Appendix to "VOSR: A Vision-Only Generative Model for Image Super-Resolution"

Rongyuan Wu[1,2*] Lingchen Sun[1,2*] Zhengqiang Zhang[1,2] Xiangtao Kong[1,2]
Jixin Zhao[1,2] Shihao Wang[1] Lei Zhang[1,2†]

[1]The Hong Kong Polytechnic University [2]OPPO Research Institute

A. Appendix

This appendix presents distillation details, ScreenSR benchmark details, training settings, ablation studies, user study results, and additional visual comparisons.

A.1. Distillation Details

We follow the notation in Sec. 3.4 of the main paper. Both distilled variants use the same student parameterization $f_\theta(z_t, t, r, \kappa)$, where κ denotes either the fully conditioned mode $(c_{\text{str}}, c_{\text{sem}})$ or the partially conditioned mode $(\alpha c_{\text{str}}, \varnothing)$. They also share the same restoration-oriented teacher target:

$$v_{\text{tea}}^{\text{guide}} = v_{\text{pcond}} + \omega\,(v_{\text{cond}} - v_{\text{pcond}}), \tag{1}$$

and the same base objective:

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{z_0, z_1, t, \kappa}\left[\left\| f_\theta(z_t, t, r{=}t, \kappa) - v_{\text{tea}}^{\text{guide}} \right\|^2\right]. \tag{2}$$

Therefore, the semantic condition and the restoration-oriented partial conditioning are kept unchanged during distillation, and the two variants differ only in how they regularize the compressed prediction with $r < t$. In all experiments, the auxiliary consistency loss weight is set to 1.

A.1.1. Shortcut-based Variant

We first study a shortcut-based variant adapted from recent shortcut distillation methods [1]. Unlike the original formulation, our goal is not to reproduce the full shortcut training pipeline, but to transplant the core idea of self-consistency into our restoration-oriented distillation framework. Specifically, we keep the student parameterization, semantic condition, and partial conditioning design in Sec. 3.4 unchanged, and use a midpoint consistency constraint to regularize the compressed prediction from t to r.

For a sampled compressed target time $r < t$, we define the midpoint $s = (t + r)/2$. Let $u_{t\to r} = f_\theta(z_t, t, r, \kappa)$ and $u_{t\to s} = f_\theta(z_t, t, s, \kappa)$. We then construct an intermediate latent using the detached first-half prediction:

$$z_s = z_t - (t - s)\,\text{sg}(u_{t\to s}), \tag{3}$$

and predict the second half as:

$$u_{s\to r} = f_\theta(z_s, s, r, \kappa). \tag{4}$$

We use the detached two-stage decomposition to regularize the direct compressed prediction:

$$\bar{u}_{t\to r} = \text{sg}\left(\frac{u_{t\to s} + u_{s\to r}}{2}\right), \tag{5}$$

$$\mathcal{L}_{\text{short-cons}} = \|u_{t\to r} - \bar{u}_{t\to r}\|_2^2. \tag{6}$$

The total objective of this variant is:

$$\mathcal{L}_{\text{short}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{short-cons}}. \tag{7}$$

This design can be viewed as a shortcut-inspired consistency regularizer tailored to our SR setting, rather than a strict reproduction of the original shortcut method.

A.1.2. Recursive-Consistency-based Variant

We further adapt the recursive-consistency (RC) based strategy [3] under the same conditioning and teacher-guidance setup. For a compressed prediction from t to r, we first compute

$$u_{t\to r} = f_\theta(z_t, t, r, \kappa). \tag{8}$$

We then perform a teacher-guided warm start from z_t to an intermediate time $t_m = \max(t - \Delta t, r)$:

$$z_{t_m} = z_t - (t - t_m)\, v_{\text{tea}}^{\text{guide}}. \tag{9}$$

Starting from z_{t_m}, we run a detached multi-step ODE rollout to time r under the same conditioning mode κ, which yields a recursive trajectory target, denoted by $u_{t_m\to r}^{\text{tar}}$. Following the RC formulation, we construct a corrected detached target as:

$$\text{corr} = \text{clip}\left(c_l\, u_{t\to r} - c_r\, u_{t_m\to r}^{\text{tar}} - v_{\text{tea}}^{\text{guide}}, [-1, 1]\right), \tag{10}$$

* Equal contribution. † Corresponding author. This research is supported by the PolyU-OPPO Joint Innovative Research Center.

Table 1. Comparison of shortcut-based and RC-based distillation on RealSR using VOSR-0.5B-ms as the teacher.

Method	LPIPS↓	MUSIQ↑
Teacher (VOSR-0.5B-ms)	0.3069	68.93
Shortcut-based distillation	0.2913	68.21
RC-based distillation	0.2856	69.78

$$\tilde{u}_{t\to r} = \mathrm{sg}(u_{t\to r}) - \mathrm{corr}, \tag{11}$$

and optimize

$$\mathcal{L}_{\text{rc-cons}} = \|u_{t\to r} - \tilde{u}_{t\to r}\|_2^2 . \tag{12}$$

The resulting objective is:

$$\mathcal{L}_{\text{rc}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{rc-cons}}. \tag{13}$$

Compared with shortcut-based regularization, the advantage of RC-based regularization is that it explicitly pulls the student trajectory toward the teacher trajectory, rather than matching a target induced by randomly sampled intermediate points. This makes the supervision more aligned with the teacher's denoising path under large temporal compression. We therefore adopt the RC variant in the main paper.

To quantitatively compare the two distilled variants, we further report their performance on RealSR using VOSR-0.5B-ms as the teacher. As shown in Table 1, both shortcut-based and RC-based distillations preserve strong perceptual performance after compression to one-step inference. Compared with the multi-step teacher, shortcut distillation improves LPIPS while maintaining competitive MUSIQ, indicating that the distilled student can inherit the teacher's perceptual restoration capability. RC-based distillation performs better overall, achieving both lower LPIPS and higher MUSIQ than the shortcut-based variant. This result is consistent with our observation that RC provides a stronger training signal under large temporal compression and is thus better suited to one-step generative SR.

A.2. ScreenSR Benchmark Details

ScreenSR is designed as a real-world paired benchmark for evaluating generative SR in practical mobile-photography scenarios. We first manually collect a diverse set of high-quality source images from the web and deliberately curate them with balanced scene categories and object/scene scales. The selected images cover indoor scenes, outdoor scenes, human subjects, animals, plants, artworks, and Chinese and English text. We also include static and dynamic scenes and intentionally keep examples at different spatial scales within each major category, so that the benchmark can better evaluate SR methods from various aspects, including semantic diversity, structural fidelity, and robustness across scale variations. A thumbnail montage of the selected 130 images is shown in Fig. 1.

Table 2. No-reference quality comparison of GT images in different real-world paired SR benchmarks. Better GT quality is indicated by lower NIQE and AFINE-NR, and higher MUSIQ, MANIQA, and TOPIQ-NR.

Benchmark	NIQE↓	MUSIQ↑	MANIQA↑	AFINE-NR↓	TOPIQ-NR↑
ScreenSR	**3.7719**	**72.1500**	**0.7187**	**-1.2093**	**0.7363**
RealSR	6.1167	57.4564	0.6016	-0.9088	0.4140
DRealSR	6.7909	50.5644	0.5588	-0.7731	0.3932

After content curation, the source images are displayed on a high-resolution screen and re-photographed by flagship smartphones to construct paired real-world LR-HR examples. The data capturing devices cover flagship models from major smartphone manufacturers, including OPPO, vivo, Xiaomi, and Huawei. To ensure reliable pairing, each source image is first placed on a white canvas with four ArUco markers near the border. After capture, the detected marker corners are used to estimate a geometric transform that warps the mobile photo back to the original image plane, producing pixel-aligned pairs at the target resolution. We further apply a wavelet-based low-frequency color alignment while preserving high-frequency details from the captured image. This screen re-photography pipeline preserves accurate pairing while introducing realistic mobile imaging degradations. Compared with purely synthetic degradations, the resulting LR images better reflect practical mobile captures, while the displayed source images provide clean and visually strong references. The final ScreenSR benchmark contains 130 paired samples, all used for zero-shot real-world evaluation in our experiments.

A key motivation for building ScreenSR is that the quality of GT images matters for real-world SR evaluation. If the GT itself has limited perceptual quality, the reliability of benchmark conclusions may be weakened. We compare the no-reference quality metrics of GT images from ScreenSR, RealSR, and DRealSR. As shown in Table 2, ScreenSR consistently achieves substantially better GT quality on all five no-reference metrics, including NIQE, MUSIQ, MANIQA, AFINE-NR, and TOPIQ-NR. These results support our claim that ScreenSR provides cleaner and more reliable references for real-world paired evaluation.

A.3. Detailed Training Settings

We train VOSR in a progressive manner. Specifically, the multi-step model is first pretrained at 256×256 resolution for 400K steps with a global batch size of 1024 and a constant learning rate of 1.0×10^{-4}, and is then further trained at 512×512 resolution for another 400K steps with a global batch size of 256 and a constant learning rate of 5.0×10^{-5}. After obtaining the multi-step teacher, we distill it into a one-step model for 50K steps using a batch size of 32 and a constant learning rate of 2.0×10^{-5}. Across all stages, we use no warm-up, set the weight decay to 0.01,

Figure 1. Thumbnail montage of the ScreenSR benchmark. The selected 130 GT images cover diverse scenarios, including indoor and outdoor scenes, humans, animals, plants, artworks, and multilingual text, with substantial variation in object and scene scales. This diversity ensures a comprehensive evaluation of generative SR methods in terms of semantic coverage, structural fidelity, and robustness across different content types.

the gradient clipping threshold to 1.0, and the EMA decay to 0.9999, and adopt AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. For the diffusion backbone, both VOSR-0.5B and VOSR-1.4B use a patch size of 2 and an MLP ratio of 4. VOSR-0.5B uses dimension 1024, depth 28, and 16 attention heads, while VOSR-1.4B uses dimension 1536, depth 36, and 24 attention heads.

A.4. Ablation Studies

Unless otherwise specified, all ablation experiments are conducted on VOSR-0.5B, and trained at 512×512 resolution for 100K steps.

A.4.1. Effect of Visual Semantic Condition

Table 4 evaluates the visual semantic condition on RealSR. Removing the semantic condition causes a clear degradation, showing that structural conditioning alone is insufficient for challenging real-world SR. Adding pretrained

Table 3. Training recipe for VOSR, including the progressive pretraining for the multi-step models and the distillation for the one-step model.

	PT 256_{px}	PT 512_{px}	Distill to One-step
Learning rate	1.0×10^{-4}	5.0×10^{-5}	2.0×10^{-5}
LR scheduler	Constant	Constant	Constant
Warm-up steps	0	0	0
Training steps	400K	400K	50K
Global batch size	1024	256	32
Weight decay	0.01	0.01	0.01
Gradient clip	1.0	1.0	1.0
EMA	0.9999	0.9999	0.9999
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		

Table 4. Ablation on visual semantic encoders on RealSR.

Method	LPIPS↓	MUSIQ↑
w/o SVE	0.3011	63.74
w/ SVE (CLIP)	0.2788	63.82
w/ SVE (SigLIPv2)	0.2817	64.81
w/ SVE (DINOv3)	0.2858	67.51
w/ SVE (DINOv2)	0.2872	68.23

visual semantic features consistently improves perceptual quality across different encoders. CLIP achieves the best LPIPS, while DINO-based encoders produce notably higher MUSIQ, indicating better perceptual realism. We use DINOv2 in the final model as a balanced choice considering both overall performance and stability.

A.4.2. Partial Conditioning

We compare three guidance designs on LSDIR: using the fully conditioned model alone without guidance, standard CFG with a fully unconditional auxiliary branch, and our restoration-oriented partial conditioning. For our method, instead of fixing the structural retention factor to a single value, we randomly sample α within $[0.05, 0.25]$ during training. We adopt this design because the partially conditioned branch is intended to represent a family of weakly conditioned, input-anchored restoration states rather than one specific conditioning strength. Randomizing α exposes the model to diverse weak-structure conditions, which improves the robustness of the auxiliary branch and avoids overfitting the guidance behavior to a narrow partial-conditioning regime. The sampled range is kept small so that the partial branch remains weaker than the fully conditioned one while still preserving minimal structural anchors from the LR input.

Table 5 shows that explicit guidance is important for high-quality generative SR. Compared with using the fully conditioned model alone, standard CFG leads to a clear performance degradation. This is due to the fact that, under limited training data and computation, the fully unconditional branch is difficult to train from scratch for SR with standard CFG. Since the unconditional branch is poorly learned, the resulting guidance becomes unstable and can even harm restoration quality. In contrast, our partial conditioning achieves the best results, suggesting that an input-anchored auxiliary branch is much easier to optimize and provides a more reliable guidance direction for balancing perceptual realism and restoration fidelity.

A.5. User Study

To further validate VOSR from a human perceptual perspective, we conduct separate user studies for the multi-step and one-step settings using 30 LR images. For each sample, participants are shown the LR input and the SR results from all compared methods, and they are asked to select the best

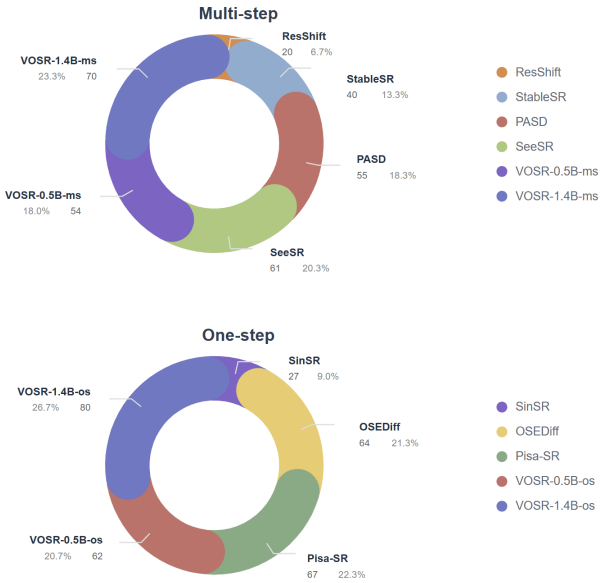

Figure 2. User study results in the multi-step and one-step settings. VOSR-1.4B-ms and VOSR-1.4B-os receive the highest numbers of votes in their respective groups, showing strong human preference in perceptual quality and consistency with the LR input.

Table 5. Ablation studies on restoration-oriented partial conditioning on LSDIR.

Guidance design	LPIPS↓	MUSIQ↑
Full condition only	0.3752	67.29
Standard CFG	0.4053	50.78
Ours (partial conditioning)	0.3772	69.26

result. The evaluation uses two equally weighted criteria: (1) perceptual quality and (2) consistency with the LR input, including structural and texture fidelity. Ten volunteers were invited and each evaluated all samples.

In the multi-step setting, we compare ResShift [9], StableSR [4], PASD [8], SeeSR [7], VOSR-0.5B-ms, and VOSR-1.4B-ms (300 votes total). As shown in Fig. 2, VOSR-1.4B-ms wins most votes (70/300, 23.3%), followed by SeeSR (61/300, 20.3%), PASD (55/300, 18.3%), and VOSR-0.5B-ms (54/300, 18.0%); StableSR and ResShift receive 40/300 (13.3%) and 20/300 (6.7%). For the one-step setting, we compare SinSR [5], OSEDiff [6], PiSA-SR [2], VOSR-0.5B-os, and VOSR-1.4B-os, resulting in 300 total votes. VOSR-1.4B-os ranks first with 80/300 votes (26.7%), followed by PiSA-SR (67/300, 22.3%), OSEDiff (64/300, 21.3%), and VOSR-0.5B-os (62/300, 20.7%), while SinSR receives 27/300 votes (9.0%). These results show that VOSR is strongly preferred by human evaluators in both multi-step and one-step settings, validating its ability to achieve better perceptual quality while preserving stronger input faithfulness.

A.6. More Visual Results

Fig. 3 provides six additional visual comparisons, including three multi-step cases (1st, 3rd and 5th) and three one-step cases (2nd, 4th and 6th). Overall, T2I-based methods often produce visually sharp results, but are less reliable in recovering fine structures that need to be faithful to the LR input. In contrast, VOSR consistently restores more local details with clearer structure and fewer hallucinations.

In the first example, the sign contains a thin symbol contour and a sharp corner structure that are difficult to recover from the degraded input. PASD [8] and SeeSR [7] produce obvious shape distortions, while StableSR [4] restores the symbol more plausibly but with inaccurate local geometry. By contrast, the VOSR variants recover clearer and more faithful symbol shapes, with VOSR-1.4B-ms producing the most complete contour and corner details. Similar trends can be observed in other multi-step examples: VOSR preserves sharper window boundaries and clearer panel edges, while T2I-based methods either blur the structures or generate less accurate local geometry. Meanwhile, compared with these T2I-based methods, VOSR runs substantially faster, as shown by the complexity comparison in the main paper.

For the one-step examples, similar conclusions can be drawn. In text and symbol regions, SinSR shows weaker generative ability, while OSEDiff and PiSA-SR sometimes produce sharper, yet less faithful shapes. By contrast, VOSR restores clearer characters, cleaner boundaries, and more stable local structures. In particular, the two VOSR one-step models recover the sign content and thin edges more faithfully while remaining highly efficient.

These visual results are consistent with the quantitative comparisons and efficiency analysis in the main paper. They further support that VOSR benefits from a restoration-oriented design: by combining spatially grounded visual semantics with input-anchored guidance, it can generate perceptually strong details without relying on the generative priors transferred from pre-trained T2I models.

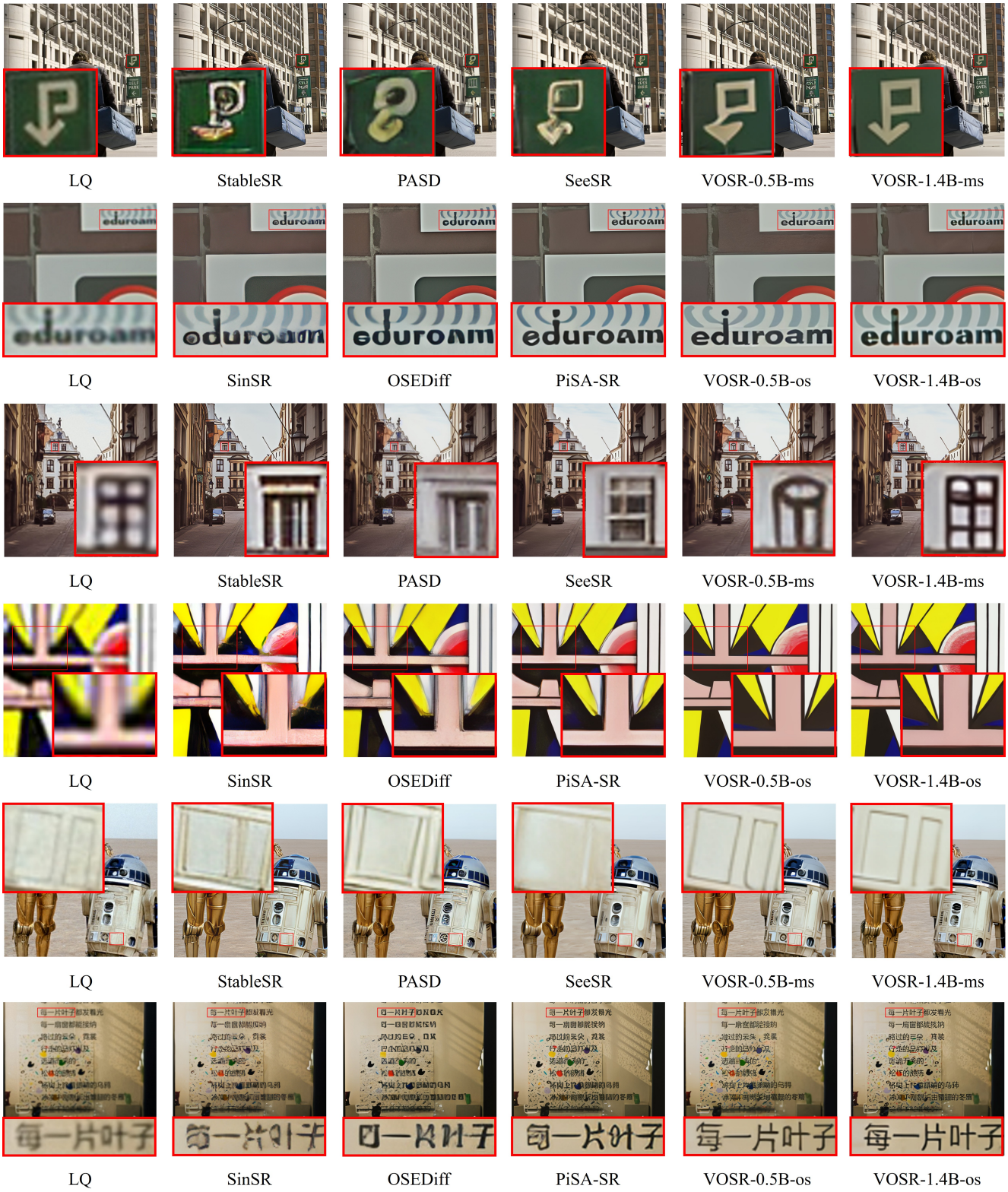

Figure 3. Additional visual comparisons of multi-step (1st, 3rd and 5th) and one-step (2nd, 4th and 6th) SR results. Compared with representative vision-only and T2I-based methods, VOSR produces perceptually more realistic details while better preserving structures that are faithful to the LR input. Please zoom in for better view.

References

[1] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024. 1

[2] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024. 4

[3] Peng Sun and Tao Lin. Any-step generation via n-th order recursive consistent velocity field estimation. In *International Conference on Learning Representations*, 2026. 1

[4] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 4, 5

[5] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. *arXiv preprint arXiv:2311.14760*, 2023. 4

[6] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024. 4

[7] 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*, pages 25456–25467, 2024. 4, 5

[8] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 4, 5

[9] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *arXiv preprint arXiv:2307.12348*, 2023. 4