UNITELE: LEARNING DEVICE-SPECIFIC DEGRADATIONS FOR UNIFIED SMARTPHONE TELEPHOTO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Super-telephoto imaging on mobile phones suffers from complex, device-dependent degradations that are poorly captured by conventional handcrafted pipelines. Collecting large-scale aligned training pairs across different phones and diverse scenes, however, is both costly and difficult to scale. We therefore study whether device-specific degradations can be learned from only a small number of physical captures and transferred for training a unified super-resolution (SR) model. Our key idea is to factor degradation learning into two learnable components: one captures the device-specific imaging characteristics under well-resolved conditions, while the other captures the additional information loss and artifacts that arise in super-telephoto imaging. We build a paired benchmark across three phone brands (Huawei, Apple, and vivo) by photographing the same content at two display sizes and zoom settings. The larger displayed image is captured at the phone's maximum optical zoom setting, while the smaller image is photographed directly at the phone's highest native zoom setting through the phone's native imaging pipeline. The benchmark contains 519 training and 60 held-out pairs. We use the benchmark to learn device-specific degradation, then apply it to external images to synthesize thousands of additional pairs to train our UniTele SR model. Relative to the pretrained Qwen and FLUX backbones before adaptation, the full UniTele training recipe substantially improves paired restoration fidelity and perceptual similarity on the held-out benchmark and yields markedly more stable and visually faithful restorations.

1 INTRODUCTION

Recovering fine detail beyond the optical zoom range is a central challenge in smartphone super-telephoto imaging. Generative super-resolution (SR) offers a promising approach by using learned image priors to reconstruct detail from severely degraded inputs (Wang et al., 2024; Wu et al., 2024b). However, visually plausible detail is not necessarily faithful to the photographed scene: ambiguous observations can lead to incorrect textures or structures. Adapting these models to smartphone telephoto restoration therefore calls for supervision that connects real degraded observations to their corresponding reference detail.

Obtaining such supervision is difficult because telephoto degradations are device dependent. The same content can exhibit different color shifts, blur, and texture loss across phones (Figure 2). Paired captures provide direct examples of these differences, but collecting and aligning them across devices and diverse scenes is costly (Cai et al., 2019; Wei et al., 2020; Chen et al., 2019). Handcrafted degradation pipelines readily generate pairs from large image collections, yet their predefined operators may not adequately capture the target device's imaging characteristics (Zhang et al., 2021; Wang et al., 2021). Learned degradation models offer a way to synthesize more realistic training data (Fritsche et al., 2019; Wolf et al., 2021). Building on this direction, we investigate how limited physical captures can support device-specific degradation synthesis on new content, extending the supervision available for generative telephoto restoration.

We propose UniTele, a framework that learns device-specific degradations from a small capture set and transfers them to external images for SR training. Our key idea is to learn the device's imaging

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

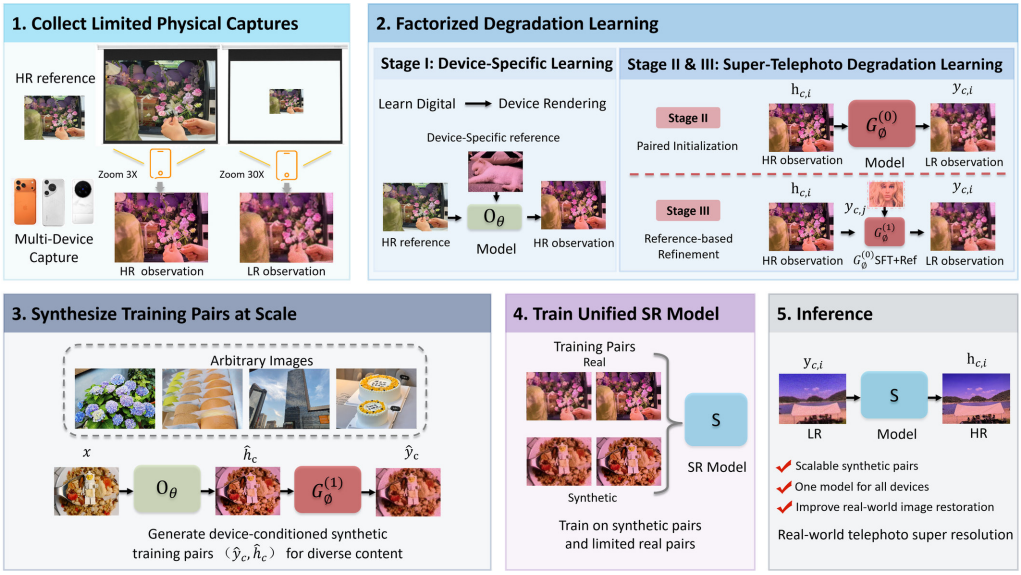

Figure 1: Overview of UniTele. Two learned mappings synthesize low-resolution (LR) images for SR training, using their corresponding synthetic high-resolution (HR) observations as targets.

characteristics and the additional degradation associated with super-telephoto imaging separately. The first mapping learns the device's imaging characteristics under well-resolved conditions; the second learns the additional detail loss and artifacts in the corresponding super-telephoto input.

To learn these mappings, we construct a controlled paired benchmark using Huawei, Apple, and vivo phones. We denote the original image displayed on the monitor as the *high-resolution reference*. Each image is photographed at two display sizes and zoom settings. The larger image is captured at the phone's maximum optical zoom setting to obtain the *high-resolution observation*; the smaller image is captured directly at the phone's highest native zoom setting to obtain the *low-resolution observation*. The benchmark contains 519 training pairs and 60 held-out pairs. The first mapping is learned from high-resolution references to high-resolution observations, and the second from high-resolution observations to low-resolution observations. We then apply the learned mappings to external images to synthesize thousands of additional pairs. These synthetic pairs are combined with real pairs to adapt pretrained generative models to telephoto SR.

Experiments on the held-out paired benchmark demonstrate improved restoration fidelity and perceptual similarity across both Qwen and FLUX backbones. Qualitative comparisons further illustrate improvements in the preservation of scene structure.

Our contributions are threefold:

- We propose a device-conditioned degradation framework that separately learns device-specific imaging characteristics and super-telephoto degradation from limited captures.
- We construct a controlled smartphone super-telephoto benchmark spanning Huawei, Apple, and vivo to characterize device-dependent degradations and evaluate restoration across different phones.
- We demonstrate the effectiveness of UniTele on Qwen and FLUX backbones, with improvements in reconstruction fidelity, perceptual similarity, and visual quality.

2 Related Work

Real-world super-resolution and optical zoom. Classical single-image super-resolution (SISR) generates LR observations through spatially invariant blur and bicubic downsampling of HR images, omitting coupled optics, sensor noise, image signal processing, sharpening, compression, and spatially varying blur. RealSR captures scenes at different focal lengths and progressively registers

pairs for supervision (Cai et al., 2019). DRealSR diversifies real degradations and identifies distinct restoration challenges in smooth regions, edges, and textures (Wei et al., 2020). Camera Lens Super-Resolution (CameraSR) studies the lens resolution–field-of-view trade-off, showing that realistic lens pairs better support digital zoom than synthetic downsampling (Chen et al., 2019). Mobile systems also align and adaptively fuse concurrent wide and telephoto captures to improve intermediate zoom (Wu et al., 2023). These studies highlight real optical observations and paired acquisition. Unlike fixed-scale SISR, our aligned inputs and targets may share a canvas but differ in effective optical detail and camera processing. Controlled, calibration-assisted captures supervise a forward model that transfers measured device characteristics to digital content.

Synthetic and learned degradation modeling. Blind SR methods enrich forward degradation to narrow the synthetic-to-real gap. BSRGAN randomly orders blur, resampling, sensor-inspired noise, and JPEG compression (Zhang et al., 2021); Real-ESRGAN adds high-order degradation and models ringing and overshoot (Wang et al., 2021). These efficient, scalable pipelines remain limited by predefined operators and manually chosen parameter ranges. Learned models relax these assumptions. DSGAN uses frequency separation and adversarial learning to transfer real-image characteristics to synthetically downsampled images while preserving low-frequency content (Fritsche et al., 2019). DeFlow learns stochastic conditional degradations in a shared latent space from unpaired clean and degraded marginals using conditional normalizing flows (Wolf et al., 2021). These methods typically learn one aggregate degradation distribution. Mobile telephoto degradations instead vary by phone, lens, zoom level, and ISP in tone, blur, sharpening, compression, and residual texture for the same scene. We condition one generator's output distribution on device identity and a same-device telephoto degradation reference, enabling explicit device switching without separate models.

Generative priors and conditional image restoration. Large pretrained generative models provide strong priors for real-world restoration. StableSR adapts pretrained text-to-image diffusion to blind SR, using controllable feature wrapping to balance perceptual quality and fidelity (Wang et al., 2024). SeeSR extracts degradation-aware semantic prompts to reduce semantic errors under severe degradation (Wu et al., 2024b). The One-Step Effective Diffusion Network (OSEDiff) distills restoration into one diffusion step for efficient inference (Wu et al., 2024a); Adversarial Diffusion Compression for Real-World Image Super-Resolution (AdcSR) further compresses it through structural pruning and adversarial distillation (Chen et al., 2025). All-in-one methods select restoration behaviors through learned prompts or language instructions: PromptIR learns degradation-specific prompts for a unified network (Potlapalli et al., 2023), while InstructIR uses human-written instructions (Conde et al., 2024). These generative or conditional priors primarily map unknown degraded images to clean outputs. Complementarily, we adapt the image-editing foundation model Qwen-Image-Edit-2511 (Wu et al., 2025; Qwen, 2025) in the forward direction, mapping digital content to phone-specific synthetic LR images.

3 METHOD

We train a unified SR model using limited paired captures and external images without phone observations. UniTele learns device-specific degradations from the captures and applies them to external content to expand paired supervision. Figure 1 shows controlled acquisition, two learned mappings for pair synthesis, and joint training on real and synthetic pairs.

3.1 MULTI-DEVICE PAIRED ACQUISITION

Learning device-specific degradations requires reliably corresponding pairs, but well-resolved counterparts to in-the-wild super-telephoto images are difficult to obtain. We therefore vary an image's display size to simulate the same subject at different distances: smaller images represent distant subjects, while larger images provide closer views.

Let x denote the original digital image, or *high-resolution (HR) reference*. At a fixed camera–monitor distance, device c captures the larger displayed image at its maximum optical zoom to obtain the *HR observation* h_c, and directly captures the smaller image at its highest native zoom to obtain the *low-resolution (LR) observation* y_c. Display sizes approximately match the subject's

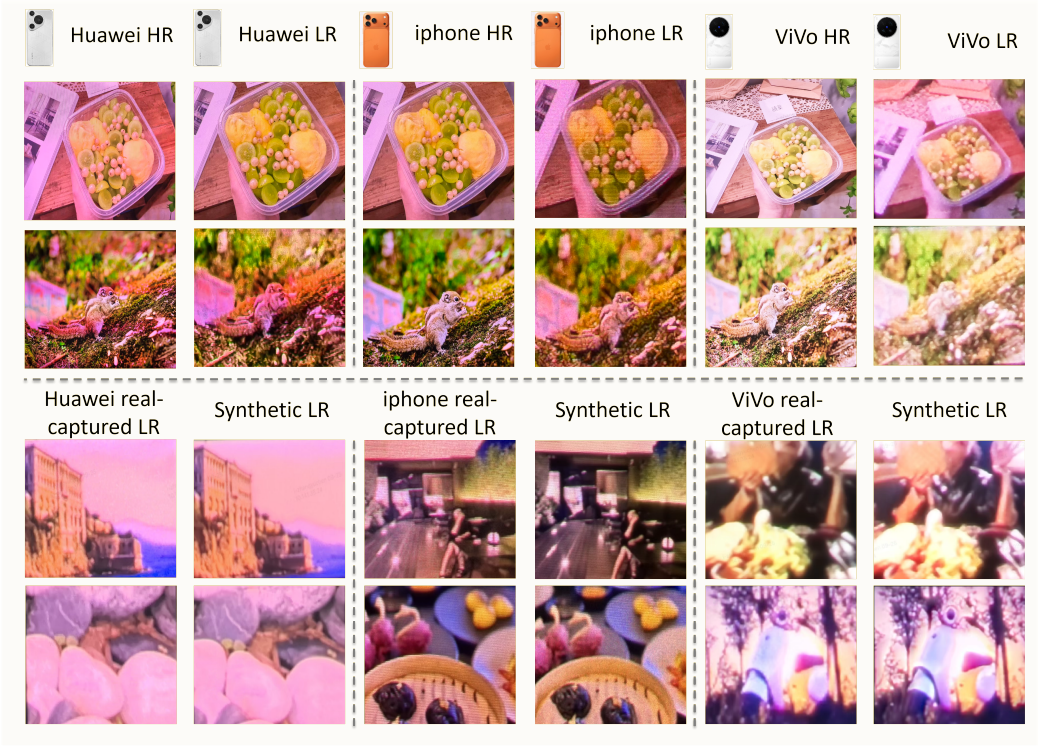

Figure 2: Paired HR (left) and LR (right) observations of the same content across Huawei, iPhone, and vivo phones.

fraction of the frame: y_c represents a distant subject at maximum native zoom, while h_c provides a well-resolved counterpart. Calibration-board captures at both zoom settings provide corresponding corners for homography-based alignment. The (x, h_c) pairs supervise well-resolved device imaging characteristics; (h_c, y_c) pairs supervise additional super-telephoto detail loss and artifacts.

The benchmark includes 519 training pairs and 60 test pairs from Huawei Pura 70 Pro+, iPhone 17 Pro Max, and vivo X300 Ultra, with 20 test pairs per device.

3.2 Factorized Degradation Learning

The goal of factorized degradation learning is to synthesize device-specific HR and LR observations from an HR reference. We formulate this process as two mappings:

$$\hat{h}_c = O_\theta(x, c, r_c^{\mathrm{HR}}), \qquad \hat{y}_c = G_\phi(\hat{h}_c, c, r_c^{\mathrm{LR}}). \tag{1}$$

Here $\hat{h}_c$ and $\hat{y}_c$ are synthetic HR and LR observations. For each device, a fixed HR–LR image pair $(r_c^{\mathrm{HR}}, r_c^{\mathrm{LR}})$ provides references for the device-specific imaging style and super-telephoto degradation style, respectively. The two mappings are trained in three stages (Figure 1).

Stage I (Device-Specific Learning) learns the phone's color response and texture rendering by training O_θ on $x \to h_c$ pairs, with x and r_c^{HR} as image inputs. Here c denotes the device name (Huawei, Apple, or vivo), which is provided through Qwen's text prompt. The same prompt instructs the model to preserve the scene content, geometry, and text in x. Stage II (Paired Initialization) establishes the HR-to-LR mapping for super-telephoto detail loss and artifacts: we initialize G_ϕ from the pretrained model and train it on $(h_c, c) \to y_c$ pairs without an LR style reference. Stage III (Reference-based Refinement) adds direct visual conditioning for the phone's degradation style by continuing to train the same G_ϕ adapter with r_c^{LR} as an additional image input, retaining the paired supervision.

We train one LoRA adapter for O_θ and another for G_ϕ on Qwen-Image-Edit-2511 (Qwen, 2025), keeping the backbone weights frozen and using the standard conditional flow-matching objective.

Each adapter is shared across devices, with balanced device sampling during training. Evaluation and synthesis use a fixed seed.

3.3 Super-Resolution Training with Real and Synthetic Data

Applying Equation 1 to external HR references x yields synthetic SR pairs $(\hat{y}_c, \hat{h}_c)$, with synthetic HR observations as targets, without additional captures. These supplement captured pairs (y_c, h_c), broadening content coverage while retaining real degradations. We LoRA-adapt pretrained Qwen and FLUX models to recover HR observations from LR inputs, first on captured pairs and then on combined real and synthetic data across devices. The Stage I ablation (Table 4) supports two-step synthesis: directly mapping x to $\hat{y}_c$ with G_ϕ alone substantially reduces synthesis fidelity.

4 Experiments

We examine real–synthetic training across SR backbones, in-the-wild transfer, the forward-model stages, and cross-device generalization.

4.1 Experimental Setup

Experiments use 519 captured training pairs, 60 paired test images (Section 3.1), and a separate in-the-wild set. We evaluate Qwen-Image-Edit-2511 (Qwen SR) (Qwen, 2025) and FLUX.2 [klein] 4B Base (FLUX.2 SR) (Black Forest Labs, 2026). We compare pretrained backbones with LoRA adapters trained on real, synthetic, or mixed pairs; the mixed models are UniTele (Qwen) and UniTele (FLUX). The Real-ESRGAN baseline trains Qwen SR on pairs synthesized with its degradation pipeline (Wang et al., 2021). Official pretrained AdcSR (Chen et al., 2025) and OSEDiff (Wu et al., 2024a) provide further comparisons.

Qwen SR training. Qwen uses conditional flow-matching with timestep-weighted MSE. LoRA (rank 32, $\alpha = 32$, dropout 0) targets attention Q/K/V and output projections, added Q/K/V and output projections, image and text MLP output projections, and modulation layers. AdamW uses constant learning rates of 10^{-4} and 5×10^{-5} for the captured- and mixed-data phases, respectively, with effective batch size 8. Dynamic resolution is capped at $786{,}432$ pixels (typically 1024×768 landscape or 768×1024 portrait). The two phases run for 3,055 and 2,340 updates, taking approximately 4.08 and 3.13 hours on 8 Ascend 910B2C NPUs.

FLUX SR training. FLUX uses conditional flow-matching with timestep-weighted MSE. LoRA (rank 32, $\alpha = 32$, dropout 0) targets dual-stream attention projections, added Q/K/V projections, input and output projections, joint attention–MLP projections, and attention output projections in 20 single-transformer blocks. AdamW uses a constant 10^{-4} learning rate and effective batch size 8. Dynamic resolution is capped at $262{,}144$ pixels (typically about 576×432). The captured- and mixed-data phases run for 2,466 and 2,531 updates, taking approximately 0.57 and 0.55 hours on 8 Ascend 910B2C NPUs.

For full-reference evaluation, we report RGB PSNR, SSIM, MAE, LPIPS (Zhang et al., 2018), and the predicted-to-target high-frequency (HF) energy ratio. We report per-device scores and their arithmetic mean across devices; an HF ratio closer to one indicates better agreement in high-frequency energy. For Human Preference (HP), data-evaluation experts compare each output with its aligned GT and assign a 1–10 score based on visual similarity, content consistency, and color shift; we report the mean score. For in-the-wild images without paired targets, we report MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023), and NIQE (Mittal et al., 2013), alongside qualitative comparisons.

4.2 Downstream Super-Resolution Evaluation

Table 1 compares untrained bases, real-only and synthetic-only training, and two real–synthetic sampling ratios, alongside Qwen trained on Real-ESRGAN degradation data and pretrained AdcSR and OSEDiff. UniTele (Qwen) at 1:2 achieves the best PSNR, SSIM, MAE, and LPIPS, gaining 0.49 dB PSNR and reducing LPIPS by 0.0109 over real-only training. UniTele (FLUX) at 2:1 also improves over real-only training. These comparisons reflect complete training recipes. Figure 3 juxtaposes

Table 1: SR performance on 60 paired test images. Ratios denote real-to-synthetic sampling; dashes indicate unavailable HP scores. AdcSR and OSEDiff use official pretrained models; Real-ESRGAN data use its degradation pipeline.

Model / training	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	HF ratio → 1	HP ↑
Input (unrestored)	16.99	0.5234	27.10	0.5704	0.680	–
AdcSR	16.90	0.5673	27.03	0.4383	0.698	2.623
OSEDiff	16.92	0.5810	26.90	0.4617	0.641	–
Qwen SR						
Base (untrained)	11.66	0.4931	54.16	0.6618	0.471	1.459
Real-ESRGAN data	16.90	0.5297	27.39	0.5288	0.740	2.262
Real-world only	18.84	0.6098	20.43	0.3187	0.740	5.180
Learned synthetic only	18.05	0.5976	22.05	0.3438	0.684	4.689
UniTele (Qwen), 1:2	**19.33**	**0.6186**	**18.58**	**0.3078**	0.729	5.689
UniTele (Qwen), 2:1	19.32	0.6174	18.81	0.3080	0.749	5.656
FLUX.2 SR						
Base (untrained)	10.59	0.3638	62.55	0.6102	1.573	3.475
Real-world only	16.53	0.5643	26.84	0.4254	0.761	5.951
Learned synthetic only	17.56	0.5810	23.73	0.3958	0.690	5.623
UniTele (FLUX), 1:2	17.56	0.6041	24.91	0.3735	0.615	6.066
UniTele (FLUX), 2:1	18.40	0.6103	21.54	0.3590	0.657	6.295



Figure 3: Qualitative SR comparison on the three-phone paired test set. GT denotes the aligned real-world HR image.

inputs, selected outputs, and real-world HR targets, showing recovered details and failures such as altered scene structure.

4.3 Generalization to In-the-Wild Images

We evaluate transfer from screen-based training to in-the-wild scenes on 16 iPhone LR images without paired HR targets, using fixed checkpoints without further adaptation. Both UniTele variants use a real-to-synthetic sampling ratio of 1:2; GPT Image 2 (OpenAI, 2026) provides an additional

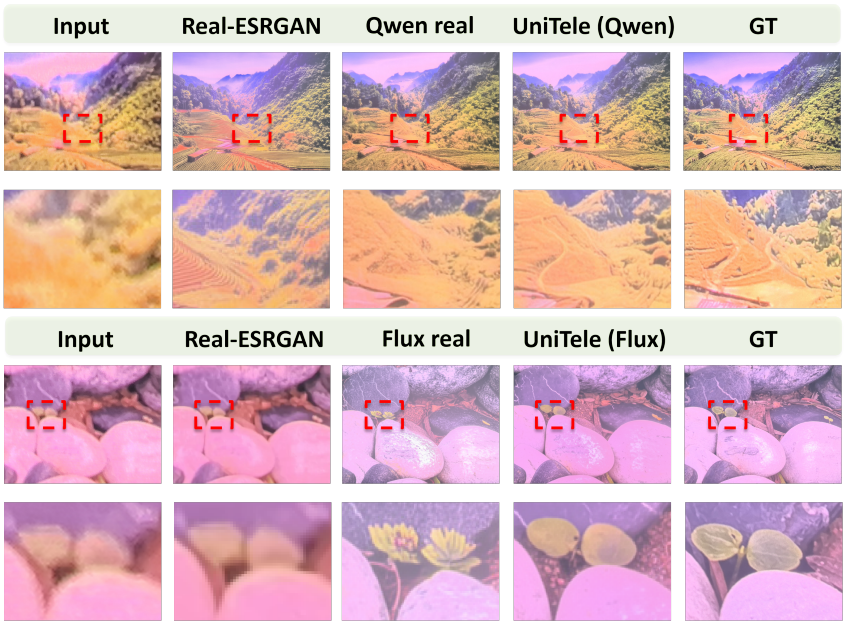

Figure 4: Effect of the real-to-synthetic sampling ratio on UniTele (Qwen) and UniTele (FLUX).

comparison. We report mean no-reference scores at 1024×768, resizing FLUX.2 outputs from their native working resolution of 576×432 before evaluation.

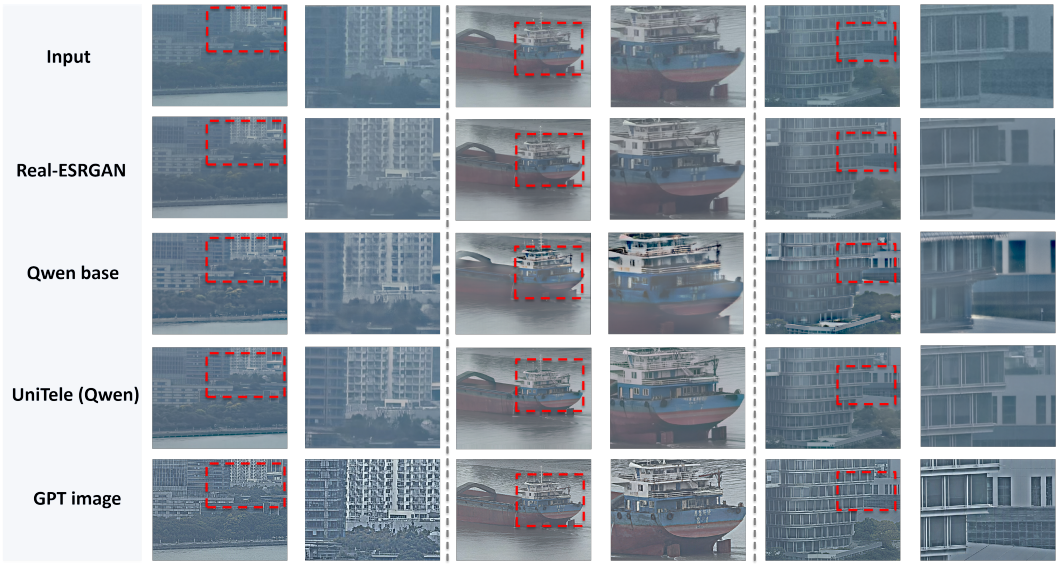

Figure 5: Qualitative SR comparison on in-the-wild images. Red boxes mark the regions shown in the adjacent close-ups.

Table 2 shows that UniTele (Qwen) improves MUSIQ and CLIP-IQA over its pretrained backbone, with a slight increase in NIQE. UniTele (FLUX) scores worse than its backbone on all three metrics, while GPT Image 2 obtains the best scores. In Figure 5, UniTele (Qwen) produces clearer window grids and ship railings than the Real-ESRGAN-trained baseline, with fewer conspicuous structural distortions than the Qwen base model. These examples illustrate transfer to real scenes, but no-reference scores and visual comparisons cannot verify the fidelity of recovered details without HR targets.

Table 2: No-reference quality on 16 in-the-wild iPhone images.

Method	MUSIQ ↑	CLIP-IQA ↑	NIQE ↓
Qwen SR base	44.33	0.2101	6.89
FLUX.2 SR base	65.60	0.5569	5.35
UniTele (Qwen), 1:2	54.49	0.2335	7.14
UniTele (FLUX), 1:2	58.27	0.3721	5.85
Qwen SR, Real-ESRGAN degradations	45.88	0.1767	8.03
GPT Image 2	67.83	0.5686	3.73

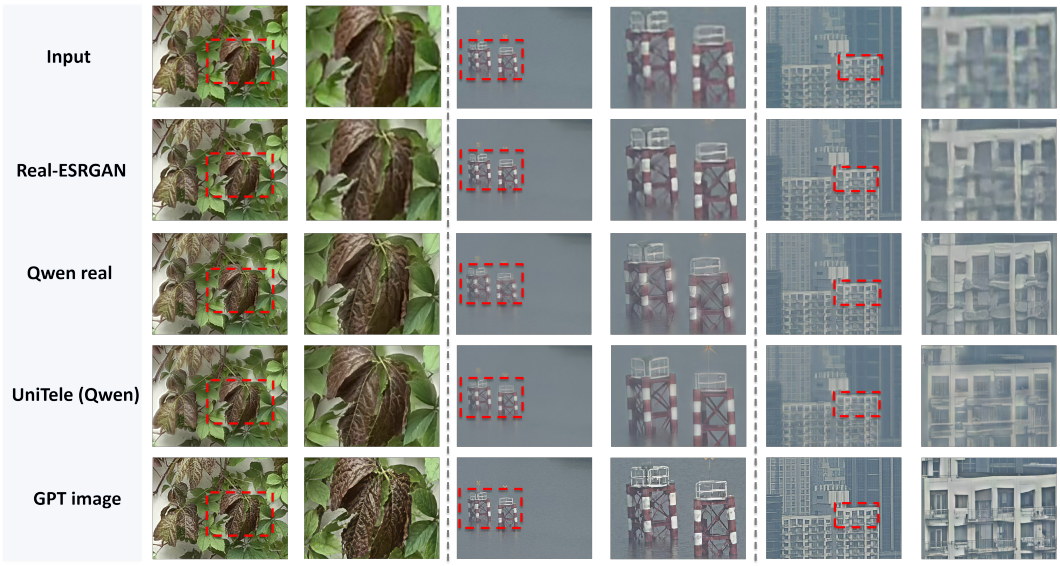

Figure 6: Additional in-the-wild SR comparisons. Rows show the input, Real-ESRGAN-trained Qwen, real-only Qwen, UniTele (Qwen), and GPT Image 2; red boxes mark the adjacent close-ups.

Figure 6 provides additional matched crops for comparing local detail and structural consistency across the real-only and mixed-data Qwen models, alongside the other in-the-wild baselines.

4.4 Ablation Studies

Training-device ablation. We compare all-device and Vivo-only training for real-only Qwen SR and UniTele (Qwen) at 1:2, evaluating all models on the complete three-device paired test set. Initialization and training settings are matched; the smaller single-device pool is sampled repeatedly. The UniTele comparison uses a controlled run separate from the main results.

Table 3: Effect of single- versus multi-device training on Qwen SR, evaluated across all three phones.

Device pool	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	HF ratio → 1
Real-world only					
All three devices	18.84	0.6098	20.43	0.3187	0.740
Vivo only	14.47	0.5451	38.99	0.4192	0.832
UniTele (Qwen), 1:2					
All three devices	18.81	0.6031	19.87	0.3199	0.766
Vivo only	13.51	0.5363	45.06	0.4617	0.815

Restricting training to Vivo alone reduces PSNR by 4.37 dB for real-world-only training and 5.30 dB for UniTele (Qwen) on the three-device test set; LPIPS also rises by 0.1005 and 0.1418, respectively. These large drops indicate limited generalization from single-device training to the multi-device benchmark, even when synthetic pairs are included.

Stage ablation. We evaluate the forward model directly, rather than only through downstream SR. Removing Stage I feeds digital images directly to the telephoto generator; removing Stage III uses the reference-free Stage II model. Both variants and the complete pipeline are evaluated against 50 source-matched real-world LR images from Huawei, iPhone, and vivo, using the same inference seed.

Table 4: Stage ablations for LR synthesis on 50 paired test images from three phones.

Forward-model variant	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	HF ratio → 1
Without Stage I	12.05	0.4657	54.39	0.6337	1.992
Without Stage III	15.20	**0.5473**	36.05	0.4001	0.950
Full three-stage model	**15.52**	0.5439	**35.01**	**0.3942**	**1.018**

Bypassing Stage I causes the largest loss, reducing PSNR by 3.47 dB and increasing LPIPS by 0.2395 relative to the full pipeline. Removing Stage III also reduces PSNR and increases LPIPS. The full model has the best PSNR and LPIPS and the HF ratio closest to one, although the Stage-II-only variant has slightly higher SSIM.

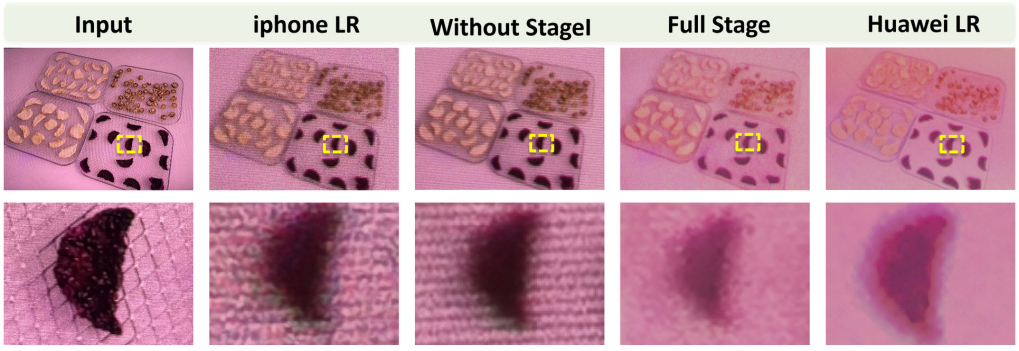

Figure 7: Stage I ablation for Huawei-targeted degradation synthesis. Columns show the digital input, iPhone LR, synthesis without Stage I, the full model, and Huawei LR; the bottom row shows yellow-box close-ups.

For shared source content, Figure 7 shows that the full pipeline better reproduces Huawei color and degradation appearance than feeding digital input directly to the telephoto generator.

5 Conclusion

UniTele learns device-conditioned camera-domain rendering and super-telephoto degradation from limited aligned captures, transferring both to external images to expand SR supervision. On the 60-pair, three-phone test set, real–synthetic training improves PSNR and LPIPS over real-only training for Qwen and FLUX. Ablations support factorized synthesis and multi-device training on this benchmark. In-the-wild results suggest qualitative transfer, but mixed no-reference scores cannot establish detail fidelity. These findings indicate that learned, device-conditioned pair synthesis can extend the utility of small capture sets for smartphone telephoto restoration.

Limitations. The capture protocol is restricted to a planar display, and residual alignment error and display-specific artifacts may affect both supervision and evaluation. Device coverage is finite, and a generative degradation model may alter content.

AI USE STATEMENT

Generative AI models were used to synthesize device-conditioned training pairs, as described in Section 3. An AI writing assistant also assisted with manuscript editing, consistency and citation checks, and LaTeX formatting. The authors are responsible for verifying AI-assisted text, references, experimental claims, and generated artifacts against the underlying sources and results, and take responsibility for the final content of this paper.

REFERENCES

Black Forest Labs. FLUX.2 [klein] 4B Base. Hugging Face model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28208–28220, 2025.

Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, 2024.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "Completely Blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

OpenAI. GPT-Image-2 model. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one blind image restoration. In *Advances in Neural Information Processing Systems*, 2023.

Qwen. Qwen-Image-Edit-2511. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen-Image-Edit-2511.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

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

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

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

Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. DeFlow: Learning complex image degradations from unpaired data with conditional flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Xiaotong Wu, Wei-Sheng Lai, YiChang Shih, Charles Herrmann, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Efficient hybrid zoom using camera fusion on mobile phones. *ACM Transactions on Graphics*, 42(6), 2023.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.