic change between the baseline and enhanced descriptions. Sentence embeddings (**all-MiniLM-L6-v2**) are used to capture high-level meaning, allowing the system to detect whether new information has been revealed by the SR process.

These components are combined using fixed weights to produce the final reward score, with greater emphasis placed on semantic improvement. This design encourages the system to prioritize meaningful gains in visual understanding rather than superficial increases in description length.

3.4. Orchestrator and Feedback Loop

The core innovation of our framework lies in its *adaptive orchestration* mechanism, whereby multiple specific agents are coordinated in an iterative process that transforms low-resolution input images into high-resolution representations consistent in terms of semantic fidelity. As shown in Figure 2, the entire process works on a feedback loop controlled not by pixel-by-pixel accuracy, but rather by semantic evaluation.

During initialization, the system begins a refinement process in which the low-quality image is first sent to the SR agent using the most computationally cost-effective set of parameters. The resulting high-quality image is then compared with the reference semantic content created in the earlier step. Importantly, throughout every iteration of the feedback loop, the original low-resolution input image is always used as the source image for the SR process, while the reward module continuously compares

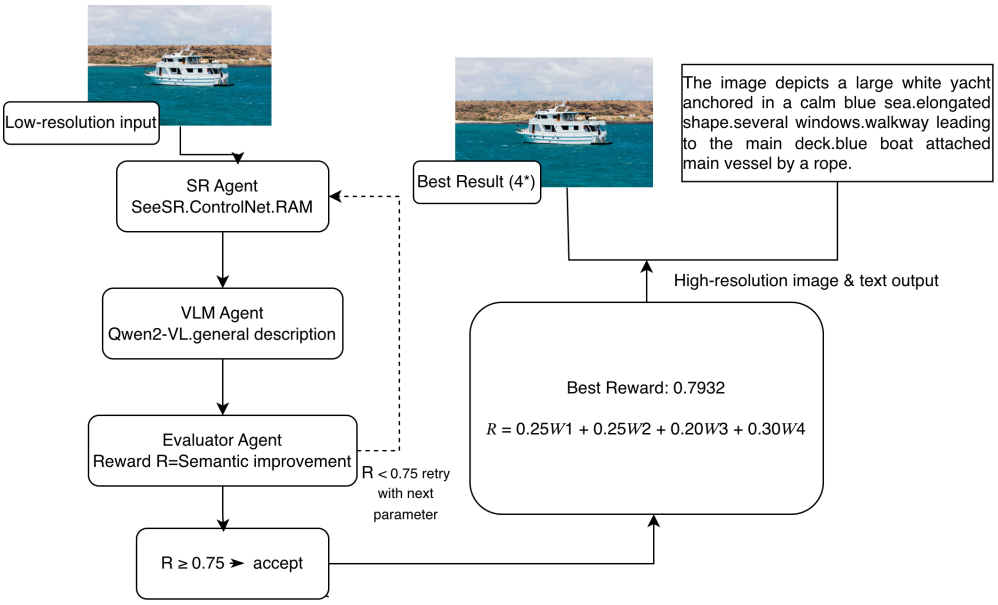

Figure 2: Workflow of the Agentic SR–VLM System

the generated high-resolution output against the semantic description extracted from the initial low-resolution input image. The evaluation module calculates how effectively the visuals from the high-quality image match the semantic content, which yields the semantic consistency score.

In case the score exceeds certain predefined criteria, then the image will be considered as the final output, thus helping in achieving optimal use of the resources by eliminating wastage of computational resources in processing unnecessary computations. But if that's not the case and the score fails to meet the criteria, then Orchestrator activates the process for a retry. In this step, the image generation process proceeds to the next level by using another set of parameters that would generate a better image.

The iteration process of improvement, analysis, and parameter fine-tuning proceeds in an iterative manner. It ends when the value of semantic similarity goes beyond the limit, or the maximum iteration count of five is achieved. This method successfully converts SR from a one-time reasoning procedure into a search operation, whereby complexity increases if the present outcome does not match the meaning of the original image.

4. Experimental Setup

All experiments were conducted on an Intel Ultra 9 275HX processor (24 cores), 32 GB of system memory, and an NVIDIA RTX 5080 GPU with 16 GB of VRAM. The CUDA version used is cu12.8.

The system was implemented in Python version 3.10.18 with GPU acceleration enabled via CUDA, allowing the iterative enhancement and evaluation loop to run efficiently despite the computational demands of diffusion inference and multimodal reasoning.

4.1. Dataset

We evaluate our method on synthetic low-resolution images generated by 4× bicubic downsampling of high-resolution sources. To demonstrate domain generalization on general photographic content, we used the Test2K dataset introduced by Kong et al. [19], which comprises 100 images (indices 1201–1300) from the DIV8K dataset [20] downsampled to 2K resolution. Representative samples are shown in Figure 3.

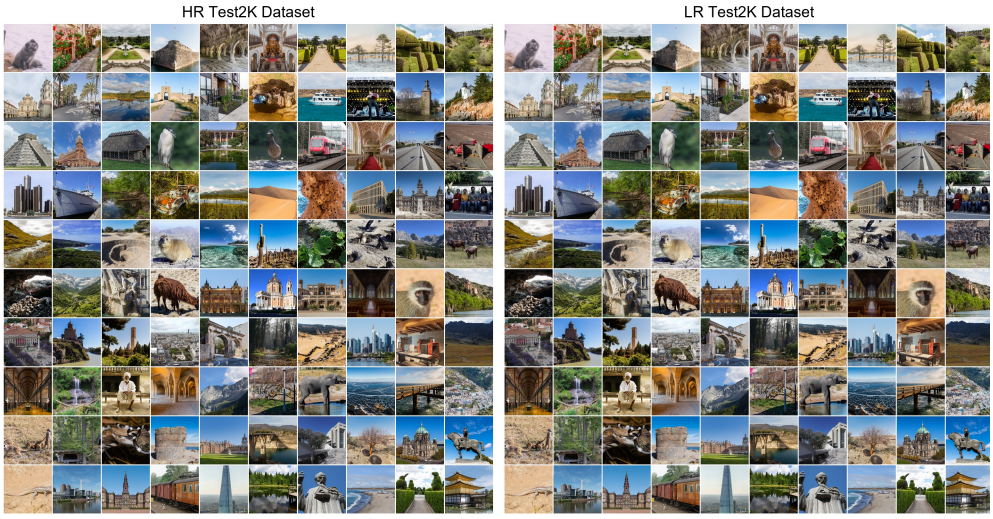

Figure 3: Visualization of Test2K Dataset

4.2. Evaluation Metrics

To comprehensively assess the proposed Agentic SR-VLM framework, we evaluate both the visual quality of the reconstructed images and the semantic fidelity of the generated descriptions. Specifically, we adopt three standard image-quality metrics (PSNR [21], SSIM [22], LPIPS [23]) and as well as semantic alignment metric (R_{HR}) tailored to our task.

Peak Signal-to-Noise Ratio (PSNR). PSNR quantifies pixel-level reconstruction fidelity between the reconstructed image $\hat{I}$ and the ground-truth image I. It is defined as

$$\mathrm{PSNR}(\hat{I}, I) = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}(\hat{I}, I)}\right) \tag{2}$$

where MAX_I is the maximum possible pixel value (255 for 8-bit images), and MSE is the mean squared error between the two images. Higher PSNR values indicate closer pixel-level agreement.

Structural Similarity Index Measure (SSIM). SSIM measures the perceptual similarity between $\hat{I}$ and I by jointly modelling luminance, contrast, and structural information. It is defined as

$$\mathrm{SSIM}(\hat{I}, I) = \frac{(2\mu_{\hat{I}}\mu_I + C_1)(2\sigma_{\hat{I}I} + C_2)}{(\mu_{\hat{I}}^2 + \mu_I^2 + C_1)(\sigma_{\hat{I}}^2 + \sigma_I^2 + C_2)} \tag{3}$$

where μ denotes the local mean, σ^2 the local variance, and $\sigma_{\hat{I}I}$ the covariance between $\hat{I}$ and I. The constants C_1 and C_2 are stabilizing terms used to avoid numerical instability. SSIM values lie in the range $[-1, 1]$, with higher values indicating greater structural similarity.

Learned Perceptual Image Patch Similarity (LPIPS). To evaluate how similar two images are based on human perception, we use LPIPS, which computes the distance between deep features extracted from a pretrained network (AlexNet). It is defined as

$$\mathrm{LPIPS}(\hat{I}, I) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\phi_l(\hat{I})_{h,w} - \phi_l(I)_{h,w}\right)\right\|_2^2 \tag{4}$$

where $\phi_l(\cdot)$ denotes the feature map at layer l, and w_l represents learned channel-wise weights. The symbols H_l and W_l denote the spatial dimensions of the feature map at layer l. Lower LPIPS values

indicate greater perceptual similarity, making LPIPS particularly suitable for evaluating generative SR outputs where pixel-accurate reconstruction is intentionally traded for perceptual realism.

Semantic Alignment Metric (R_{HR}). To evaluate how accurately the generated descriptions preserve the semantic content of the ground-truth HR images, we employ the R_{HR} metric defined as the cosine similarity. The higher the value, the better is the alignment between the semantics of the generated descriptions and the reference ones.

$$R_{HR} = \cos(\mathbf{e}_g, \mathbf{e}_r) = \frac{\mathbf{e}_g \cdot \mathbf{e}_r}{\|\mathbf{e}_g\| \, \|\mathbf{e}_r\|} \tag{5}$$

where $\mathbf{e}_g$ and $\mathbf{e}_r$ denote the embedding vectors of the generated description and the reference (ground-truth HR image) description, respectively. The embeddings are extracted using the `all-MiniLM-L6-v2` sentence-transformer model. Higher values of R_{HR} indicate stronger semantic alignment between the generated and reference descriptions.

5. Results

As shown in Table 1, the average reward achieved by our Agentic SR-VLM is 80.41%, outperforming the baseline approach that takes low-resolution inputs without SR significantly (42.98%). The latter Section 5.1 shows how poor the quality of visual information and semantic understanding will be without any enhancement. Our system manages to satisfy the set reward criterion in 95% of the time, showing consistent and reliable operation in terms of output quality.

Although the LR input achieves higher PSNR and SSIM values (24.85 and 0.71) than the proposed Agentic SR-VLM (23.11 and 0.63), this is expected due to the perception–distortion tradeoff [24], where diffusion-based SR approaches sacrifice pixel-level fidelity for perceptual realism. The significantly lower LPIPS score of the proposed method (0.20) confirms improved perceptual quality and semantic fidelity.

Table 1
Performance comparison between the baseline input and the proposed Agentic SR-VLM framework. **Bold** values denote the best result in each column

Configuration	SR Backbone	Avg. R ↑	Threshold Met	PSNR ↑	SSIM ↑	LPIPS ↓
LR input (no SR)	–	0.4298	–	**24.85**	**0.71**	0.43
Agentic SR-VLM	SeeSR	**0.8041**	**95.00%**	23.11	0.63	**0.20**

As illustrated in Figure 4, in almost all of the tested samples, our framework achieves good results, as 54% have a reward between 0.75 and 0.80, and 41% receive more than 0.81 reward, showing that our method continuously improves beyond minimum performance. It must also be noted that only 5% of tests were unable to pass the reward threshold; in 2% of these cases, the reward was less than 0.75, while in 3% CUDA Out of Memory (OOM) occurred.

5.1. Qualitative Analysis

Further, to investigate the performance of the proposed Agentic SR-VLM, a comparative study is conducted at the initial stage of the Agentic SR-VLM versus the final stage of the Agentic SR-VLM, low-resolution image-generated text versus the SR images-generated text. Figure 5 shows three example scenarios, including an urban street scene, a wildlife close-up, and a natural landscape.

In each case, the initial stage of Agentic SR-VLM (left side), constrained by low-resolution artifacts and blur, produces generic descriptions that often miss fine-grained details or misinterpret contextual cues. For instance, in the street scene, the baseline fails to distinguish specific architectural features, while the lizard description lacks precise morphological details. In contrast, the final stage of Agentic

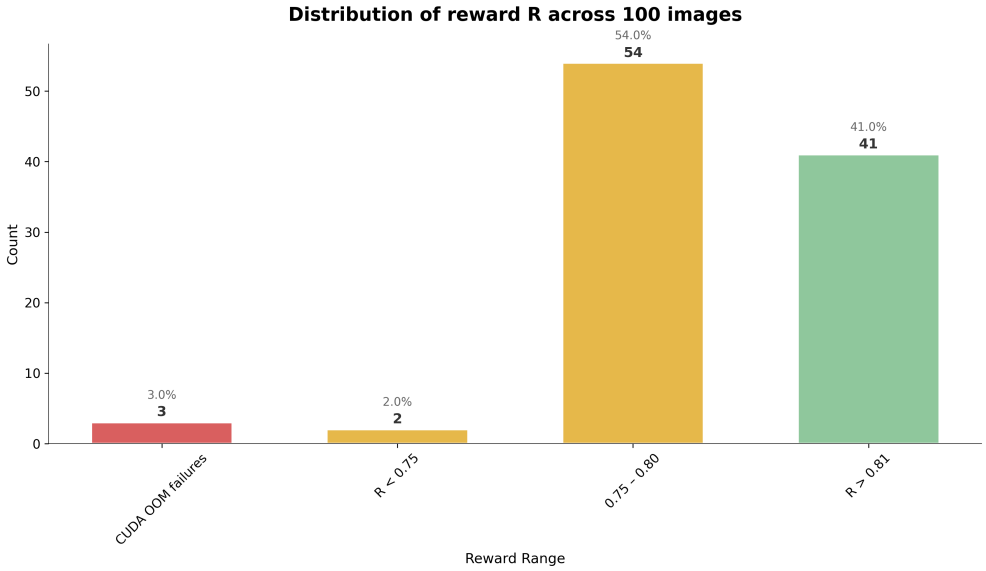

Figure 4: Distribution of test samples under different reward thresholds

SR-VLM (right side), leveraging high-resolution reconstructed inputs, generates significantly richer and more accurate textual descriptions.

Key improvements include:

- **Enhanced Detail Recognition:** The Agentic SR-VLM correctly identifies specific elements such as *hanging flower pots*, *cobblestone patterns*, and *weathered stone facades* in the urban scene.
- **Accurate Attribute Extraction:** In the wildlife examples, the agent captures subtle visual traits, such as the lizard's *smooth yet textured skin* and the duck's *speckled brown plumage*, and *orange legs*.
- **Contextual Grounding:** The generated text better reflects the environmental context, noting the *desert-like habitat* for the lizard and the *natural sunlight* conditions for the duck, which were ambiguous in the low-resolution counterparts.

These results demonstrate that SR preprocessing enables the VLM to access high-frequency visual information, leading to more semantically grounded and descriptive outputs.

5.2. Semantic Fidelity Analysis

Table 2 presents the metric R_{HR} between the baseline LR image descriptions and the proposed Agentic SR-VLM framework-generated descriptions, both compared against the ground-truth HR image descriptions. The LR descriptions achieve a score of 0.6937 when compared with the ground-truth HR descriptions. In contrast, the Agentic SR-VLM descriptions obtain a significantly higher score of 0.8638, resulting in an absolute improvement of 0.1701.

Table 2
Comparison of LR and Agentic SR-VLM description against the ground-truth HR description

Method	R_{HR}
LR description vs ground-truth HR description	0.6937
Agentic SR-VLM description vs ground-truth HR description	**0.8638**
Improvement (Δ)	+0.1701

This shows that the introduced Agentic SR-VLM can yield more accurate descriptions semantically since it is able to retrieve more fine-grained visual details that are often not available in LR input

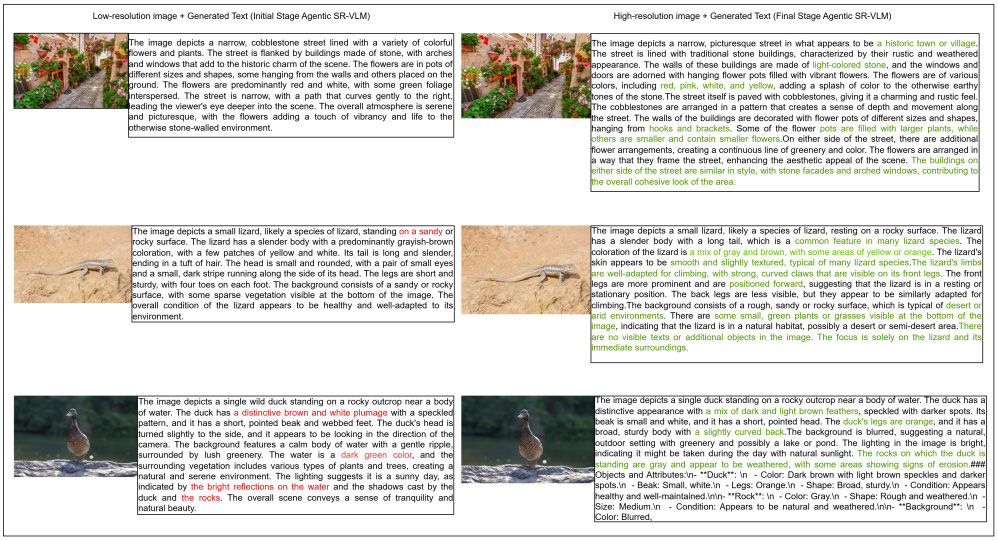

Figure 5: Comparison of generated textual **initially from low-resolution images** (left) versus **after high-resolution images** (right) processed by Agentic SR-VLM. The after SR-VLM outputs demonstrate significantly richer detail, contextual accuracy, and semantic grounding due to enhanced visual perception

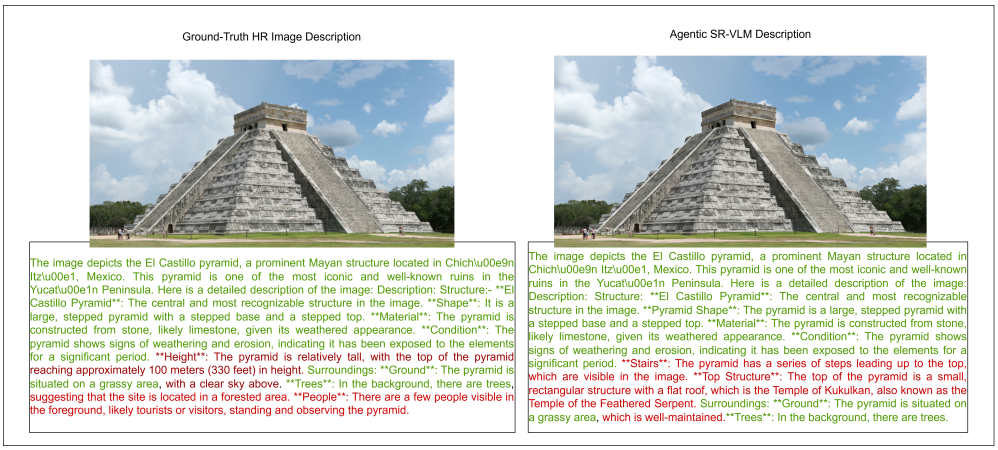

Figure 6: Semantic comparison of the ground-truth HR image description and the Agentic SR-VLM-generated description. Sentences shown in **green** represent semantically aligned content between the two descriptions, whereas sentences shown in **red** denote distinct semantic information unique to each description

images. As illustrated in Figure 6, the description generated by the Agentic SR-VLM more closely aligns with the ground-truth HR image description. In addition, the Agentic SR-VLM successfully captures fine-grained details that are missing in the baseline descriptions, including the geometry of the stairs and the rectangular structure with a flat roof.

5.3. Processing Time and System Performance Analysis

The performance cost of the proposed Agentic SR-VLM was estimated using a set of 100 images using an *NVIDIA RTX 5080*. As shown in Table 3, the average inference time is 137.2 seconds (~2.3 minutes), while the median values 120.6 seconds. The execution time depends on the complexity of the image; simple images can be processed within 17.99 seconds, while difficult cases that require all five agentic steps take up to 748.8 seconds. It should be noted that 88% of the images are processed within 13 minutes, while the model is highly stable, with only a 3.0% error due to memory issues (CUDA OOM). The total execution time for the entire dataset was approximately 3.7 hours.

Table 3
Summary of processing time and system performance across 100 images

Metric	Value	Interpretation
Mean processing time	137.2s	~2.3 min per image on RTX 5080
Median processing time	120.6s	Most images processed in ~2 min
Min processing time	17.99s	Simple images accepted very fast
Max processing time	748.8s	Hard image, ran all 5 iterations
Images processed in <60s	9 / 100	9.0% very fast acceptance
Images processed in 60*s* <800s	88 / 100	88.0% under ~13 min
CUDA OOM failures	3 / 100	3.0% failure rate memory issue
Total dataset time	~3.7 hrs	Full experiment runtime on single GPU

6. Comparison Between Systems

To determine the advantage of our proposed Agentic SR-VLM over standard LLMs in terms of additional information provided, we perform a direct comparison on a low-resolution input image, as shown in Figure 7. As a baseline, we use ChatGPT (Free Tier, chatgpt.com), where it generates text output after performing image enhancement. Our Agentic SR-VLM, on the other hand, uses its capabilities to enhance the input image through SR and further applies VLM to generate a descriptive summary of the enhanced image.

From Figure 7, it can be observed that although ChatGPT (Free Tier) generates a generally plausible description, its interpretation remains generic and partially inaccurate. In particular, the free-tier incorrectly infers that the yacht is *cruising* across the water, whereas the low-resolution input image clearly indicates that the vessel is stationary and anchored. This demonstrates how the generated high-resolution reconstruction can mislead the textual interpretation and introduce semantic hallucinations. In contrast, our Agentic SR-VLM produces a more reliable and semantically rich understanding of the scene by accurately recognizing structural components (*canopy*, *walkway*, *observation deck*), human-related activities (*crew/passengers engaged in activities*), environmental context (*rocky shoreline*, *sparse shrubs/trees*, *calm water surface*), and even meta-information (*no visible text in the image*). Furthermore, by generating a faithful enhanced high-resolution image, the proposed Agentic SR-VLM enables downstream applications that rely on precise spatial understanding, such as object counting, pose estimation, and fine-grained scene reasoning.

7. Proof of Concept

We implemented an end-to-end PoC system to demonstrate the practical feasibility of the proposed Agentic SR-VLM framework as a web application. A screenshot of our application is shown in Figure 8, which depicts the low-resolution image of a cobblestone street processed by our framework. All important parameters used in our pipeline, such as reward threshold, number of iterations, and optional hints for prompts have been exposed in the sidebar.

Once the user uploads the low-resolution image and starts the operation by using the *Run Agentic Pipeline* command, the program automatically runs the agent loop involving SR, VLM, and Evaluator agents in sequence until either the threshold for the rewards is met or until the maximum number of iterations is reached. After that, the main dashboard offers a complete analysis of the results: the initial low-resolution input image (left side) and enhanced output image (right side) are shown for comparison, allowing the user to visually assess the enhancement. Importantly, the system offers a comparison of the semantics: the description created by the VLM model for the low-resolution image input (Baseline Description) is compared with the more detailed description obtained from the enhanced image (Best Description). This allows for a qualitative demonstration of the connection between visual enhancement and semantic quality. Finally, a separate metrics panel shows the efficiency of the run by providing the following information: Best Reward received (for example, 0.886), Iterations Run (for

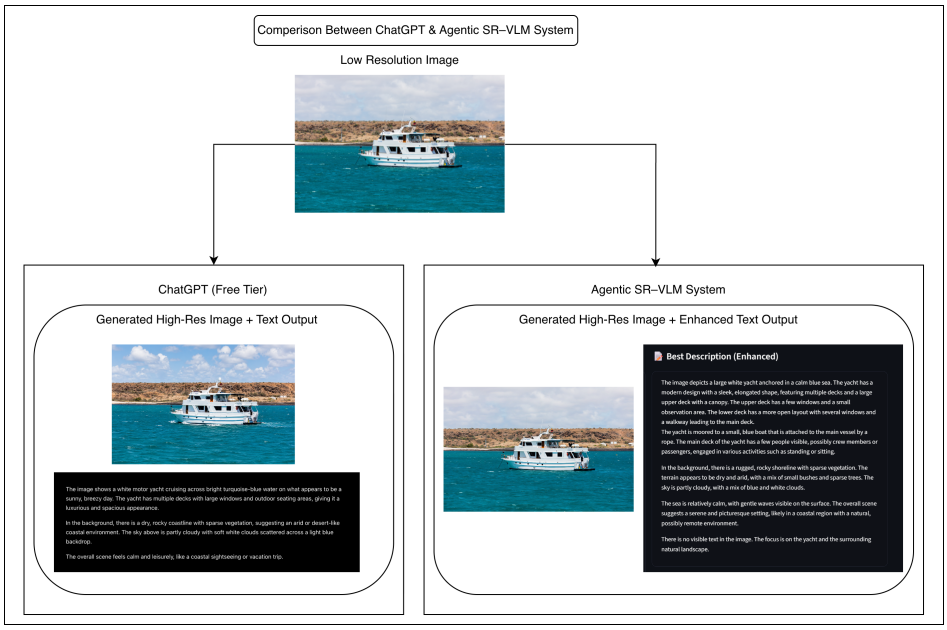

Figure 7: Comparison of output between **ChatGPT (Free Tier)** and our proposed **Agentic SR-VLM** given identical low-resolution image input

example, 2), Time elapsed (for example, 1043.2 seconds), and a boolean variable Threshold Was Met.

8. Discussion & Limitation

The experimental results demonstrate that integrating adaptive SR with semantic feedback significantly enhances the descriptive capability of VLMs. By shifting the optimization objective from pixel-level fidelity to semantic usefulness, the proposed Agentic SR-VLM framework enables more meaningful visual understanding, as reflected in both quantitative reward improvements and qualitative gains in descriptive richness. The multi-agent design further allows the system to iteratively refine outputs, ensuring that computational effort is allocated only when semantic improvement is necessary. Moreover, the multi-agent paradigm provides a flexible and extensible framework for integrating diverse models and evaluation strategies.

Despite these advantages, several limitations remain. First, the framework introduces additional computational overhead due to its iterative multi-agent feedback loop. During experimentation, we observe hardware resource constraints, particularly in GPU memory. In a small number of cases (3% of the test dataset), the system failed to generate both the SR image and the corresponding textual description due to CUDA OOM. These failures occurred primarily in computationally intensive scenarios requiring multiple iterations with high inference steps. This highlights a scalability challenge of the proposed multi-agent framework, especially when deployed on resource-limited hardware.

9. Conclusion

In this paper, we introduced Agentic SR-VLM, a novel multi-agent framework designed to enhance VLM perception through adaptive SR guided by semantic feedback. Unlike conventional approaches that focus solely on pixel-level reconstruction, our method formulates SR as a semantic optimization problem, where improvements are driven by the quality of generated textual descriptions. Our experimental results on the test set demonstrate that this approach significantly improves semantic understanding, achieving higher descriptive richness and contextual accuracy compared to baseline methods operating on low-resolution inputs. Those findings highlight the importance of task-aware image enhancement, showing that optimizing for downstream semantic performance can lead to more meaningful and

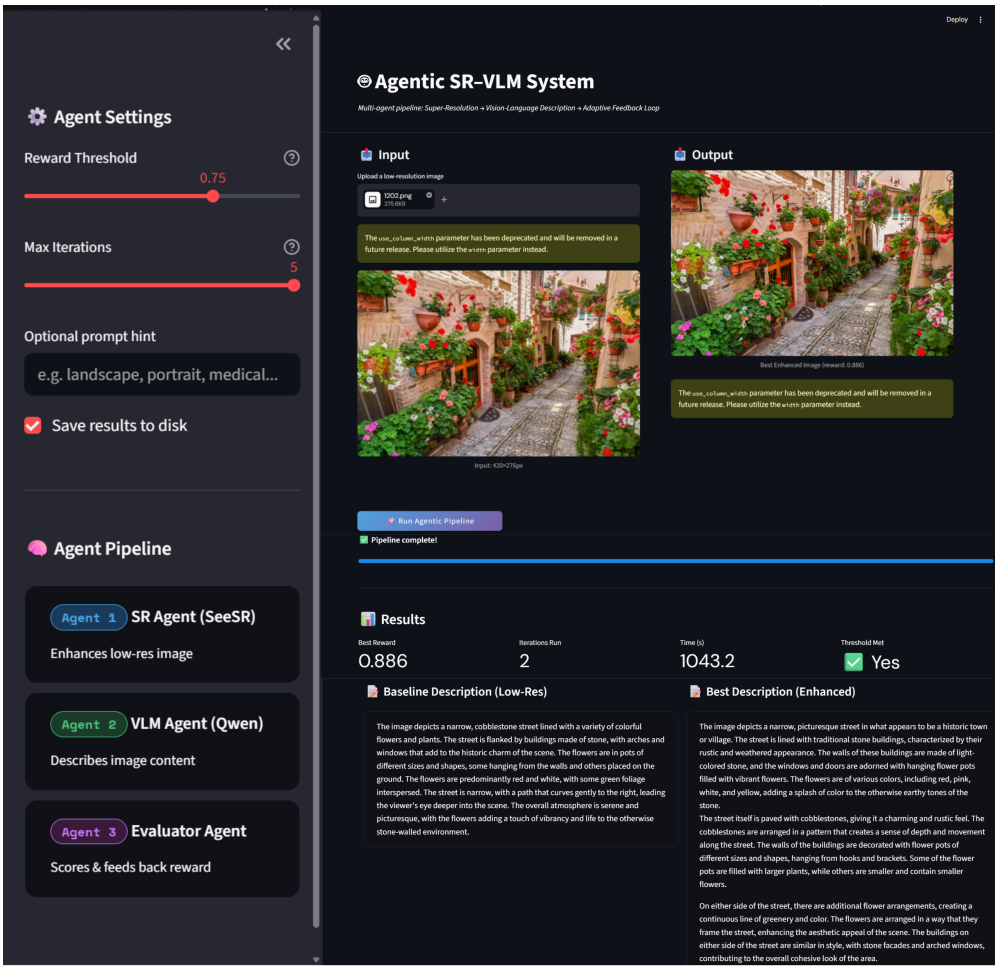

Figure 8: Proof of Concept of the Agentic SR-VLM System

interpretable results. Moreover, the multi-agent paradigm provides a flexible and extensible framework for integrating diverse models and evaluation strategies.

Overall, this work bridges the gap between low-level visual processing and high-level multimodal reasoning, offering a new perspective on how image enhancement techniques can be aligned with semantic objectives. We believe that this approach opens promising directions for future research in agentic multimodal AI, particularly in the agricultural domain, where robust perception under challenging visual conditions is required, such as low-resolution crop imagery, disease detection, and field monitoring.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[2] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, Proceedings of Machine Learning Research, PMLR, 2022, pp. 12888–12900. URL: https://proceedings.mlr.press/v162/li22n.html.

[3] R. Timofte, E. Agustsson, L. V. Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, W. Bae, J. Yoo, Y. Han, J. C. Ye, J.-S. Choi, M. Kim, Y. Fan, J. Yu, W. Han, D. Liu, H. Yu, Z. Wang, H. Shi, X. Wang, T. S. Huang, Y. Chen, K. Zhang, W. Zuo, Z. Tang, L. Luo, S. Li, M. Fu, L. Cao, W. Heng, G. Bui, T. Le, Y. Duan, D. Tao, R. Wang, X. Lin, J. Pang, J. Xu, Y. Zhao, X. Xu, J. Pan, D. Sun, Y. Zhang, X. Song, Y. Dai, X. Qin, X.-P. Huynh, T. Guo, H. S. Mousavi, T. H. Vu, V. Monga, C. Cruz, K. Egiazarian, V. Katkovnik, R. Mehta, A. K. Jain, A. Agarwalla, C. V. S. Praveen, R. Zhou, H. Wen, C. Zhu, Z. Xia, Z. Wang, Q. Guo, Ntire 2017 challenge on single image super-resolution: Methods and results, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1110–1121. doi:10.1109/CVPRW.2017.149.

[4] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, 2017. URL: https://arxiv.org/abs/1609.04802. arXiv:1609.04802.

[5] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, L. Zhang, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution , in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2024, pp. 25456–25467. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02405. doi:10.1109/CVPR52733.2024.02405.

[6] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[7] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[8] F. Amato, F. Colace, L. Greco, V. Moscato, A. Picariello, Semantic processing of multimedia data for e-government applications, Journal of Visual Languages and Computing 32 (2016) 35 – 41. URL: https://www.scopus.com/pages/publications/84949036166?origin=resultslist. doi:10.1016/j.jvlc.2015.10.012, cited by: 34.

[9] D. Khaledyan, A. Amirany, K. Jafari, M. H. Moaiyeri, A. Z. Khuzani, N. Mashhadi, Low-cost implementation of bilinear and bicubic image interpolation for real-time image super-resolution, in: 2020 IEEE Global Humanitarian Technology Conference (GHTC), 2020, pp. 1–5. doi:10.1109/GHTC46280.2020.9342625.

[10] K. O'Shea, R. Nash, An introduction to convolutional neural networks, 2015. URL: http://arxiv.org/abs/1511.08458. arXiv:1511.08458.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144. URL: https://doi.org/10.1145/3422622. doi:10.1145/3422622.

[12] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3626235. doi:10.1145/3626235.

[13] A. Ferraro, A. Galli, V. L. Gatta, M. Postiglione, A Deep Learning pipeline for Network Anomaly Detection based on Autoencoders, in: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), IEEE, 2022, p. 75. doi:10.1109/MetroXRAINE54828.2022.9967598.

[14] A. Ferraro, A. Galli, V. La Gatta, M. Postiglione, D. Russo, G. M. Orlando, G. Riccio, A. Romano, V. Moscato, From explanation to exploration: Promoting diversity in recommendation systems, in: L. Boratto, A. De Filippo, E. Lex, F. Ricci (Eds.), Recommender Systems for Sustainability and Social Good, Springer Nature Switzerland, Cham, 2025, pp. 135–150. doi:10.1007/978-3-031-87654-7_

13.
[15] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, 2023. Publisher Copyright: © 2023 11th International Conference on Learning Representations, ICLR 2023. All rights reserved.; 11th International Conference on Learning Representations, ICLR 2023 ; Conference date: 01-05-2023 Through 05-05-2023.
[16] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. rong Wen, A survey of large language models, 2023. URL: https://api.semanticscholar.org/CorpusID:257900969.
[17] F. Amato, F. Moscato, V. Moscato, F. Pascale, A. Picariello, An agent-based approach for recommending cultural tours, Pattern Recognition Letters 131 (2020) 341 – 347. URL: https://www.scopus.com/pages/publications/85078235683?origin=resultslist. doi:10.1016/j.patrec.2020.01.005, cited by: 27.
[18] A. Ferraro, G. M. Orlando, D. Russo, Generative agent-based modeling with large language models for insider threat detection, Engineering Applications of Artificial Intelligence 157 (2025) 111343. URL: https://www.sciencedirect.com/science/article/pii/S0952197625013454. doi:https://doi.org/10.1016/j.engappai.2025.111343.
[19] X. Kong, H. Zhao, Y. Qiao, C. Dong, Classsr: A general framework to accelerate super-resolution networks by data characteristic, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12011–12020. doi:10.1109/CVPR46437.2021.01184.
[20] S. Gu, A. Lugmayr, M. Danelljan, M. Fritsche, J. Lamour, R. Timofte, Div8k: Diverse 8k resolution image dataset, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3512–3516. doi:10.1109/ICCVW.2019.00435.
[21] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electronics Letters 44 (2008) 800. doi:10.1049/el:20080522.
[22] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.
[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595. doi:10.1109/CVPR.2018.00068.
[24] Y. Blau, T. Michaeli, The perception-distortion tradeoff, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6228–6237. doi:10.1109/CVPR.2018.00652.

A. Online Resources

The resources and pretrained models used in our pipeline are available at:

- Test2K dataset
- SeeSR
- Qwen2-VL-2B-Instruct
- all-MiniLM-L6-v2