VHA: VQA-BASED HALLUCINATION ASSESSMENT FOR GENERATIVE IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image super-resolution produces visually realistic high-resolution images, but may also introduce hallucinations by adding unsupported content or altering existing details. Because existing super-resolution metrics fail to measure whether generated details remain grounded in the reference image, we present a hallucination assessment technique based on bidirectional visual question answering (VQA), which we term VQA-based Hallucination Assessment (VHA). The proposed approach uses a vision-language model (VLM) to determine whether the semantic content and low-level visual evidence in a super-resolved image are supported by its ground-truth counterpart. To this end, it constructs question–answer pairs from each image and evaluates them on both images. Discrepancies between the resulting scores in both directions reveal missing or altered ground-truth information and unsupported additions. However, obtaining these scores requires computationally expensive VLM inference. We therefore develop VHA+, a learned, lightweight model that directly predicts VHA scores from ground-truth and super-resolved image pairs. We construct a hallucination benchmark, VHA-bench, through a human study and show that both VHA and VHA+ correlate highly with human judgments of hallucination. We will publicly release our code, benchmark, and model weights to support reproducibility.

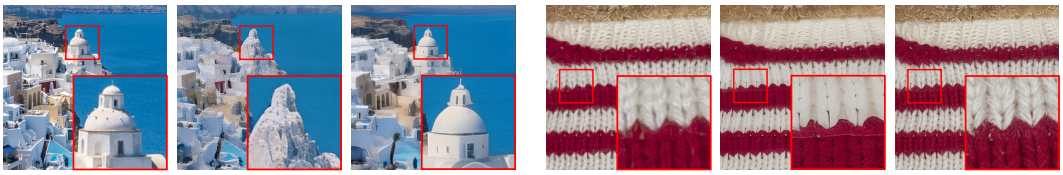
1 INTRODUCTION

Generative image super-resolution (SR) can synthesize visually realistic details from low-resolution inputs by drawing on learned image priors (Wang et al., 2018; 2024; Wei et al., 2025; Wu et al., 2026). Since a low-resolution input can correspond to multiple high-resolution images, a plausible reconstruction may differ from the ground truth in its semantic content or visual attributes. An SR output may alter an object's identity, remove meaningful details, or introduce a texture absent from the original image (Ren et al., 2025; Li et al., 2025b). Such discrepancies illustrate hallucinations in generative SR, which we define as the loss or alteration of ground-truth evidence in the super-resolved image, or the addition of content unsupported by the ground truth.

Existing SR evaluation metrics do not adequately reflect whether reconstructed content is faithful to the ground truth. Figure 1 illustrates these limitations with visual examples in which conventional metrics favor less faithful reconstructions. Distortion metrics such as PSNR and perceptual similarity metrics such as LPIPS (Zhang et al., 2018) quantify discrepancies in pixel values or learned visual features, but these differences do not necessarily reflect the significance of a content change. Perceptual quality metrics (Ke et al., 2021; Wang et al., 2023; You et al., 2025) can assign favorable scores to visually realistic outputs even when their content is unsupported by the ground truth.

We introduce VQA-based Hallucination Assessment (VHA), which compares how strongly the ground truth and the SR output support individual visual evidence. Prior work (Ren et al., 2025) produces a hallucination rating and its explanation, but does not explicitly quantify how individual content or attribute changes contribute to the rating. We represent these evidences as question–answer pairs about specific image content and attributes. If the image content is preserved, the same answer should remain supported under both images. Inspired by VQAScore (Lin et al., 2024b), we measure the changes by comparing the VLM's answer log-probabilities across the two images on same question-answer pairs. We aggregate these probability differences into the VHA score.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



PSNR ↑	**17.63**	17.11	PSNR ↑	**23.24**	23.14
MUSIQ ↑	74.14	**75.18**	MUSIQ ↑	**70.82**	67.51
VHA (**Ours**) ↓	0.6382	**0.4954**	VHA (**Ours**) ↓	0.5872	**0.5126**
VHA+ (**Ours**) ↓	0.8641	**0.3488**	VHA+ (**Ours**) ↓	1.5106	**0.4791**
Human eval.	–	✓	Human eval.	–	✓

Figure 1: Limitation of conventional evaluation metrics. In the left example, PSNR favors a reconstruction that incorrectly depicts the building as a rock formation. In the right example, even the perceptual quality metric, MUSIQ (Ke et al., 2021) favors a reconstruction that alters the knitted texture present in the ground truth. In contrast, VHA and VHA+ assess the hallucinations, which conventional metrics fail to capture.

VHA assesses hallucinations using question–answer pairs generated from both the ground-truth and SR output. Questions from the ground truth assess whether original content is preserved in the SR output, while questions from the SR output assess whether its content is supported by the ground truth. In both directions, taxonomy-guided questions cover high-level semantics, such as objects and text, and low-level visual attributes, such as color and texture. To reduce the cost of repeated VLM inference, we introduce VHA+, a lightweight model trained on image pairs labeled with VHA scores. VHA+ predicts these scores directly from ground-truth and SR image pairs without VLM inference.

We evaluate the alignment of VHA and VHA+ with human hallucination judgments. We construct VHA-bench, a human-annotated benchmark covering outputs from multiple generative SR methods. On this benchmark, VHA and VHA+ achieve higher correlations with human judgments than conventional SR metrics and direct VLM ratings using the same backbones. We also show that fine-tuning an SR model on VHA+-based supervision with an image-quality objective both reduces the VHA score and improves image quality metrics. Our main contributions are as follows:

- We introduce a bidirectional VQA-based hallucination assessment method (VHA) that compares answer log-probabilities for same question–answer pairs across ground-truth and SR outputs. Taxonomy-guided question–answer generation enables VHA to assess hallucinations across diverse semantic and visual aspects.
- We propose VHA+, a lightweight model trained to predict VHA scores without VLM inference, and demonstrate that its learned representations can guide the fine-tuning of a super-resolution model.
- We construct a human-annotated VHA-bench and demonstrate that VHA and VHA+ correlate stronger with human judgments than conventional SR metrics and direct VLM ratings.

2 Related Works

Evaluation Metrics for Super-Resolution. SR evaluation commonly considers distortion, perceptual similarity, and perceptual quality. Distortion metrics such as PSNR and SSIM (Wang et al., 2004) summarize pixel-wise reconstruction errors and local structural differences against a reference, respectively. LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) compare deep image representations, with DISTS explicitly combining structure and texture similarity. No-reference image quality assessment methods, including MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023), and DeQA-Score (You et al., 2025), estimate perceptual quality from the test image alone. A-FINE (Chen et al., 2025) combines reference-based comparisons with image naturalness estimates, extending quality prediction to settings where the reference itself may be imperfect.

Hallucination in Generative Image Super-Resolution. Generative SR models can produce visually realistic details while introducing semantic errors and spurious structures (Wu et al., 2024b;

2026). PURE (Wei et al., 2025) addresses this issue by integrating image understanding with autoregressive restoration to preserve semantic content and local structures. However, its evaluation does not separately quantify hallucinations. To directly assess hallucinations, Ren et al. (2025) introduce Hallucination Score, which prompts GPT-4o with low-resolution, super-resolved, and ground-truth images to produce an overall hallucination rating and an accompanying explanation. Although the explanation can identify specific discrepancies, their individual contributions to the overall rating are not explicitly quantified. VHA constructs its assessment signal by aggregating measured changes in support for individual visual evidences.

VQA-based Evaluation of Generated Images. Visual question answering has been used to assess the semantic alignment and factual correctness of generated images. TIFA (Hu et al., 2023) and VQ^2 (Yarom et al., 2023) generate questions from textual descriptions and use VQA models to evaluate whether the specified content appears in an image. Davidsonian Scene Graph (DSG) (Cho et al., 2024) organizes atomic questions into dependency graphs to improve semantic coverage and answer consistency. VQAScore (Lin et al., 2024b) measures image–text alignment through the probability of a "Yes" answer to a question asking whether the image depicts the supplied text. I-HallA (Lim et al., 2025) evaluates factual correctness in text-to-image generation using VQA accuracy on curated multiple-choice questions informed by factual knowledge and visual evidence. Building on this question-based evaluation paradigm, VHA compares answer log-probabilities across GT and SR outputs using same question-answer pairs derived independently from each image.

3 VQA-BASED HALLUCINATION ASSESSMENT

We introduce VQA-based Hallucination Assessment (VHA) to quantify hallucinations in a super-resolved image (I_{SR}) using its ground-truth counterpart (I_{GT}) as the reference. To examine diverse semantic content and visual attributes, we generate taxonomy-guided question–answer (Q/A) pairs from both images (Section 3.1) using a vision-language model (VLM). For each Q/A pair, we compare the VLM's answer log-probabilities under the two images on same Q/A pairs. GT-derived pairs assess the loss or alteration of original evidence, while SR-derived pairs assess unsupported additions; we aggregate these probability differences into a bidirectional hallucination score (Section 3.2). Figure 2 illustrates our assessment pipeline.

3.1 TAXONOMY-GUIDED QUESTION–ANSWER GENERATION

We first generate Q/A pairs covering both high-level semantics and low-level visual attributes using a VLM. Without explicit guidance on what to examine, a VLM may focus its questions on certain types of content and attributes while overlooking others. We therefore use a taxonomy to guide Q/A generation across different types of visual evidence. The high-level categories comprise scene semantics, object semantics, and text or logos, while the low-level categories comprise color and tone, edges and shapes, and texture.

We organize Q/A generation around explicitly identified visual targets to limit repeated questions about the same content. Within each category, we first prompt the VLM to identify visible content or attributes. We then prompt it to generate one question and a short answer for each identified item. We apply this procedure independently to I_{GT} and I_{SR} to obtain the corresponding Q/A sets, $\mathcal{Q}_{\mathrm{GT}}$ and $\mathcal{Q}_{\mathrm{SR}}$, respectively. The full prompts for both stages are provided in Appendix A.

3.2 BIDIRECTIONAL HALLUCINATION SCORING

We assess hallucinations by evaluating whether GT content is preserved in SR and whether SR content is supported by GT. Forward and reverse scoring use $\mathcal{Q}_{\mathrm{GT}}$ and $\mathcal{Q}_{\mathrm{SR}}$, respectively, to compare each same answer's likelihood under both images. For a question (q) and an answer consisting of T tokens ($a = (a_1, \ldots, a_T)$), we compute the mean token-level log-probability under a fixed VLM (p_ϕ) in a teacher-forced manner:

$$\ell(I; q, a) = \frac{1}{T} \sum_{t=1}^{T} \log p_\phi(a_t \mid I, q, a_{<t}). \tag{1}$$

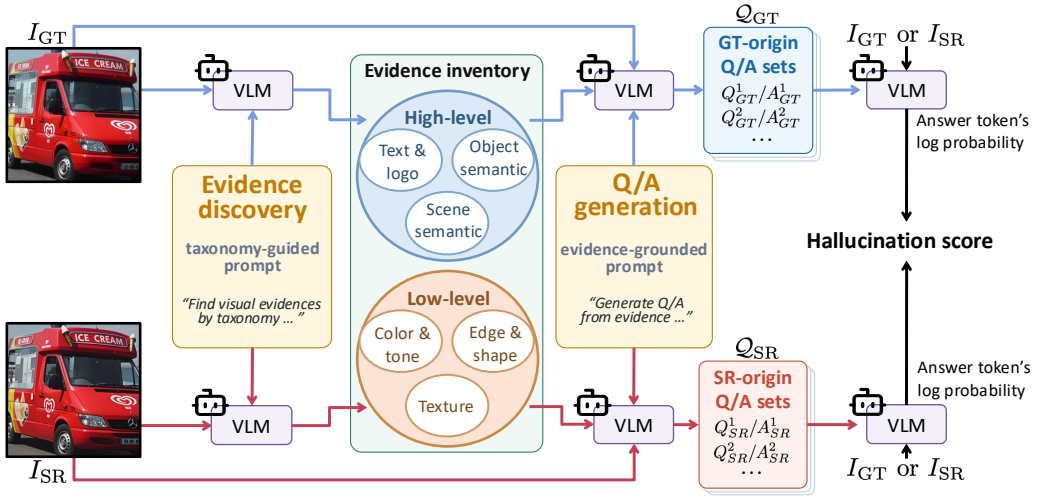


Figure 2: Overview of our bidirectional VQA-based hallucination assessment. Taxonomy-guided question–answer pairs are generated from both ground truth and super-resolved images, and each answer is scored under both images to assess evidence preservation and unsupported content.

We define the answer score as the difference between the image-conditioned and text-only answer log-probabilities:

$$s(I; q, a) = \ell(I; q, a) - \ell(\varnothing; q, a), \tag{2}$$

where $\varnothing$ denotes an input without an image. We then use each Q/A pair's score on its source image to gate its contribution to the hallucination score.

To assess the preservation of original ground-truth content and attributes, we define the forward score F by aggregating score differences over GT-derived Q/A pairs:

$$F(I_{\mathrm{GT}}, I_{\mathrm{SR}}) = \frac{1}{|\mathcal{Q}_{\mathrm{GT}}|} \sum_{(q,a)\in\mathcal{Q}_{\mathrm{GT}}} \mathbf{1}[s(I_{\mathrm{GT}}; q, a) > \tau_g] \cdot \big[s(I_{\mathrm{GT}}; q, a) - s(I_{\mathrm{SR}}; q, a)\big]. \tag{3}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function, and τ_g is the GT answer-score threshold. For pairs passing the threshold, a decrease in the answer score from GT to SR contributes positively to F. These contributions can offset one another.

GT-derived questions may overlook newly introduced content in the SR output. We therefore define the reverse score R to assess potentially unsupported additions by measuring decreases in answer scores from SR to GT for SR-derived Q/A pairs:

$$R(I_{\mathrm{GT}}, I_{\mathrm{SR}}) = \frac{1}{|\mathcal{Q}_{\mathrm{SR}}|} \sum_{(q,a)\in\mathcal{Q}_{\mathrm{SR}}} \mathbf{1}[s(I_{\mathrm{SR}}; q, a) > \tau_s] \cdot \big[s(I_{\mathrm{SR}}; q, a) - s(I_{\mathrm{GT}}; q, a) - \tau_m\big]_+, \tag{4}$$

where $[x]_+ = \max(x, 0)$. The SR answer-score threshold τ_s suppresses contributions from answers weakly supported by the SR output. The reverse score then counts only the portion of the SR–GT score difference exceeding the margin τ_m.

The final hallucination score is computed as follows:

$$H(I_{\mathrm{GT}}, I_{\mathrm{SR}}) = F(I_{\mathrm{GT}}, I_{\mathrm{SR}}) + R(I_{\mathrm{GT}}, I_{\mathrm{SR}}). \tag{5}$$

4 Learning a Lightweight Hallucination Assessment Model

VHA requires repeated VLM inference, and its full pipeline is not end-to-end differentiable. We therefore train VHA+, a lightweight, differentiable model that predicts VHA scores directly from GT–SR image pairs. Figure 3 illustrates dataset construction (Section 4.1) and model training (Section 4.2).

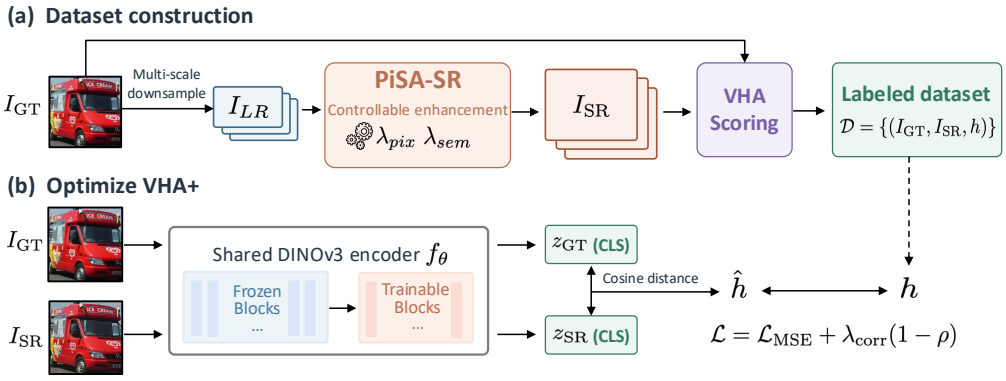


Figure 3: Overview of our VHA+ training pipeline. (a) Diverse GT–SR pairs are constructed using PiSA-SR and labeled with our VHA scores. (b) VHA+ is trained to predict these scores directly from image pairs.

4.1 Training Dataset Construction

We construct GT-SR pairs spanning different degrees of content preservation and alteration by varying the input resolution and enhancement settings. Each GT image is downsampled at multiple scales and restored with PiSA-SR (Sun et al., 2025), whose pixel-level and semantic-level controls vary fidelity and perceptual detail. We label each pair with its VHA score to obtain the training dataset $\mathcal{D} = \{(I_{\text{GT}}^{(i)}, I_{\text{SR}}^{(i)}, h_i)\}_{i=1}^{N}$, where $h_i = H(I_{\text{GT}}^{(i)}, I_{\text{SR}}^{(i)})$ and N is the number of image pairs. Additional dataset construction details are provided in Section 5.1.

4.2 Model Architecture and Training Objective

VHA+ predicts the VHA score from the distance between learned representations of the GT image and SR outputs. Following the DINO-based scoring approach of Hallucination Score (Ren et al., 2025), we use a shared DINOv3 encoder f_θ (Siméoni et al., 2025) to extract the final CLS tokens, $z_{\text{GT}} = f_\theta(I_{\text{GT}})$ and $z_{\text{SR}} = f_\theta(I_{\text{SR}})$. An affine transformation maps their cosine distance to the VHA score scale:

$$\hat{h}(I_{\text{GT}}, I_{\text{SR}}) = \alpha \left(1 - \frac{{z_{\text{GT}}}^{\mathsf{T}} z_{\text{SR}}}{\|z_{\text{GT}}\|_2 \|z_{\text{SR}}\|_2}\right) + \beta, \tag{6}$$

where α and β are learned scale and offset parameters.

We fine-tune the final transformer blocks and learn the affine parameters, while keeping the earlier blocks frozen. Supervision from VHA encourages the learned feature distance to reflect content loss, alteration, and unsupported additions. The training objective combines mean squared error (MSE) on standardized scores with a Pearson correlation loss:

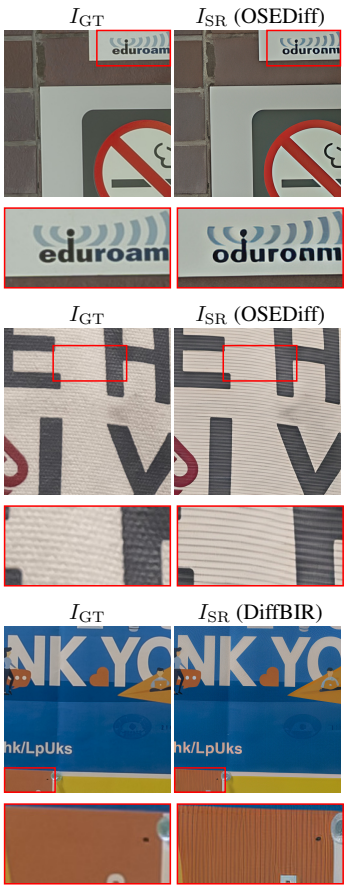
$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(\hat{\mathbf{h}}, \mathbf{h}) + \lambda_{\text{corr}}\big(1 - \rho(\hat{\mathbf{h}}, \mathbf{h})\big), \tag{7}$$

where $\hat{\mathbf{h}}$ and $\mathbf{h}$ contain the predicted and target VHA scores, ρ is the Pearson correlation, and λ_{corr} controls the correlation term. The MSE term encourages accurate prediction of target values, while the correlation term encourages the alignment in score distribution. Additional training details are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Benchmark construction. We construct VHA-bench, a human-annotated hallucination benchmark, by collecting human judgments of hallucination in generative SR outputs. For this study, we use 90 ground-truth images from the RealSR validation set (Cai et al., 2019). Each 512×512 GT image is downsampled to 128×128, and the low-resolution input is restored to 512×512 using



Taxonomy / Source	Question / Answer	$s(I_{GT})$	$s(I_{SR})$
Text (high) $\mathcal{Q}_{GT}$	**Q.** What text is on the upper sign? **A.** eduroam	10.58	−0.04
Text (high) $\mathcal{Q}_{SR}$	**Q.** What text is on the upper sign? **A.** oduronni	−2.15	19.75
Object (high) $\mathcal{Q}_{GT}$	**Q.** What is the lower sign? **A.** no smoking sign	5.04	6.66
Scene (high) $\mathcal{Q}_{GT}$	**Q.** What is the background surface? **A.** brick wall	9.12	9.25

Taxonomy / Source	Question / Answer	$s(I_{GT})$	$s(I_{SR})$
Text (high) $\mathcal{Q}_{GT}$	**Q.** What letter is visible in upper left quadrant? **A.** E	13.00	13.00
Object (high) $\mathcal{Q}_{GT}$	**Q.** What letter is in upper right? **A.** H	11.90	11.90
Texture (low) $\mathcal{Q}_{GT}$	**Q.** What does the black ink surface look like? **A.** Slightly grainy texture	1.97	0.64
Texture (low) $\mathcal{Q}_{SR}$	**Q.** What is the background texture? **A.** Fine horizontal ribs	−3.69	2.90

Taxonomy / Source	Question / Answer	$s(I_{GT})$	$s(I_{SR})$
Object (high) $\mathcal{Q}_{GT}$	**Q.** What is on left side of banner? **A.** Walking person	3.72	3.66
Object (high) $\mathcal{Q}_{GT}$	**Q.** What is on right side of banner? **A.** Person on paper airplane	6.59	4.33
Tone (low) $\mathcal{Q}_{GT}$	**Q.** What contrast exists between text and background? **A.** High contrast white on blue	2.13	1.90
Texture (low) $\mathcal{Q}_{SR}$	**Q.** What texture is on orange panel? **A.** Vertical wood grain	−4.32	2.85

Figure 4: Examples of Q/A generation and scoring on VHA-bench. $\mathcal{Q}_{GT}$ and $\mathcal{Q}_{SR}$ denote Q/A sets generated from the GT image and SR outputs, respectively. For each Q/A pair, $s(I)$ denotes the answer score $s(I; q, a)$ computed by Eq. 2.

six generative SR methods: StableSR (Wang et al., 2024), DiffBIR (Lin et al., 2024a), SeeSR (Wu et al., 2024b), OSEDiff (Wu et al., 2024a), PURE (Wei et al., 2025) and VOSR (Wu et al., 2026), yielding 540 GT–SR pairs. We conduct a human study with 200 participants, each randomly assigned 54 GT–SR pairs for evaluation. Using the GT image as reference, participants rate hallucination on a five-point scale, from 1 (no hallucination) to 5 (severe hallucination), focusing on changes in semantic content and visual attributes while ignoring differences in sharpness or noise alone. Each GT–SR pair receives approximately 20 ratings on average, which we aggregate into a mean opinion score (MOS). The study interface is shown in Figure 7 in the appendix.

Evaluation metrics. We use existing evaluation metrics to examine their correlations with human hallucination judgments on VHA-bench and to assess the outputs of SR methods. The metrics include PSNR and SSIM for pixel-level similarity, and LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2020), DINO, and CLIP for feature-level similarity. For DINO and CLIP, we compute cosine distances between the final-layer CLS features of GT and SR images using frozen pretrained `DINOv3 ViT-B/16` and `CLIP ViT-B/16` encoders (Siméoni et al., 2025; Radford et al., 2021), respectively. For no-reference image quality assessment (NR-IQA), we include MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023), Q-Insight (Li et al., 2025a), VisualQuality-R1 (Wu et al., 2025), and DeQA-Score (You et al., 2025). We also include A-FINE (Chen et al., 2025) as a generalized full-reference IQA (GFR-IQA) metric.

Implementation details. We use `Qwen3-VL-8B-Instruct` (Bai et al., 2025) for Q/A generation and answer scoring. For bidirectional VHA scoring, we set $\tau_g = -2$, $\tau_s = 0$ and $\tau_m = 0.5$. For VHA+ training dataset construction, we select 20,000 images each from LSDIR (Li et al., 2023)

Table 1: Spearman (ρ) and Pearson (r) correlations with human mean opinion scores (MOS) for hallucination, computed on 540 SR outputs from six methods and rated by 200 participants. Higher correlations indicate closer agreement with human judgments. VHA denotes our bidirectional VQA-based assessment method, while VHA+ is its learned predictor that operates without VLM inference.

Category	Method	$\rho \uparrow$	$r \uparrow$	Category	Method	$\rho \uparrow$	$r \uparrow$
Pixel sim.	PSNR	0.162	0.181	*Feature sim.*	LPIPS	0.083	0.089
	SSIM	0.010	0.012		DISTS	0.066	0.069
NR-IQA	MUSIQ	-0.008	-0.014		DINO	0.081	0.115
	CLIP-IQA	0.168	0.159		CLIP	0.282	0.273
	Q-Insight	0.080	0.036	*Hallucination Assessment*	Direct rating	0.497	0.482
	VQ-R1	0.175	0.126		HS prompting	0.543	0.563
	DeQA	0.145	0.156		**VHA (Ours)**	<u>0.618</u>	<u>0.645</u>
GFR-IQA	A-FINE	0.173	0.299		**VHA+ (Ours)**	**0.659**	**0.675**

and TextVQA (Singh et al., 2019) and use their 512×512 center crops as GT images. We downsample each GT image by factors of 0.5, 0.375, and 0.25 and restore the degraded inputs using PiSA-SR (Sun et al., 2025), with and without additional $2\times$ input upsampling. We set the pixel-level and semantic-level control weights of PiSA-SR $(\lambda_{\text{pix}}, \lambda_{\text{sem}})$ to $(1, 0)$, $(0.5, 1)$, and $(1, 1)$.

5.2 Results

Examples of Q/A generation and scoring. Figure 4 illustrates how generated Q/A pairs capture high-level semantic content and low-level visual attributes, and how their score differences reflect changes between the GT image and SR outputs. In the first example, the Q/A pair about the original text in $\mathcal{Q}_{\text{GT}}$ yields a higher $s(I_{\text{GT}})$ than $s(I_{\text{SR}})$, whereas the pair about the hallucinated text in $\mathcal{Q}_{\text{SR}}$ yields a lower $s(I_{\text{GT}})$ than $s(I_{\text{SR}})$. In the second, the Q/A pair about texture in $\mathcal{Q}_{\text{GT}}$ yields a higher $s(I_{\text{GT}})$ than $s(I_{\text{SR}})$, whereas the pair about background texture in $\mathcal{Q}_{\text{SR}}$ yields a lower $s(I_{\text{GT}})$ than $s(I_{\text{SR}})$. The Q/A pairs about unchanged letters receive nearly identical scores across the two images. In the third, a Q/A pair in $\mathcal{Q}_{\text{SR}}$ captures newly introduced wood-grain texture, with $s(I_{\text{GT}})$ much lower than $s(I_{\text{SR}})$.

Comparison of correlations with existing metrics. Table 1 shows that VHA and VHA+ achieve higher correlations with human hallucination judgments than the existing metrics evaluated on VHA-bench. VHA achieves Spearman and Pearson correlations of 0.618 and 0.645, respectively, compared with at most 0.282 and 0.299 among the existing metrics. The low correlations of these baselines suggest that general image similarity and quality measures alone are insufficient to capture human judgments of hallucination severity. VHA+ achieves correlations of 0.659 and 0.675, respectively, similar to those of VHA. These results demonstrate that the lightweight predictor retains strong agreement with human judgments without requiring VLM inference at evaluation time.

Comparison with hallucination assessment methods. Table 1 compares hallucination assessment methods using the same VLM backbone (`Qwen3-VL-8B-Instruct`). Direct rating provides the VLM with GT images and SR outputs and prompts it to assign a hallucination rating from 1 to 5. The full prompt is provided in Appendix D. HS prompting uses exactly the same prompt as Hallucination Score (Ren et al., 2025), with GT images, low-resolution inputs, and SR outputs. Both baselines show reasonable correlations with human judgments, but their discrete ratings provide limited granularity. Also, neither baseline explicitly quantifies how visual elements or attributes contribute to the hallucination score. Our method achieves higher correlations while providing continuous scores from answer log-probabilities and explicit contributions from individual Q/A pairs.

5.3 Ablation studies and analysis

Ablations of VHA scoring. Table 2 presents ablation results for VHA scoring on VHA-bench. Table 2(a) shows that bidirectional scoring achieves higher correlations with human judgments than either forward or reverse scoring alone. Table 2(b) shows that scores from high-level Q/A pairs align more closely with human hallucination judgments than those from low-level pairs. Combining both further improves correlations, suggesting that high-level semantic content and low-level visual at-

Table 2: Ablations of VHA scoring on VHA-bench. ρ and r denote Spearman and Pearson correlations.

(a) Bidirectional scoring

Scoring	$\rho \uparrow$	$r \uparrow$
Forward	0.516	0.562
Reverse	0.465	0.560
Bidirectional	**0.618**	**0.645**

(b) High- and low-level Q/A

Q/A set	$\rho \uparrow$	$r \uparrow$
High only	0.539	0.584
Low only	0.204	0.197
High + Low	**0.618**	**0.645**

(c) Forward and reverse thresholding

Forward	Reverse	$\rho \uparrow$	$r \uparrow$
✓		0.570	0.614
	✓	0.607	0.635
✓	✓	**0.618**	**0.645**

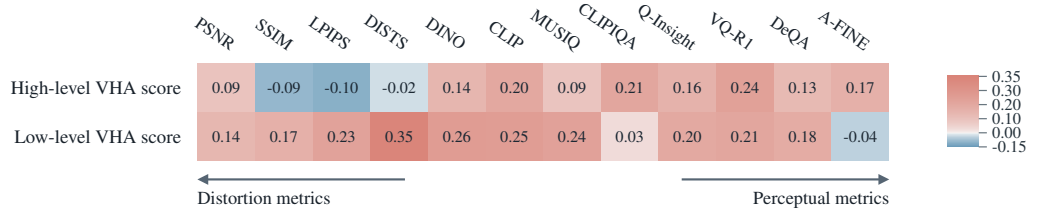


Figure 5: Spearman correlations of our high-level and low-level bidirectional Q/A scores with the image evaluation metrics on VHA-bench.

tributes provide complementary information for assessing hallucination. Finally, Table 2(c) evaluates forward thresholding with τ_g in Eq. 3 and reverse thresholding with τ_s and τ_m in Eq. 4. Both help suppress unreliable Q/A contributions and improve correlations with human judgments.

Correlations of high- and low-level VHA scores with existing metrics. We compute high- and low-level VHA scores using only high- and low-level Q/A pairs, respectively. Figure 5 shows their Spearman correlations with existing metrics on VHA-bench. The high-level score correlates weakly with pixel-level similarity metrics (PSNR and SSIM) and perceptual similarity metrics (LPIPS and DISTS), while showing relatively higher correlations with semantic feature distances (DINO and CLIP) and most NR-IQA metrics. Compared with the high-level score, the low-level score shows higher correlations with most existing metrics, especially for pixel-level and perceptual similarity metrics. This pattern suggests that many existing metrics reflect low-level visual changes more closely than high-level semantic changes.

Computational efficiency of VHA+. Table 3 compares the parameter counts and inference times of VHA and VHA+. We measure the average inference time per image pair over 50 pairs on a single NVIDIA RTX A6000. VHA+ uses 85M parameters, compared with 8,767M for VHA, and reduces the average inference time from 101 s to 17 ms per pair. This corresponds to approximately $5{,}900\times$ faster inference under the measured configuration. By directly predicting scores without VLM inference, VHA+ substantially reduces evaluation cost while retaining strong agreement with human judgments.

Table 3: Parameter counts and average inference time of VHA and VHA+.

Method	Parameters (M)	Inference time (ms)
VHA	8,767	101,000
VHA+	85	17

5.4 Fine-tuning Super-Resolution with VHA+

The lightweight and differentiable design of VHA+ allows it to guide generative SR training. We use the hallucination score predicted by the frozen VHA+ model as an auxiliary loss. Following Hallucination Score (Ren et al., 2025), we combine this loss with a MUSIQ-based quality objective (Ke et al., 2021) to discourage oversmoothing and preserve perceptual quality. We fine-tune pretrained VOSR (Wu et al., 2026) by adding these terms to its original distillation objective while retaining its training pipeline. We denote the fine-tuned model as VOSR*. Training details, including datasets, loss weights, and optimization settings, are provided in Appendix F.1.

Table 4 compares VOSR* with pretrained generative SR models and bicubic upsampling on the RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) test sets provided by StableSR (Wang et al., 2024). Although bicubic upsampling achieves high pixel-level similarity score, it produces blurry outputs which results in low perceptual quality score. Also, it does not synthesize new fine details, thereby limiting hallucinations arising from unsupported additions. Consistently, bicubic

Table 4: Quantitative comparison of SR methods on RealSR and DRealSR. VOSR* denotes VOSR fine-tuned with auxiliary losses based on our VHA+ score and MUSIQ.

Dataset	Method	*Pixel sim.*		*Feature sim.*		*NR-IQA*			*GFR-IQA*	*Ours*	
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	MUSIQ ↑	CLIPIQA ↑	VQ-R1 ↑	A-FINE ↓	**VHA ↓**	**VHA+ ↓**
RealSR	Bicubic	25.48	0.7200	0.4564	0.2633	25.81	0.3098	1.2210	77.22	0.9861	0.8055
	StableSR	23.08	0.6611	0.3107	0.2186	65.00	0.6142	3.5362	27.48	1.4394	0.9863
	DiffBIR	23.27	0.6140	0.3617	0.2419	68.30	0.6921	3.9381	31.87	1.3234	1.0500
	SeeSR	23.86	0.6879	0.2980	0.2222	69.49	0.6725	3.8458	25.84	1.3439	1.0230
	OSEDiff	23.59	0.6923	0.2921	0.2128	69.08	0.6688	4.0439	20.96	1.2806	1.0086
	PURE	21.52	0.5755	0.3829	0.2450	67.05	0.6943	3.7221	41.68	1.3613	1.1788
	VOSR	23.77	0.6816	0.2858	0.2097	70.00	0.5761	3.7426	21.39	1.0190	0.9283
	VOSR* (Ours)	**24.29**	**0.6914**	**0.2820**	**0.1985**	**70.98**	**0.6074**	**3.8557**	**11.24**	**0.9494**	**0.6413**
DRealSR	Bicubic	27.87	0.7869	0.4608	0.2788	22.59	0.3190	1.0280	74.54	0.7310	0.6948
	StableSR	25.78	0.6985	0.3379	0.2295	58.32	0.6250	3.1017	37.97	0.8453	0.7859
	DiffBIR	23.52	0.5466	0.4923	0.2913	65.99	0.7213	3.7510	33.10	1.2837	1.1691
	SeeSR	26.09	0.7284	0.3205	0.2304	64.86	0.6841	3.5276	22.73	1.1364	0.9278
	OSEDiff	25.85	0.7374	0.2967	0.2163	64.69	0.6961	3.7727	21.38	0.8258	0.8162
	PURE	23.34	0.5762	0.4526	0.2719	61.08	0.6773	3.3473	41.07	1.0318	1.0200
	VOSR	25.44	0.7029	0.3479	0.2362	65.80	0.5906	3.4163	29.46	0.7516	0.9781
	VOSR* (Ours)	**25.83**	**0.7076**	**0.3473**	**0.2263**	**66.97**	**0.6422**	**3.5343**	**21.63**	**0.7161**	**0.6473**

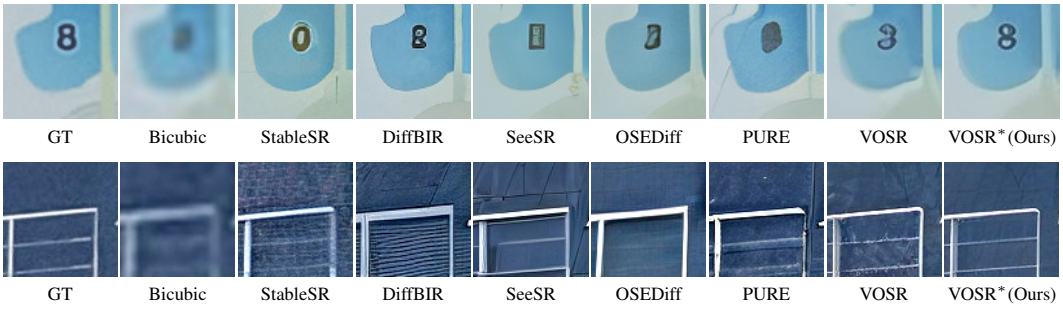


Figure 6: Qualitative comparisons of bicubic upsampling and generative SR methods on RealSR testset. VOSR*, fine-tuned using our VHA+ score and MUSIQ as auxiliary losses, shows fewer hallucinated details and more faithfully preserves ground-truth content than other methods.

upsampling achieves better VHA and VHA+ scores than the generative SR models. VOSR* improves all reported metrics over VOSR on both datasets, including VHA and VHA+. These results suggest that combining VHA+-based supervision with a perceptual quality objective can reduce hallucination while maintaining perceptual quality.

Figure 6 illustrates examples in which VOSR* better preserves text and suppresses spurious textures compared with VOSR. In the top example, VOSR* more faithfully reconstructs the digit "8", whose shape is distorted in the VOSR output. In the bottom example, VOSR* reduces spurious textures, more closely matching the smoother appearance of the GT image. Additional qualitative examples are provided in Appendix G.

6 Conclusion

We propose a bidirectional VQA-based method (VHA) for evaluating hallucination in generative image super-resolution. Taxonomy-guided questions cover high-level semantic content and low-level visual attributes, enabling assessment of diverse forms of hallucination. Bidirectional comparisons of answer score assess both the loss or alteration of ground-truth evidence and unsupported additions in SR outputs. This addresses a gap in conventional SR evaluation, where visually realistic details can receive favorable quality scores despite misrepresenting the original content. We also train VHA+, a lightweight model that directly predicts hallucination score without VLM inference. On our human-annotated benchmark, VHA and VHA+ achieve high correlations with human hallucination judgments. Furthermore, our VHA+-based supervision extends this assessment to SR model training. We believe this work adds a complementary dimension to SR evaluation and encourages future research to consider hallucination alongside perceptual quality when developing SR models.

AI USE STATEMENT

In this work, we used generative AI tools to support research execution, including developing code for our methods and experiments and building the human-study website used to collect ratings for VHA-bench. We also used these tools to obtain feedback on experimental methodology and assist with result interpretation. Additionally, we used them for literature search, translation, drafting sections of the manuscript, polishing the writing, and preparing figures and tables. We reviewed AI-assisted suggestions during method development and manuscript preparation, and checked reported numerical results against experimental outputs. We take responsibility for the final content of this work, including AI-assisted text, code, claims, and artifacts.

ETHICS STATEMENT

This work aims to improve the assessment and reduction of hallucinations in generative super-resolution. In the rating data used for analysis, participants are represented by study-specific IDs rather than email addresses or platform identifiers. We report human judgments as mean opinion scores for image pairs, without participant identifiers.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe VHA scoring and VHA+ training, and Section 5.1 specifies benchmark construction and experimental settings. The full Q/A generation and direct-rating prompts are provided in Appendices A and D. Additional VHA+ training and SR fine-tuning details are provided in Appendices B and F.1, and Figure 7 documents the human study interface. We will publicly release our code, benchmark, and model weights to support reproduction of the results.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *CVPR*, 2025.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *NeurIPS*, 2025a.

Yachao Li, Dong Liang, Tianyu Ding, and Sheng-Jun Huang. StructSR: Refuse spurious details in real-world image super-resolution. In *AAAI*, 2025b.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR Workshops*, 2023.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering. In *AAAI*, 2025.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024a.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Weiming Ren, Raghav Goyal, Zhiming Hu, Tristan Ty Aumentado-Armstrong, Iqbal Mohomed, and Alex Levinshtein. Hallucination score: Towards mitigating hallucinations in generative image super-resolution. *arXiv preprint arXiv:2507.14367*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *CVPR*, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. In *ICCV*, 2025.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024b.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *CVPR*, 2026.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *NeurIPS*, 2025.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In *NeurIPS*, 2023.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

A PROMPTS FOR Q/A GENERATION

Tables 5–8 provide the full prompt templates used for taxonomy-guided evidence discovery and evidence-conditioned question–answer generation in Section 3. We use separate templates for high-level semantics and low-level visual attributes, and apply the same templates independently to ground truth and super-resolved images.

Table 5: High-level taxonomy-guided evidence discovery prompt.

```
Look at this HR image ({image_id}) and identify high-level semantic
evidence targets for short-answer Q/A generation.

Taxonomy definitions:

- scene_semantic: The overall scene, environment, or image-level context.
  This taxonomy must describe the image as a whole or a broad scene region,
  not a single object.
- object_semantic: A distinct object or structure that contributes
  meaningfully to the image content, such as a tree, grass, tower, wall,
  house, vehicle, bridge, or signboard; or a recognizable visual property
  or part within an object or local region, such as texture, style,
  material, decoration, or a visible object part.
- text_logo: Clearly readable text, words, numbers, signs, brands, or logos.

Generate up to 1 scene_semantic target, up to 4 object_semantic targets,
and up to 4 text_logo targets.

A target is suitable when it is clearly visible in the HR image,
supports a specific short-answer question, and may be altered by
downsampling or super-resolution.

Generate as many valid targets as are clearly supported by the image,
within these limits.

Reply with valid JSON only:

[
  {
    "evidence_id": "he01",
    "taxonomy": "scene_semantic",
    "region_scope": "whole_scene or localized region",
    "region_hint": "brief visible location",
    "target": "specific high-level semantic content",
    "hallucination_risk": "how this content may be altered or invented"
  }
]
```

Table 6: High-level evidence-conditioned question–answer generation prompt.

```
Look at this HR image ({image_id}) and generate one high-level
short-answer Q/A pair for each provided {taxonomy}
evidence item.

Taxonomy definition:

{taxonomy_definition}

The question must identify the relevant image region unambiguously and
ask only about the specified high-level semantic content.

Use a precise semantic label or short noun phrase.

Do not ask about overall image quality.

Do not create yes/no questions.

Each answer must be specific and 1-3 words only.

Evidence items:

{evidence_json}

Reply with valid JSON only:

[
  {"evidence_id": "{example_evidence_id}", "q": "...", "a": "..."}
]
```

Table 7: Low-level taxonomy-guided evidence discovery prompt.

```
Look at this HR image ({image_id}) and identify low-level visual evidence
targets for short-answer Q/A generation.

Taxonomy definitions:

- color_tone: Visible color, brightness, contrast, highlight, shadow, or
  tone relationship within an image region.
- edge_shape: Visible orientation, curvature, thickness, continuity, or
  shape of a boundary or small structural part.
- texture: Visible surface grain, material appearance, fine pattern,
  line structure, or texture density.

Generate 3 color_tone targets, 3 edge_shape targets, and 3 texture targets
for this image. A target is suitable when it is clearly visible in the HR
image, supports a specific short-answer question, and may be altered by
downsampling or super-resolution.

Reply with valid JSON only:
[
  {
    "evidence_id": "le01",
    "taxonomy": "color_tone",
    "region_scope": "local",
    "region_hint": "brief visible location",
    "target": "specific low-level visual property",
    "hallucination_risk": "how this property may be altered or invented"
  }
]
```

Table 8: Low-level evidence-conditioned question–answer generation prompt.

```
Look at this HR image ({image_id}) and generate one highly specific
low-level short-answer Q/A pair for each provided evidence item.

The question must identify the relevant image region and ask only about
the specified low-level visual property.

Describe the finest reliably visible appearance. Avoid broad categorical
answers when the image supports a more precise description.

Use the following attribute structure:

- color_tone: describe the visible hue together with its lightness,
  saturation, or local tonal character.
- edge_shape: describe the visible boundary using its thickness,
  continuity, smoothness, geometry, or orientation.
- texture: describe the visible texture using its scale, direction,
  regularity, density, or structural appearance.

Combine two or three complementary visual attributes when they are
clearly supported by the image. Do not add an attribute merely to make
the answer longer or more specific.

Use visually grounded everyday terms. Do not use uncommon or technical
terminology solely for greater specificity.

Do not ask about overall image quality or sharpness preference.
Do not create yes/no questions.

Each answer must be specific and 2-3 words long.

Evidence items:
{evidence_json}

Reply with valid JSON only:
[
  {"evidence_id": "le01", "q": "...", "a": "..."}
]
```

B VHA+ TRAINING DETAILS

We train the last four transformer blocks of the `DINOv3 ViT-B/16` encoder and the affine scale and offset parameters, while keeping the earlier layers frozen. We set $\lambda_{\text{corr}} = 3$ in Eq. 7 and train VHA+ using AdamW (Loshchilov & Hutter, 2019) for 20,000 steps with a batch size of 32. The learning rates are 10^{-6} for the encoder and 10^{-4} for the affine parameters.

Score standardization and calibration. Let μ_H and σ_H denote the mean and standard deviation of the training-set target scores. We standardize each target score h_i as

$$\tilde{h}_i = \frac{h_i - \mu_H}{\sigma_H}. \tag{8}$$

We train the affine head of VHA+ to predict the standardized target $\tilde{h}_i$. Both loss terms in Eq. 7 are computed between the model outputs and the standardized targets $\tilde{h}_i$. For numerical stability, we add 10^{-8} to the denominator when computing the Pearson correlation. At inference, we convert the affine head output to the original score scale using the stored training-set mean and standard deviation. For our training dataset, $\mu_H = 0.7545$ and $\sigma_H = 1.0254$.

C HUMAN STUDY INTERFACE

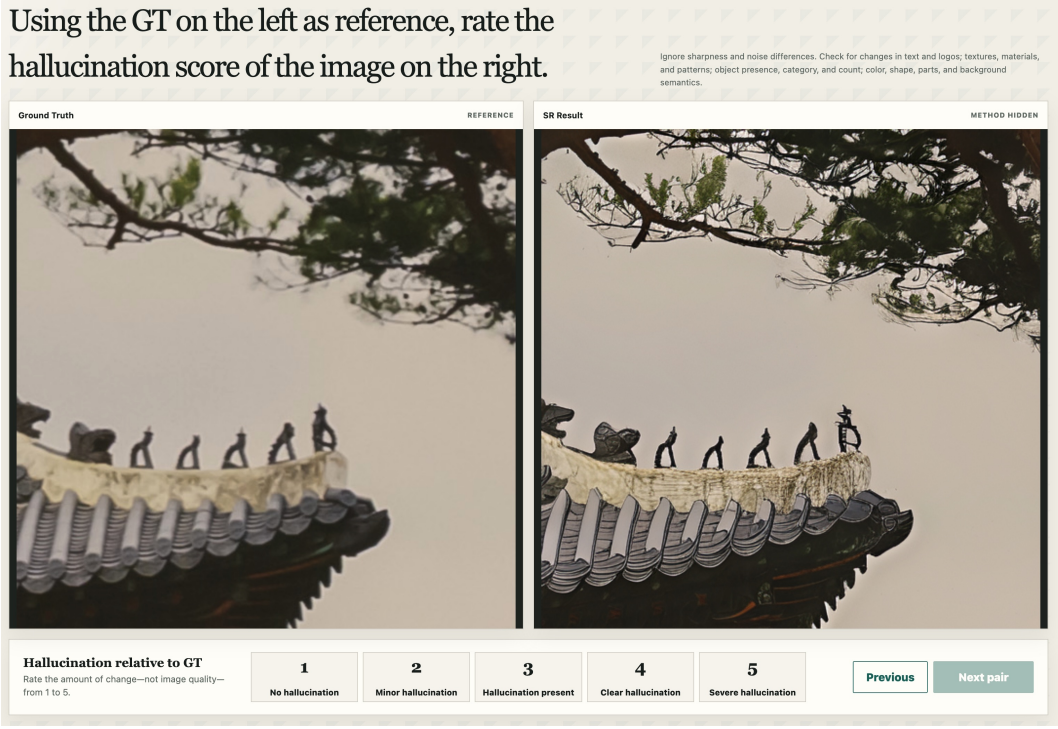


Figure 7: Web interface used for the human hallucination study. Participants compare the ground-truth reference (left) with a super-resolution result (right) and assign a hallucination score from 1 (no hallucination) to 5 (severe hallucination), while ignoring differences in sharpness and noise.

D PROMPTS FOR DIRECT VLM RATING

Table 9 provides the full prompt used for the Direct rating baseline in the comparison of VLM-based hallucination assessment methods in Section 5.2 (Table 1).

Table 9: Full prompt for Direct rating.

```
You will receive two images:
1.  Ground Truth (GT): The reference high-resolution image.
2.  Super-Resolved Image (SR): An image restored by an AI super-resolution
   model.
We define hallucinations as the loss or alteration of ground-truth evidence in SR,
or the addition of content unsupported by GT.
Rate the amount of hallucination in SR compared with GT on a scale from 1 to 5.
A lower score means less hallucination, and a higher score means more
hallucination.
Output only the following JSON:
{
  "score":  <integer from 1 to 5>
}
```

E ADDITIONAL ABLATIONS

E.1 VHA SCORING THRESHOLD ABLATIONS

Table 10: Ablations of VHA scoring thresholds on VHA-bench. ρ and r denote Spearman and Pearson correlations.

(a) Forward score threshold

τ_g	ρ	r
0	0.5919	0.6295
-1	**0.6190**	0.6427
-2	0.6179	**0.6449**
-3	0.6103	0.6408
-4	0.6085	0.6396
-5	0.6075	0.6355
-6	0.6079	0.6354

(b) Reverse score threshold

τ_s	ρ	r
-3	0.6160	0.6426
-2	0.6164	0.6431
-1	0.6159	0.6437
0	**0.6179**	**0.6449**
1	0.6144	0.6430
2	0.6087	0.6383
3	0.5941	0.6295

(c) Reverse margin threshold

τ_m	ρ	r
-1	0.5909	0.6366
-0.5	0.6045	0.6407
0	0.6173	0.6443
0.5	**0.6179**	**0.6449**
1	0.6133	0.6423

Table 10 shows the effects of the GT answer-score threshold τ_g in Eq. 3 and the SR answer-score threshold τ_s and margin τ_m in Eq. 4. VHA is relatively insensitive to variations in these hyper-parameters. However, overly strict or permissive thresholds can reduce correlations with human judgments. These results suggest that appropriate thresholds help suppress noisy Q/A contributions while preserving informative ones.

E.2 VHA+ TRAINING ABLATIONS

Table 11: VHA+ training ablations on VHA-bench. ρ and r denote Spearman and Pearson correlations, respectively; bold indicates the selected settings.

(a) Correlation loss weight

λ_{corr}	ρ	r
0	0.6370	0.6680
1	0.6540	0.6771
2	0.6586	0.6763
3	**0.6594**	**0.6753**
4	0.6593	0.6741
5	0.6582	0.6735

(b) Trainable blocks

Last K blocks	ρ	r
2	0.5962	0.6212
4	**0.6594**	**0.6753**
6	0.6556	0.6638
8	0.6494	0.6660
10	0.6389	0.6631

Table 11 presents VHA+ training ablations. Table 11(a) shows the results of varying the Pearson correlation loss weight λ_{corr} in Eq. 7. An appropriate correlation loss weight improves agreement with human judgments compared with MSE alone. Table 11(b) shows the results for different numbers of trainable transformer blocks. Updating fewer or more than four blocks yields lower correlations, which may reflect underfitting and overfitting, respectively.

Table 12: Ablations of VHA+ and MUSIQ loss weights for VOSR fine-tuning.

Dataset	γ_{vha}	γ_{musiq}	*Pixel sim.*		*Feature sim.*		*NR-IQA*			*GFR-IQA*	*Ours*	
			PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	MUSIQ ↑	CLIPIQA ↑	VQ-R1 ↑	A-FINE ↓	VHA ↓	VHA+ ↓
RealSR	0	0.001	23.04	0.6418	0.3601	0.2643	75.89	0.6820	3.723	11.29	1.372	1.172
	0.05	0.001	23.96	0.6791	0.2925	0.2089	72.17	0.6373	3.930	11.03	0.9164	0.6856
	0.10	0.001	24.29	0.6914	0.2820	0.1985	70.98	0.6074	3.856	11.24	0.9494	0.6413
	0.15	0.001	24.25	0.6914	0.2698	0.1907	70.25	0.5910	3.769	13.17	0.9557	0.6304
	0.20	0.001	24.38	0.6988	0.2616	0.1892	69.39	0.5610	3.791	13.94	0.9355	0.6211
	0.10	0	25.78	0.7431	0.2237	0.1737	51.31	0.3457	2.709	48.03	0.8744	0.5645
	0.10	0.0005	24.53	0.7026	0.2618	0.1892	68.87	0.5710	3.761	16.08	0.9410	0.6341
	0.10	0.0010	24.29	0.6914	0.2820	0.1985	70.98	0.6074	3.856	11.24	0.9494	0.6413
	0.10	0.0015	23.98	0.6816	0.2899	0.2026	71.83	0.6255	3.943	10.10	1.0331	0.6536
	0.10	0.0020	23.70	0.6653	0.3042	0.2087	73.02	0.6391	3.945	8.811	0.9854	0.6641
DRealSR	0	0.0010	24.51	0.6482	0.4357	0.3002	73.57	0.6899	3.184	12.27	1.271	1.256
	0.05	0.0010	25.51	0.6954	0.3573	0.2349	68.52	0.6545	3.678	18.59	0.7262	0.6788
	0.10	0.0010	25.83	0.7076	0.3473	0.2263	66.97	0.6422	3.534	21.63	0.7161	0.6473
	0.15	0.0010	25.84	0.7060	0.3349	0.2175	66.01	0.6237	3.471	25.25	0.6745	0.6101
	0.20	0.0010	25.95	0.7139	0.3266	0.2146	65.11	0.6042	3.459	27.07	0.6342	0.6043
	0.10	0	27.24	0.7714	0.2539	0.1837	48.03	0.4260	2.489	52.57	0.6347	0.5812
	0.10	0.0005	26.08	0.7173	0.3247	0.2140	64.51	0.6123	3.462	28.58	0.7118	0.6473
	0.10	0.0010	25.83	0.7076	0.3473	0.2263	66.97	0.6422	3.534	21.63	0.7161	0.6473
	0.10	0.0015	25.52	0.6947	0.3592	0.2310	68.30	0.6541	3.666	20.93	0.6843	0.6469
	0.10	0.0020	25.30	0.6799	0.3733	0.2386	69.49	0.6648	3.673	19.24	0.6466	0.6527

F Fine-Tuning SR Models with VHA+

F.1 Additional Training Details

VOSR (Wu et al., 2026) first trains a multi-step SR model and then distills it into a one-step student. We fine-tune the pretrained one-step student with the additional VHA+ and MUSIQ losses while keeping the released multi-step model fixed as the teacher. We use the full LSDIR training set (Li et al., 2023) and all 800 DIV2K training images (Agustsson & Timofte, 2017). Following VOSR's training setup, we use random 512×512 crops and its degradation pipeline. The fine-tuning objective is

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \gamma_{\text{vha}}\mathcal{L}_{\text{VHA+}} + \gamma_{\text{musiq}}\mathcal{L}_{\text{MUSIQ}}, \tag{9}$$

where $\mathcal{L}_{\text{distill}}$ is VOSR's original distillation loss. The auxiliary losses $\mathcal{L}_{\text{VHA+}}$ and $\mathcal{L}_{\text{MUSIQ}}$ are the score predicted by the frozen VHA+ model and the negative MUSIQ score (Ke et al., 2021), respectively. We set $\gamma_{\text{vha}} = 0.1$ and $\gamma_{\text{musiq}} = 0.001$. We fine-tune the student for 1,000 updates using AdamW (Loshchilov & Hutter, 2019), a learning rate of 10^{-5}, and a batch size of 16. We use an exponential moving average (EMA) with decay 0.99.

F.2 SR fine-tuning ablations

Table 12 shows the results of varying the VHA+ and MUSIQ loss weights in Eq. 9. Increasing the VHA+ loss weight tends to lower perceptual-quality scores, whereas increasing the MUSIQ loss weight can reduce image similarity and increase VHA+ scores. These results suggest that appropriate loss weights are needed to reduce hallucinations while maintaining perceptual quality.

G ADDITIONAL QUALITATIVE RESULTS

Figures 8 and 9 present additional qualitative comparisons of generative SR models and bicubic upsampling on the RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) test sets provided by StableSR (Wang et al., 2024), respectively.



Figure 8: Qualitative comparisons of bicubic upsampling and generative SR methods on RealSR testset. VOSR*, fine-tuned using our VHA+ score and MUSIQ as auxiliary losses, shows fewer hallucinated details and more faithfully preserves ground-truth content than other methods.

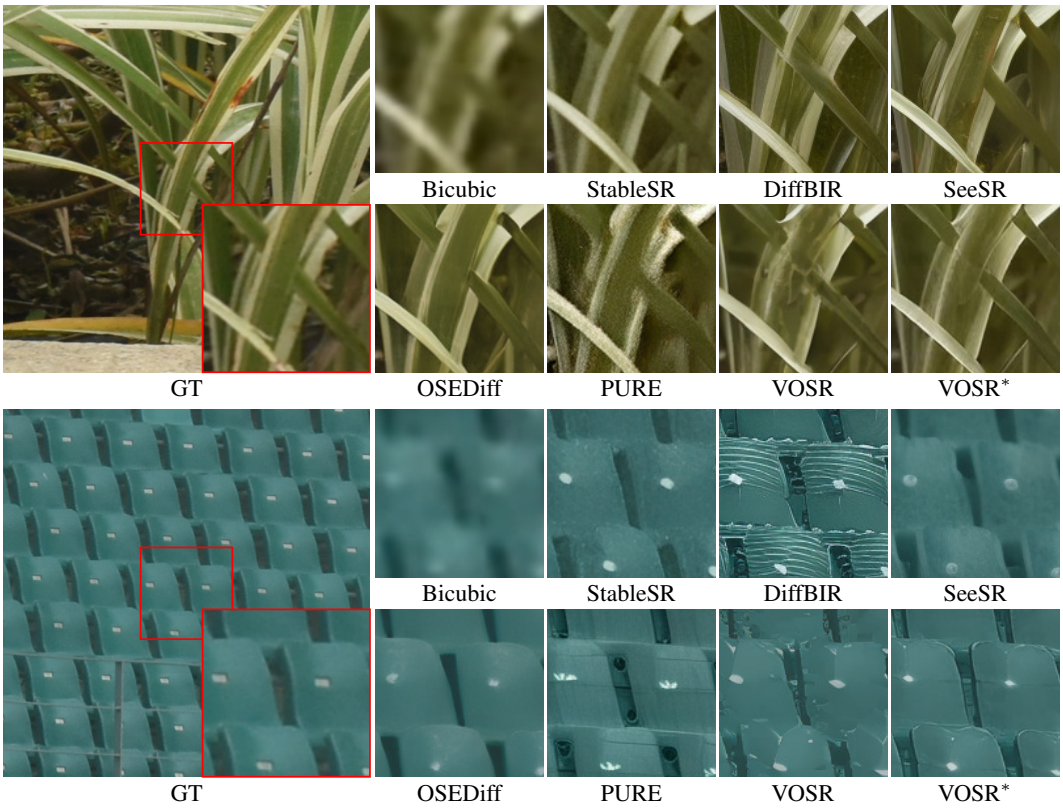


Figure 9: Qualitative comparisons of bicubic upsampling and generative SR methods on DRealSR testset. VOSR*, fine-tuned using our VHA+ score and MUSIQ as auxiliary losses, shows fewer hallucinated details and more faithfully preserves ground-truth content than pretrained VOSR.