LOCALIZING FAILURES: HALLUCINATION-GROUNDED PREFERENCE LEARNING FOR FAITHFUL GENERATIVE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Super-Resolution (SR) can improve perceptual quality but often introduces **SR hallucinations**—semantically unsupported objects, textures, structures, or text that are inconsistent with the paired Ground Truth (GT). Existing preference optimization methods rely on either image-level signals without spatial information or predefined spatial units that do not directly localize actual SR hallucinations. To address this limitation, we propose a framework comprising Hallucination-Grounded-Marks (HGM) and Hallucination-Selective Preference Optimization (HSPO). HGM uses a VLM to localize candidate regions on a 16×16 grid and verifies them against the GT to produce a cell-level binary hallucination mask, without requiring precomputed object regions or boxes. Since existing perceptual metrics struggle to quantify hallucination severity, we further introduce the Area-Weighted Hallucination Score (AHS) as a composite evaluation metric. HSPO exchanges hallucinated and faithful regions among SR candidates generated from the same Low-Resolution (LR) input to construct preference pairs for region-level optimization. This alleviates global preference ambiguity caused by inconsistent hallucination locations while preventing faithful regions from dominating the optimization signal. Experiments on Syn-TestSet and RealSR-4x show lower hallucination scores than C-FLUX and DP^{2}O while maintaining competitive perceptual quality.

1 INTRODUCTION

Single-Image Super-Resolution (SISR) has evolved from Peak Signal-to-Noise Ratio (PSNR)-driven methods (Dong et al., 2016; Lim et al., 2017) to generative approaches, such as diffusion (Wu et al., 2024b; Yang et al., 2024) and flow-matching models (Lipman et al., 2023; Esser et al., 2024), with steadily improving perceptual realism. By synthesizing plausible high-frequency details, generative SR can produce high-quality results even at large magnification factors. However, the inherent ambiguity of the degraded Low-Resolution (LR) image may cause models to generate semantically unsupported content that is inconsistent with the Ground Truth (GT), namely, hallucinations (Cohen et al., 2024; Ren et al., 2025). As shown in Figure 1, our method reduces such hallucinations. Cohen et al. (2024) reveal an inherent trade-off between perceptual quality and reliability, showing that improved perceptual quality is generally accompanied by greater uncertainty and hallucination risk. Our objective is therefore to suppress hallucinations while preserving perceptual quality, rather than allowing the model to regress toward over-smoothed outputs.

Reward- or preference-based optimization methods for generative SR can be divided into image-level and region-level approaches. At the image level, DP^{2}O (Wu et al., 2025) aggregates multiple Image Quality Assessment (IQA) rewards for preference optimization, RefReward-SR (Song et al., 2026) trains an LR-conditioned reward model via Group Relative Policy Optimization (GRPO), and Ren et al. (2025) fine-tune SR models using Vision-Language Model (VLM)-generated hallucination scores. At the region level, FinPercep-RM (Liu et al., 2026) generates perceptual degradation maps to localize defects, while DSPO (Cai et al., 2025) performs preference optimization over Segment Anything Model (SAM)-segmented regions. These methods mainly optimize perceptual quality and fidelity rather than detecting and selectively optimizing hallucination regions. Moreover, region-level methods such as DSPO rely on precomputed segments, which may miss hallucinations that

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.

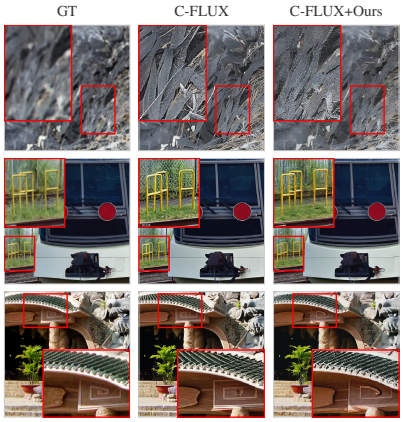

Figure 1: Qualitative examples comparing GT, C-FLUX, and C-FLUX+Ours. Red boxes indicate enlarged regions for closer inspection.

Category	Definition
Object Modification	Objects added, removed, replaced by a different semantic category, or count changed without plausible cause
Structural Distortion	Scene elements misplaced, occlusion inverted, geometric structures (lines, grids, contours) collapsed
Texture Artifacts	Local surface material property altered (e.g., wood grain → fur)
Text Distortion	Glyphs warped into unreadable scribbles, text rewritten in a different language, or legible text vanished
Excluded: mild SR artifacts like blur, over-smoothing, mild detail distortion, color shifts and noise	

Table 1: Hallucination taxonomy for VLM detection prompt.

cross or lie outside semantic boundaries while including faithful content within selected regions. Visual examples are provided in the supplementary material. We therefore directly localize hallucination regions against the GT and build a detection and optimization pipeline.

We make two observations that motivate our approach. First, hallucinations are spatially sparse: in our annotated data, they cover only approximately 10–20% of an image on average, so global preference signals are dominated by faithful content. Optimization should therefore follow localized failures rather than the entire image or precomputed instances. Second, stochastic SR candidates exhibit diverse hallucination locations: one candidate may be faithful in a region where another hallucinates, making it difficult to assign unambiguous image-level chosen/rejected labels when constructing Direct Preference Optimization (DPO) training pairs. We exploit this diversity to exchange faithful and hallucinated regions, thereby constructing synthetic preference pairs and restricting their learning space to the detected failure regions.

Our main contributions are:

(i) We design **Hallucination-Grounded-Marks (HGM)**, a two-stage pipeline that provides regional information beyond direct VLM scoring. Without relying on precomputed segmentation masks or object boxes, HGM localizes failures directly through GT-grounded verification; its area-weighted output serves as a hallucination evaluation metric.

(ii) We introduce a region-exchange strategy for dataset construction and build approximately 56K preference pairs, addressing the spatial inconsistency among stochastic SR candidates.

(iii) We introduce **Hallucination-Selective Preference Optimization (HSPO)** for finer-grained learning than global optimization, preventing the learning signal from being overwhelmed by irrelevant regions.

2 Related Work

2.1 Generative Super-Resolution

Diffusion-based SR models (Wu et al., 2024b; Yang et al., 2024; Sun et al., 2024) and flow-matching-based SR models (Lipman et al., 2023; Esser et al., 2024) have significantly improved perceptual quality. However, Cohen et al. (2024) and Ren et al. (2025) find that greater generative capacity can also produce semantically unfaithful details. For example, SRSR (Chen et al., 2025a) and DTPSR (Jiang et al., 2026) improve semantic fidelity through spatial text guidance and disentangled textual priors, respectively.

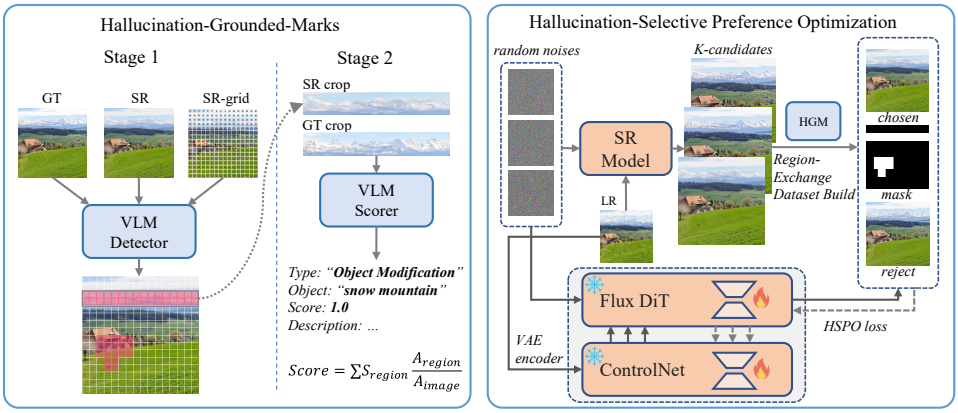

Figure 2: Overview of HGM and HSPO. HGM obtains hallucination regions through two-stage localization and scoring. HSPO then samples K candidates with different random seeds and exchanges regions to construct preference pairs, which are used to train LoRA on the diffusion transformer (DiT) and ControlNet.

2.2 Hallucination Assessment

Full-Reference/No-Reference (FR/NR) IQA metrics (Wang et al., 2004; Ke et al., 2021; Wu et al., 2024a) and hallucination-specific scores (Ren et al., 2025; Tivnan et al., 2024) measure image fidelity or overall quality but do not localize semantic hallucinations. Existing spatial detection methods (Xie et al., 2023; Li et al., 2025a) primarily target generic artifacts or reconstruction uncertainty. In contrast, our HGM adapts Set-of-Mark prompting (Yang et al., 2023) for region-level semantic verification of SR grids against GT.

2.3 Hallucination Mitigation

Global preference optimization (Wallace et al., 2024; Wu et al., 2025; Song et al., 2026; Ren et al., 2025) provides no spatial information about where hallucinations occur. For Real-ISR, DSPO (Cai et al., 2025) uses SAM to identify semantic instances and performs preference optimization at the instance level. FinPercep-RM (Liu et al., 2026) uses SAM semantic masks to exchange regions between GT and SR to synthesize localized degradations for a fine-grained reward model. In addition, SRSR (Chen et al., 2025a) improves local semantic fidelity through spatial guidance at inference time. Most of these methods do not focus on hallucinations, and their localized variants rely on precomputed semantic regions or synthetic degradations rather than regions determined by actual SR hallucinations. In contrast, HGM localizes actual hallucinations by comparing SR outputs with GT and uses the resulting masks to guide preference-pair construction and selective flow-matching optimization.

3 Method

Our failure-localized pipeline first uses **Hallucination-Grounded-Marks (HGM)** to identify GT-grounded hallucinations, then constructs region-level preference pairs and applies **Hallucination-Selective Preference Optimization (HSPO)** only to the detected failures.

3.1 Problem Definition

Following Ren et al. (2025), we define an SR hallucination as a difference between the SR output and the GT in which the generated visual elements form a recognizable anomaly. This definition is distinct from general image quality (e.g., CLIPIQA, NR-IQA) and does not require the SR output to exactly match the GT. For example, a blurry regressive output may have low quality but no hallucinations, whereas a sharp generative output may exhibit hallucination levels ranging from negligible to severe.

For detection, we use the four-category VLM taxonomy defined in Table 1, which contains Object Modification, Structural Distortion, Texture Artifacts, and Text Distortion. From top to bottom, Figure 1 illustrates the first three categories, respectively.

3.2 HALLUCINATION-GROUNDED-MARKS

HGM aims to localize and assess SR hallucinations without relying on precomputed semantic regions. In the two-stage detection pipeline (Figure 2), Stage 1 localizes candidate cells on a 16×16 grid. The cells associated with the same detected anomaly form a candidate region, and Stage 2 verifies each region against the corresponding GT crop to remove false positives and assess hallucination severity. The verified regions form the cell-level binary hallucination mask M.

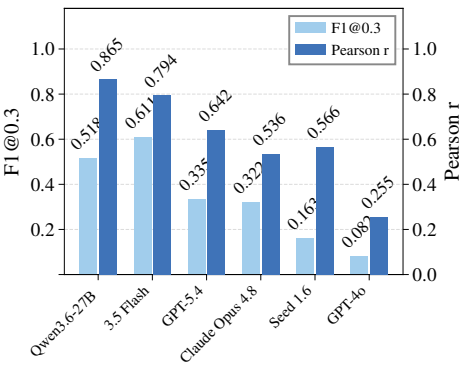

Figure 3: Performance of different VLMs on the calibration set. Bars show area correlation with human annotations and per-image F1 at a 0.3 annotation-overlap threshold.

3.2.1 GRID-OF-MARKS DETECTION

Stage 1 follows Set-of-Mark (Yang et al., 2023) but partitions I_{SR} into an indexed 16×16 grid, producing $I_{\text{SR-grid}}$. The triplet $(I_{\text{GT}}, I_{\text{SR}}, I_{\text{SR-grid}})$ is then provided in a single multi-image VLM query. The prompt includes the taxonomy (Table 1), marking instructions, and examples, and requires output in JSON format. For each detected anomaly, the VLM returns its category, description, and cell indices; these indices define a candidate region in the candidate cell mask. The full prompts are provided in the supplementary material.

3.2.2 STAGE 2: CANDIDATE VERIFICATION AND SEVERITY SCORING

Each candidate region detected in Stage 1 is expanded with a small margin, cropped into a rectangle, and provided to the VLM alongside the corresponding GT crop. *Verification* distinguishes semantic hallucinations from false positives such as blur and over-smoothing, while *scoring* assigns a severity score $s_{\text{region}} \in [0, 1]$ based on the anomaly's noticeability and deviation from the GT, where 0 indicates negligible impact and 1 indicates a severe semantic error. Verified regions form the binary hallucination mask, while false positives are discarded. The severity score is used only for evaluation, whereas HSPO uses the binary hallucination mask for optimization. Inference with Qwen3.6-27B takes approximately 2 seconds per image on two A800 GPUs.

3.2.3 AREA-WEIGHTED HALLUCINATION SCORE

Ren et al. (2025) directly use a VLM to score hallucinations in each image, making the assessment subjective and lacking region-level interpretability. To address this limitation, we propose the Area-Weighted Hallucination Score (AHS):

$$\underbrace{\text{AHS}}_{\text{Area-Weighted Hallucination Score}} = \sum\nolimits_{\text{regions}} s_{\text{region}} \times \frac{A_{\text{region}}}{A_{\text{image}}}, \tag{1}$$

where s_{region} is the Stage 2 severity and $A_{\text{region}}/A_{\text{image}}$ is the region-to-image area ratio. This continuous, evaluation-only score captures varying perceptual impacts across regions within the same hallucination category.

We denote AHS (Eq. 1) computed with Qwen and Gemini as **AHS-Q** and **AHS-G**, respectively. Because Qwen constructs binary hallucination masks and computes AHS-Q, we additionally report AHS-G computed by the independent Gemini model. **Hal.Img-Q/G** count images with at least one region detected by the respective VLM.

3.2.4 VLM DETECTION BENCHMARK

To validate HGM localization and select VLMs for mask construction and evaluation, after optimizing the Qwen prompt we evaluate all VLMs on the same 100-image human-annotated SeeSR set from DIV2K-val (Agustsson & Timofte, 2017). These calibration crops are distinct from the 300 crops used in Syn-TestSet. Figure 3 reports Pearson r between predicted and annotated hallucination areas and per-image F1; each predicted cell is evaluated separately and matches an annotation when their intersection covers at least 30% of the annotated region. Both Qwen and Gemini align well with human judgments. We use locally deployable Qwen3.6-27B (Qwen Team, 2026) for dataset building and evaluation, and Gemini 3.5 Flash (Google DeepMind, 2026) as the independent evaluator.

3.3 HALLUCINATION-SELECTIVE PREFERENCE OPTIMIZATION

HSPO constructs region-level preference pairs from SR candidates generated with random noise to address the issue of hallucination-region specificity in SR images, then employs region-level loss computation to restrict learning to regions where failures occur.

3.3.1 PREFERENCE DATASET BUILD

Different noise inputs produce diverse hallucination patterns across stochastic SR candidates, naturally supervising region-level preference learning. Image-level ranking remains ambiguous when candidates exhibit hallucinations in different regions; region exchange instead isolates these complementary differences, producing pairs in which regions in the chosen image are faithful while the corresponding regions in the rejected image contain hallucinations. Using **C-FLUX** (ControlNet-Flux (Wu et al., 2025) with flow matching (Lipman et al., 2023)), we generate $K = 4$ candidates $\{I_{\mathrm{SR}}^{(k)}\}_{k=1}^{K}$ with different seeds and obtain each rasterized cell-level binary hallucination mask $M^{(k)}$.

Chosen/rejected construction. We randomly select $I^{(a)}$ and $I^{(b)}$ from the $K = 4$ candidates, retaining their original content outside $M = M^{(a)} \cup M^{(b)}$. Within M, faithful and hallucinated content are assigned to I_{chosen} and I_{rejected}, respectively. If both candidates hallucinate in a region, the remaining candidate with the lowest regional severity is chosen; ties favor higher PSNR, and GT is used only when none is faithful. This fallback affects only the chosen region, while the rejected image retains hallucinated SR content. Gaussian feathering ($\sigma = 16$ px) reduces boundary discontinuities. On average, M covers 17% of the image, with GT fallback accounting for 5.9% of the full image area, yielding approximately 56K pairs.

3.3.2 HSPO LOSS

Global preference optimization applies the same pairwise signal to failures and faithful content. We replace its flow-matching Mean Squared Error (MSE) with a masked objective, restricting learning to failures and avoiding irrelevant updates that can degrade faithful perceptual content. Let $v_{\mathrm{target}} \in \mathbb{R}^{N_{\mathrm{seq}} \times C}$ be the flow-matching target velocity (the straight-line noise-to-data path) and $\hat{v}_\theta$ the prediction. Average-pooling M to the packed Flux latent resolution gives $W \in [0,1]^{N_{\mathrm{seq}}}$, with a $16\times$ effective reduction from Variational Autoencoder (VAE) encoding and 2×2 latent packing, N_{seq} positions, and C channels. The masked MSE

$$\ell_{\mathrm{masked}}(\hat{v}_\theta, v_{\mathrm{target}}) = \Big(\sum_{i=1}^{N_{\mathrm{seq}}} W_i\Big)^{-1} \sum_{i=1}^{N_{\mathrm{seq}}} \frac{W_i}{C} \sum_{c=1}^{C} (\hat{v}_{\theta,i,c} - v_{\mathrm{target},i,c})^2 \tag{2}$$

If $\sum_{i=1}^{N_{\mathrm{seq}}} W_i = 0$, neither candidate has a detected hallucination, so the pair has no learnable hallucination signal and is skipped. Gaussian feathering reduces exchange-boundary discontinuities.

Using this masked error, HSPO adopts Diffusion-DPO's pairwise logistic objective (Wallace et al., 2024). Superscripts w and l denote chosen and rejected composites. They share LR conditioning, timestep t, and the same sampled noise ϵ; within each branch, the trainable and frozen reference models receive the same noisy state. All four loss terms use the same weights W obtained from the pair's union mask M. The factor $1/2$ follows the noise-aware Diffusion-DPO objective; constants induced by timestep weighting and mean reduction are absorbed into the effective preference temperature, which we set to $\beta = 1000$. Here, σ is the logistic sigmoid,

Syn-TestSet (300 images)

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	Q-Align↑	FID↓	FD-DINOv3↓	AHS-Q↓	Hal.Img-Q	AHS-G↓	Hal.Img-G
Bicubic	24.95	0.6627	0.5921	24.82	0.3689	1.320	123.11	641.76	0.0004	1/300	0.0001	1/300
CCSR	24.10	0.6471	0.3526	55.86	0.5453	2.824	119.18	1055.57	0.0171	44/300	0.0083	62/300
FiDeSR	23.70	0.6227	0.2537	65.64	0.6815	3.711	73.62	750.63	0.0197	58/300	0.0122	91/300
StableSR	23.71	0.6337	0.3149	51.48	0.4810	2.622	82.19	834.80	0.0226	50/300	0.0111	72/300
StructSR	23.46	0.6105	0.3229	60.66	0.6899	3.190	105.08	1011.74	0.0310	56/300	0.0123	81/300
ODTSR	23.06	0.6144	0.2609	65.40	0.5939	3.615	74.32	713.81	0.0327	71/300	0.0151	94/300
SeeSR	23.30	0.6113	0.2927	64.83	0.6666	3.769	83.63	902.89	0.0450	95/300	0.0236	132/300
PASD	21.20	0.5086	0.3784	69.04	0.7410	4.007	113.82	1227.73	0.1323	142/300	0.0589	162/300
C-FLUX	22.09	**0.5577**	0.3192	**66.20**	0.6591	**3.805**	91.36	902.41	0.0444	101/300	0.0423	121/300
+ Ours	22.09	0.5466	0.3294	65.91	**0.6950**	3.766	84.48	852.02	0.0286	81/300	0.0208	110/300
+ Ours $_{\times 1.5}$	**22.44**	0.5574	**0.3070**	63.73	0.6578	3.504	**81.04**	**784.09**	**0.0238**	**67/300**	**0.0145**	**87/300**
DP^{2}O	22.33	0.5646	0.3005	**69.89**	0.7426	**4.188**	90.50	885.99	0.0476	92/300	0.0406	97/300
+ Ours	22.54	0.5642	0.3027	69.66	**0.7562**	4.172	**86.67**	860.20	0.0408	78/300	0.0277	103/300
+ Ours $_{\times 1.5}$	**22.67**	**0.5717**	**0.2972**	68.47	0.7383	3.882	86.76	**833.14**	**0.0324**	**71/300**	**0.0179**	**88/300**

RealSR-4x (100 images)

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	Q-Align↑	FID↓	FD-DINOv3↓	AHS-Q↓	Hal.Img-Q	AHS-G↓	Hal.Img-G
Bicubic	24.97	0.7175	0.4757	24.90	0.3118	1.617	144.16	698.54	0.0010	2/100	0.0000	0/100
CCSR	24.05	0.7255	0.3034	61.83	0.5636	3.065	145.58	1161.23	0.0589	21/100	0.0244	35/100
FiDeSR	23.91	0.7051	0.2723	69.91	0.6920	3.667	112.97	862.26	0.0629	33/100	0.0306	34/100
StableSR	23.87	0.7144	0.2612	55.52	0.4516	2.731	129.19	1060.33	0.0917	28/100	0.0258	39/100
StructSR	23.25	0.6851	0.3459	64.95	0.6568	3.285	152.01	1183.68	0.0772	35/100	0.0310	41/100
ODTSR	22.96	0.6946	0.2430	67.83	0.5727	3.588	102.11	803.69	0.0706	20/100	0.0254	34/100
SeeSR	23.51	0.6904	0.2919	68.12	0.6441	3.588	124.37	998.99	0.0851	24/100	0.0392	50/100
PASD	21.37	0.6002	0.3661	70.45	0.7089	3.812	149.18	1381.10	0.1735	42/100	0.0557	60/100
C-FLUX	23.03	**0.6698**	**0.3048**	67.67	0.5834	3.430	**117.38**	**881.30**	0.0747	38/100	0.0295	41/100
+ Ours	23.16	0.6582	0.3264	**67.98**	0.6408	**3.556**	120.46	917.98	0.0617	35/100	**0.0260**	**35/100**
+ Ours $_{\times 1.5}$	**23.25**	0.6588	0.3219	67.43	**0.6496**	3.521	120.42	928.89	**0.0551**	**30/100**	0.0282	37/100
DP^{2}O	23.06	0.6652	0.3032	**72.08**	0.7044	3.947	120.58	**893.54**	0.0954	**28/100**	0.0399	42/100
+ Ours	23.41	0.6643	0.3065	72.00	**0.7288**	**3.987**	**117.79**	931.57	0.0846	37/100	**0.0267**	**35/100**
+ Ours $_{\times 1.5}$	**23.44**	**0.6680**	**0.2946**	70.49	0.7107	3.698	121.36	894.82	**0.0676**	34/100	0.0341	39/100

Table 2: Performance comparison on **Syn-TestSet** (300 images) and **RealSR-4x** (100 images). ↑/↓: higher/lower is better. **AHS-Q↓/AHS-G↓**: Area-Weighted Hallucination Scores evaluated by Qwen3.6-27B and Gemini 3.5 Flash; **Hal.Img-Q/Hal.Img-G**: the numbers of images containing at least one detected hallucination region. Bold: best; underlined: second best across our baselines and methods.

$\mathcal{D}_{\text{pref}}$ is the constructed preference dataset, the reference is the initial backbone checkpoint, and $\Delta = (\ell^w_{\text{masked}} - \ell^l_{\text{masked}}) - (\ell^w_{\text{masked,ref}} - \ell^l_{\text{masked,ref}})$:

$$\mathcal{L}_{\text{HSPO}} = -\mathbb{E}_{(I^w, I^l)\sim\mathcal{D}_{\text{pref}},\, t\sim p(t),\, \epsilon\sim\mathcal{N}(0,I)} \log\sigma\left(-\frac{\beta}{2}\Delta\right) \tag{3}$$

Total objective. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{HSPO}} + \lambda_{\text{mse}}\mathcal{L}_{\text{MSE}}$, where each step computes $\mathcal{L}_{\text{MSE}}$ on paired GT under the same LR conditioning ($\lambda_{\text{mse}} = 0.1$). Both terms update the ControlNet and DiT LoRA adapters; the reference remains frozen.

4 Experiments

We evaluate HSPO from C-FLUX and DP^{2}O initializations, examining hallucination suppression across random seeds and at comparable perceptual quality.

4.1 Experimental Setup

Baselines. We evaluate HSPO from two initializations: C-FLUX and its boosted version DP^{2}O (Wu et al., 2025). We include CCSR (Chen et al., 2025b), FiDeSR (Kim et al., 2026), StableSR (Wang et al., 2024), StructSR (Li et al., 2025b), ODTSR (Fang et al., 2026), SeeSR (Wu et al., 2024b), and

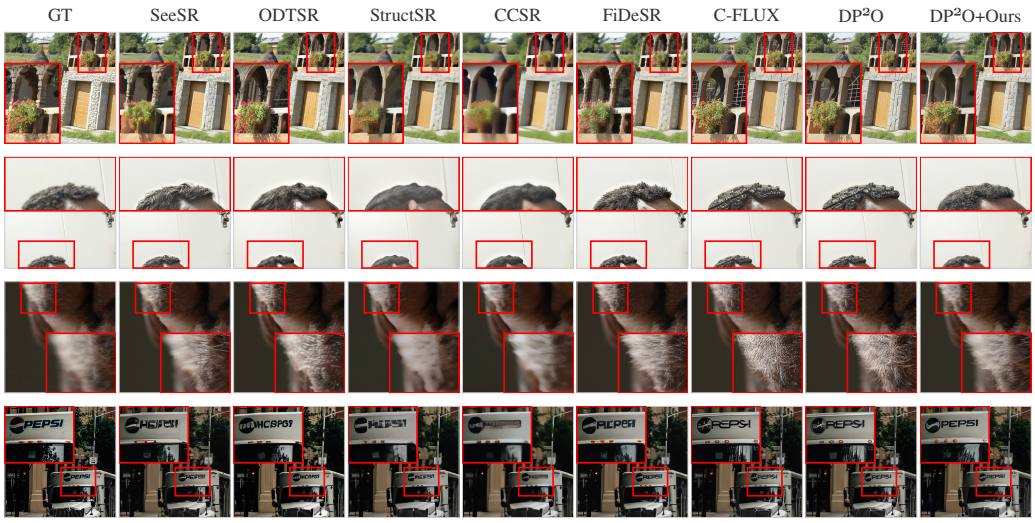

Figure 4: Qualitative comparison of generative SR methods. Highlighted boxes indicate enlarged regions for closer inspection.

PASD (Yang et al., 2024). DSPO (Cai et al., 2025) and FinPercep-RM (Liu et al., 2026) could not be included in the comparison because their code or model weights are unavailable.

Data. For training, we generate four C-FLUX outputs per LSDIR crop (Li et al., 2023), yielding approximately 38K SR images and 56,320 preference pairs. To construct Syn-TestSet, we randomly sample three 512×512 crops from each of the 100 DIV2K-val images (Agustsson & Timofte, 2017), resulting in 300 GT crops, and generate their LR inputs using the StableSR degradation pipeline (Wang et al., 2024). RealSR-4x (Cai et al., 2019) uses 512×512 center crops from 100 images.

Metrics. As reported in Table 2, we evaluate pixel-level fidelity using metrics such as PSNR, perceptual quality using metrics such as MUSIQ, and hallucinations using AHS-Q and AHS-G. Both AHS variants are reported for all methods, with AHS-G serving as an independent evaluation by a separate VLM. All VLM-based scores are computed three times and averaged.

Inference. All methods use the same fixed random seed and an empty text prompt as the text conditioning input to ensure fair comparison.

Training. We train ControlNet and DiT LoRA adapters (Hu et al., 2022) (ranks 32 and 64) for 500 steps using AdamW with a learning rate of 5×10^{-5}. Training uses two A800 GPUs with a per-GPU micro-batch of 1 and gradient accumulation over 64 steps, yielding an effective batch size of 128.

4.2 Main Results

Table 2 gives a broad SR comparison for the perception–hallucination trade-off and controlled HSPO comparisons initialized from C-FLUX and DP^{2}O. Across the broader comparison, Qwen and Gemini show consistent overall trends: Bicubic has the lowest hallucination scores, while PASD has the highest on both test sets. AHS should be interpreted alongside perceptual metrics because detail-poor regressive outputs can avoid hallucinations. **Ours** uses a LoRA fusion weight of 1.0; the optional **Ours**$_{\times 1.5}$ uses 1.5× LoRA and 1.2× ControlNet weights for the first 10 sampling steps, then 1.0× for the remaining 15.

At default fusion, HSPO reduces AHS-Q by 35.6%/17.4% from C-FLUX and 14.3%/11.3% from DP^{2}O on Syn-TestSet/RealSR-4x; corresponding AHS-G reductions are 50.8%/11.9% and 31.8%/33.1%. Across all four controlled comparisons, PSNR is maintained or improved, CLIPIQA increases, and both AHS variants decrease. The $\times 1.5$ setting further lowers hallucination scores at a measurable perceptual cost.

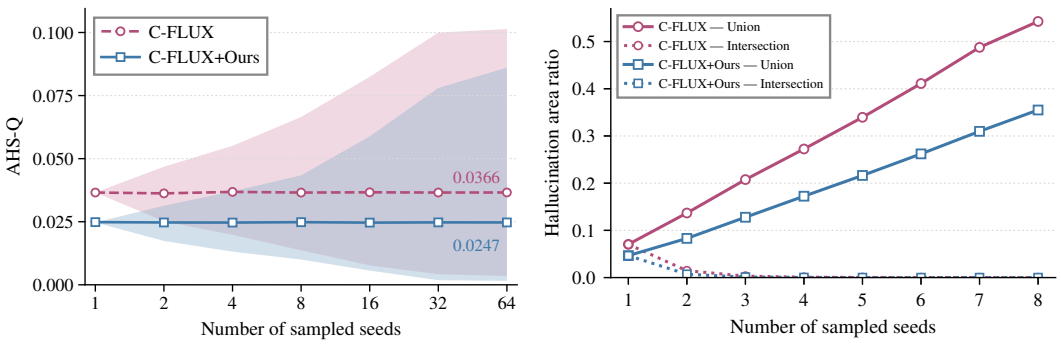

Figure 5: Multi-seed analysis on Syn-TestSet: (left) AHS-Q mean across K, with shaded bounds showing the average maximum and minimum; (right) intersection and union areas of hallucination masks across K.

On RealSR-4x, DP^{2}O+Ours lowers AHS-Q but raises Hal.Img-Q, indicating more frequent yet milder hallucinations. Under Gemini, **Ours**$_{\times 1.5}$ raises AHS-G and Hal.Img-G over **Ours**; stronger scaling is thus non-monotonic and may add blur or artifacts, but offers inference-time trade-off control without re-training. Its enlarged regions show that our method suppresses localized GT-inconsistent content while retaining faithful structures and perceptual details outside those failures. Thus, the score reductions reflect corrections of localized semantic inconsistencies rather than uniform smoothing. Figure 4 provides a qualitative comparison.

4.3 Quantitative Experiments

In the quantitative experiments, we validate the rationale behind multi-seed sampling for dataset construction and the region-exchange strategy, demonstrate the consistent improvements achieved across different seeds, and evaluate the effects of different optimization methods.

Quantitative Experiments on Multi-seed Sampling. Our dataset construction and training build on multi-seed sampling, which produces hallucinations of different types and in different regions. To validate this premise, we generate $N = 64$ candidates with different random seeds for each of the first 20 images in Syn-TestSet. For each $K \in \{1, 2, 4, 8, 16, 32, 64\}$, we randomly sample K candidates from this pool for 5,000 trials and compute the mean, minimum, and maximum AHS-Q. Figure 5 compares C-FLUX and C-FLUX+Ours.

As shown in Figure 5, increasing K raises the maximum and lowers the minimum AHS-Q, indicating that different random seeds produce hallucinations of varying severity that can be used to construct chosen images with fewer hallucinations and rejected images with more hallucinations. Ours also lowers the mean, minimum, and maximum AHS-Q at every K, demonstrating consistent hallucination suppression across different random seeds rather than improvement from a particular seed.

We further measure the image-area ratios of their intersection and union. As shown in Figure 5, the intersection approaches zero while the union grows with K. This result validates the hypothesis of hallucination-region specificity and supports the rationale of our region-exchange strategy for preference dataset construction. Balancing hallucination diversity and computational cost, we use $K = 4$.

Quantitative Experiments on Optimization Methods.

To analyze the effects of region exchange and HSPO, we conduct the following experiments while keeping training settings fixed. During inference, we adjust the LoRA weights to ensure comparable perceptual quality across the outputs.

The variants use (i) **Ours**: the original HSPO; (ii) **Random Mask**: an area-matched random masked loss; (iii) **Global**: a full-image loss; and (iv) **Global**$_{\text{w/o exchange}}$: a full-image loss with a ranked preference dataset. Across variants with comparable perceptual quality, Ours achieves the lowest AHS-Q and Hal.Img-Q. Hallucination suppression degrades progressively when the selective loss

Method	PSNR	MUSIQ	CLIPIQA	AHS-Q↓	Hal.Img-Q
Ours	22.09	65.91	0.695	**0.0286**	**81/300**
Random Mask	22.01	64.83	0.636	0.0372	85/300
Global	21.89	65.27	0.634	0.0389	89/300
$\text{Global}_{\text{w/o exchange}}$	21.65	66.07	0.640	0.0455	92/300

Table 3: Comparison of different training strategies on Syn-TestSet.

in HSPO is replaced with a random masked loss, when supervision is expanded to the full image, and when region exchange is further removed. This trend shows that region exchange reconciles the conflict of hallucination heterogeneity within preference pairs, while selective optimization concentrates learning on localized hallucinations. HSPO benefits from them.

4.4 Human Evaluation

To test whether AHS reductions match human perception, six users provide 360 judgments over 180 comparisons between DP^{2}O and DP^{2}O+Ours on Syn-TestSet and RealSR-4x (60 per user; two per comparison). Each comparison presents GT and the two SR outputs in randomized left–right order. Users judge hallucinations rather than generic quality and choose "Left is better", "Tie", or "Right is better". As shown in Figure 6, DP^{2}O+Ours is preferred more often than DP^{2}O. Frequent ties arise because neither output contains hallucinations in a substantial proportion of the comparisons. Overall, the trend supports the perceptual relevance of the AHS reduction from DP^{2}O.

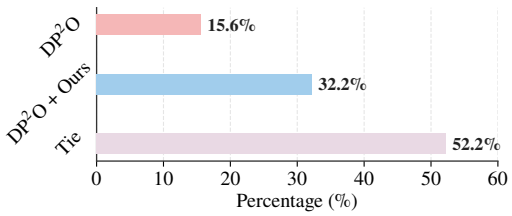

Figure 6: Human preference results comparing DP^{2}O with DP^{2}O+Ours.

4.5 Limitations

HGM depends on paired GT for detection and training, limiting fully blind SR and its applicability to generative tasks without reference images, such as text-to-image generation. Its reliance on VLM grounding may also propagate localization errors and model biases into preference construction, training supervision, and evaluation.

5 Conclusion

We presented a failure-localized framework in which HGM identifies GT-grounded hallucination regions, region exchange constructs spatially aligned preference pairs, and HSPO selectively optimizes these regions. AHS evaluates hallucinations by their severity and spatial extent. Across C-FLUX and DP^{2}O initializations, HSPO consistently lowers AHS-Q and AHS-G on Syn-TestSet and RealSR-4x while retaining competitive perceptual quality.

6 Reproducibility Statement

We provide implementation and training details in Sec. 4.1, including datasets, optimization settings, and inference configurations. Additional details of the HGM detection pipeline and VLM prompts are provided in the supplementary material. Code and model checkpoints will be released upon publication.

7 AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, implement methods, assist with translation, and clean and reformat the dataset. We have not used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative or thematic data analysis, or interpret results. Developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs were not applicable to this work. We have reviewed all AI-assisted work. We manually reviewed the dataset and code to check their correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3086–3095, 2019. doi: 10.1109/ICCV.2019.00318. URL https://arxiv.org/abs/1904.00523.

Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. DSPO: Direct semantic preference optimization for real-world image super-resolution, 2025. URL https://arxiv.org/abs/2504.15176. arXiv:2504.15176.

Chen Chen, Majid Abdolshah, Violetta Shevchenko, Hongdong Li, Chang Xu, and Pulak Purkait. SRSR: Enhancing Semantic Accuracy in Real-World Image Super-Resolution with Spatially Re-Focused Text-Conditioning, 2025a. URL https://arxiv.org/abs/2510.22534v1.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging Fidelity-Reality with Controllable One-Step Diffusion for Image Super-Resolution, December 2025b. URL http://arxiv.org/abs/2512.14061. arXiv:2512.14061 [cs].

Regev Cohen, Idan Kligvasser, Ehud Rivlin, and Daniel Freedman. Looks Too Good To Be True: An Information-Theoretic Analysis of Hallucinations in Generative Restoration Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, October 2024. doi: 10.52202/079017-0711. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2847d43f17410c5beb25b2736c3ae778-Abstract-Conference.html.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. doi: 10.48550/arXiv.2403.03206. URL http://arxiv.org/abs/2403.03206.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-Step Diffusion Transformer for Controllable Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23440–23450, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fang_One-Step_Diffusion_Transformer_for_Controllable_Real-World_Image_Super-Resolution_CVPR_2026_paper.html.

Google DeepMind. Gemini 3.5 Flash model documentation. https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash, 2026. Accessed 2026-07-27.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2106.09685. URL http://arxiv.org/abs/2106.09685.

Lei Jiang, Xin Liu, Xinze Tong, Zhiliang Li, Jie Liu, Jie Tang, and Gangshan Wu. DTPSR: Disentangled textual priors for diffusion-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38228–38237, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_Disentangled_Textual_Priors_for_Diffusion-based_Image_Super-Resolution_CVPR_2026_paper.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. FiDeSR: High-Fidelity and Detail-Preserving One-Step Diffusion Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38270–38280, 2026.

Jianfei Li, Ines Rosellon-Inclan, Gitta Kutyniok, and Jean-Luc Starck. CHEM: Estimating and Understanding Hallucinations in Deep Learning for Image Processing, December 2025a. URL http://arxiv.org/abs/2512.09806. arXiv:2512.09806 [cs].

Yachao Li, Dong Liang, Tianyu Ding, and Sheng-Jun Huang. StructSR: Refuse spurious details in real-world image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5022–5030, 2025b. doi: 10.1609/aaai.v39i5.32532. URL https://ojs.aaai.org/index.php/AAAI/article/view/32532.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1775–1787, 2023. doi: 10.1109/CVPRW59228.2023.00178.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2210.02747. URL http://arxiv.org/abs/2210.02747.

Yidi Liu, Zihao Fan, Jie Huang, Jie Xiao, Dong Li, Wenlong Zhang, Lei Bai, Xueyang Fu, and Zheng-Jun Zha. FinPercep-RM: A Fine-grained Reward Model and Co-evolutionary Curriculum for RL-based Real-world Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4839–4849, June 2026. doi: 10.48550/arXiv.2512.22647. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_FinPercep-RM_A_Fine-grained_Reward_Model_and_Co-evolutionary_Curriculum_for_RL-based_CVPR_2026_paper.html.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, April 2026.

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, July 2025. URL http://arxiv.org/abs/2507.14367. arXiv:2507.14367 [cs].

Yushuai Song, Weize Quan, Weining Wang, Jiahui Sun, Jing Liu, Meng Li, Pengbin Yu, Zhentao Chen, Wei Shen, Lunxi Yuan, and Dong-ming Yan. RefReward-SR: LR-Conditioned Reward Modeling for Preference-Aligned Super-Resolution, March 2026. URL https://arxiv.org/abs/2603.24198. arXiv:2603.24198 [cs.CV].

Lingchen Sun, Rongyuan Wu, Jie Liang, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the Stability and Efficiency of Diffusion Models for Content Consistent Super-Resolution, September 2024. URL http://arxiv.org/abs/2401.00877. arXiv:2401.00877 [eess].

Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination Index: An Image Quality Metric for Generative Reconstruction Models. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15010, pp. 449–458. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72116-8 978-3-031-72117-5. doi: 10.1007/978-3-031-72117-5_42. URL https://link.springer.com/10.1007/978-3-031-72117-5_42. Series Title: Lecture Notes in Computer Science.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion Model Alignment Using Direct Preference Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, June 2024. doi: 10.1109/CVPR52733.2024.00786. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7.

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.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Liu, et al. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_SeeSR_Towards_Semantics-Aware_Real-World_Image_Super-Resolution_CVPR_2024_paper.html.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Shihao Wang, Tianhe Wu, Qiaosi Yi, Shuai Li, and Lei Zhang. DP^{2}O-SR: Direct Perceptual Preference Optimization for Real-World Image Super-Resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, October 2025. doi: 10.48550/arXiv.2510.18851. URL https://neurips.cc/virtual/2025/poster/119975.

Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. DeSRA: Detect and Delete the Artifacts of GAN-based Real-World Super-Resolution Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38204–38226, 2023. URL https://proceedings.mlr.press/v202/xie23c.html.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V, November 2023. URL http://arxiv.org/abs/2310.11441. arXiv:2310.11441 [cs.CV].

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 74–91, July 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/3436_ECCV_2024_paper.php.

SUPPLEMENTARY MATERIAL

A GRID-BASED VS. SAM-BASED DETECTION

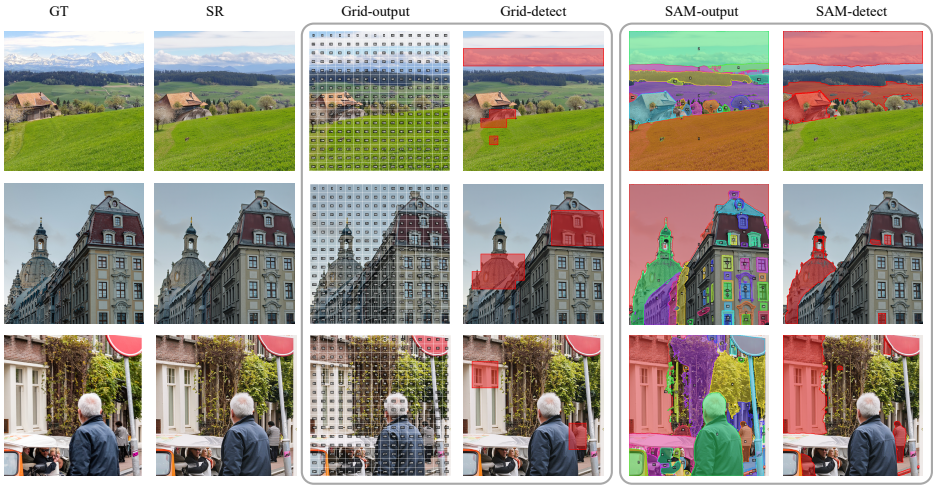

Figure 7: Comparison between our grid-based method and the SAM-based method. The middle two columns show the grid-based representation and detection result; the right two columns show the SAM-segmented regions and corresponding VLM detection results. Hallucinated regions are marked in red.

Our method relies on a Set-of-Mark-based visual prompting approach that lets the VLM automatically output the hallucination type and region. Earlier local optimization methods, such as DSPO (Cai et al., 2025) and FinPercep-RM (Liu et al., 2026), instead extract instances and rank them manually to construct their datasets. We abstract these prior methods as SAM-based instance detection, in which a VLM judges each instance; the prompt used for this judgment is identical to that of the scoring stage. Figure 7 shows qualitative results.

Compared with our grid-based method, the SAM-based approach has the following limitations:

- **Hallucination-region expansion.** SAM may merge semantically different areas into one region. In the case where a cloud is hallucinated in the mountain scene, it may group the mountain, cloud, and sky as one background, although only the cloud is hallucinated through an object modification. Depending on the SAM settings, the small animal that is fabricated out of thin air on the grass, as in the figure, may either be detected or filtered out. However, as shown in the figure, when the surrounding-grass crop is sent to the VLM, the crop-level judgment may attribute the hallucination to the entire grass region. If an irregularly shaped mask, rather than a crop, is fed to the VLM, the missing edge semantics or background information may also lead to misjudgment.
- **Hallucination-region omission.** SAM has inherent difficulty delineating small or fragmented objects. For example, only part of the large stone in front of the house may be covered by a SAM mask, causing part of the hallucinated region to be omitted from the result.
- **Misjudgment amplification.** The SAM-based approach incurs many misjudgments. For example, the trees in the figure are sent to the VLM as small regions and thus observed at a finer granularity, which exacerbates the misjudgment.
- **VLM call count.** Per image, our grid-based method uses $1 + n$ VLM calls, where n is the number of candidate regions: one global grid-detection call plus n candidate-region verification calls. The SAM-based method uses N calls, where N is the number of SAM regions, with one call per region. Although the fine-grained SAM crops can make local hallucinations easier for the VLM to identify, they require a separate VLM call for every region.

B VLM DETECTION PROMPTS

This section details the Vision-Language Model (VLM) prompts used in the two-stage Hallucination-Grounded-Marks (HGM) pipeline (Sec. 3.2 of the main paper). Stage 1 performs coarse Grid-of-Marks detection using a 16×16 indexed grid overlay. The cells associated with the same detected anomaly form a candidate region, and Stage 2 verifies each region against the corresponding GT crop and assigns a per-region severity score.

B.1 STAGE 1: GRID-OF-MARKS VLM DETECTION PROMPT

The Stage 1 prompt tasks a Vision-Language Model with detecting hallucinated cells on a 16×16 indexed grid. The model receives three images: the Ground Truth (I_{GT}), the Super-Resolution output (I_{SR}), and the grid-overlaid SR image ($I_{\mathrm{SR\text{-}grid}}$).

Compare the GT (Ground Truth reference image) with the SR (Super-Resolution output image), and use the SR-grid (the SR image with a grid overlay, size: {`grid_rows`}x{`grid_cols`}, cell numbers: 1–{`total_cells`}) to locate and detect "Super-Resolution hallucinations" in the SR image.

You will receive three images:

- **GT (Ground Truth reference):** the original High-Resolution reference image (no grid annotation).
- **SR (Super-Resolution output):** the image produced by the Super-Resolution algorithm (no grid annotation).
- **SR-grid (SR with grid):** the exact same image as SR, but divided into a grid, with a black-background white-text cell number at the center of each cell.

Hallucination Definition and Categories

Detect only the following four hallucination types.

Object Modification

Given that GT contains clear object outlines, categories, and counts, any of the following is judged a hallucination:

1. SR fabricates an independent object or partial entity that does not exist in GT;
2. An object that exists in GT disappears for no reason;
3. An object is entirely replaced by an object of a different category;
4. The number of objects increases or decreases without a reasonable cause.

Exclusions: slight edge rounding of an object, slight detail wear, or small brightness/contrast changes do not constitute an object-modification hallucination. Changes limited to surface texture, patterns, or leaf details, with the object's broad category unchanged, do not constitute object modification and may be classified as texture artifacts. An object that clearly exists in both GT and SR, even if rendered with slightly different shape, color, or detail, is NOT a fabricated/disappeared object and must not be reported.

Structural Distortion

Using GT's spatial layout, geometric perspective, hierarchy, and physical constraints as the baseline, any of the following is judged a hallucination:

1. The spatial position of scene elements is misplaced in a way that violates the original relative-layout logic;
2. Foreground/background, occlusion relationships, or front-back layering are reversed or distorted, violating real-world physical occlusion rules;
3. Geometric structures such as buildings, lines, grids, and straight outlines collapse: straight lines become irregular bends, or parallel/perpendicular relationships are destroyed.

Exclusions: slight line curvature, small perspective deviations, or minor layout shifts are quality distortions and are not judged as structural-distortion hallucinations.

Texture Artifacts

Any of the following is judged a hallucination:

1. Texture distortion changes local semantic attributes, altering the physical property of an object's surface, e.g. wood becomes fur, bricks become stone. Physical property refers to the essential material that constitutes the object.

Exclusions: slight texture blur, slight texture mismatch, over-smoothing, or small texture loss are only quality defects and do not constitute texture-artifact hallucinations.

Text Distortion

Any of the following is judged a hallucination:

1. The glyphs or strokes of text, numbers, or symbols are distorted into unreadable gibberish lines;

2. Text content is incorrectly rewritten into another language;
3. Clearly legible text disappears or is replaced by meaningless graphics.

Exclusions: slight text blur or slightly soft text edges, provided that the text remains legible, are quality defects and are not judged as text-distortion hallucinations.

Notes

The following are NOT judged as hallucinations.

Common Super-Resolution artifacts, blur/softening, slight detail distortion, texture loss/over-smoothing, color shift, compression artifacts, noise/grain, slight edge blur, slight texture mismatch, and slight brightness deviations.

Descriptions must be objective and factual; subjective or exaggerated wording is prohibited (completely, thoroughly, dramatically, totally unreadable, semantically entirely destroyed, etc.). Do not subjectively exaggerate the degree of distortion.

Avoid:

1. Using absolute, subjective, exaggerated descriptive words forbidden by this prompt;
2. Wrong hallucination-judgment logic: misjudging explicitly excluded quality defects (over-smoothing, color shift, detail softening) as texture-artifact / object-modification hallucinations;
3. Description paragraphs lacking key argumentation: failing to explain why the distortion is not an ordinary quality defect;
4. Wrong granularity of analysis: vague global descriptions throughout instead of focusing on the minimal affected local grid cells.

Spatial Localization and Grid-Marking Principle

To ensure localization precision, the **minimal-marking principle** must be strictly followed:

1. **Precise marking:** mark only the **minimal set of cells that directly contain the hallucinated content**, not the entire object or macro region containing the hallucination.
2. Do not mark the entire image or nearly the entire image; such cases are usually treating Super-Resolution defects as hallucinations.

Output Format

Output only valid JSON, with no leading/trailing narration or explanation.

```
{
  "gt_description": "One sentence describing the overall content of the GT image
     (scene, main objects, presence of text, key materials). Describe GT only.",
  "has_hallucination": true/false,
  "hallucinations": [
    {
      "object": "affected target object description (e.g. license-plate text,
    wall edge)",
      "hallucination_type": "Object Modification/Structural Distortion/Texture
    Artifacts/Text Distortion",
      "description": "specific description of the difference based on GT vs SR,
    e.g. explicitly state which hallucination rule is met (e.g. object-
    modification rule 1), and why it is not an ordinary quality defect",
      "grid_numbers": [12, 13, 28, 29]
    }
  ]
}
```

B.2 Stage 2: Candidate Verification and Severity Scoring Prompt

For each candidate region detected in Stage 1, we crop the rectangular area covering all of its cells, with a small margin, from both I_{SR} and I_{GT} and feed the cropped pair to the VLM *without* grid overlay, so the model judges purely on visual content.

Compare the GT and SR crops, judge whether the marked region is hallucinated and its severity, and output a score from 0 to 1.

Core Concepts

SR Artifact: a known limitation of Super-Resolution algorithms. Edge distortion, texture dissolution, detail loss, noise on flat regions, etc. **Does not change the semantic category.**

Semantic Hallucination: the SR output semantically deviates from GT. An object appears or disappears out of nowhere, a category is replaced, a spatial relation is reversed, or text turns into legible but wrong content.

Scoring Rules

Assign the marked region a score from 0 to 1 based on the severity of its semantic deviation from the GT.

Score	Meaning
0–0.3	SR quality anomaly only; no clear semantic change.
0.3–0.7	Uncertain whether a semantic hallucination is present; use 0.5 when unclear.
0.7–1	Clear semantic hallucination; higher scores indicate greater noticeability and deviation from the GT.

Judgment Guidelines

- Edge aliasing, waviness, breaks → edge distortion, low score.
- Blurred texture details / oil-painting effect / dissolution → quality issue, low score.
- Overall object shape severely distorted but broad category unchanged → structural issue, medium-low score.
- Object category replaced, or appears/disappears out of nowhere → semantic hallucination, high score.
- Text going from legible to unreadable gibberish → semantic hallucination, medium-high score.
- Distant small objects blurred → resolution limitation, low score.
- Noise/grid on flat-color regions → flat-region artifact, low score.
- When uncertain, use a score near 0.5.

Output Format

```
{
  "score": [0.0, 1.0],
  "rationale": "One-sentence reason"
}
```