Reference-to-Image (R2I) Aware Latent Diffusion Model for Image Super-Resolution

Byeonghun Lee[1] Hyunmin Cho[1]
Sunghoon Im[2⋆] Kyong Hwan Jin[1⋆]

[1] Korea University, Seoul, Republic of Korea
{byeonghun_lee, hyun_cho, kyong_jin}@korea.ac.kr
[2] KAIST, Daejeon, Republic of Korea
sunghoonim27@gmail.com

Abstract. Recent diffusion-based super-resolution (SR) methods often employ large text-to-image (T2I) backbones as generic priors. While this yields strong perceptual quality, restoration performance can become overdependent on text-driven semantics. T2I pipelines typically rely on classifier-free guidance (CFG), doubling the number of function evaluations at inference time. To address these limitations, we propose the **R**eference-to-Image **A**ware **L**atent Diffusion Model for Image Super-Resolution (**ReAL**), a purely Reference-to-Image (R2I) model conditioned only on the LR input and a retrieved reference. The Reference Fusion Module encodes the reference once, caches its key–value tensors, and injects them into the denoising model's self-attention layers at every timestep, providing strong feature-level guidance for texture and structure recovery. The R2I design substantially reduces dependence on T2I priors and improves semantic consistency, achieving state-of-the-art performance compared to T2I-based SR baselines.

Keywords: Reference-based Super-Resolution · Reference-to-Image Generation · Retrieval-Augmented Generation

1 Introduction

Single image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from a single low-resolution (LR) observation, an ill-posed inverse problem that remains central in low-level vision [31]. In realistic settings, where degradations are complex and often unknown, the goal is not only to optimize distortion-oriented metrics, but also to recover *perceptually convincing* details. Early deep SR methods focusing on fidelity [10, 28, 31] or adversarial training [43, 49, 58] often struggle with over-smoothing or unstable artifacts. More recently, diffusion models [20] have emerged as a dominant generative framework for image synthesis [4, 15, 39] and restoration [8, 35, 53].

Large text-to-image (T2I) diffusion methods [15, 39, 41] are widely reused as generic priors. However, specifying informative prompts for every restoration

⋆ Co-corresponding authors.

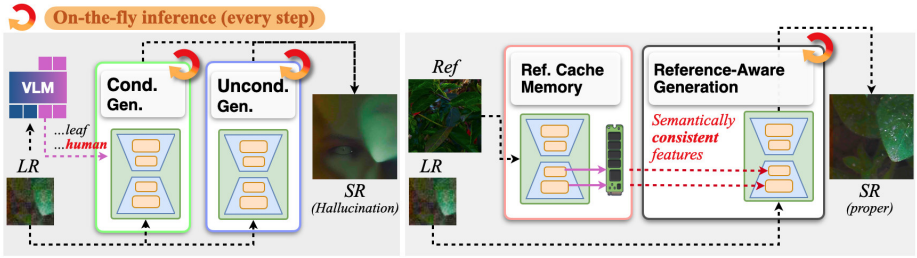

Fig. 1: Prompt vs. reference-conditioned super-resolution stability. Given the same low-resolution input, SR with a text-to-image prior ((a), T2I) can hallucinate semantically unrelated content (e.g., generating a **human face** from a leaf patch). In contrast, reference-conditioned SR with a reference-to-image prior ((b), R2I) exploits a concrete visual exemplar and tends to reuse existing structures rather than invent new ones, yielding more stable and semantically consistent reconstructions.

case is *highly non-trivial*. While recent approaches [24,53] employ vision-language models [3, 19, 46] to automatically generate prompts, the generation process remains governed by a broad, text-driven prior. When the degraded observation is weak, models can *over-rely* on this prior, causing SR to drift toward open-ended generation and hallucinate semantically inconsistent details.

To ground the generation in verifiable evidence, reference-based SR (RefSR) provides a concrete high-resolution reference image [7, 14, 55]. Recent diffusion-based RefSR approaches, such as CoSeR [45] and iRAG [27], have incorporated reference images to improve restoration. However, these methods typically adapt frozen T2I backbones, relying on intermediate textual embeddings, adapter modules, or abstracted semantic guidance to fuse the reference. This indirect conditioning dilutes the direct visual signal, leaving the pipeline vulnerable to the semantic drift and rigidity inherent in the underlying T2I architecture.

We propose a **Re**ference-to-Image **A**ware **L**atent Diffusion Model for Image Super-Resolution (**ReAL**), which shifts the paradigm from text-guided generation to pure Reference-to-Image (R2I) restoration. Instead of adapting a *frozen* T2I model, ReAL repurposes a pretrained latent-diffusion backbone and retrains it into a pure R2I model in the latent space, conditioned *solely* on the LR observation and a retrieved HR reference, without text prompts or classifier-free guidance (CFG). The novelty of our R2I approach lies in its direct, visual-to-visual feature integration, completely bypassing intermediate text representations. By injecting reference features into the denoising process, ReAL explicitly forces the network to learn a task-specific, instance-level prior. This substantially curtails the hallucination problem, ensuring that generated textures strictly adhere to the provided visual blueprint. Our main contributions are as follows:

– We introduce **ReAL**, a pure Reference-to-Image (R2I) diffusion model for reference-based SR that learns a task-specific, reference-aware prior without any text-to-image (T2I) conditioning.

- We design a **Reference Fusion Module** that encodes the reference once and injects its cached features into self-attention, enabling efficient, feature-level guidance for detail restoration.
- We demonstrate that ReAL achieved state-of-the-art or competitive performance while substantially reducing semantic drift and improving consistency over baselines using T2I.

2 Related Works

Diffusion Probabilistic Model. Diffusion probabilistic models [20] are a dominant class of generative models for high-fidelity image synthesis. A diffusion model defines a forward noising process

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\alpha_t}\,\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}\big), \quad t = 1 \dots T,$$

and learns a reverse process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ by training a neural network ϵ_θ to predict the additive Gaussian noise. While early diffusion models operated in pixel space [20], latent diffusion models (LDMs) [15, 39, 41] operate the diffusion process in a compressed latent space, substantially reducing cost while preserving visual quality. This latent formulation facilitates diverse conditional and unconditional image generation tasks, including image SR [8, 53, 56].

Image Super-Resolution for human perception. Image SR for human perception aims to recover perceptually convincing, artifact-free images under complex degradations. Early deep SR methods [10,13,28] primarily optimized distortion metrics, often producing over-smoothed outputs. Adversarial approaches [18, 26,51] improved sharpness but suffered from unstable training [30]. Recently, diffusion models [20,41] have been adopted as expressive priors for Real-ISR [8,48, 53], offering finer textures and stable training. While diffusion-based SR achieves strong performance, mitigating semantic drift and improving condition alignment remain open challenges.

Reference-based Image Super-Resolution. Reference-based image super-resolution (RefSR) exploits auxiliary high-resolution images to provide high-frequency details. Early methods [55, 61, 62] focused on aligning reference features via patch matching or selective aggregation. Recently, inspired by retrieval-augmented generation (RAG) [17, 29], diffusion-based RefSR models such as CoSeR [45] and iRAG [27] dynamically retrieve and condition on relevant exemplars. However, these methods condition a *frozen* T2I backbone on the reference indirectly through textual or semantic embeddings, retaining a strong dependence on T2I priors. While we adopt the hash-based retrieval of iRAG, our Reference-to-Image (R2I) paradigm differs fundamentally in how the reference is *used*: it injects the retrieved exemplar *directly* as self-attention key/value pairs and retrains the backbone without any text branch or classifier-free guidance.

Data Retrieval via Neural Hash Network. Neural hashing has advanced large-scale visual retrieval by learning compact binary codes that capture se-

Fig. 2: Qualitative comparison on the DRealSR dataset. The right panels show $\times 4$ SR results from baseline models and our ReAL. In this flipped regime, text-driven methods, including those using VLM-generated prompts, exhibit noticeable semantic drift, whereas ReAL recovers sharper and more legible text on the bottle label with fewer artifacts, yielding reconstructions closer to the ground-truth image.

mantic and structural information. Supervised methods [16, 36, 54, 57] typically exploit label information or pairwise similarity constraints so that semantically related images are mapped to nearby Hamming codes. In parallel, unsupervised hashing methods [32, 40] dispense with manual annotations and instead infer codes from the data distribution, aiming to uncover *intrinsic visual structure* and discover neighborhoods that reflect underlying semantics. Such hashing-based retrieval provides an efficient mechanism for indexing and searching large image repositories.

Hallucination in Diffusion Models. Hallucination in diffusion models refers to samples that deviate from the true data distribution or contradict the conditioning signal, failing to produce semantically meaningful outputs. In image synthesis, this manifests as entangled objects [37] or implausible structures. Recent studies [2, 11, 25] propose mitigation strategies by better aligning the generative prior or constraining the sampling procedure. However, most works target intrinsic aspects of the diffusion model. Hallucinations driven by the ambiguity of extrinsic conditions—such as text prompts in T2I pipelines—remain underexplored. Our R2I framework directly tackles this by substituting abstract text prompts with concrete visual references, fundamentally bypassing the ambiguous text-to-image mapping that typically induces semantic drifts.

3 Avoiding T2I priors for semantic consistency

Score-based diffusion models can be interpreted as learning the score (i.e., gradient of the log-density) of a target distribution [44]. For a conditional generative task with observation y (e.g., a low-resolution image) and high-resolution image x, the ideal target is the posterior

$$p(x \mid y) \propto p(y \mid x)\, p(x), \tag{1}$$

where $p(y \mid x)$ encodes the degradation process and $p(x)$ is a prior over natural images. In the score-based perspective, the gradient of the log-posterior

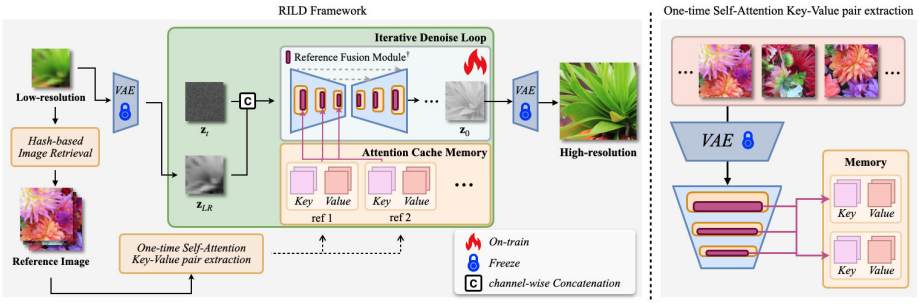

Fig. 3: An overview of our proposed ReAL model. Our approach efficiently integrates reference information via a two-stage process. First, the reference image (I_{ref}) is encoded once to extract and cache its self-attention Key (K_{ref}) and Value (V_{ref}) tensors. Second, during the iterative denoising of the low-resolution input's latent (z_t), these cached tensors are concatenated with the main path's Key and Value tensors within each self-attention block. This direct feature fusion mechanism allows the model to leverage detailed reference textures throughout the entire generation process to produce the final high-quality output I_{SR}.

decomposes as

$$\nabla_x \log p(x \mid y) = \nabla_x \log p(x) + \nabla_x \log p(y \mid x), \tag{2}$$

a sum of a *prior term* and a *likelihood term*. Modern guidance methods [12, 21] can be understood as controlling the relative strength of such conditional components, effectively trading off between the learned prior and the conditioning signal at sampling time.

T2I diffusion models instantiate a strong semantic prior $p(x \mid c)$ conditioned on a text prompt c. When such models are repurposed for SR, the generative process is better thought of as approximating $p(x \mid (y, c))$, where y is the LR observation and c is a text description of the scene. In the extreme SR regime, however, y is heavily aliased and thus *only weakly informative* about the fine-scale content of x. The likelihood term $p(y \mid x)$ is therefore broad and potentially even *misleading* over a large set of candidate HR images that are all consistent with the LR evidence.

In this underdetermined regime, we idealize the extreme SR setting by assuming that the LR observation y is *approximately independent* of the fine-scale semantics captured by the T2I model. Under this approximation, such weakly informative conditions can be treated as effectively ignorable [42], so that the posterior is governed primarily by the T2I prior:

$$p(x \mid (y, c)) \;\approx\; p(x \mid c), \tag{3}$$

so that the model primarily reflects generic knowledge from c rather than the instance-specific information y. This imbalance explains why the T2I prior often "fills in" textures and structures that are globally plausible but factually incorrect for the given input, leading to semantic drift and hallucinated details even when they contradict the LR observation.

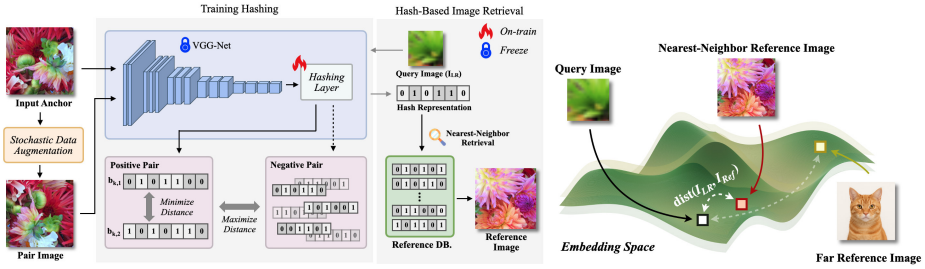

Fig. 4: (a) Overview of the Hash-Based Data Retrieval Module. Contrastive objective minimizes the distance between hash codes of positive image pairs. At inference, the low-resolution query (I_{LR}) is mapped to a hash representation for efficient nearest-neighbor retrieval of the optimal high-resolution reference (I_{ref}). **(b) Illustration of nearest-neighbor retrieval in the learned embedding space.** The low-resolution query image $\mathbf{I}_{\text{LR}}$ is mapped to an embedding (*white square*) and compared against embeddings of all reference images. The closest reference $\mathbf{I}_{\text{ref}}$ (*red square*) under the distance measure d is selected as the guidance exemplar, while more distant references are discarded.

4 Method

To mitigate the issues discussed in Sec. 3, we present the Reference-to-Image Aware Latent (**ReAL**) for semantically consistent super-resolution using reference images. As illustrated in Fig. 1, we adopt the standard RefSR setting, where for each low-resolution input I_{LR} we assume access to an auxiliary high-resolution reference image I_{ref} that is semantically related to the target scene [7, 55, 62]. In contrast to approaches that reuse large T2I diffusion models as frozen priors, we learn a restoration-focused generative prior that is explicitly conditioned on $(I_{\text{LR}}, I_{\text{ref}})$ without additional guidance such as classifier-free guidance (CFG) and text prompts.

At a high level, as depicted in the overall framework in Fig. 3, ReAL samples from $p_\phi(I_{\text{SR}} \mid I_{\text{LR}}, I_{\text{ref}})$ by combining three components. First, a neural hash-based retrieval selects reference patches that are most relevant to the current LR input, enabling scalable reference usage from large image collections. Second, a Reference Fusion Module injects reference features into the network at each inference step, allowing the model to reuse instance-specific textures. Third, the model is optimized with a standard diffusion noise loss, relying on the injected reference features to reconstruct both LR-consistent structures and realistic fine details.

4.1 Hashing for Data Retrieval

Following the retrieval-augmented framework introduced in iRAG [27], we employ a compact, hash-based *retrieval* module [40]. We adopt this retrieval from iRAG as an off-the-shelf component; our contribution is *orthogonal* to retrieval and lies in *how* the retrieved reference is used—the direct key/value fusion of

our Reference Fusion Module and the text-free R2I training—rather than in the retrieval mechanism itself. As outlined in Fig. 4a, this module identifies a high-resolution exemplar that maximizes the visual correlation with the degraded input. For each input image I_{LR}, the module selects a reference image I_{ref} from an external database of candidates, denoted by ref.DB. Retrieval is performed in a learned embedding space where every image I is encoded into a hash vector $\mathbf{h}$ with learned mapping function $H_\phi(\cdot)$. As illustrated in Fig. 4b, we then perform nearest-neighbor search in the embedding space to select the exemplar. Let $\{I_{\mathrm{DB}}^{(k)}\}_{k=1}^{|\mathrm{ref.DB}|}$ denote the images in the reference database and $\mathbf{h}^{(k)} = H_\phi(I_{\mathrm{DB}}^{(k)})$ their corresponding hash codes. Given the query code $\mathbf{h}_{\mathrm{LR}} = H_\phi(I_{\mathrm{LR}})$, we identify the closest neighbor

$$k^\star = \underset{1\le k\le|\mathrm{ref.DB}|}{\arg\min}\ \mathrm{dist}\big(\mathbf{h}_{\mathrm{LR}}, \mathbf{h}^{(k)}\big), \quad I_{\mathrm{ref}} \leftarrow I_{\mathrm{DB}}^{(k^\star)}, \tag{4}$$

where $\mathrm{dist}(\cdot,\cdot)$ denotes a distance function defined on the embedding space.

The hash function H_ϕ is learned with a contrastive objective so that images with similar semantics are assigned nearby binary codes. For each database image x_k, we form two augmented views $\tilde{x}_{k,i} = \mathcal{A}_i(x_k)$ $(i \in \{1,2\})$ and encode them with a fixed VGG-based extractor $f(\cdot)$ followed by a learnable hashing head $g_\phi(\cdot)$, yielding binary codes $\mathbf{b}_{k,i} = g_\phi(f(\tilde{x}_{k,i}))$. With a temperature-scaled cosine score $s(\mathbf{b},\mathbf{b}') = \exp(\mathcal{C}(\mathbf{b},\mathbf{b}'))/\tau$ [9], we take $\mathbf{b}_{k,1}$ as anchor and $\mathbf{b}_{k,2}$ as its positive and adopt an InfoNCE [38]-style loss:

$$\mathcal{L}_{k,1} \triangleq -\log \frac{s(\mathbf{b}_{k,1}, \mathbf{b}_{k,2})}{\displaystyle\sum_{\ell=1}^{N}\sum_{j\in\{1,2\}} \mathbb{1}_{(\ell,j)\neq(k,1)}\, s(\mathbf{b}_{k,1}, \mathbf{b}_{\ell,j})}. \tag{5}$$

A symmetric term $\mathcal{L}_{k,2}$ swaps the anchor and positive, and minimizing the batch-averaged loss $\mathcal{L}_{\mathrm{cl}} = \frac{1}{N}\sum_k(\mathcal{L}_{k,1} + \mathcal{L}_{k,2})$ $(N = |\mathrm{ref.DB}|)$ clusters semantically related images, enabling accurate and efficient exemplar retrieval.

4.2 ReAL: Reference-to-Image Aware Latent Diffusion

Our model is built upon the Latent Diffusion Model (LDM) framework, which performs the diffusion process within the compact latent space of a pre-trained Variational Autoencoder (VAE) for enhanced computational efficiency [22, 41]. The core of the LDM is a time-conditional U-Net, ϵ_θ, which is trained to predict the noise added to a latent representation z_t at a given timestep t. We introduce two primary conditioning mechanisms to guide this denoising process: one for the LR input and another for the HR reference.

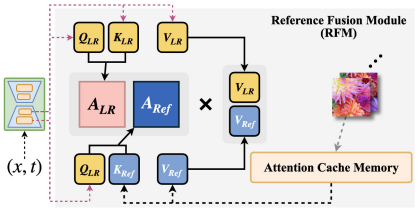

Fig. 5: Reference Fusion Module (RFM). LR features are updated on-the-fly at each denoising step (plum dashed arrows). Conversely, the reference is encoded only once into an *Attention Cache Memory*.

To ensure the generated output remains faithful to the source content, the LR image, I_{LR}, is first encoded into its latent representation $z_{LR} = E(I_{LR})$. This latent vector is then concatenated channel-wise with the noisy latent z_t at each step of the reverse diffusion process, providing a strong structural condition for the U-Net.

The key innovation of our model lies in the Reference Fusion Module (RFM), a mechanism inspired by the CacheKV architecture, designed for the efficient and effective integration of the reference image, I_{ref} [14, 22]. Instead of repeatedly processing the reference image at each of the hundreds of denoising steps, the RFM operates as follows: at the beginning of inference, I_{ref} is passed through the denoising U-Net a single time. During this pass, the key (K_{ref}) and value (V_{ref}) tensors from every self-attention layer are extracted and stored in a cache. Subsequently, during the main denoising loop for the LR image, the query (Q_{main}), key (K_{main}), and value (V_{main}) are computed from the current noisy latent z_t. The cached reference tensors are then concatenated with the main key and value tensors along the token dimension:

$$K_{fused} = \text{concat}(K_{main}, K_{ref})$$
$$V_{fused} = \text{concat}(V_{main}, V_{ref})$$

The attention is then computed as:

$$\mathit{Attention}(Q_{main}, K_{fused}, V_{fused}). \tag{6}$$

We inject the reference on the key/value side rather than the query: this keeps the LR query set—and hence the denoising trajectory—intact, while letting each LR token freely decide how much to borrow from the reference. Concatenating the reference on the query instead would generate reference-driven content and break the LR-anchored reconstruction.

This design allows every patch of the LR latent representation to directly attend to all features within the HR reference at every layer and every timestep of the generation process, facilitating an effective transfer of fine-grained textures and details without the computational overhead of repeated feature extraction.

4.3 Training Objectives

The training of our U-Net ϵ_θ is optimized using the standard objective from Denoising Diffusion Probabilistic Models (DDPM) and Latent Diffusion Models (LDM) [20, 41].

The Diffusion Noise Loss is the sole objective function for the U-Net. It trains the model to predict the noise ϵ that was added to the clean latent z_0 (derived from the ground truth HR image) to create the noisy latent z_t at timestep t. We formulate this as the standard ℓ_2 (mean-squared error) noise-prediction loss, the de facto objective for diffusion training [20, 41]:

$$\mathcal{L} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t, c} \left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right] \tag{7}$$

where c is the set of conditions provided to the U-Net. In our ReAL model, c is a combination of the current timestep t, the LR image latent representation

Table 1: Quantitative comparison on real-world benchmarks. We report distortion-oriented metrics (PSNR, SSIM, LPIPS) and perceptual metrics (CLIP-IQA+, MUSIQ) on various benchmarks. All models are trained on the union of Flickr2K [31], DIV2K [1], and OST [50]. ReAL achieves consistently strong performance, attaining the best or second-best perceptual scores across all benchmarks while maintaining competitive distortion metrics. Entries formatted as **red**, blue, and *green* indicate the best, second, and third results among all models excluding the GAN baselines (ESRGAN, BSRGAN). For the GAN baselines, the **best** result between them is bolded.

Training Loss		GAN Loss		Diffusion Loss								
Using References		✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
Benchmarks	Metrics	ESRGAN [49]	BSRGAN [58]	LDM [41]	StableSR [48]	PASD [56]	DiffBIR [33]	SeeSR [53]	FaithDiff [8]	CoSeR [45]	iRAG [27]	***Ours***
DIV2K Valid [1]	PSNR↑	**20.067**	19.858	18.727	18.956	*19.453*	18.747	19.157	18.375	19.579	19.280	**19.650**
	SSIM↑	**0.525**	0.504	0.438	0.435	*0.497*	0.418	0.484	0.443	0.498	0.458	**0.507**
	LPIPS↓	**0.395**	0.422	0.428	0.415	0.462	0.451	0.392	0.433	0.415	*0.410*	**0.391**
	CLIP-IQA+↑	**0.603**	0.559	0.541	0.628	0.625	**0.727**	0.710	0.652	0.599	*0.676*	0.646
	MUSIQ↑	**58.397**	58.199	58.743	62.083	60.877	**69.379**	*68.594*	69.316	56.543	66.480	63.325
RealSR [6]	PSNR↑	**21.401**	21.189	20.338	20.263	20.706	20.132	20.609	18.986	21.007	*20.860*	**21.044**
	SSIM↑	**0.616**	0.594	0.516	0.486	0.593	0.486	0.571	0.508	*0.587*	0.551	**0.608**
	LPIPS↓	**0.423**	0.431	0.464	0.481	0.410	0.487	*0.401*	0.424	0.407	**0.372**	0.375
	CLIP-IQA+↑	**0.601**	0.565	0.513	0.624	0.628	**0.728**	0.711	0.653	0.596	0.666	*0.667*
	MUSIQ↑	58.447	**59.657**	56.391	62.072	63.758	69.263	**70.613**	70.372	56.745	*70.280*	66.529
DRealSR [52]	PSNR↑	**26.010**	25.352	23.518	23.397	**24.898**	22.507	23.999	21.988	*24.734*	24.080	24.836
	SSIM↑	**0.757**	0.723	0.592	0.551	**0.731**	0.500	0.692	0.535	0.698	*0.720*	0.725
	LPIPS↓	**0.350**	0.377	0.466	0.507	*0.388*	0.579	0.390	0.498	0.395	0.367	**0.359**
	CLIP-IQA+↑	**0.557**	0.528	0.482	0.580	0.579	**0.689**	0.672	0.609	0.550	*0.621*	0.612
	MUSIQ↑	51.285	**53.040**	51.938	57.299	57.112	65.606	*64.829*	**65.900**	52.345	64.630	60.368

z_{LR}, and the reference features $\{K_{ref}, V_{ref}\}$ injected via the RFM. Because the retrieved reference already supplies high-frequency detail, we do not add auxiliary perceptual or adversarial losses common in GAN-based SR; the simple ℓ_2 noise objective suffices and avoids their training instability.

5 Experiments

Training Dataset. Our training set for the restoration network is derived from the DF2K-OST dataset, which is a combination of the DIV2K [1], Flickr2K [31] and OST [50] datasets. We follow a partitioning strategy where 75% of this combined dataset is allocated for training the main restoration model.

Reference Database. The reference database is constructed from the remaining 25% of the DF2K-OST dataset. To enrich the diversity of available references and improve the robustness of our retrieval mechanism, we augment this reference set by adding synthetic images in a one-to-one ratio with the real image patches [27].

Evaluation. We evaluate ReAL's performance on three standard real-world super-resolution benchmarks: the DIV2K validation [1], the RealSR [5], and the DRealSR [52] dataset. For evaluation, we use several standard metrics, including PSNR, SSIM, LPIPS [59], CLIP-IQA+ [47], and MUSIQ [23], to provide a comprehensive assessment of both fidelity and perceptual quality.

Training Detail. Our ReAL model is initialized using the pre-trained weights of Stable Diffusion 2-base [41]. We keep the VAE frozen and train the entire denoising U-Net together with the learned reference embedding. The network is optimized using the Adam optimizer with a constant learning rate of 5×10^{-5}, a batch size of 1, and gradient accumulation over 4 steps, on 512×512 resolution

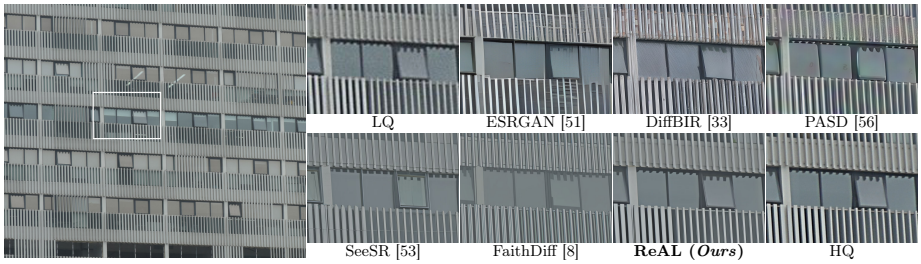

Fig. 6: Qualitative comparison on a DRealSR example. Baseline methods create unnatural textures and introduce severe, distracting color artifacts that distort the image's original tones. In contrast, our method (ReAL) restores the complex linear structures.

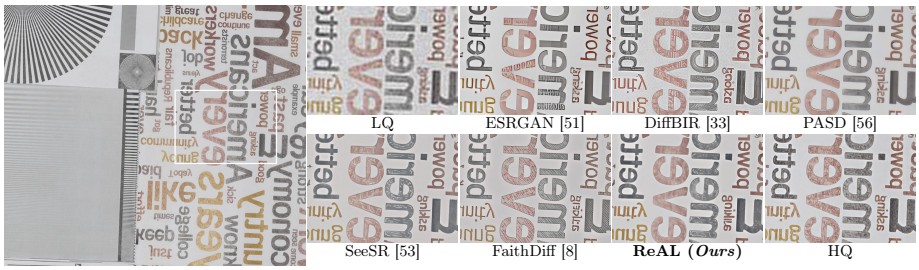

Fig. 7: Qualitative comparison on a RealSR example. Baseline methods tend to produce globally *over-sharpened* outputs, especially around text, where strokes become unnaturally crisp or distorted relative to the ground-truth image. In contrast, ReAL recovers sharp yet faithful characters and cleaner edges with fewer artifacts, yielding reconstructions that more closely match the original scene.

image patches. All experiments are conducted on NVIDIA 4090 GPUs. During the inference phase, we employ a DDIM sampler with 50 timesteps to generate the final results.

Quantitative & Qualitative Results. As shown in Tab. 1, ReAL achieves state-of-the-art performance in distortion-oriented metrics (PSNR, SSIM, LPIPS), securing the best or second-best results across all three benchmarks and indicating strong fidelity under real-world degradations. For example, ReAL leads all three distortion metrics on DIV2K simultaneously and attains the best PSNR and SSIM on RealSR; among reference-based diffusion baselines, it improves PSNR over iRAG by 0.37,dB on DIV2K and lowers LPIPS over CoSeR by 0.036 on DRealSR.

Qualitative comparisons in Fig. 2, Fig. 6, and Fig. 7 further support this analysis. Compared to GAN-based baselines that often introduce unnatural textures and distracting color artifacts (Fig. 2), ReAL yields much cleaner outputs. Generic diffusion-based methods produce globally *over-sharpened* outputs that distort fine details (Fig. 6), and because they are heavily governed by text-driven semantics, they frequently exhibit noticeable *semantic drift*, failing to recover legible text (Fig. 7). Furthermore, compared to recent reference-based diffusion models (CoSeR [45], iRAG [27]), ReAL faithfully restores complex linear struc-

(a) RAM [60] **(b)** LLaVa [34] **(c)** Reference **(d)** LQ-GT

Fig. 8: Failure of T2I-driven SR and benefit of reference-aware conditioning. SR results from RAM (SeeSR [53]), LLaVa (FaithDiff [8]), and ReAL. Their corresponding conditioning signal cues are used to form text prompts, retrieved HR reference and the LQ input. Despite access to powerful VLM prompts, T2I-based methods exhibit semantic drift, whereas the ReAL reconstruction remains faithful to the GT content.

Table 2: Ablation on ReAL components. ReAL(-L) removes the LR–latent concatenation path; **ReAL(-R)** removes reference conditioning; **ReAL(-F)** disables the Reference Fusion Module; and **ReAL(-P)** utilizes a random reference image. The best results are highlighted in **bold**, and the second-best are underlined. The full model attains the best overall balance of fidelity, while **-F** yields slightly higher CLIP-IQA+ and MUSIQ, indicating a mild perception–fidelity trade-off.

SR (×4)	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-IQA+ ↑	MUSIQ ↑
ReAL (-L)	26.689	0.774	0.309	0.608	60.544
ReAL (-R)	27.084	0.799	0.250	0.614	60.294
ReAL (-F)	26.837	0.782	0.248	**0.633**	**61.802**
ReAL (-P)	**27.385**	0.780	0.263	0.602	60.250
ReAL (*full*)	27.352	**0.803**	**0.246**	0.625	61.565

tures (Fig. 2) and recovers sharp, legible characters with cleaner edges and fewer artifacts, proving the effectiveness of our pure R2I guidance.

6 Discussion

In this section, we further analyze the core advantages of our methodology. We validate the robustness of our feature alignment mechanism through an ablation study, highlight ReAL's contributions to content fidelity by mitigating T2I-induced hallucinations, and explore its practical applicability by assessing the efficiency and scalability of our reference retrieval module.

6.1 Ablation study

To validate the contribution of each key component in our ReAL framework, we conduct an ablation study as shown in Tab. 2. We analyze four variants: **ReAL(-L)**, which removes the LR-latent concatenation path; **ReAL(-R)**, which removes all reference conditioning; **ReAL(-F)**, which disables our core Reference Fusion Module (RFM) by removing the reference KV caching and fusion; and **ReAL(-P)**, which replaces the properly retrieved reference image with a randomly selected one to assess the impact of reference relevance.

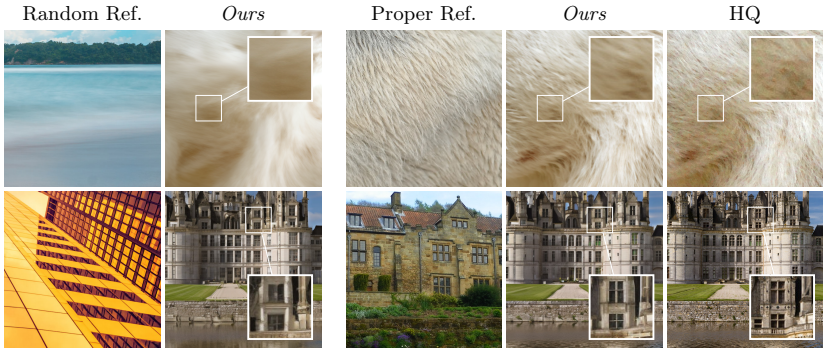

Fig. 9: Qualitative effect of reference relevance. Qualitative comparison of our ReAL model conditioned on a *random* reference versus a *proper* reference. Given a proper reference, ReAL effectively leverages the relevant high-frequency cues to restore realistic details.

Table 3: Impact of reference database scale and source. Performance consistently improves as the reference database size increases. ReAL also generalizes robustly when using a completely external database.

Ref. DB Type	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-IQA+ ↑	MUSIQ ↑
Small (30%)	20.812	0.564	0.401	0.615	61.240
Medium (60%)	20.953	0.589	0.386	0.642	64.105
Large (100%)	21.044	**0.608**	**0.375**	**0.667**	**66.529**
External: CUFED [61]	**21.060**	0.597	0.385	0.656	65.927

The results demonstrate that the **ReAL (full)** model achieves the best overall performance in structure and perception. Disabling the RFM (**ReAL(-F)**) severely degrades fidelity, confirming the necessity of its cross-attention mechanism for the *implicit alignment* of LR and reference features. While **ReAL(-F)** yields slightly higher CLIP-IQA+ and MUSIQ scores, it incurs a significant perception-fidelity trade-off.

Crucially, replacing the retrieved reference with a random image in **ReAL(-P)** decreases SSIM and increases LPIPS. As shown in Fig. 9, a random reference makes the model fall back on the LR structure, yielding a plausible but over-smoothed output, whereas a *proper* reference lets ReAL transfer the correct high-frequency details.

6.2 Mitigating T2I-Induced Hallucinations

A primary challenge in recent generative SR models is the phenomenon of 'semantic drift' or 'content hallucination,' which is particularly prevalent in models leveraging T2I backbones like Stable Diffusion. Because these models rely on highly generalized world knowledge, their broad priors often conflict with the specific, low-level details of the input image. Consequently, they may generate structures that are globally plausible but factually incorrect for the specific input, such as the wrong letter on a sign or an incorrect clothing pattern.

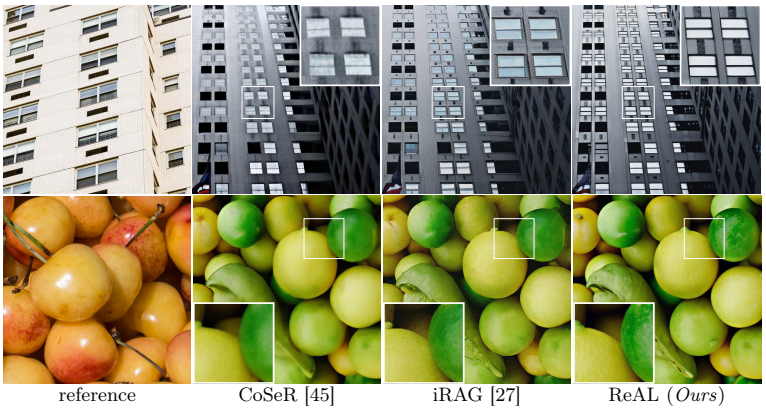

Fig. 10: Qualitative comparison with diffusion-based RefSR models using identical references. Given the same reference image, ReAL is compared against contemporary diffusion-based methods CoSeR [45] and iRAG [27]. In both examples, our model's direct feature fusion mechanism demonstrates superior texture transfer.

Table 4: Comparison of reference matching methods. While VGG-Retrieval yields slightly higher fidelity, our learned Hashing drastically reduces the matching cost, enabling highly efficient retrieval.

Matching Metric	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-IQA+ ↑	MUSIQ ↑	Matching Cost
ReAL (w/ VGG-Retrieval)	**27.385**	**0.806**	0.249	0.620	61.450	3830 ms
ReAL (w/ Hash)	27.352	0.803	**0.246**	**0.625**	**61.565**	**62 ms**

Our ReAL model fundamentally circumvents this problem by discarding the T2I prior. By replacing abstract semantic guidance with concrete, pixel-level information from a reference image, the task shifts from generating a plausible semantic fit to reconstructing a direct visual blueprint. As Tab. 1 and Fig. 8 show, this leads to a reduction in hallucinations and a significant improvement in content fidelity.

6.3 Superiority in Reference Feature Utilization

While contemporary diffusion-based RefSR models like CoSeR [45] and iRAG [27] also leverage reference images, they often rely on intermediate representations or abstracted semantic guidance, which can dilute the reference information. Although ReAL reuses iRAG's retrieval module, it departs from iRAG by discarding the text branch and integrating the reference *directly* via the Reference Fusion Module (RFM)—a significantly stronger and more precise signal for texture transfer. As shown in Fig. 10, given the exact same reference, ReAL faithfully restores intricate details such as the fruit's skin texture and the building's fine linear patterns.

To further understand *why* this direct fusion is effective, we analyze the RFM's internal behavior in Fig. 11. We measure the attention probability that the LR queries place on the cached reference key–value tokens, and visualize the principal components of the fused self-attention features. When a proper

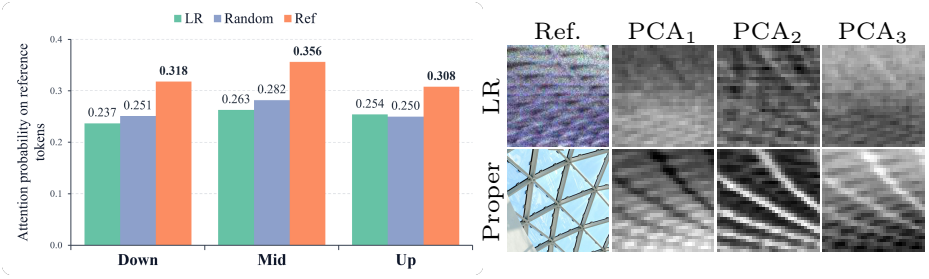

Fig. 11: RFM feature analysis. *Left*: attention mass from the LR queries to the cached reference tokens—strong for a proper reference but near zero for a random or mismatched one, so the shared softmax falls back to the LR tokens without explicit gating. *Right*: PCA of the fused features; a proper reference injects high-frequency, SR-relevant structure absent when the LR is used as the reference.

Table 5: Inference cost and CFG overhead. ReAL is CFG-free and caches the reference once, attaining the lowest latency and memory among reference-conditioned diffusion baselines.

Method	Steps	Ref.	Time (s) ↓	Mem. (GB) ↓
DiffBIR [33]	50	✗	5.06	9.10
SeeSR [53]	50	✗	3.23	9.24
CoSeR [45]	50	✓	16.85	14.94
iRAG [27]	200	✓	22.35	22.48
ReAL (Ours)	50	✓	**2.65**	**5.83**

reference is provided, the LR queries attend strongly to it and the fused features acquire high-frequency, SR-relevant structure. In contrast, a random or mismatched reference receives almost no attention mass, so the single shared softmax effectively ignores it and the model falls back to the LR signal—an implicit, gating-free robustness to irrelevant references.

6.4 Efficiency and Scalability of Reference Retrieval

First, Tab. 4 compares our learned Hashing against a standard VGG-based retrieval approach. Although VGG-Retrieval yields slightly superior fidelity, it incurs a massive computational overhead. In contrast, our Hashing drastically reduces this cost, making our approach highly scalable to large datasets.

Second, Tab. 3 demonstrates that performance consistently improves as the reference database scales from 30% to 100%, confirming that ReAL effectively capitalizes on a broader pool of exemplars. Furthermore, evaluating ReAL with CUFED [61] as a completely external database yields highly competitive results. This demonstrates strong generalization capabilities, proving that ReAL robustly extracts and fuses high-frequency details without overfitting to the training distribution.

Beyond retrieval, ReAL is also efficient at the denoising stage (Tab. 5). T2I-based pipelines rely on classifier-free guidance and evaluate the network *twice* per step even with a fixed prompt, whereas ReAL is CFG-free and encodes the

reference only once for reuse across all steps. Consequently, ReAL attains the lowest latency and memory among reference-conditioned diffusion methods.

7 Limitations

ReAL delivers strong results but entails several practical considerations. First, end-to-end optimization of the denoiser and fusion modules is heavier than fine-tuning; the burden is partly offset at inference by using a single cached reference pass, but the training cost remains higher. Second, performance depends on the relevance of retrieved exemplars, so gains attenuate when database coverage is weak, with two characteristic failure modes. Under *heavy blur*, the low-resolution structure is too degraded to anchor the fusion, so fine textures may remain over-smoothed even with a good reference. Under *cross-domain references*, the softmax fallback suppresses most irrelevant tokens, yet a small amount of foreign texture can still leak into flat regions. These risks can be mitigated with multi-reference fusion, retrieval-confidence gating, LR-only fallbacks, and jointly learned retrieval.

8 Conclusion

This paper introduces a new perspective on diffusion guidance for image super-resolution. Rather than treating large text-to-image (**T2I**) backbones as the default prior, we argue that strong dependence on text-driven semantics can make restoration pipelines vulnerable to semantic drift. This motivates reducing T2I dependency and instead exploiting concrete visual guidance from alternative sources. To explore this alternative design point, we proposed **ReAL** for Image SR, a purely reference-to-image (**R2I**) guided diffusion model. ReAL employs a Reference Fusion Module that encodes the reference *only once*, and provides direct feature-level guidance without repeated reference processing. Our experiments show that this reference-aware design improves *semantic consistency* and achieves SOTA performance compared to T2I-based SR baselines, suggesting that reference-guided diffusion is a promising direction for high-fidelity image restoration.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00335741) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2025-02219277, AI Star Fellowship Support Project (DGIST)), both funded by the Korea government (MSIT), and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (Project Name: International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, Project Number: RS-2024-00345025).

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Aithal, S.K., Maini, P., Lipton, Z.C., Kolter, J.Z.: Understanding hallucinations in diffusion models through mode interpolation. Advances in neural information processing systems **37**, 134614–134644 (2024)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv e-prints pp. arXiv–2506 (2025)
5. Cai, J., Gu, S., Timofte, R., Zhang, L.: Ntire 2019 challenge on real image super-resolution: Methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
6. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: 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. pp. 3086–3095 (2019)
7. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer. In: European conference on computer vision. pp. 325–342. Springer (2022)
8. Chen, J., Pan, J., Dong, J.: FaithDiff: Unleashing diffusion priors for faithful image super-resolution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28188–28197 (2025)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
10. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021)
11. Cho, H., Ahn, D., Hong, S., Kim, J.E., Kim, S., Jin, K.H.: TAG:tangential amplifying guidance for hallucination-resistant diffusion sampling (2025), `https://arxiv.org/abs/2510.04533`
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
14. Dong, R., Yuan, S., Luo, B., Chen, M., Zhang, J., Zhang, L., Li, W., Zheng, J., Fu, H.: Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 27684–27694 (2024)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)

16. Gao, H., Hu, C., Han, G., Mao, J., Huang, W., Wan, K.: Hashneck is a boosting tool for deep learning to hashing. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 83–91 (2024)
17. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
19. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022), `https://arxiv.org/abs/2207.12598`
22. Hsiao, C.W., Liu, Y.L., Yang, C.K., Kuo, S.P., Jou, Y.K., Chen, C.P.: ReF-LDM: A latent diffusion model for reference-based face image restoration. Advances in Neural Information Processing Systems **37**, 74840–74867 (2024)
23. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
24. Kim, B.S., Kim, J., Ye, J.C.: Chain-of-Zoom: Extreme super-resolution via scale autoregression and preference alignment (2026)
25. Kim, S., Jin, C., Diethe, T., Figini, M., Tregidgo, H.F., Mullokandov, A., Teare, P., Alexander, D.C.: Tackling structural hallucination in image translation with local diffusion. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
26. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
27. Lee, B., Cho, H., Choi, H.G., Kang, S.M., Ahn, I., Jin, K.H.: Reference-based super-resolution via image-based retrieval-augmented generation diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10764–10774 (2025)
28. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1929–1938 (2022)
29. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
30. Liang, J., Zeng, H., Zhang, L.: Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5657–5666 (2022)
31. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)

32. Lin, K., Lu, J., Chen, C.S., Zhou, J.: Learning compact binary descriptors with unsupervised deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1183–1192 (2016)
33. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior (2024)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
35. Liu, X., Tang, H.: DiffFNO: Diffusion fourier neural operator. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 150–160 (2025)
36. Nie, X., Liu, X., Guo, J., Wang, L., Yin, Y.: Supervised discrete multiple-length hashing for image retrieval. IEEE Transactions on Big Data **9**(1), 312–327 (2022)
37. Okawa, M., Lubana, E.S., Dick, R., Tanaka, H.: Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. Advances in Neural Information Processing Systems **36**, 50173–50195 (2023)
38. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
40. Qiu, Z., Su, Q., Ou, Z., Yu, J., Chen, C.: Unsupervised hashing with contrastive information bottleneck. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 959–965. International Joint Conferences on Artificial Intelligence Organization (8 2021). https://doi.org/10.24963/ijcai.2021/133, https://doi.org/10.24963/ijcai.2021/133, main Track
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
42. Sadat, S., Kansy, M., Hilliges, O., Weber, R.: No training, no problem: Rethinking classifier-free guidance for diffusion models. In: International Conference on Learning Representations. vol. 2025, pp. 76833–76858 (2025)
43. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS
45. Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y.: CoSeR: Bridging image and language for cognitive super-resolution. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 25868–25878 (2024)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
47. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)

48. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
49. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
50. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018)
51. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
52. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: European conference on computer vision. pp. 101–117. Springer (2020)
53. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: SeeSR: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
54. Yan, C., Gong, B., Wei, Y., Gao, Y.: Deep multi-view enhancement hashing for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(4), 1445–1451 (2020)
55. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5791–5800 (2020)
56. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European conference on computer vision. pp. 74–91. Springer (2024)
57. Yuan, L., Wang, T., Zhang, X., Tay, F.E., Jie, Z., Liu, W., Feng, J.: Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3083–3092 (2020)
58. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4791–4800 (2021)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
60. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize Anything: A strong image tagging model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1724–1732 (2024)
61. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7982–7991 (2019)
62. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: CrossNet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European conference on computer vision (ECCV). pp. 88–104 (2018)