REFGC-SR^2: REFERENCE-GUIDED SUPER-RESOLUTION AND REFINEMENT OF AI GENERATED CONTENT

Anonymous authors
Paper under double-blind review

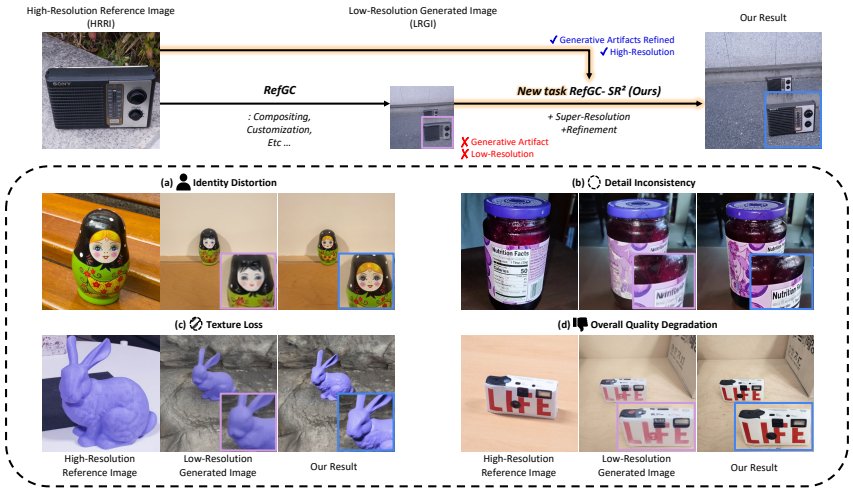


Figure 1: Reference-guided Generated Content Super-Resolution-Refinement (RefGC-SR^2). Existing reference-guided generation pipelines first produce a low-resolution generated image (LRGI) from a high-resolution reference image (HRRI), but the output often suffers from four common generative artifacts: *identity distortion*, *detail inconsistency*, *texture loss*, and *overall quality degradation*. We introduce RefGC-SR^2, a post-processing task that takes the LRGI and the HRRI as inputs, and produces a high-resolution, artifact-refined output. *Best viewed in zoom.*

ABSTRACT

Reference-guided generation (e.g., object compositing, customization) has progressed rapidly, yet current pipelines share a fundamental limitation: the object-centric high-resolution reference image (HRRI) provided by users is downsampled to a fixed low-resolution (LR) before being fed into the model, so the fine-grained details are discarded before the output is even produced. In addition, the generation step then introduces its own artifacts (e.g., identity distortion) on top of this loss. Existing reference-guided generated content refinement (RefGCR) methods can correct some of these artifacts but still operate in the LR domain; reference-guided super-resolution (RefSR) methods recover resolution but assume natural-image degradations and ignore the artifact distribution of generative pipelines. To address both gaps in a single formulation, we introduce a new task: *reference-guided generated content super-resolution-refinement* (RefGC-SR^2), where the original HRRI is reused at the post-processing stage to recover lost details, refine generative artifacts, and upscale the output simultaneously. We construct the first real-world triplet data generation pipeline for this RefGC-SR^2 task, training a diptych-conditioned generator to synthesize paired low-quality anchors that public pretrained models cannot provide. We further present a frequency-aware diffusion transformer model for RefGC-SR^2 that selectively injects fine details from the HRRI while removing generative artifacts. Extensive experiments demonstrate that our RefGC-SR^2 model successfully (i) refines the object identity faithfully with respect to the reference, and (ii) recovers high-resolution details, so that the final result is significantly higher quality and practically more usable compared to existing RefGCR and RefSR baselines.

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.

Table 1: Comparison of related image enhancement tasks. RefGC-SR^2 is the first task that jointly satisfies all four criteria: a generated content (GC) input and an HR reference image (HRRI) input, performing super-resolution (SR) and artifact refinement simultaneously.

Task	Super-Resolution Category				Refinement Category				**RefGC-SR^2 (Ours)**
	ISR	Ref**SR**	GC**SR**	RefGC**SR**	**IR**	Ref**R**	GC**R**	RefGC**R**	
GC	✗	✗	✓	✓	✗	✗	✓	✓	✓
HRRI	✗	✓	✗	✓	✗	✓	✗	✓	✓
SR	✓	✓	✓	✓	✗	✗	✗	✗	✓
Refinement	✗	✗	✗	✗	✓	✓	✓	✓	✓

1 INTRODUCTION

Image generation has rapidly evolved into a user-conditioned process, where users provide not only text prompts but also masks, structural controls, and reference images to personalize and control the output (Brooks et al., 2023; Zhang et al., 2023; Ye et al., 2023). In particular, reference-guided generation has become a practical paradigm for user-driven visual creation, supporting applications such as image editing (Liu et al., 2025c; Yang et al., 2023; Chen et al., 2024a), customization (Feng et al., 2026; Zhang et al., 2025; Tarrés et al., 2025; Tan et al., 2025; Wu et al., 2025b), and compositing (Huang et al., 2025; Chen et al., 2024b; Song et al., 2026; Li et al., 2025; Wang et al., 2025; Song et al., 2023). However, existing reference-guided generated content (RefGC) pipelines remain fundamentally limited in leveraging high-resolution reference images (HRRIs), which significantly limits practical adoption. Specifically, in real-world scenarios, users often possess HRRIs that contain rich visual information. However, most RefGC pipelines downsample these references to a fixed low-resolution (LR) before injecting them into the generation process, e.g., 224^2 visual tokens (Ye et al., 2023; Chen et al., 2024b) or 512^2 diffusion inputs (Tan et al., 2025). As a result, high-frequency (HF) information in the HRRI is largely discarded before generation begins. Then as shown in Fig. 1, the generated content (GC) further suffers from recurring generative artifacts: *identity distortion*, *detail inconsistency*, *texture loss*, and *overall quality degradation*. In other words, although users provide rich HRRIs, current RefGC pipelines fail to fully transfer their fine-grained information to the final GC output.

Positioning of Our Work To position our work in the right context, we first categorize a wide range of related image enhancement tasks in Table 1. Image Super-Resolution (ISR) (Choi et al., 2025; Duan et al., 2025; Dong et al., 2025) recovers high-resolution (HR) images from LR inputs, but assumes natural-image degradations such as bicubic downsampling, blur, noise, compression, or realistic camera degradations. Reference-guided SR (RefSR) (Lee et al., 2025; 2024) further exploits an external HRRI, but is still formulated for natural-image SR rather than GC with generative artifacts. Generated content super-resolution (GCSR) (Jeong et al., 2025; Du et al., 2024; Tragakis et al., 2024) aims to overcome the native resolution limits of generative models, but operates on the GC alone without any HRRI. Reference-guided generated content SR (RefGCSR), which is yet to be explored in literature, can further use an HRRI for GCSR, but it is not designed to jointly refine RefGC-specific generative artifacts. On the other hand, image restoration (IR) (Ye et al., 2024), generated content restoration (GCR) (Lin et al., 2024), and reference-guided refinement (RefR) (Zhang et al., 2024; Guo et al., 2024) refine degraded or artifact-contaminated images, but do not jointly perform SR with an HRRI for LR RefGC inputs. Most closely related, reference-guided generated content refinement (RefGCR) (Song et al., 2024; Ouyang et al., 2025; Zhou et al., 2026; Liu et al., 2025b) uses a reference image to correct generative artifacts in GC, but it is primarily a fixed-resolution refinement task and does not recover an HR output from an LR GC input. Detailed discussions of these related tasks are provided in Appendix B.

As shown in Table 1, existing tasks satisfy only a subset of the four criteria required by our setting: a GC image and an HRRI as inputs, and SR and generative artifact refinement as joint target operations. In contrast, our newly defined *Reference-guided Generated Content Super-Resolution-Refinement* (RefGC-SR^2) is the first task that jointly satisfies all four criteria, making it a distinct post-processing task well-suited for contemporary RefGC pipelines. RefGC-SR^2 takes (i) a low-resolution generated content image (LRGI) from a RefGC pipeline and (ii) the original HRRI, and generates an HR image that preserves the HRRI's fine-grained information while refining generative artifacts in the LRGI. It targets a practical yet underexplored stage of modern reference-guided generation pipelines,

converting an initial GC output into an HR, reference-faithful final image that preserves object identity in HRRI and recovers fine-grained details, helping to achieve the last mile of personalized image editing for practical adoption.

A challenge in training a RefGC-SR^2 model is acquiring paired triplets. Note that such datasets do not exist in standard SR or refinement datasets (Liu et al., 2025b; Zhou et al., 2026). In each triplet, HRGT denotes the desired high-resolution ground-truth target, HRRI provides reference-specific details of the same object instance (corresponds to object-centric user-provided reference image), and LRGI serves as the LR input containing generative artifacts. This construction is non-trivial: existing SR datasets use hand-crafted degradations rather than real RefGC artifacts, while existing GC datasets usually lack paired HRGTs and HRRIs. Moreover, naively using off-the-shelf RefGC models (Huang et al., 2025; Zhang et al., 2025; Feng et al., 2026; Song et al., 2026) to synthesize LRGIs often produces outputs whose object pose differs from HRGT. Such pose mismatch is unsuitable for our post-processing task as users simply wish to improve the quality and resolution of the generated result *as is* without *changing the pose* of the generated object. Thus, the object pose should be aligned between LRGI and HRGT, allowing the model to focus on artifact refinement and resolution recovery rather than learning an unintended pose correction task.

To this end, we construct RefGC-SR^2 Dataset, the first real-world LRGI-HRRI-HRGT triplet dataset tailored to this task. The HRRI and HRGT depict the same object instance under different views, poses, or scene contexts, and are collected from real-world datasets (Kim et al., 2025; Liu et al., 2025a; Xue et al., 2025) and curated with a vision-language model (VLM) (Bai et al., 2025). To synthesize an aligned LRGI for each HRRI-HRGT pair, we introduce DipRefGC (Diptych-Conditioned RefGC Generator). DipRefGC generates an LRGI that inherits the object appearance from HRRI while matching the object pose of HRGT, thereby producing a realistic artifact-containing RefGC output for supervised training.

Furthermore, we propose the *first* RefGC-SR^2-targeted model that is trained on RefGC-SR^2 Dataset. Our RefGC-SR^2 Model consists of two main components: (i) frequency-adaptive mixture of LoRA experts (FreqMoLE) and (ii) frequency-adaptive loss ($\mathcal{L}_f$). FreqMoLE is motivated by the layer-wise frequency hierarchy of FLUX-Kontext (Labs et al., 2025), where early DiT blocks capture global structure and later blocks refine fine details. It introduces two LoRA experts: a low-frequency (LF) expert and an HF expert. A routing gate adaptively controls their weights, assigning higher weight to the LF expert in early layers and to the HF expert in later layers. Our $\mathcal{L}_f$ is designed based on the frequency relationship among LRGI, HRRI, and HRGT. It guides the model to learn global structure aligned with HRGT during training, while enabling the transfer of fine details from HRRI at inference. Our contributions can be summarized as follows:

- We formulate RefGC-SR^2 problem, the first post-processing task that reuses the user-provided object-centric HRRI as a recovery source to jointly upscale and refine RefGCs.
- We construct RefGC-SR^2 Dataset, the first real-world LRGI-HRRI-HRGT triplet dataset supporting supervised training and evaluation of the RefGC-SR^2 task.
- We propose the *first* model for RefGC-SR^2 with frequency-aware modules inserted into DiT blocks to effectively remove artifacts in LRGI while increasing resolution.

2 REFGC-SR^2 DATASET

We define three key requirements for constructing RefGC-SR^2 data triplets (HRRI, HRGT, LRGI) such that they are closely aligned to the reference-guided generation and refinement pipeline in real-world scenarios: First, HRRI and HRGT should depict the *same* object instance in *real-world* images, while *differing* in view, pose, or scene context. Second, the input LRGI should be an LR GC image containing generative artifacts as shown in Fig. 1. Third, LRGI should be aligned (with object pose preserved) with the corresponding HRGT, ensuring that the model focuses on SR and refinement tasks rather than unintended pose correction. Note that no existing datasets satisfy these requirements simultaneously. SR datasets rely on hand-crafted degradations and lack generative artifacts, while recent refinement datasets (Liu et al., 2025b; Zhou et al., 2026) construct corrupted inputs using image editing model (Wu et al., 2025a), which do not capture the true distribution of RefGCs. Off-the-shelf compositing or customization models cannot be reused as data generators, since they alter the object pose. As a result, they are insufficient for generating LRGIs that both

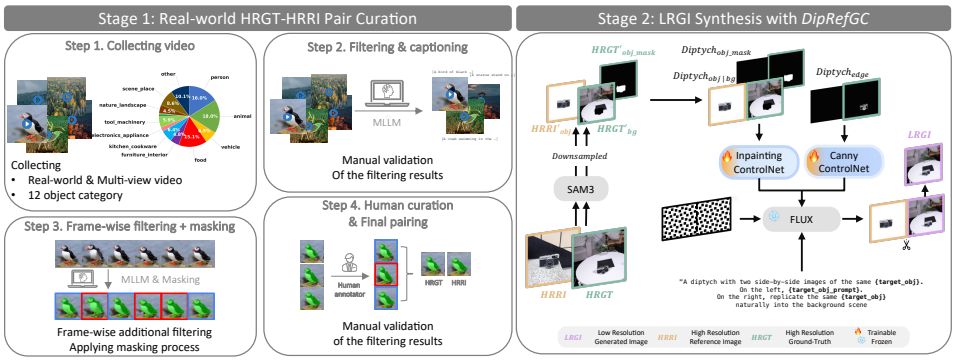


Figure 2: Construction pipeline of our RefGC-SR^2 Dataset. Stage 1 (Sec. 2.1) curates real-world HRRI-HRGT pairs from object-centric multi-view videos through VLM-based filtering (Bai et al., 2025), masking, and human validation. Stage 2 (Sec. 2.2) synthesizes the corresponding LRGI with DipRefGC, which combines HRRI appearance with HRGT-derived pose conditions using Inpainting (Team, 2024) and Canny ControlNets (InstantX, 2024). This produces pose-consistent, generative artifact-containing LRGIs paired with HRRI and HRGT, enabling supervised training for RefGC-SR^2. *Best viewed in zoom.*

contain realistic generative artifacts and preserve the object pose of the HRGT, as required by our task. We construct the dataset in two stages: (i) real-world HRRI-HRGT pair curation and (ii) LRGI synthesis with our DipRefGC.

2.1 Stage 1: Real-world HRRI-HRGT Pair Curation

We collect HRRI-HRGT pairs from higher-resolution (1024^2 to 2048^2) real-world image and video datasets that cover diverse object categories, scene contexts, and viewpoints. ORIDa (Kim et al., 2025) provides real compositing pairs of objects in different backgrounds. uCO3D (Liu et al., 2025a) offers multi-view captures of objects across everyday categories. For the video dataset uCO3D, we apply additional processing to construct HRRI-HRGT pairs, as shown in Fig. 2-Stage 1. We first collect the videos (Step 1). Next, we filter object-centric videos using Qwen3-VL (Bai et al., 2025) (Step 2). We then perform frame-wise filtering with Qwen3-VL and generate object masks using SAM3 (Carion et al., 2025) (Step 3). Finally, human annotators conduct manual verification (Step 4). Through this process, we obtain high-quality HRRI-HRGT pairs for training. See Appendix E for further analysis.

2.2 Stage 2: LRGI Synthesis with DipRefGC Using HRRI-HRGT Pairs

Given an HRRI-HRGT pair, a seemingly straightforward way to obtain an LRGI is to feed the pair into off-the-shelf RefGC models (Feng et al., 2026; Song et al., 2026) and downsample the outputs. However, since the goal of our RefGC-SR^2 task is to jointly recover resolution and refine generative artifacts of the RefGCs, this naive pipeline is insufficient as the LRGI will unavoidably modify the object pose of the corresponding HRGT. We therefore impose a pose constraint on LRGI synthesis so that the model correctly learns SR and artifact refinement rather than an unintended pose correction. Note that existing RefGC models (Huang et al., 2025; Feng et al., 2026; Song et al., 2026; Zhang et al., 2025; Chen et al., 2024b) have no explicit control over object pose. Thus, we propose DipRefGC (Diptych-Conditioned RefGC Generator), built on a frozen FLUX backbone (Labs, 2024) adapted with LoRA-tuned (Hu et al., 2022) dual ControlNets (Zhang et al., 2023). DipRefGC synthesizes LRGIs that inherit the object appearance from HRRI while following the object pose of HRGT, thereby producing LRGIs with generative artifacts suitable for triplet supervision in RefGC-SR^2. Before DipRefGC synthesis, we downsample the curated HRRI-HRGT

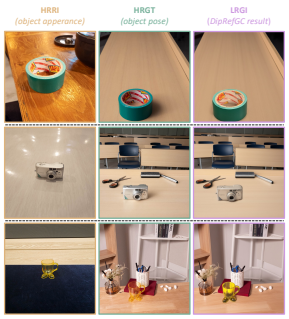


Figure 3: RefGC-SR^2 triplet examples (HRRI, HRGT and LRGI).

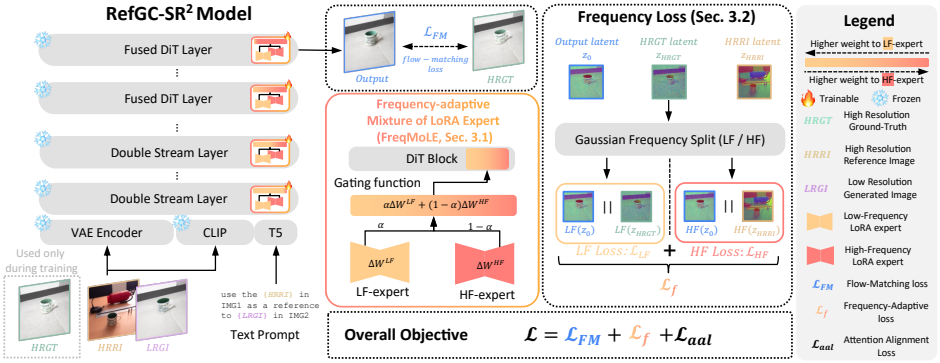


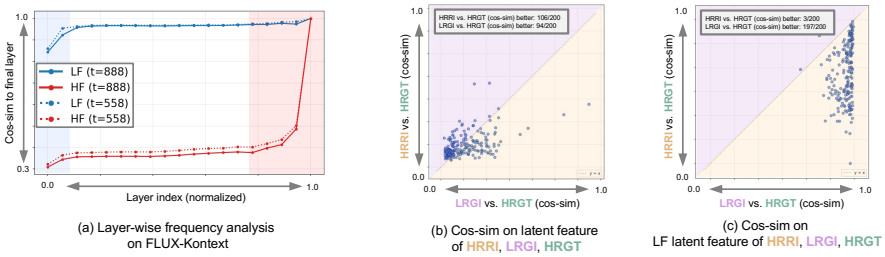
Figure 4: Overview of our RefGC-SR^2 model. The frozen VAE encodes the inputs (LRGI, HRRI, HRGT (train only)) and T5 encodes the text prompt. We insert trainable FreqMoLE modules into a frozen FLUX-Kontext (Labs et al., 2025) backbone, supervised by our frequency-adaptive loss $\mathcal{L}_f$.

pairs to 512^2 resolution, while retaining the original HR images as HRRI and HRGT in the final triplets. DipRefGC adopts a diptych (Shin et al., 2025) formulation with two specialized ControlNets (Zhang et al., 2023) to disentangle appearance and pose control. As shown in Fig. 2-Stage 2, we construct two diptych-style conditioning inputs, each composed of a left reference panel and a right generation panel. For appearance control, the *Inpainting ControlNet* (Team, 2024) receives a diptych input whose left panel contains the segmented object from HRRI, while the right panel contains the masked HRGT background and a tight object mask. We mask out the object in *HRGT* so that the model needs to rely on the object in HRRI. This set-up guides the model to resynthesize the object inside the target region in HRGT based on the object appearance in *HRRI*, effectively mimicking the generative artifacts (e.g., object identity distortion) of RefGC approaches. For pose control, the *Canny ControlNet* (InstantX, 2024) receives a second diptych input whose right panel contains the HRGT edge map restricted to the object region. This condition enforces the generated object to be aligned to the HRGT. This separation allows DipRefGC to independently control *what* to generate (appearance) and *how* to generate it (pose and structure). We fine-tune DipRefGC on the curated HRRI-HRGT pairs using the standard flow-matching objective (Lipman et al., 2022) of the FLUX backbone, where the reconstruction loss is applied only to the object region. We then run the trained DipRefGC on the curated HRRI-HRGT pairs to construct the final RefGC-SR^2 Dataset, consisting of 40K training triplets and a 200-sample evaluation split. As shown in Fig. 3, each triplet provides an object pose-consistent, generative artifact-containing 512^2 LRGI, its corresponding HRRI, and the target HRGT, enabling supervised training of RefGC-SR^2 model.

3 RefGC-SR^2 Model

We propose the RefGC-SR^2 Model addressing the RefGC-SR^2 task, which jointly removes generative artifacts and increases resolution. RefGC-SR^2 is particularly challenging because the model must not only recover HR details from an LRGI, but also refine generative artifacts while selectively exploiting a clean reference image (HRRI) captured under different conditions. Our method consists of two key components: (1) a *frequency-adaptive Mixture-of-LoRA Experts* (FreqMoLE) applied to all DiT blocks, and (2) a *frequency-adaptive loss* ($\mathcal{L}_f$) that supervises the experts with band-specific objectives. The overall pipeline is illustrated in Fig. 4.

Motivations We design our RefGC-SR^2 model based on two key observations from FLUX-Kontext (Labs et al., 2025) and our RefGC-SR^2 Dataset. First, to better understand the behavior of the backbone model (FLUX-Kontext) when applied to the RefGC-SR^2 task, we analyze the frequency characteristics of our backbone. We measure the cosine similarity between the LF/HF components of intermediate-layer latent activations and their counterparts in the final layer. As shown in Fig. 5-(a), LF quickly saturates within the first $\sim$5% of layers, while HF emerges sharply in the last $\sim$10%. This indicates that global structure (LF) is formed in early layers, whereas fine details (HF) are formed in later layers. This observation motivates our design of FreqMoLE (Sec. 3.1), which combines two band-specific LoRA experts, ΔW^{LF} and ΔW^{HF}, using a coarse-to-fine, layer-depth-dependent gating α. Second, to effectively exploit the relationship among LRGI, HRRI and HRGT in our



(a) Layer-wise frequency analysis on FLUX-Kontext

(b) Cos-sim on latent feature of HRRI, LRGI, HRGT

(c) Cos-sim on LF latent feature of HRRI, LRGI, HRGT

Figure 5: Motivation experiments for our method: (a) shows the layer-wise frequency characteristics of the FLUX-Kontext model (Labs et al., 2025). (b) compares the similarity between latent features of LRGI, HRRI, and HRGT. (c) shows the low-frequency similarity between their latent features.

model design, we analyze their relation in the latent space. We compute cosine similarity to measure which input is closer to HRGT in Fig. 5-(b), where the results show no clear trend across all features. However, as shown in Fig. 5-(c), when applying a frequency cut-off to isolate LF components, LRGI features are consistently closer to HRGT than HRRI. This observation motivates the use of frequency-aware supervision, leading to our frequency-adaptive loss ($\mathcal{L}_f$, Sec. 3.2).

Overall Pipeline RefGC-SR^2 model takes three inputs: an LRGI of the target viewpoint, an HRRI of the same instance from a different viewpoint, and a text instruction that specifies the task of refining and super-resolving the LRGI using the HRRI. Our goal is to refine and upscale ($4\times$) the LRGI with the guidance of HRRI, so that the result is close to HRGT. As shown in Fig. 4, LRGI and HRRI are encoded using a frozen VAE (Kingma & Welling, 2013), and the prompt encoder follows ImageCritic (Ouyang et al., 2025). To exploit the layer-wise frequency hierarchy of the FLUX-Kontext (Labs et al., 2025) model described above, we freeze the backbone and apply our proposed Frequency-adaptive Mixture of LoRA Experts (FreqMoLE) to all MMDiT layers. FreqMoLE injects frequency information in a layer-dependent manner, emphasizing LF components in early layers and HF components in later layers. Furthermore, we introduce a frequency-adaptive loss (L_f) as a guidance term to effectively utilize information from both HRRI and HRGT. It uses LF information from HRGT to guide global structure, and HF information from HRRI to restore fine details.

3.1 Frequency-adaptive Mixture of LoRA Experts (FreqMoLE)

Based on the observation on layer-wise frequency hierarchy in Sec. 3-**Motivations**, we replace a single LoRA in each DiT block with two *band-specialized* experts: an LF expert ΔW^{LF} and a HF expert ΔW^{HF}. The outputs of the two experts are combined using a depth-dependent gating function α: $W^{FreqMoLE} = \alpha\,\Delta W^{LF} + (1-\alpha)\,\Delta W^{HF}$. The gating value α is parameterized as a learnable scalar for each layer and initialized with a coarse-to-fine prior. It is set close to 1.0 in earlier layers to favor LF components and approaches 0 in later layers to emphasize HF components (Sec. 3-**Motivations**). The gating function is frozen in the early stage of training to keep LF dominance in early layers and HF dominance in later layers, and is then jointly optimized with the experts.

3.2 Frequency-adaptive Loss ($\mathcal{L}_f$)

To encourage band-specific behavior of the experts, we introduce a frequency-adaptive loss ($\mathcal{L}_f$) applied at the *latent* level. The goal of this loss is to promote frequency decomposition between the LF and HF experts, while guiding the LRGI input to learn LF global structure from the HRGT and HF details from the HRRI. We first decompose each latent z into LF and HF components using Gaussian band split (Burt & Adelson, 1983): $\mathrm{LF}(z) = G_\sigma * z$, and $\mathrm{HF}(z) = z - \mathrm{LF}(z)$. The LF term aligns the global structure with the HRGT: $\mathcal{L}_{LF} = \big\|\mathrm{LF}(z_0) - \mathrm{LF}(z_{HRGT})\big\|_{\mathcal{M}}$, where $\|\cdot\|_{\mathcal{M}}$ denotes the ℓ_1 norm restricted by an object mask. The HF term transfers fine detail from the HRRI. Since the HRRI and HRGT differ in viewpoint, we avoid pixel-wise alignment and instead match channel-wise statistics (Li et al., 2017): $\mathcal{L}_{HF} = \Big|\mu_{\mathcal{M}}\big(\mathrm{HF}(z_0)\big) - \mu_{\mathcal{M}}\big(\mathrm{HF}(z_{HRRI})\big)\Big| + \Big|\sigma_{\mathcal{M}}\big(\mathrm{HF}(z_0)\big) - \sigma_{\mathcal{M}}\big(\mathrm{HF}(z_{HRRI})\big)\Big|$. The frequency-adaptive loss ($\mathcal{L}_f$) is defined as: $\mathcal{L}_f = \lambda_{LF}\,\mathcal{L}_{LF} + \lambda_{HF}\,\mathcal{L}_{HF}$. $\mathcal{L}_{LF}$ mainly updates the LF expert through the α-weighted path, while $\mathcal{L}_{HF}$ mainly updates the

Table 2: Quantitative comparison on RefGC-SR^2. The best and second-best results are highlighted in **bold** and underline, respectively. The Reference column indicates whether the model uses an additional HRRI. Asterisk (*) denotes that the model has been finetuned on our dataset.

Task	Model	Reference	CLIP-I↑	DINO↑	PSNR↑	SSIM↑	LPIPS↓
SR	DiT4SR (ICCV'25)	✗	0.8156	0.6555	15.3762	0.4932	0.4282
	DiT4SR* (ICCV'25)	✗	0.8186	0.6545	15.1005	0.5726	0.3884
	TSD-SR (CVPR'25)	✗	0.8224	0.6593	15.3725	0.5318	0.3766
	TSD-SR* (CVPR'25)	✗	0.8238	0.6599	17.3098	0.6165	0.2849
RefSR	ReFIR (NeurIPS'24)	✓	0.8199	0.6689	15.5310	0.5343	0.4259
	AdaRefSR (ICLR'26)	✓	0.8311	0.6858	15.6200	0.5629	0.3523
	AdaRefSR* (ICLR'26)	✓	0.8435	0.6828	17.1132	0.6260	0.2842
RefGCR	OmniPaint (ICCV'25)	✓	0.7831	0.5720	13.4472	0.5486	0.5166
	ImageCritic (CVPR'26)	✓	0.8536	0.7156	17.4391	0.6193	0.2991
	ImageCritic* (CVPR'26)	✓	0.8542	0.7165	17.2090	0.6060	0.3039
RefGC-SR^2	Ours	✓	**0.8696**	**0.7474**	**17.5148**	**0.6335**	**0.2746**

Table 3: In-the-wild evaluation, where outputs from compositing (Huang et al., 2025; Song et al., 2026) and customization (Zhang et al., 2025; Feng et al., 2026) models are treated as LRGI given HRRI and HRGT. **Compositing** denotes results using outputs from compositing models as LRGI, and **Customization** denotes results using outputs from customization models as LRGI.

	Compositing					Customization	
Model	CLIP-I↑	DINO↑	PSNR↑	SSIM↑	LPIPS↓	CLIP-I^T ↑	DINO^T ↑
DiT4SR*	0.8101	0.6057	13.6421	0.4781	0.4751	0.7184	0.3565
TSD-SR*	0.8188	0.5770	14.8691	0.4901	0.4253	0.7132	0.3160
ReFIR	0.8026	0.5506	14.0969	0.4834	0.4805	0.7106	0.3260
AdaRefSR*	0.8290	0.6081	14.2231	0.4840	0.4341	0.7206	0.3406
OmniPaint	0.7945	0.5570	14.1785	0.4873	0.4947	0.7106	0.3213
ImageCritic*	0.8407	0.6446	14.9504	0.5045	0.4335	0.7240	0.3659
Ours	**0.8476**	**0.6626**	**15.1028**	**0.5063**	**0.4170**	**0.7280**	**0.3771**

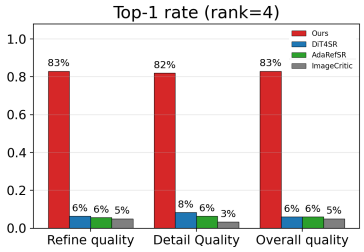


Figure 6: Top-rank-1 user study results. The user study is conducted with 16 participants comparing one model from each SR, RefSR and RefGCR task and RefGC-SR^2 (Ours).

HF expert through the $(1-\alpha)$ path. Combined with the layer-wise prior, this enables smooth specialization without hard routing. For the full loss (objective function) including this term, see the Appendix D and Overall Objective in Fig. 4.

4 Experiments

Implementation Details Details of the training setting on the RefGC-SR^2 Dataset and the evaluation protocols for Tables 2 and 3 are provided in Appendix C.

Training Datasets We denote the final triplet dataset generated using DipRefGC, described in Sec. 2.2, as the RefGC-SR^2 Dataset. Our training set contains 40K triplets. For training, we downsample the images generated by DipRefGC to 256×256 using bicubic interpolation to obtain the LRGIs, while retaining the corresponding HRGT images at 1024×1024.

Evaluation Datasets We further propose the RefGC-SR^2 Benchmark for evaluation. This benchmark consists of 200 triplets generated from HRRI and HRGT pairs that are not included in the RefGC-SR^2 Dataset. In addition, to evaluate in-the-wild setting, we use extra HRRI and HRGT pairs that are not used in either the RefGC-SR^2 Dataset or Benchmark. In-the-wild benchmark is constructed with HRRI-HRGT pairs that are not included in the RefGC-SR^2 Dataset or Benchmark and we generate the corresponding LRGIs by applying two recent compositing models, DreamFuse (Huang et al., 2025) and InsertAnything (Song et al., 2026), and two customization models, FreeCus (Zhang et al., 2025) and PersonalizeAnything (Feng et al., 2026). With 50 triplets, we obtain 200 samples after applying these four models and report the results for each RefGC task. For customization, the background and target object often undergo substantial changes. Therefore, metrics requiring pixel-aligned comparison with the GT or region-based evaluation using the reference object location are not meaningful. We instead report full-image metrics, denoted as CLIP-I^T and DINO^T.

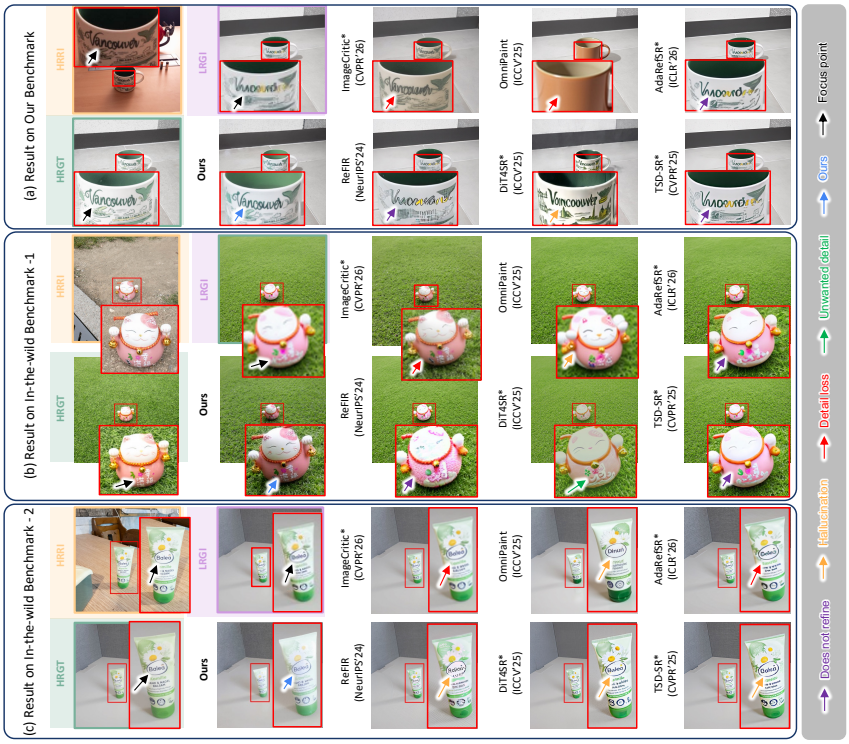


Figure 7: Qualitative comparison between our RefGC-SR^2 model and other models. (a) shows qualitative results on the RefGC-SR^2 Benchmark, while (b) and (c) present results in the in-the-wild benchmark. Colored arrows mark failure cases of competing methods (see legend), and blue arrows indicate our results. Our RefGC-SR^2 model better preserves fine details from HRRI and achieves high-quality upscaling compared to competing methods. *Best viewed in zoom.*

Quantitative Results We conduct quantitative evaluation on our RefGC-SR^2 Benchmark and in-the-wild benchmark. Table 2 compares our RefGC-SR^2 model with ImageCritic (Ouyang et al., 2025), OmniPaint (Yu et al., 2025), AdaRefSR (Wang et al., 2026), ReFIR (Guo et al., 2024), DiT4SR (Duan et al., 2025), and TSD-SR (Dong et al., 2025). Except for training-free methods (Guo et al., 2024) or those without open-source training code (Yu et al., 2025), we report additional results after fine-tuning on our training dataset for fair comparison (marked with * in Table 2). As shown in Table 2, our model achieves state-of-the-art performance across all metrics, demonstrating its effectiveness for the proposed RefGC-SR^2 task. To demonstrate the generalization ability of our model, we provide quantitative results on the in-the-wild benchmark described in Sec. 4-Dataset in Table 3. As shown in Table 3, our model achieves the best performance across all metrics, demonstrating the practical applicability of our RefGC-SR^2 model.

Qualitative Results Fig. 7 presents qualitative results on both the proposed RefGC-SR^2 Benchmark and the practical in-the-wild benchmark. We use colored arrows to highlight important regions, as described in the gray box on the right side of Fig. 7 (Please follow the arrows for detailed comparisons). (a) shows results on the RefGC-SR^2 Benchmark, while (b) and (c) present results in the in-the-wild setting. As shown in these examples, our RefGC-SR^2 model effectively refines generative artifacts faithfully to HRRI (e.g., bring back identity loss) while successfully increasing image resolution. We further provide additional qualitative results on other samples in the *suppl.*, and present qualitative evaluations on the outputs of both commercial and open-source models in Appendix I to demonstrate the additional generalization capability of our RefGC-SR^2 model.

User study To further demonstrate the advantage of our model, we conduct a user study under the in-the-wild setting. These outputs are then used as inputs for SR (DiT4SR (Duan et al., 2025)), RefSR (AdaRefSR (Wang et al., 2026)), RefGCR (ImageCritic (Ouyang et al., 2025)), and RefGC-SR^2 (Ours) for comparison. We evaluate the results using three criteria: refinement faithfulness (artifact

Table 4: Ablation study of the proposed components on RefGC-SR^2 benchmark. The best and second-best results are highlighted in **bold** and underline, respectively.

Setting	(a)	(b)	(c)	(d)
FreqMoLE	✗	✗	✓	✓
$\mathcal{L}_f$	✗	✓	✗	✓
CLIP-I↑	0.8437	0.8654	0.8673	**0.8696**
DINO↑	0.6870	0.7386	0.7317	**0.7474**
PSNR↑	16.3956	17.1981	17.3893	**17.5148**
SSIM↑	0.6068	0.6221	0.6221	**0.6335**
LPIPS↓	0.3538	0.2835	0.2896	**0.2746**

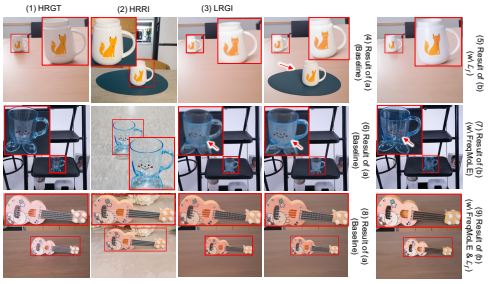


Figure 8: Qualitative ablation results of the proposed model. *Best viewed in Zoom.*

removal and restoration based on HRRI), detail restoration (SR ability), and overall quality. As shown in Fig. 6, our model outperforms all others across all criteria, achieving top-1 rates of 83%, 82%, and 83% on refine, detail, and overall quality, respectively, compared to at most 8% for the competing methods, demonstrating its ability to effectively perform both refinement and super-resolution. For more details on the user study protocol and results, please refer to Appendix J.

Ablation Study on RefGC-SR^2 Model Table 4 reports the ablation results on RefGC-SR^2 Benchmark. The baseline (a), removing both FreqMoLE and $\mathcal{L}_f$, shows the lowest performance across all metrics. Adding $\mathcal{L}_f$ alone in (b) significantly improves identity preservation (DINO $+7.5\%$) and perceptual quality (LPIPS -19.9%), confirming that decomposing supervision into LF alignment with HRGT and HF statistical matching with HRRI promotes faithful detail restoration. This is also reflected in Fig. 8: without $\mathcal{L}_f$ (4), HRRI is directly injected into the output, whereas with $\mathcal{L}_f$ (5), the model leverages HRRI while preserving the structure of HRGT. Adding FreqMoLE alone in (c) improves reconstruction fidelity (PSNR $+0.99$ dB, CLIP-I $+2.8\%$), as layer-wise mixing of LF/HF experts aligns with the coarse-to-fine frequency hierarchy of FLUX-Kontext analyzed in Sec. 3. In Fig. 8-(6), without FreqMoLE, the cup remains opaque (red arrow) and does not match the transparency in HRRI, while in (7) with FreqMoLE, the model correctly restores its transparency, consistent with both HRRI and HRGT. Combining both in (d) achieves the best performance across all metrics, confirming that FreqMoLE and $\mathcal{L}_f$ play complementary roles (structural modulation and band-specific supervision) that are both essential to RefGC-SR^2. As shown in Fig. 8-(9), the final model effectively transfers fine details from HRRI while correcting mismatched regions such as the star and ear (red arrows). Further analyses of high-frequency supervision and robustness to large viewpoint differences are provided in Appendix G.1 and Appendix G.2, respectively.

Ablation Study on DipRefGC We further conduct an ablation study of DipRefGC to validate the role of pose conditioning and RefGC-oriented fine-tuning. The results show that Canny guidance improves pose consistency between LRGI and HRGT, while fine-tuning on both compositing and customization settings improves reference identity preservation and reduces the distribution gap to real RefGC outputs (Feng et al., 2026; Zhang et al., 2025; Huang et al., 2025; Song et al., 2026). See Fig. 3 for qualitative triplet examples. See Appendix F for further DipRefGC ablations.

5 Conclusion

In this paper, we propose **RefGC-SR^2**, a new task that jointly performs super-resolution and artifact refinement on generated images guided by HRRI. To support this task, we introduce the first real-world triplet dataset (RefGC-SR^2 Dataset) and the first dedicated model (RefGC-SR^2 Model). For dataset construction, we propose DipRefGC, a FLUX-based generator that synthesizes HRGT-aligned artifact-corrupted LRGIs, yielding triplets suitable for supervised training. After training on this dataset, our RefGC-SR^2 model is able to leverage HRRI to faithfully refine generative artifacts while recovering high-resolution details. Extensive experiments show that our model consistently outperforms SR, RefSR, and RefGCR baselines, preserving reference identity and restoring fine details, on our RefGC-SR^2 Benchmark as well as on real in-the-wild RefGC inputs.

6 AI USE STATEMENT

Generative AI tools were used in several parts of this work. First, Qwen3-VL (Bai et al., 2025) was used for object-centric video filtering and captioning during dataset curation, as described in Sec. 2.1. Second, generative image models were used to construct synthetic training and evaluation data, as detailed in Sec. 2.2. Third, ChatGPT (OpenAI, 2022) was used to assist with translation and grammar checking of this manuscript. ChatGPT was also used to assist with development-environment configuration. Generative AI tools were not used to formulate the research hypotheses, design the proposed methodology or experiments, perform the scientific analysis, or interpret the experimental results. All AI-assisted text and code were reviewed and verified by the authors, while AI-assisted data construction followed the filtering and validation procedures described in the paper. We take full responsibility for the final content of this work, including all text, code, data, claims, and artifacts produced with the aid of generative AI.

7 REPRODUCIBILITY STATEMENT

The dataset construction and model formulation are described in Sec. 2 and Sec. 3, respectively. Detailed training configurations, hyperparameters, and evaluation protocols are provided in the appendix. We plan to release the dataset, benchmark, code, and checkpoints.

REFERENCES

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Peter J. Burt and Edward H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. doi: 10.1109/TCOM.1983.1095851.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *Advances in Neural Information Processing Systems*, 37:84010–84032, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602, 2024b.

Seungho Choi, Jeahun Sung, and Jihyong Oh. Framer: Frequency-aligned self-distillation with adaptive modulation leveraging diffusion priors for real-world image super-resolution. *arXiv preprint arXiv:2512.01390*, 2025.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104. IEEE, 2022.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23174–23184, 2025.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6159–6168, 2024.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3921–3929, 2026.

Google DeepMind. Gemini 2.5 flash image (nano banana). `https://deepmind.google/models/gemini-image/flash/`, 2025. Accessed: 2026-06-01.

Google DeepMind. Gemini 3.1 Flash Image model card, February 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-flash-image/`. Accessed: 2026-09-26.

Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yaohua Zha, Bin Chen, and Shu-tao Xia. Refir: Grounding large restoration models with retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:46593–46621, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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 *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Junjia Huang, Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamfuse: Adaptive image fusion with diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17292–17301, 2025.

InstantX. Flux.1-dev-controlnet-canny. `https://huggingface.co/InstantX/FLUX.1-dev-Controlnet-Canny`, 2024.

Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2355–2365, 2025.

Jinwoo Kim, Sangmin Han, Jinho Jeong, Jiwoo Choi, Dongyeoung Kim, and Seon Joo Kim. Orida: Object-centric real-world image composition dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3051–3060, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Black Forest Labs. Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. 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.

Hongjae Lee, Jun-Sang Yoo, and Seung-Won Jung. Refqsr: Reference-based quantization for image super-resolution networks. *IEEE Transactions on Image Processing*, 33:2823–2834, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haowen Li, Zhenfeng Fan, Zhang Wen, Zhengzhou Zhu, and Yunjin Li. Aicomposer: Any style and content image composition via feature integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16840–16850, 2025.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y Zhang, Natalia Neverova, et al. Uncommon objects in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14102–14113, 2025a.

Yaoli Liu, Ziheng Ouyang, Shengtao Lou, and Yiren Song. Omnirefiner: Reinforcement-guided local diffusion refinement. *arXiv preprint arXiv:2511.19990*, 2025b.

Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13072–13082, 2025c.

OpenAI. Introducing ChatGPT, November 2022. URL `https://openai.com/index/chatgpt/`. Accessed: 2026-09-26.

OpenAI. Gpt image 1.5. `https://openai.com/index/image-generation-api/`, 2025. gpt-image-1.5, introduced December 2025. Accessed: 2026-06-01.

Openverse Project. Openverse - extensive library of free stock photos, images, and audio. `https://openverse.org`, 2026. Accessed: 2026-08-29.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ziheng Ouyang, Yiren Song, Yaoli Liu, Shihao Zhu, Qibin Hou, Ming-Ming Cheng, and Mike Zheng Shou. The consistency critic: Correcting inconsistencies in generated images via reference-guided attentive alignment. *arXiv preprint arXiv:2511.20614*, 2025.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4047–4056, October 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7986–7996, 2025.

Wensong Song, Hong Jiang, Zongxin Yang, Zheqiao Cheng, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9097–9105, 2026.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18310–18319, 2023.

Yizhi Song, Liu He, Zhifei Zhang, Soo Ye Kim, He Zhang, Wei Xiong, Zhe Lin, Brian Price, Scott Cohen, Jianming Zhang, et al. Refine-by-align: Reference-guided artifacts refinement through semantic alignment. *arXiv preprint arXiv:2412.00306*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, He Zhang, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Multitwine: Multi-object compositing with text and layout control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8094–8104, 2025.

AlimamaCreative Team. Flux.1-dev controlnet inpainting - beta. `https://huggingface.co/alimama-creative/FLUX.1-dev-Controlnet-Inpainting-Beta`, 2024. Hugging Face model checkpoint.

Athanasios Tragakis, Marco Aversa, Chaitanya Kaul, Roderick Murray-Smith, and Daniele Faccio. Is one gpu enough? pushing image generation at higher-resolutions with foundation models. In *NeurIPS*, 2024.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18325–18334, 2025.

Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. *arXiv preprint arXiv:2602.01864*, 2026.

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.

Wikimedia Foundation. Wikimedia commons - a database of freely usable media files. `https://wikimedia.org`, 2026. Accessed: 2026-08-29.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025b.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, and Lei Zhu. Learning diffusion texture priors for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2524–2534, 2024.

Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17324–17334, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Yanbing Zhang, Zhe Wang, Qin Zhou, and Mengping Yang. Freecus: Free lunch subject-driven customization in diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15531, 2025.

Yi Zhang, Qixue Yang, Damon M Chandler, and Xuanqin Mou. Reference-based multi-stage progressive restoration for multi-degraded images. *IEEE Transactions on Image Processing*, 33: 4982–4997, 2024.

Dewei Zhou, You Li, Zongxin Yang, and Yi Yang. Refineanything: Multimodal region-specific refinement for perfect local details. *arXiv preprint arXiv:2604.06870*, 2026.

APPENDIX CONTENTS

A Technical appendices and supplementary material **16**

B Related Works **16**

C Implementation Details **18**

D Details on Auxiliary Loss Terms **18**

D.1 Flow-Matching Loss ($\mathcal{L}_{FM}$) 18

D.2 Attention Alignment Loss ($\mathcal{L}_{aal}$) 19

D.3 Overall Objective 19

E Dataset Statistics and Diversity **19**

E.1 Category and Instance Coverage 19

E.2 Geometric Diversity 19

F DipRefGC Ablations and LRGI Validation **20**

F.1 Design and Training Ablations 20

F.2 Validation of DipRefGC-Generated LRGIs 21

G Further Analysis and Results of Our RefGC-SR^2 Model **22**

G.1 Analysis of High-Frequency Supervision 22

G.2 Robustness to Large Viewpoint Samples 22

G.3 Sensitivity to Reference Quality 23

G.4 Evaluation on Independent Sources and RefGC Outputs 25

H Multi-Object Evaluation **26**

I Qualitative Results on Outputs from Commercial and Open-Source Models **28**

J User Study Protocols and Analysis **32**

J.1 Protocol 33

J.2 Detailed Analysis 33

K Limitations **33**

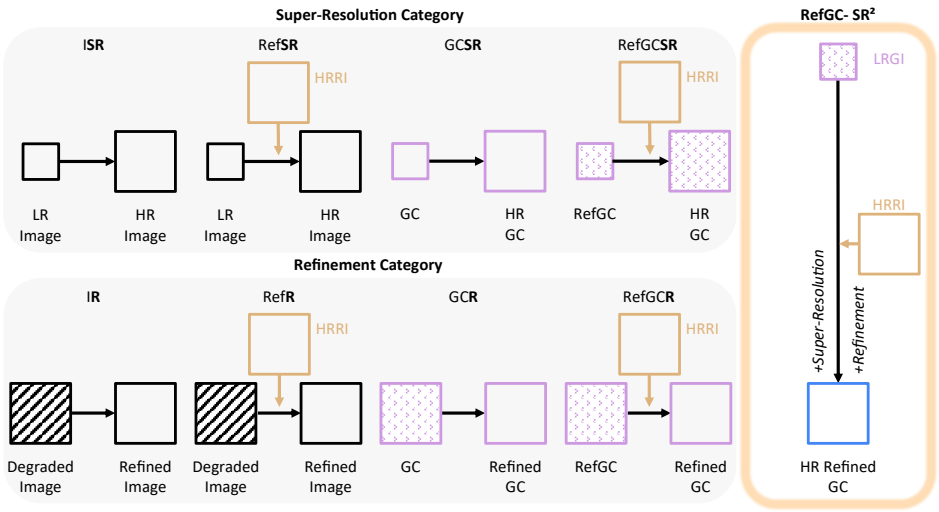


Figure 9: **Visual taxonomy of related image enhancement tasks.** We illustrate the input-output structure of all eight related tasks (Sec. B, Table 1) alongside our RefGC-SR^2. Tasks are organized into two categories: the *Super-Resolution Category* (top), which enlarges spatial resolution, and the *Refinement Category* (bottom), which removes artifacts or degradations at a fixed resolution. **Box size** denotes spatial resolution: smaller boxes are low-resolution (LR) inputs and larger boxes are high-resolution (HR) outputs, so a size change between input and output indicates super-resolution while equal sizes indicate resolution-preserving. **Box color** denotes the image type: *black* for natural images, *purple* for generated content (GC), *yellow* for the user-provided high-resolution reference image (HRRI), and *blue* for our final HR refined output. **Box pattern** denotes the degradation type: diagonal stripes indicate natural degradations such as blur and noise, while hatched fill patterns indicate generative artifacts such as identity distortion, detail inconsistency, texture loss, and overall quality degradation. Among all tasks, our RefGC-SR^2 is the only one that *simultaneously* consumes LRGI and an HRRI as inputs and jointly performs super-resolution and generative-artifact refinement, producing an HR refined output.

A Technical Appendices and Supplementary Material

We provide part of the qualitative results from the proposed RefGC-SR^2 Benchmark, along with additional comparisons, in the supplementary material. Due to space limitations, only a subset of the benchmark and comparison results is included. We plan to release the full qualitative results, RefGC-SR^2 Dataset, Benchmark, code, and checkpoints in the future.

B Related Works

Image Super-Resolution (ISR) Image Super-Resolution (ISR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) observation. As shown in Fig.9 (**ISR**), ISR takes a single LR image as input and produces an HR image, without using any external reference. Recent diffusion-based ISR methods(Choi et al., 2025; Duan et al., 2025; Dong et al., 2025) exploit pretrained generative priors and have achieved strong performance under natural-image degradation assumptions such as bicubic downsampling, blur, noise, compression, and realistic camera degradations. However, ISR is not designed for LR generated content images (LRGIs), whose degradations arise from generative pipelines rather than camera-side or hand-crafted degradation processes. Moreover, ISR cannot recover identity-specific information that is absent from the LRGI but still available in the user-provided HRRI. RefGC-SR^2 differs from ISR by explicitly targeting generated content with artifacts and by reusing the HRRI as an external recovery source.

Reference-guided Super-Resolution (RefSR) Reference-guided SR (RefSR) extends ISR by exploiting an external HR reference image (HRRI) to recover fine-grained details (Lee et al., 2025;

2024). As shown in Fig. 9 (Ref**SR**), RefSR takes an LR image and an HRRI as inputs, and produces an HR image. This formulation is related to ours in that it leverages a reference image as a source of high-frequency information. However, existing RefSR methods are still formulated for natural-image SR, where the LR input is assumed to be degraded from a real HR image. They do not model RefGC-specific artifacts such as identity distortion, detail inconsistency, hallucinated structures, or texture loss. Therefore, directly applying RefSR to LRGIs may sharpen the image but does not explicitly refine generative artifacts. RefGC-SR^2 inherits the reference-guided recovery concept of RefSR, but extends it to the generated-content domain and jointly performs generative artifact refinement.

Generated Content Super-Resolution (GCSR) Generated Content Super-Resolution (GCSR) aims to overcome the native resolution limits of generative models (Jeong et al., 2025; Du et al., 2024; Tragakis et al., 2024). As shown in Fig. 9 (GC**SR**), GCSR takes generated content as input and increases its resolution without using any HRRI. Existing GCSR methods rely on the internal priors of pretrained generative models to hallucinate missing details at higher resolution. This is effective when the goal is to upscale a generated image itself, but it is insufficient for reference-guided generated content (RefGC). In RefGC pipelines, the generated output may have already lost reference-specific identity, logo, texture, or material details due to low-resolution reference conditioning. Since GCSR has no access to the original HRRI, it cannot explicitly recover such missing reference information. In contrast, RefGC-SR^2 reintroduces the HRRI at the post-generation stage and uses it as an explicit source of high-frequency reference details.

Reference-guided Generated Content Super-Resolution (RefGCSR) We use RefGCSR to denote a possible setting that performs super-resolution on generated content with the help of an HRRI. As shown in Fig. 9 (RefGC**SR**), RefGCSR would take both GC and HRRI as inputs and produce an HR generated output. However, this setting has not been systematically established as a dedicated task in the literature, and a straightforward implementation would be to apply existing RefSR methods to generated content. Such a formulation still focuses on resolution recovery and does not explicitly address generative artifacts inherited from RefGC pipelines. As a result, the output may become higher-resolution, but artifacts such as identity distortion, detail inconsistency, and texture loss can remain unresolved. RefGC-SR^2 is stricter and more comprehensive: it requires not only HRRI-guided super-resolution, but also generative-artifact refinement.

Image Restoration (IR) Image restoration (IR) maps a degraded image to a clean image by removing degradations such as noise, blur, compression artifacts, or other natural corruptions (Ye et al., 2024). As shown in Fig. 9 (I**R**), IR refines a degraded natural image at the same spatial resolution, without using an HRRI. Although modern diffusion-based restoration methods provide strong natural-image priors, their assumptions differ from those of RefGC-SR^2. They do not consume generated content as a specific input domain, do not use the original HRRI, and do not perform super-resolution as part of the task. Therefore, IR cannot recover reference-specific details missing from LRGIs, nor can it produce an HR output. RefGC-SR^2 extends beyond IR by jointly using HRRI guidance, generated-content artifact modeling, and resolution recovery.

Generated Content Restoration (GCR) Generated Content Restoration (GCR) focuses on improving generated images by removing visible artifacts or low-quality regions (Lin et al., 2024). As shown in Fig. 9 (GC**R**), GCR takes generated content as input and produces a refined generated image at the same resolution. This task is closer to ours than IR because it recognizes that generated images have a different artifact distribution from natural degraded images. However, GCR still operates without an HRRI. Consequently, it must infer missing details from the generated image alone, and cannot restore identity-specific information that was discarded during the upstream reference-guided generation process. Moreover, GCR is typically a fixed-resolution refinement problem. RefGC-SR^2 differs by using the HRRI as an external recovery source and by jointly performing super-resolution.

Reference-guided Refinement (RefR) Reference-guided refinement (RefR) uses an external reference image to refine a degraded image (Zhang et al., 2024; Guo et al., 2024). As shown in Fig. 9 (Ref**R**), RefR augments image restoration with an HRRI, but still focuses on natural-image degradations and fixed-resolution refinement. This setting shares with ours the use of an HRRI, but its target domain and output requirement are different. RefR is not formulated for generated

content from RefGC pipelines, and it does not address the coupled problem of LR input, generative artifacts, and missing reference-specific details. RefGC-SR^2 extends reference-guided refinement to the RefGC setting and additionally requires HR output recovery.

Reference-guided Generated Content Refinement (RefGCR) Reference-guided Generated Content Refinement (RefGCR) is the most closely related task to ours. RefGCR methods (Song et al., 2024; Ouyang et al., 2025; Zhou et al., 2026; Liu et al., 2025b) use a reference image to correct visible artifacts or inconsistencies in generated content. As shown in Fig. 9 (RefGC**R**), RefGCR consumes GC and HRRI as inputs, and produces a refined GC output. Representative methods perform local alignment, reference-guided detail correction, or region-specific refinement to improve consistency between generated content and the reference. However, existing RefGCR methods are primarily fixed-resolution refinement methods. They do not formulate the problem of recovering an HR output from an LR generated input. This distinction is critical for modern RefGC pipelines, where the user-provided HRRI is often downsampled before generation, causing fine-grained reference details to be discarded before the generated output is produced. RefGCR can refine artifacts at the current resolution, but cannot recover the lost HR information. RefGC-SR^2 resolves this limitation by treating artifact refinement and super-resolution as a single coupled post-generation task.

C Implementation Details

RefGC-SR^2 Dataset To construct the RefGC-SR^2 Dataset, we build DipRefGC on a frozen FLUX (Labs, 2024) backbone with two LoRA-tuned (Hu et al., 2022) ControlNets (InstantX, 2024; Team, 2024). DipRefGC is trained on a single NVIDIA B200 GPU with a batch size of 16 and a learning rate of 10^{-4} for 19,000 iterations.

RefGC-SR^2 Model We build the RefGC-SR^2 Model on a frozen FLUX-Kontext (Labs et al., 2025) backbone. For both DipRefGC and the RefGC-SR^2 Model, the VAE (Kingma & Welling, 2013) and T5 encoders are kept frozen, and we follow ImageCritic (Ouyang et al., 2025) for the text encoder configuration. The RefGC-SR^2 Model is trained on a single NVIDIA B200 GPU for $4\times$ upscaling with a batch size of 16 and a learning rate of 10^{-4}. The frequency-adaptive loss ($\mathcal{L}_f$) weights are set to $\lambda_{LF} = 0.1$ and $\lambda_{HF} = 0.2$.

In-the-wild Benchmark For compositing, CLIP-I and DINO are computed on the masked foreground object regions, while PSNR, SSIM, and LPIPS are computed on crops enclosing the object masks obtained using SAM3. For customization, we report only whole-image CLIP-I and DINO because customization models may synthesize the subject in a different layout, making region-restricted and pixel-aligned metrics unreliable. Therefore, the compositing and customization results follow separate evaluation protocols and are not directly comparable across the two settings.

D Details on Auxiliary Loss Terms

The overall training objective of our RefGC-SR^2 Model combines our proposed frequency-adaptive loss $\mathcal{L}_f$ (Sec. 3) with two auxiliary loss terms: the flow-matching loss $\mathcal{L}_{FM}$ inherited from the FLUX-Kontext backbone (Labs et al., 2025), and the attention alignment loss $\mathcal{L}_{aal}$ adopted from ImageCritic (Ouyang et al., 2025). For completeness, we briefly review these two terms here.

D.1 Flow-Matching Loss ($\mathcal{L}_{FM}$)

FLUX-Kontext (Labs et al., 2025) is trained under the rectified flow-matching framework, where the diffusion transformer is supervised to predict the velocity field that transports a Gaussian prior to the data distribution along a linear trajectory. Following this formulation, given a clean target latent z_1 (encoded from HRGT) and a Gaussian noise sample $z_0 \sim \mathcal{N}(0, I)$, we define a linear interpolant

$$z_t = (1 - t)\, z_0 + t\, z_1, \quad t \in [0, 1],$$

whose ground-truth velocity along the path is $z_1 - z_0$. The model v_θ predicts the velocity at z_t conditioned on the task inputs c (the LRGI latent, the HRRI latent, and the text instruction). The

flow-matching loss is then defined as

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\, z_0,\, z_1,\, c}\left[\left\|v_\theta(z_t, t, c) - (z_1 - z_0)\right\|_2^2\right].$$

This is the same objective used when fine-tuning FLUX-Kontext with LoRA in standard adaptation pipelines, and it serves as the primary supervision signal that drives our RefGC-SR^2 Model to produce outputs aligned with HRGT.

D.2 Attention Alignment Loss ($\mathcal{L}_{aal}$)

To encourage the model to faithfully transfer reference information from HRRI to the refined output, we adopt the attention alignment loss proposed by ImageCritic (Ouyang et al., 2025). ImageCritic introduces a *reference-guided attentive alignment* mechanism that supervises the cross-attention behavior between the model's output tokens and the HRRI tokens, encouraging the model to concentrate its attention on the object region of the HRRI rather than diffusing it across irrelevant background. This regularization is particularly important in our setting because the HRRI and HRGT differ in viewpoint and scene context, so the model must learn to attend to the object identity in the HRRI while ignoring its background. We adopt the same formulation as in the original ImageCritic work (Ouyang et al., 2025), and refer the reader to it for the full mathematical specification.

D.3 Overall Objective

Combining the two auxiliary terms above with our proposed $\mathcal{L}_f$, the full training objective is

$$\mathcal{L} = \mathcal{L}_{FM} + \mathcal{L}_f + \mathcal{L}_{aal}.$$

The three terms play complementary roles: $\mathcal{L}_{FM}$ provides the dominant generative supervision toward HRGT under the rectified-flow framework of FLUX-Kontext, $\mathcal{L}_{aal}$ controls *where* the model attends in the HRRI, and our $\mathcal{L}_f$ (Sec. 3) controls *which frequency bands* are transferred from each reference (low-frequency from HRGT, high-frequency from HRRI). This decomposition is consistent with the design of our FreqMoLE module, which routes information through frequency-specialized LoRA experts in a layer-depth-dependent manner.

E Dataset Statistics and Diversity

We provide additional statistics on the semantic coverage and geometric and photometric diversity of the RefGC-SR² Dataset. These statistics characterize the variation between the HRRI and the HRGT, which depict the same object instance under different capture conditions.

E.1 Category and Instance Coverage

The RefGC-SR² Dataset contains 16 super-categories, 577 distinct object labels, and approximately 3,500 distinct object identities. Object labels are extracted from the source data with assistance from Qwen3-VL (Bai et al., 2025) and then normalized to obtain the final category counts. The 12 labels used during video collection describe broad source-level groups rather than the final semantic categories of the curated dataset. For example, the collection-stage label *kitchen cookware* encompasses objects such as bottles, mugs, bowls, jars, and cups. These collection labels should therefore be distinguished from the category and instance counts of the final dataset.

E.2 Geometric Diversity

We quantify HRRI–HRGT variation in the 40K training pairs from ORIDa (Kim et al., 2025) and uCO3D (Liu et al., 2025a) using three measures. Viewpoint difference is the relative camera rotation estimated with MASt3R (Leroy et al., 2024) on foreground object regions. The scale ratio is defined as $\sqrt{a_{\mathrm{HRGT}}/a_{\mathrm{HRRI}}}$, where $a = A/(WH)$ is the foreground mask area normalized by the corresponding image area. As a proxy for differences in lighting conditions and other photometric changes, we compute the Euclidean distance between the mean foreground color vectors of HRRI and HRGT in 8-bit encoded LAB space.

Table 5: Geometric diversity and lighting-condition variation of HRRI–HRGT pairs. Values are reported as median [interquartile range]. Viewpoint statistics are computed from a stratified sample of 3,000 pairs, while scale and foreground-color statistics use all 40,000 valid pairs. Foreground-color distance is computed as the Euclidean distance between mean foreground colors in 8-bit encoded LAB space.

Source	Viewpoint difference Δ (degrees)	Scale ratio $\sqrt{a_{\mathrm{HRGT}}/a_{\mathrm{HRRI}}}$	Foreground-color distance (encoded LAB)
uCO3D	**82.9** [54.8–133.1]	1.00 [0.90–1.11]	32.0 [16–62]
ORIDa	29.8 [12.1–149.9]	1.00 **[0.68–1.47]**	**88.0** [46–166]

For viewpoint estimation, we draw a stratified sample of 3,000 pairs: 750 ORIDa compositing pairs, 750 ORIDa customization pairs, and 1,500 uCO3D pairs. Scale and foreground-color statistics use 40,000 valid pairs, comprising 20,000 ORIDa pairs and 20,000 uCO3D pairs.

Table 5 reports the median and interquartile range for each measure. uCO3D exhibits a larger median viewpoint difference, whereas ORIDa exhibits wider scale variation and larger foreground-color differences. Although both sources have a median scale ratio of 1.00, their interquartile ranges show variation in individual pairs. These statistics demonstrate the diversity of HRRI–HRGT pairs in viewpoint, scale, and lighting conditions.

F DipRefGC Ablations and LRGI Validation

F.1 Design and Training Ablations

Table 6: Ablation of DipRefGC on the RefGC-SR^2 Benchmark. We evaluate identity preservation using DINOv2 (Oquab et al., 2023) and CLIP-I (Radford et al., 2021) similarity to the HRRI, pose consistency using object-mask IoU with the HRGT, and image-feature distribution similarity using FID (Heusel et al., 2017) against the outputs of external RefGC generators. Rows (a)–(e) report the core ablation, and the shaded row (e) is our final model. Row (f) further fine-tunes (e) with extended data to examine the potential benefit of additional training data. Bold and underline indicate the best and second-best distinct values in each column, respectively; ties share the same formatting.

	Compositing				**Customization**			
Variant	DINO ↑	CLIP-I ↑	IoU ↑	FID ↓	DINO ↑	CLIP-I ↑	IoU ↑	FID ↓
(a) DiptychPrompting (Shin et al., 2025) (base)	0.539	0.823	0.480	106.92	0.484	0.796	0.311	**104.94**
(b) DiptychPrompting + Canny	0.622	0.841	0.601	99.34	0.634	0.845	0.601	<u>107.64</u>
Phase 1								
(c) Compo-only	**0.691**	**0.863**	**0.641**	97.52	0.680	<u>0.854</u>	**0.630**	109.18
(d) Custom-only	0.664	0.854	0.624	100.45	<u>0.683</u>	**0.856**	<u>0.621</u>	109.69
(e) Jointly fine-tuned (final)	0.684	0.860	<u>0.637</u>	<u>97.24</u>	0.677	0.853	0.602	109.27
Phase 1 → Phase 2								
(f) + Extended data fine-tuning	<u>0.687</u>	<u>0.861</u>	**0.641**	**95.97**	**0.684**	**0.856**	0.605	107.91

Table 6 reports the ablation study of DipRefGC. Adding Canny ControlNet (InstantX, 2024) (b) improves mask IoU over the base DiptychPrompting model (Shin et al., 2025) from 0.480 to **0.601** in compositing and from 0.311 to **0.601** in customization, verifying its role in enforcing the pose constraint between the LRGI and HRGT. RefGC-oriented fine-tuning improves identity preservation, with higher DINO and CLIP-I scores than the Canny-only variant in both compositing-only and customization-only settings. Joint fine-tuning provides a practical balance across the two RefGC modes, achieving competitive identity and pose scores while improving FID in compositing.

For training, Phase 1 uses 40K triplets constructed from ORIDa (Kim et al., 2025) and uCO3D (Liu et al., 2025a), consisting of 20K samples from each dataset. Phase 2 further fine-tunes the final model (e) on 12K additional triplets from UltraVideo (Xue et al., 2025), which contains more challenging videos with richer subject motion. This additional fine-tuning reduces FID in both compositing and customization, suggesting that DipRefGC can benefit from more challenging motion-rich data while preserving the pose consistency required for RefGC-SR^2 training.

Table 7: Comparison of artifact metric distributions on the 200-pair RefGC-SR^2 Benchmark. We compare 200 outputs from the final DipRefGC model with 200 external outputs, selecting 50 per external generator and one DipRefGC output for the same pair and task. Each benchmark pair appears once in each pool. We report the one-dimensional 1-Wasserstein distance between the pooled distributions and the mean and maximum of the six pairwise distances among the four external output sets. Lower values indicate closer distributions within each metric.

Metric (score range)	DipRefGC vs. pooled external outputs	Among external generators Mean / Max
Identity similarity (DINO, $[-1, 1]$)	**0.0542**	0.2139 / 0.3649
Detail similarity (mSSIM, $[-1, 1]$)	0.0716	0.0649 / 0.1173
Texture preservation (HF ratio, ≥ 0)	**0.2406**	0.3916 / 0.5186
Overall quality (VLM, $[0, 4]$)	**0.1600**	0.3400 / 0.5200

F.2 Validation of DipRefGC-Generated LRGIs

We further examine whether the LRGIs generated by DipRefGC exhibit artifact characteristics similar to those of outputs from existing RefGC generators.

Evaluation Setup All comparisons use the 200 HRRI–HRGT pairs of the RefGC-SR^2 Benchmark. For every benchmark pair, DipRefGC generates one compositing LRGI and one customization LRGI. As external generators, we use DreamFuse (Huang et al., 2025) and InsertAnything (Song et al., 2026) for compositing, and FreeCus (Zhang et al., 2025) and PersonalizeAnything (Feng et al., 2026) for customization; each external generator also produces one output for every benchmark pair.

We characterize four aspects of RefGC artifacts. Identity similarity is measured by DINOv2 (Oquab et al., 2023) cosine similarity to the HRRI on foreground object crops. Detail similarity is measured by masked SSIM (Wang et al., 2004) against the HRGT, averaging the SSIM map over the HRGT foreground mask. Texture preservation is measured by the high-frequency energy ratio relative to the HRGT: within the crop defined by the HRGT object mask, we compute the fraction of Fourier power above a normalized radial frequency of 0.25 and divide the output fraction by the corresponding HRGT fraction. Overall quality is assessed on the full output image by Qwen3-VL-8B-Instruct (Bai et al., 2025), which assigns an integer score from 0 to 4, with higher scores indicating more pronounced artifacts.

For each metric, we compute the one-dimensional 1-Wasserstein distance (Ramdas et al., 2017) between two empirical score distributions, where lower values indicate closer distributions. DINOv2 cosine similarity and SSIM lie in $[-1, 1]$, so the corresponding distances lie in $[0, 2]$. The VLM score lies in $[0, 4]$, giving distances in $[0, 4]$. The high-frequency energy ratio is nonnegative and has no fixed upper bound. Because the metrics have different scales, distances are interpreted only within each metric.

Comparison with Existing RefGC Generators Table 7 compares the final DipRefGC model (e) with the external generators on the benchmark. We partition the 200 benchmark pairs into four disjoint groups of 50, stratified by data source with a fixed random seed, and assign one group to each external generator. This yields 50 outputs per external generator and 200 external outputs in total. For each assigned pair, we take the DipRefGC output of the same task, giving 200 DipRefGC outputs (100 compositing and 100 customization). Both pools therefore cover every benchmark pair exactly once with the same task composition. As a descriptive reference for variation among existing generators, we report the mean and maximum of the six pairwise distances among the four external output sets. Because these sets are defined on disjoint pairs, the reference also reflects differences in content and sampling, and we use it for descriptive comparison rather than as a statistical threshold.

The distances between DipRefGC and the pooled external outputs are **0.0542**, 0.0716, **0.2406**, and **0.1600** for identity similarity, detail similarity, texture preservation, and overall artifact severity, respectively. The identity, texture, and overall-artifact distances are smaller than the corresponding mean distances among the external generators (0.2139, 0.3916, and 0.3400). The detail distance of 0.0716 exceeds the corresponding mean of 0.0649 but remains below the maximum of 0.1173. Thus, on three of the four measured dimensions, DipRefGC outputs differ from the external outputs by less

than the external generators typically differ from one another, and on the remaining dimension the difference stays within the observed range among them.

G Further Analysis and Results of Our RefGC-SR^2 Model

In this section, we provide additional analyses and experimental results for our RefGC-SR^2 model to further examine its design choices and robustness under various conditions.

G.1 Analysis of High-Frequency Supervision

In this section, we present the motivation for high-frequency supervision and provide empirical evidence supporting its design. The HRRI and HRGT depict the same object but are generally not spatially registered because of differences in viewpoint, pose, and object placement. Direct pixel-wise supervision against the HRRI can therefore impose incorrect constraints between spatial locations that do not correspond to each other. To avoid this problem, our method retains the HRGT-anchored flow-matching and low-frequency objectives, $\mathcal{L}_{\mathrm{FM}}$ and $\mathcal{L}_{\mathrm{LF}}$, while using $\mathcal{L}_{\mathrm{HF}}$ as an auxiliary distributional constraint.

We conduct an ablation using the RefGC-SR^2 model to examine four supervision strategies, as summarized in Table 8. Setting (a) removes $\mathcal{L}_{\mathrm{f}}$ and serves as the baseline for measuring its contribution. Setting (b) replaces $\mathcal{L}_{\mathrm{HF}}$ with a pixel-wise ℓ_1 loss against the HRRI to test whether direct reference supervision is effective. Setting (c) applies a full-band ℓ_1 loss against the HRGT to determine whether HRGT-only reconstruction supervision is sufficient. Setting (d) is our proposed approach, which matches the channel-wise mean and standard deviation of the high-frequency components between the output and HRRI. All settings use the same FreqMoLE backbone, training schedule, Gaussian frequency decomposition and $\mathcal{L}_{\mathrm{FM}}$.

Table 8: Ablation of frequency-adaptive supervision in the RefGC-SR^2 model. All settings share the same FreqMoLE backbone, training schedule, and flow-matching loss $\mathcal{L}_{\mathrm{FM}}$, while varying the auxiliary supervision. The best and second-best results are highlighted in **bold** and underline, respectively.

Auxiliary supervision	CLIP-I	DINO	PSNR	SSIM	LPIPS
(a) None (w/o $\mathcal{L}_f$)	<u>0.8673</u>	0.7317	17.3893	<u>0.6221</u>	<u>0.2896</u>
(b) Pixel-wise ℓ_1 to HRRI	0.8656	<u>0.7453</u>	16.9123	0.6202	0.2960
(c) Full-band ℓ_1 to HRGT	0.8578	0.7243	**17.7273**	0.6191	0.2929
(d) Distributional $\mathcal{L}_{\mathrm{HF}}$ to HRRI (Ours)	**0.8696**	**0.7474**	<u>17.5148</u>	**0.6335**	**0.2746**

Setting (a) provides a strong reconstruction baseline without $\mathcal{L}_f$. Setting (b) improves DINO over (a), but degrades CLIP-I and all reconstruction metrics, suggesting that pointwise reference supervision can impair reconstruction fidelity when the HRRI and HRGT are not spatially aligned. Setting (c) achieves the highest PSNR, but produces the lowest CLIP-I and DINO scores. Its SSIM and LPIPS are also worse than those of (a). This indicates that full-band HRGT supervision encourages pixel fidelity but does not explicitly associate the output high-frequency content with reference-specific details from the HRRI. Setting (d) achieves the best CLIP-I, DINO, SSIM, and LPIPS scores while maintaining the second-best PSNR. Its improvements over (a) support the effectiveness of the proposed frequency-adaptive supervision. Although channel-wise statistics admit multiple spatial configurations, this many-to-one property applies only to the auxiliary regularizer. In setting (d), the output remains anchored to the HRGT through $\mathcal{L}_{\mathrm{FM}}$ and $\mathcal{L}_{\mathrm{LF}}$, allowing reference-based high-frequency supervision without imposing invalid pixel-level correspondence.

G.2 Robustness to Large Viewpoint Samples

In this section, we further evaluate whether RefGC-SR^2 can effectively leverage information from the HRRI under substantial HRRI-HRGT viewpoint discrepancies. Since the HRRI is an independent observation of the same object instance, substantial viewpoint differences can arise between the HRRI and HRGT. We quantify this variation by estimating the relative HRRI-to-HRGT camera rotation Δ using MASt3R (Leroy et al., 2024) on the foreground object regions. The median rotation across the

Table 9: Quantitative comparison on the large-viewpoint subset of the RefGC-SR^2 Benchmark. The subset contains the 25% of samples with the largest HRRI-HRGT viewpoint differences, corresponding to $\Delta \geq 132.5^\circ$ and $N = 50$. The best and second-best results are highlighted in **bold** and underline, respectively. An asterisk ($*$) indicates a baseline finetuned on the RefGC-SR^2 Dataset.

Model	CLIP-I↑	DINO↑	PSNR↑	SSIM↑	LPIPS↓
OmniPaint	0.7999	0.5870	12.99	0.5365	0.4958
DiT4SR*	0.8291	0.6529	15.03	0.5034	0.4349
TSD-SR*	0.8199	0.6277	15.70	0.4490	0.4627
ReFIR	0.8252	0.6518	15.58	0.5457	0.4149
AdaRefSR*	0.8355	0.6504	**16.56**	0.5918	**0.3144**
ImageCritic*	0.8444	0.6755	15.70	0.5854	0.3964
RefGC-SR^2 (Ours)	**0.8742**	**0.7451**	16.32	**0.6009**	0.3196

RefGC-SR^2 Benchmark is 72°. Moreover, 55%, 41%, and 22.5% of the image pairs exhibit rotations greater than 60°, 90°, and 135°, respectively. These statistics indicate that the benchmark already contains substantial viewpoint variation. For a more challenging evaluation, we select the 25% of benchmark samples with the largest viewpoint differences. This subset contains 50 samples satisfying $\Delta \geq 132.5^\circ$. Table 9 presents the quantitative results on this subset.

RefGC-SR^2 achieves the best CLIP-I, DINO, and SSIM scores, together with the second-best PSNR and LPIPS. Compared with its performance on the complete benchmark in Table 2, CLIP-I increases by 0.0046, while DINO decreases by only 0.0023. In addition, its DINO margin over the strongest baseline increases from 0.0309 on the complete benchmark to 0.0696 on the large-viewpoint subset. These results demonstrate that RefGC-SR^2 maintains strong identity preservation and reconstruction quality under substantial viewpoint discrepancies. If the reference depicts a completely different side of the object, appearance information that is entirely absent from the HRRI cannot be faithfully recovered. This limitation is inherent to reference-guided restoration methods.

G.3 Sensitivity to Reference Quality

We evaluate how degradation of the HRRI affects refinement and super-resolution performance. Although the RefGC-SR^2 task assumes a high-quality, high-resolution reference image, references available in practical applications may be degraded. We consider one condition for each of three degradation types: Gaussian blur with $\sigma = 1$, JPEG compression with $q = 15$, and $4\times$ spatial downsampling from 1024×1024 to 256×256.

Quantitative Results We evaluate a randomly selected subset of 100 samples. Each degradation is applied only to the HRRI, while the LRGI remains unchanged. All methods use the same samples, masks, and reference conditions, and none is retrained for the degraded-reference conditions. We report DINO for identity preservation, PSNR for reconstruction fidelity, and LPIPS for perceptual similarity. Values in parentheses in Table 10 indicate the change relative to the clean-reference result of the same model. The results in Table 10 show that HRRI degradation affects performance even when the LRGI is unchanged. For RefGC-SR^2, DINO decreases by 0.001 to 0.040, PSNR decreases by 0.83 to 1.63 dB, and LPIPS increases by 0.026 to 0.054 across the three conditions. These changes indicate that degraded reference information can adversely affect identity preservation, reconstruction fidelity, and perceptual similarity. The magnitude and direction of the changes vary across models and metrics, including a slight LPIPS improvement for AdaRefSR under blur. Our model achieves the highest DINO score under all three degraded conditions, but its performance remains affected by HRRI degradation.

Qualitative Results Fig. 10 presents example outputs obtained with degraded HRRIs under the three evaluated conditions. These examples complement the quantitative comparisons with clean references by showing the resulting object appearance and fine details when degraded reference information is used.



Figure 10: Qualitative comparison under degraded HRRI conditions. Rows (a) and (b) show Gaussian blur with $\sigma = 1$, rows (c) and (d) show $4\times$ downsampling, and rows (e) and (f) show JPEG compression with $q = 15$. Each degradation is applied only to the HRRI, while the LRGI remains unchanged. We compare RefGC-SR^2 with ImageCritic (Ouyang et al., 2025) and AdaRefSR (Wang et al., 2026). The legend on the right explains the arrows. Best viewed zoomed in.

Table 10: Effects of HRRI degradation on a randomly selected subset of 100 samples from the RefGC-SR^2 Benchmark. Only the HRRI is degraded, and no model is retrained. Parentheses show the change relative to the same model under the clean condition ($\Delta = \text{degraded} - \text{clean}$). Negative changes in DINO and PSNR and positive changes in LPIPS indicate worse performance. The best and second-best metric values within each condition are shown in **bold** and with underlining, respectively. An asterisk ($*$) denotes a baseline fine-tuned on the RefGC-SR^2 Dataset.

HRRI condition	Model	DINO↑	PSNR↑	LPIPS↓
Clean	AdaRefSR*	0.692	17.34	0.289
	ImageCritic*	<u>0.742</u>	<u>17.62</u>	<u>0.285</u>
	RefGC-SR^2 (Ours)	**0.751**	**18.54**	**0.256**
Blur	AdaRefSR*	0.688 (-0.004)	<u>17.09</u> (-0.25)	<u>0.287</u> (-0.002)
	ImageCritic*	<u>0.697</u> (-0.045)	16.11 (-1.51)	0.444 ($+0.159$)
	RefGC-SR^2 (Ours)	**0.750** (-0.001)	**17.69** (-0.85)	**0.282** ($+0.026$)
JPEG	AdaRefSR*	0.665 (-0.027)	<u>17.15</u> (-0.19)	<u>0.299</u> ($+0.010$)
	ImageCritic*	<u>0.688</u> (-0.054)	16.13 (-1.49)	0.428 ($+0.143$)
	RefGC-SR^2 (Ours)	**0.729** (-0.022)	**17.71** (-0.83)	**0.283** ($+0.027$)
downsampling	AdaRefSR*	0.684 (-0.008)	**16.98** (-0.36)	**0.290** ($+0.001$)
	ImageCritic*	<u>0.695</u> (-0.047)	16.00 (-1.62)	0.461 ($+0.176$)
	RefGC-SR^2 (Ours)	**0.711** (-0.040)	<u>16.91</u> (-1.63)	<u>0.310</u> ($+0.054$)

Table 11: Evaluation on 100 triplets from independent sources. HRRI–HRGT pairs are collected from sources excluded from the training set and original benchmark, while LRGIs are generated using the frozen DipRefGC model. An $*$ denotes a baseline fine-tuned on the RefGC-SR^2 Dataset. The best results are shown in bold.

Method	CLIP-I↑	DINO↑	PSNR↑	SSIM↑	LPIPS↓
DiT4SR*	0.8311	0.6456	14.32	0.4683	0.4850
TSD-SR*	0.8490	0.7334	16.30	0.4148	0.4835
ReFIR	0.8715	0.7604	16.38	0.5561	0.3897
AdaRefSR*	0.8715	0.7675	17.07	0.5634	0.3366
OmniPaint	0.8580	0.7006	14.24	0.5405	0.4659
ImageCritic*	0.8895	0.7871	16.19	0.5765	0.4022
RefGC-SR^2 (Ours)	**0.8911**	**0.8059**	**17.53**	**0.5956**	**0.3062**

G.4 Evaluation on Independent Sources and RefGC Outputs

Our original benchmark contains non-overlapping HRRI–HRGT pairs with RefGC-SR^2 Dataset, but shares source datasets and the DipRefGC generator with training. We therefore evaluate two additional settings: independent reference pairs with DipRefGC, and independent reference pairs with off-the-shelf RefGC generators (Huang et al., 2025; Song et al., 2026; Zhang et al., 2025; Feng et al., 2026).

Independent Reference Pairs We evaluate 100 triplets constructed from HRRI–HRGT pairs collected from Wikimedia Commons (Wikimedia Foundation, 2026), ABO (Collins et al., 2022), DreamBench++ (Peng et al., 2024), and DreamBooth (Ruiz et al., 2023). These sources are excluded from both our training set and original benchmark. We generate the corresponding LRGIs using the frozen DipRefGC model and follow the evaluation protocol of Table 2. As shown in Table 11, RefGC-SR^2 achieves the best results across all five metrics. This evaluation supports generalization to pairs from unseen sources.

Independent Reference Pairs with Off-the-shelf RefGC Outputs For this evaluation, we select 80 HRRI–HRGT pairs from a 200-pair independent-source pool (Wikimedia Foundation, 2026; Collins et al., 2022; Peng et al., 2024; Ruiz et al., 2023; Openverse Project, 2026). We generate LRGIs using DreamFuse (Huang et al., 2025) and InsertAnything (Song et al., 2026) for compositing,

Table 12: Evaluation on 80 independent-source HRRI–HRGT pairs with off-the-shelf RefGC outputs. Each pair is processed by two generators per setting, yielding 160 compositing and 160 customization inputs. CLIP-I and DINO compare outputs with the HRRI using object bounding-box crops for compositing and full images for customization. Scores are averaged over the pooled outputs within each setting. An asterisk denotes a baseline fine-tuned on the RefGC-SR^2 Dataset. The best results are shown in bold.

	Compositing		Customization	
Method	CLIP-I↑	DINO↑	CLIP-I↑	DINO↑
DiT4SR*	0.8324	0.6653	0.8085	0.6173
TSD-SR*	0.8547	0.7483	0.8196	0.6543
ReFIR	0.8700	0.7671	0.8359	0.6633
AdaRefSR*	0.8708	0.7634	0.8346	0.6617
OmniPaint	0.8376	0.6836	0.8269	0.6604
ImageCritic*	0.8921	0.8024	0.8605	0.7140
RefGC-SR^2 (Ours)	**0.8928**	**0.8056**	**0.8610**	**0.7168**

and FreeCus (Zhang et al., 2025) and PersonalizeAnything (Feng et al., 2026) for customization. Each generator processes all 80 pairs, yielding 160 inputs per setting. Both the source datasets and input generators differ from those used to construct our training triplets.

Because these generators do not enforce pose alignment with the HRGT, we report CLIP-I (Radford et al., 2021) and DINO (Oquab et al., 2023) similarity to the HRRI rather than pixel-aligned reconstruction metrics. For compositing, we use separate object bounding-box crops for the output and HRRI, with a 10% margin on each side. The output crop is defined by the HRGT object mask, while the reference crop is defined by the HRRI mask; background pixels within the crops are retained. For customization, we compare the full output and HRRI images because the generated layout may differ from the target. Within each setting, we report the mean of the per-image scores over all 160 outputs.

Table 12 shows that RefGC-SR^2 achieves the highest CLIP-I and DINO scores in both settings. The margins over ImageCritic* (Ouyang et al., 2025), which shares the same backbone (Labs et al., 2025) with ours, are modest: 0.0007 CLIP-I and 0.0032 DINO for compositing, and 0.0005 CLIP-I and 0.0028 DINO for customization. These results demonstrate that RefGC-SR^2 generalizes well beyond the training distribution, including both unseen reference sources and inputs generated by off-the-shelf RefGC models.

H Multi-Object Evaluation

Benchmark Construction To evaluate generalization beyond single-object scenes, we construct a multi-object benchmark using EntitySeg (Qi et al., 2023). We select 36 real images containing three to five distinct objects and use each original image as the HRGT. For each HRGT, we generate an HRRI using Gemini 3.1 Flash Image (Google DeepMind, 2026) by varying the viewpoint and scene context while preserving the identities of all objects. We manually inspect the generated pairs and exclude cases in which the HRRI deviates substantially from the HRGT or fails to preserve the object identities. The corresponding LRGI is synthesized using DipRefGC following the same procedure used for our main dataset.

Evaluation Protocol All models are evaluated without retraining or finetuning on multi-object data, following the same evaluation protocol as Table 2.

Quantitative Result As shown in Table 13, the RefGC-SR^2 model achieves state-of-the-art performance across all five metrics, demonstrating strong generalization to scenes containing multiple object instances without multi-object training.

Qualitative Results Fig. 11 provides additional qualitative comparisons under the multi-object setting. As shown in Fig. 11(a), when refining a scene containing multiple spoons, the competing methods fail to recover the black appearance of the two leftmost spoons, whereas our method

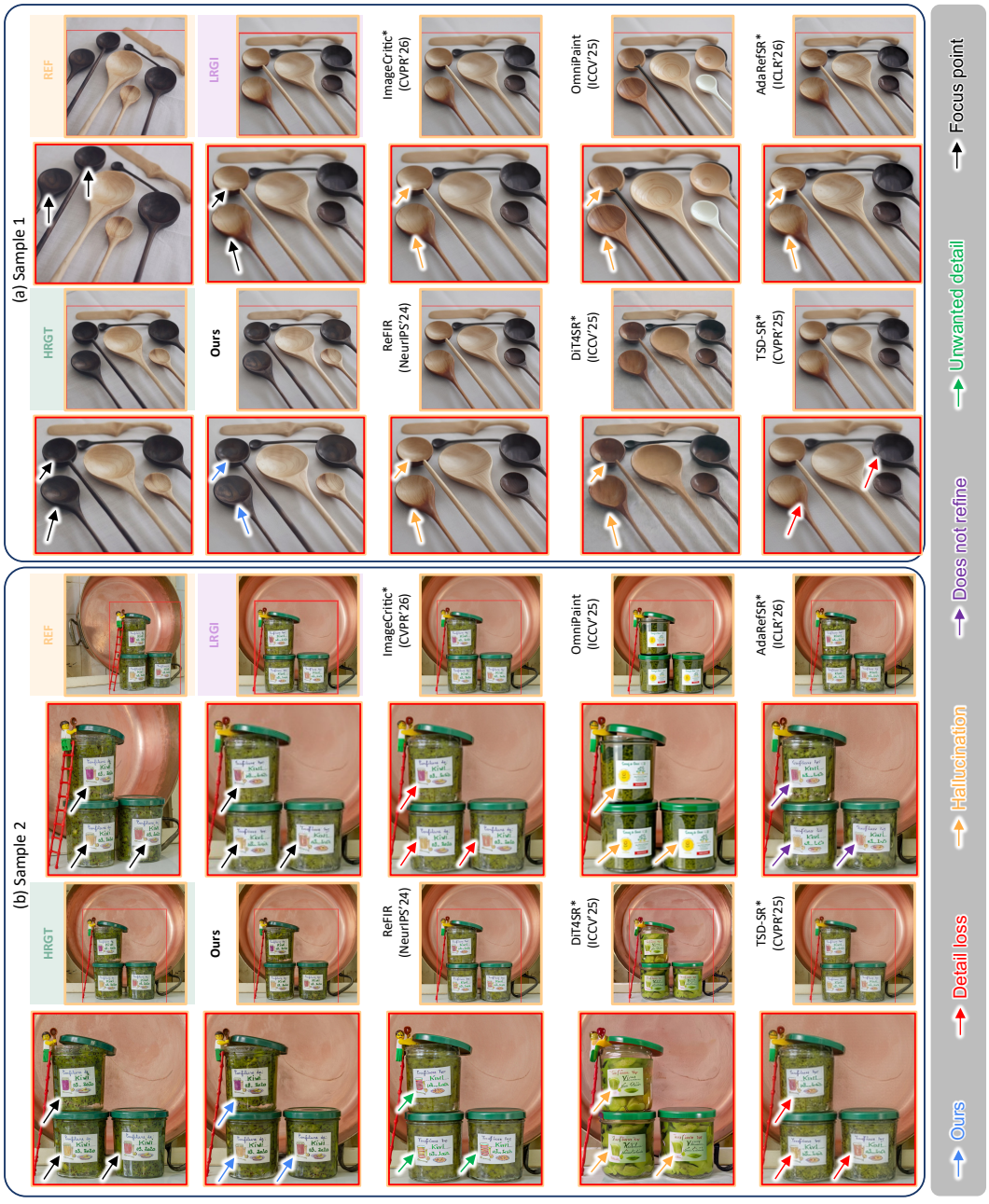


Figure 11: Qualitative comparison in the multi-object setting. We compare our RefGC-SR^2 against ImageCritic (Ouyang et al., 2025), OmniPaint (Yu et al., 2025), AdaRefSR (Wang et al., 2026), ReFIR (Guo et al., 2024), TSD-SR (Dong et al., 2025), and DiT4SR (Duan et al., 2025). Please refer to the legend below the figure for the meaning of each arrow. Best viewed zoomed in.

Table 13: Quantitative comparison on the multi-object benchmark with 36 scenes containing three to five objects each. All methods are evaluated without multi-object training. The best and second-best results are shown in **bold** and underline, respectively. An asterisk (∗) indicates a baseline finetuned on the RefGC-SR^2 Dataset.

Model	CLIP-I	DINO	PSNR	SSIM	LPIPS
DiT4SR*	0.8920	0.7668	16.0412	0.5940	0.3617
TSD-SR*	0.8996	0.8031	16.7320	0.5759	0.3876
ReFIR	0.9280	0.8483	15.7616	0.5494	0.3422
AdaRefSR*	0.9234	0.8336	16.9844	0.5992	0.2889
OmniPaint	0.8905	0.7722	14.6862	0.5493	0.5020
ImageCritic*	0.9188	0.8424	16.2762	0.5931	0.4158
RefGC-SR^2 (Ours)	**0.9378**	**0.8720**	**18.6452**	**0.6416**	**0.2194**

successfully restores their reference-consistent appearance. In Fig. 11(b), the competing methods introduce hallucinated content and unwanted details. In contrast, the proposed RefGC-SR^2 restores the multiple objects without introducing these artifacts.

I Qualitative Results on Outputs from Commercial and Open-Source Models

Fig. 12 provides a qualitative comparison across seven models, including our RefGC-SR^2 Model, when the outputs of Gemini 2.5 Flash Image are treated as LRGIs. Similarly, Fig. 13 reports the qualitative comparison when the outputs of GPT-Image 1.5 are treated as LRGIs, and Fig. 14 presents the qualitative comparison when the outputs of Qwen-Image-Edit are treated as LRGIs. As shown in Figs. 12 - 14, our proposed RefGC-SR^2 faithfully restores fine details and enhances resolution while simultaneously removing generative artifacts, whereas the competing models commonly suffer from either hallucinating undesired details or failing to recover fine details. These results demonstrate that our proposed RefGC-SR^2 Model generalizes well even to image pairs drawn from distributions different from those seen during training.

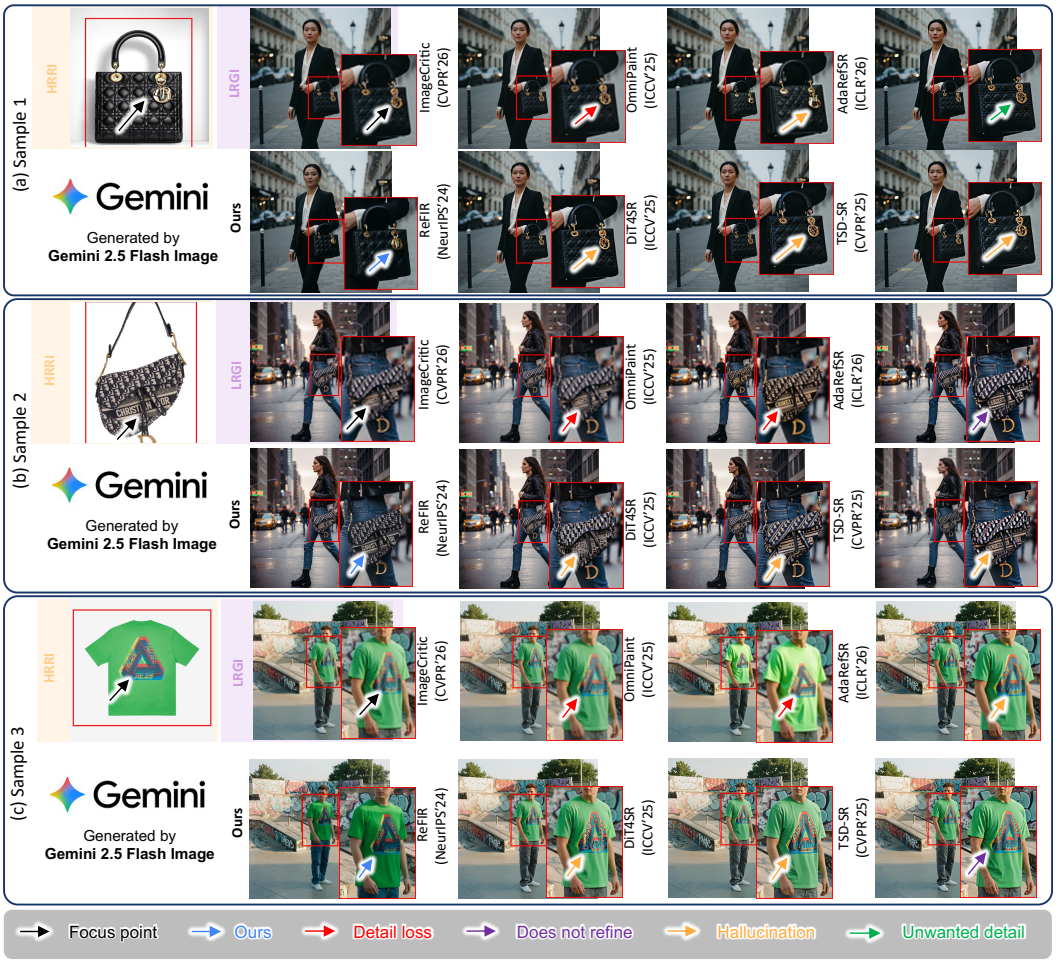


Figure 12: Qualitative comparison on LRGIs generated by the commercial model Gemini 2.5 Flash Image (Google DeepMind, 2025). We compare our RefGC-SR^2 against ImageCritic (Ouyang et al., 2025), OmniPaint (Yu et al., 2025), AdaRefSR (Wang et al., 2026), ReFIR (Guo et al., 2024), TSD-SR (Dong et al., 2025), and DiT4SR (Duan et al., 2025). Please refer to the legend below the figure for the meaning of each arrow. Best viewed zoomed in.



Figure 13: Qualitative comparison on LRGIs generated by the commercial model GPT-Image 1.5 (OpenAI, 2025). We compare our RefGC-SR^2 against ImageCritic (Ouyang et al., 2025), OmniPaint (Yu et al., 2025), AdaRefSR (Wang et al., 2026), ReFIR (Guo et al., 2024), TSD-SR (Dong et al., 2025), and DiT4SR (Duan et al., 2025). Please refer to the legend below the figure for the meaning of each arrow. Best viewed zoomed in.

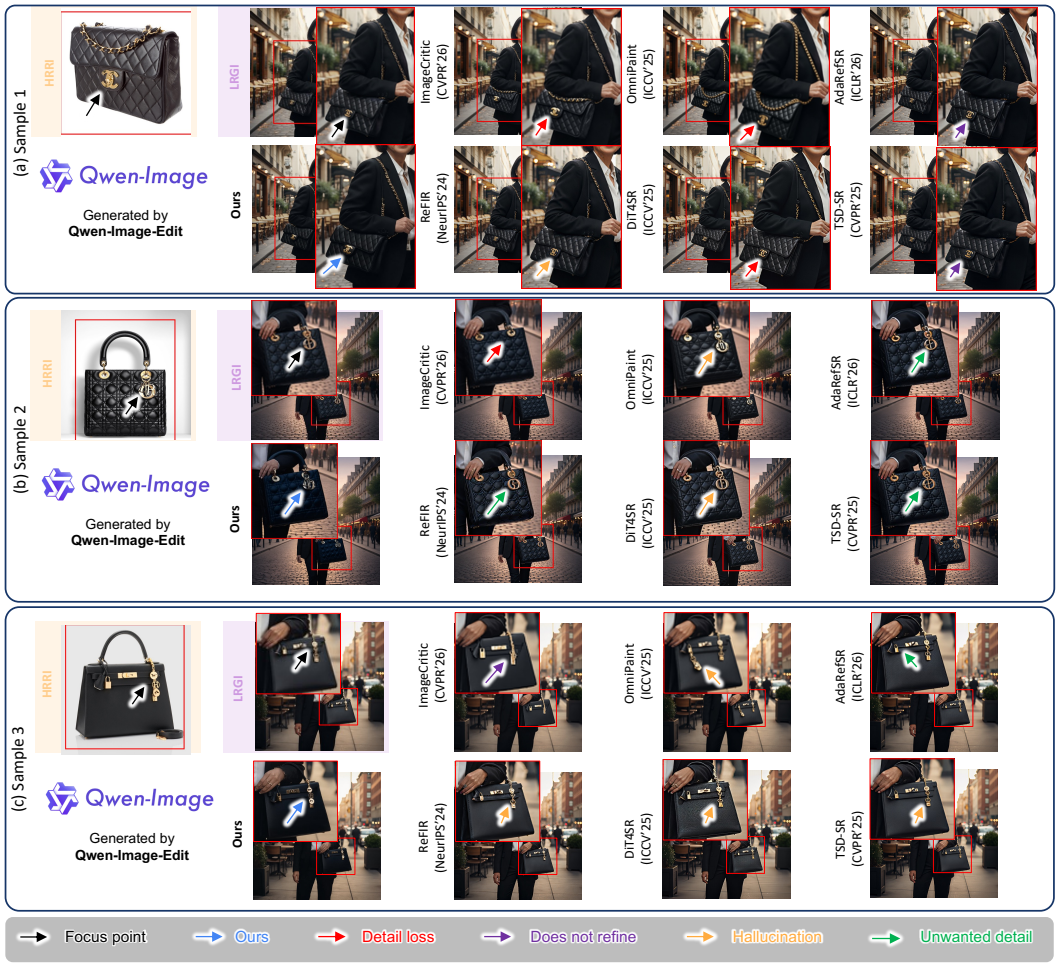


Figure 14: Qualitative comparison on LRGIs generated by the open-source model Qwen-Image-Edit (Wu et al., 2025a). We compare our RefGC-SR^2 against ImageCritic (Ouyang et al., 2025), OmniPaint (Yu et al., 2025), AdaRefSR (Wang et al., 2026), ReFIR (Guo et al., 2024), TSD-SR (Dong et al., 2025), and DiT4SR (Duan et al., 2025). Please refer to the legend below the figure for the meaning of each arrow. Best viewed zoomed in.

J User Study Protocols and Analysis



Figure 15: User study questionnaire interface (Google Forms). Left: task instructions and three evaluation criteria. Center: a sample question showing HRRI, HRGT, LRGI, and four anonymized model outputs ("Model 1"–"Model 4") with zoomed-in object crops. Right: 4-point Likert scale rating for refine/detail/overall quality.

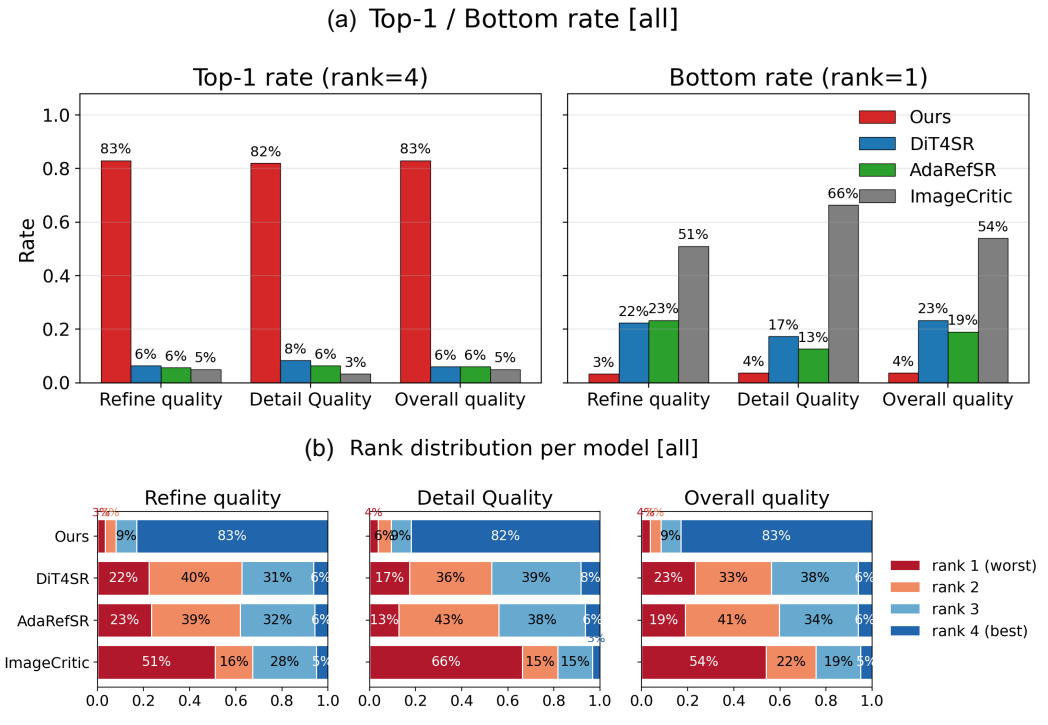


Figure 16: Detailed user study results on the In-the-wild Benchmark. (a) Top-1 (rank=4) and bottom (rank=1) rates per model across three evaluation criteria. (b) Full rank distribution per model. Our model dominates the best rank (82–83%) and is rarely ranked worst (3–4%), while ImageCritic incurs the highest worst-rank rates (51%/66%/54%) due to its oversmoothing tendency.

J.1 PROTOCOL

We provide additional details on the user study reported in Sec. 4-**User Study**. The study was conducted as a blind, side-by-side comparison via a web-based questionnaire (Google Forms) and took approximately 20 minutes per participant. Participation was voluntary, with no monetary compensation; participants were informed about the purpose and approximate duration of the study before consenting to participate, and could withdraw at any time. No personally identifiable information was collected.

Question Composition The questionnaire consisted of 20 questions, each corresponding to one in-the-wild test case. As shown in Fig. 15, each question presents three reference images at the top (HRRI, HRGT, and LRGI) followed by four model outputs anonymized as "Model 1"–"Model 4". The four outputs correspond to one method from each of SR (DiT4SR (Duan et al., 2025)), RefSR (AdaRefSR (Wang et al., 2026)), RefGCR (ImageCritic (Ouyang et al., 2025)), and our RefGC-SR^2 Model. To facilitate fine-grained comparison, every image is accompanied by a zoomed-in crop of the object region, with the relevant region highlighted by a red bounding box.

Evaluation Criteria For each question, participants rate all four model outputs on three independent criteria using a 4-point Likert scale (1: Very Poor, 2, 3, 4: Excellent): (i) *Refine quality*: how well the model removes undesired structures or artifacts that deviate from the HRGT or HRRI; (ii) *Detail quality*: how effectively the model restores fine details from the HRGT and HRRI when enhancing the low-resolution image; (iii) *Overall quality*: the overall visual quality of the image, with emphasis on alignment with the HRGT and HRRI. The full text of the instructions provided to participants is shown on the left panel of Fig. 15.

Scoring From the collected ratings, we compute two complementary statistics: (i) *top-1 rate*, defined as the proportion of (participant, question) pairs in which the model received the highest score (rank 4) among the four candidates, and (ii) *bottom rate*, defined analogously for rank 1. The aggregated results across 16 participants $\times$ 20 questions are reported in the main paper (Fig. 6) and in further detail in Fig. 16.

J.2 DETAILED ANALYSIS

Fig. 16 provides two complementary views of the results: (a) top-1 and bottom rates per model under each criterion, and (b) the full rank distribution per model.

Overall Preference As reported in the main paper, our model achieves a top-1 rate of 82–83% across all three criteria, while baselines remain below 9% (Fig. 16-(a), left). Symmetrically, our model is rarely ranked worst (3–4%), in contrast to baselines that incur substantially higher worst-rank rates (Fig. 16-(a), right). The full rank distribution in Fig. 16-(b) further confirms that our model dominates rank 4 (best) across all criteria, while DiT4SR and AdaRefSR are predominantly distributed across ranks 2 and 3, indicating mid-tier performance.

Discrepancy with Quantitative Metrics for ImageCritic A particularly notable observation is the contrast between ImageCritic's quantitative and human evaluation results. While ImageCritic ranks second-best on the RefGC-SR^2 Benchmark in Table 2, Fig. 16-(b) shows that it receives the worst rank in 51%, 66%, and 54% of cases for refine, detail, and overall quality, respectively, the worst among all compared models. We attribute this discrepancy to ImageCritic's tendency to produce *oversmoothed outputs*: while smoothing tends to align well with HRGT in pixel-level metrics (PSNR, SSIM) and feature-level similarity (CLIP-I, DINO), human evaluators perceive the resulting loss of fine-grained texture and high-frequency detail as a clear quality degradation. This finding underscores the importance of complementing quantitative metrics with human evaluation for tasks involving fine detail recovery, and supports the design of our frequency-aware FreqMoLE module, which is explicitly aimed at preserving high-frequency information from the HRRI.

K LIMITATIONS

Although trained only on object-centric single-object data, RefGC-SR^2 achieves the best performance across all five metrics on the multi-object benchmark (Appendix H, Table 13). Nevertheless, its outputs are not always free of artifacts. In Fig. 17, competing methods introduce errors in the

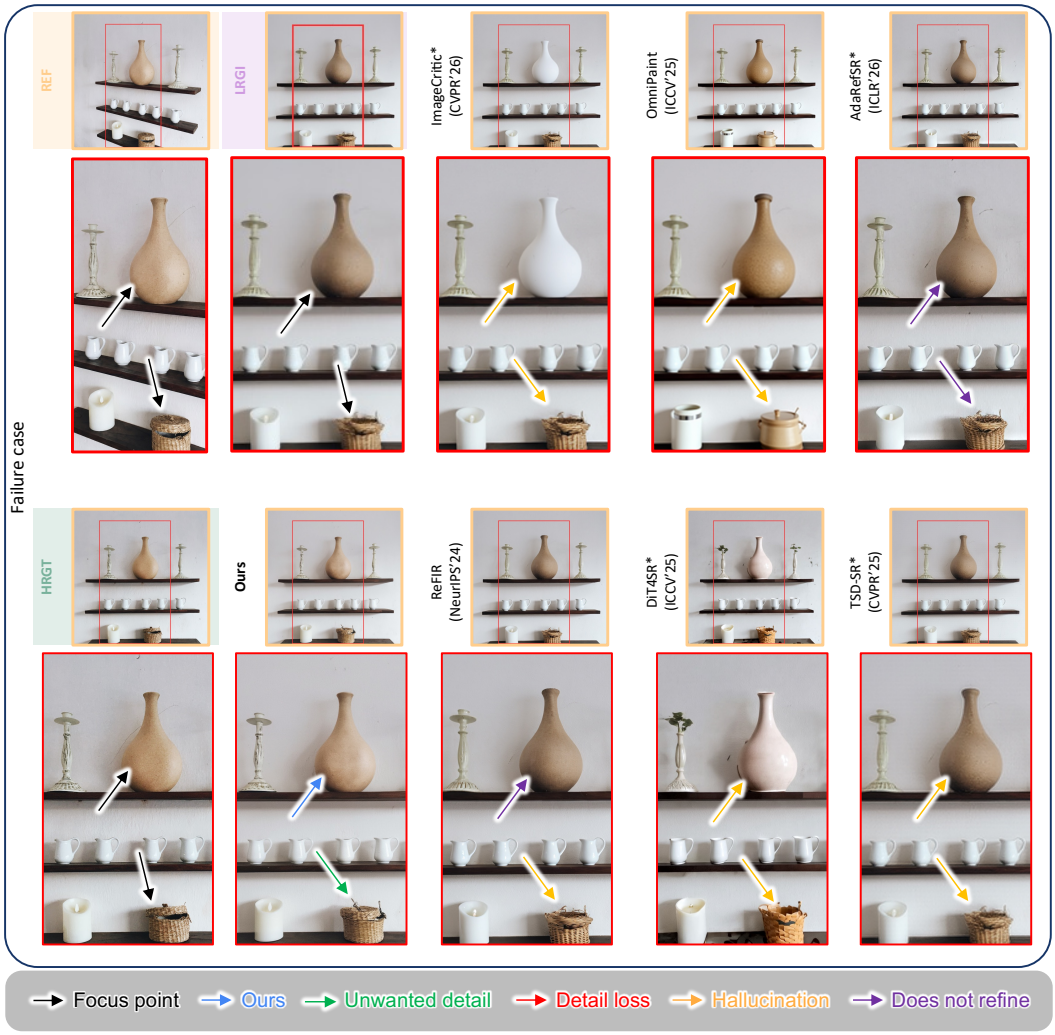


Figure 17: Failure case in a multi-object scene. Our RefGC-SR^2 model restores the reference-consistent appearance of the vase on the upper shelf but still introduces unwanted details around the basket on the lower shelf, illustrating its limitations in multi-object refinement.

clay vase on the upper shelf, the basket on the lower shelf, or both. Our model restores the vase's reference-consistent appearance but still produces unwanted details around the basket. Thus, while RefGC-SR^2 performs better overall, it can still produce local artifacts in multi-object scenes. We leave the construction of a multi-object dataset and subsequent training of RefGC-SR^2 on that dataset to future work.