Supplemental Material to "Perceive, Understand and Restore: Real-World Image Super-Resolution with Autoregressive Multimodal Generative Models"

Hongyang Wei[1,3,*], Shuaizheng Liu[2,3,*], Chun Yuan[1,†], Lei Zhang[2,3,†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]The Hong Kong Polytechnic University [3]OPPO Research Institute

In this supplementary file, we provide additional visual comparisons of competing Real-ISR methods, present comprehensive results from our ablation study, demonstrate the inference acceleration performance of Speculative Jacobi Decoding (SJD) [2] approach, and include comparative analyses with the SUPIR [3].

A. More Visual Comparisons

We present more visual comparisons with state-of-the-art diffusion-based Real-ISR methods on real-world images. As illustrated in Fig. 1 and Fig. 2, our proposed PURE demonstrates superior restoration quality across diverse scenarios, including animals, plants, humans, and landscapes. The PURE framework not only effectively preserves the structural and textural integrity of LQ input images, but also generates more abundant and plausible details, achieving exceptional visual quality.

B. Visual Results for Ablation Study

We provide visual comparison results for the ablation studies conducted in the main paper.

Perception-Understanding Guidance. We first present visual comparisons on perception-understanding guidance, as illustrated in Fig. 3. The *No Guidance* method demonstrates suboptimal restoration results with noticeable degradation in both cases, mainly due to its sole reliance on the LQ input image as the conditioning factor. Both *Perception Guidance* and *No Guidance* fail to generate rich details due to the absence of semantic description guidance. The *Understanding Guidance* produces erroneous details (*e.g.*, eyelashes), which can be attributed to the lack of adaptive degradation awareness. The *Full Guidance* achieves the best restoration quality among all variants.

Entropy-based Top-k Sampling. To demonstrate the effectiveness of our proposed entropy-based Top-k sampling strategy, we present visual comparison results by applying both fixed and dynamic Top-k sampling approaches. As illustrated in Fig. 4, the *Top-1* strategy yields the smoothest yet most blurry restoration results, as its generative capability is constrained by selecting only the most probable image token at each step. In contrast, while the *Top-2000* approach generates more details, it produces inferior visual perception quality with more erroneous textures. Notably, our entropy-based Top-k sampling achieves the best restoration quality, maintaining superior fidelity while generating more realistic and authentic details.

C. Inference acceleration results of SJD

Regarding the efficiency issues mentioned in the main paper, how to accelerate AR models has been actively studied. Recently, using Speculative Jacobi Decoding (SJD) plus quantization, Lumina-mGPT2's [1] speed and memory have been much improved (694s→304s, 80GB→33.8GB). We found that by applying SJD to PURE, the inference speed can be improved from 256.87s to 188.43s with similar results. We believe the rapid advancement of AR models will make them deployable for Real-ISR tasks.

D. Comparisons with SUPIR

To ensure fair comparisons among methods trained on datasets of similar scale, we have excluded SUPIR's results [3] from our main paper. Here, we include the results of SUPIR to provide a broader context for our evaluation. Note that SUPIR employs LLaVA for scene understanding to aid image reconstruction. We report quantitative comparisons with

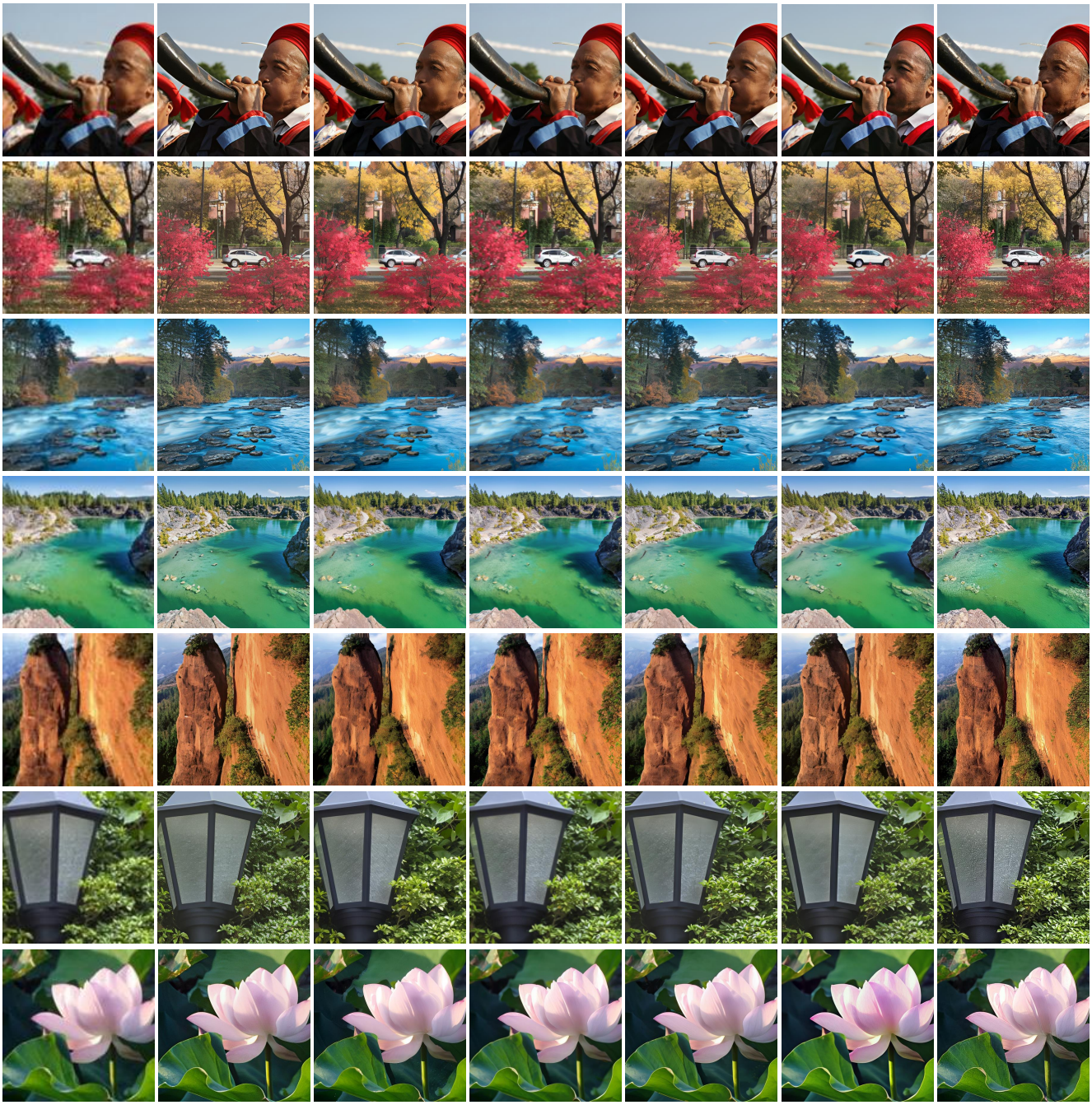

Figure 1. Visual comparisons of different Real-ISR methods on real-world images. Please zoom in for a better view.

SUPIR [3] in Table 2. Furthermore, as evidenced in Fig. 5, SUPIR still exhibits limitations in rendering natural local details despite employing LLaVA-generated captions, a stronger Stable Diffusion backbone, and training on more extensive, higher-quality datasets. For perception of local structure, diffusion-based methods model the overall distribution of global images, whereas AR approaches explicitly capture the joint distribution between adjacent tokens (via next-token prediction): $P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|\mathbf{c}_{<t}, \mathbf{x})$. So AR models can better avoid generating unnatural textures.

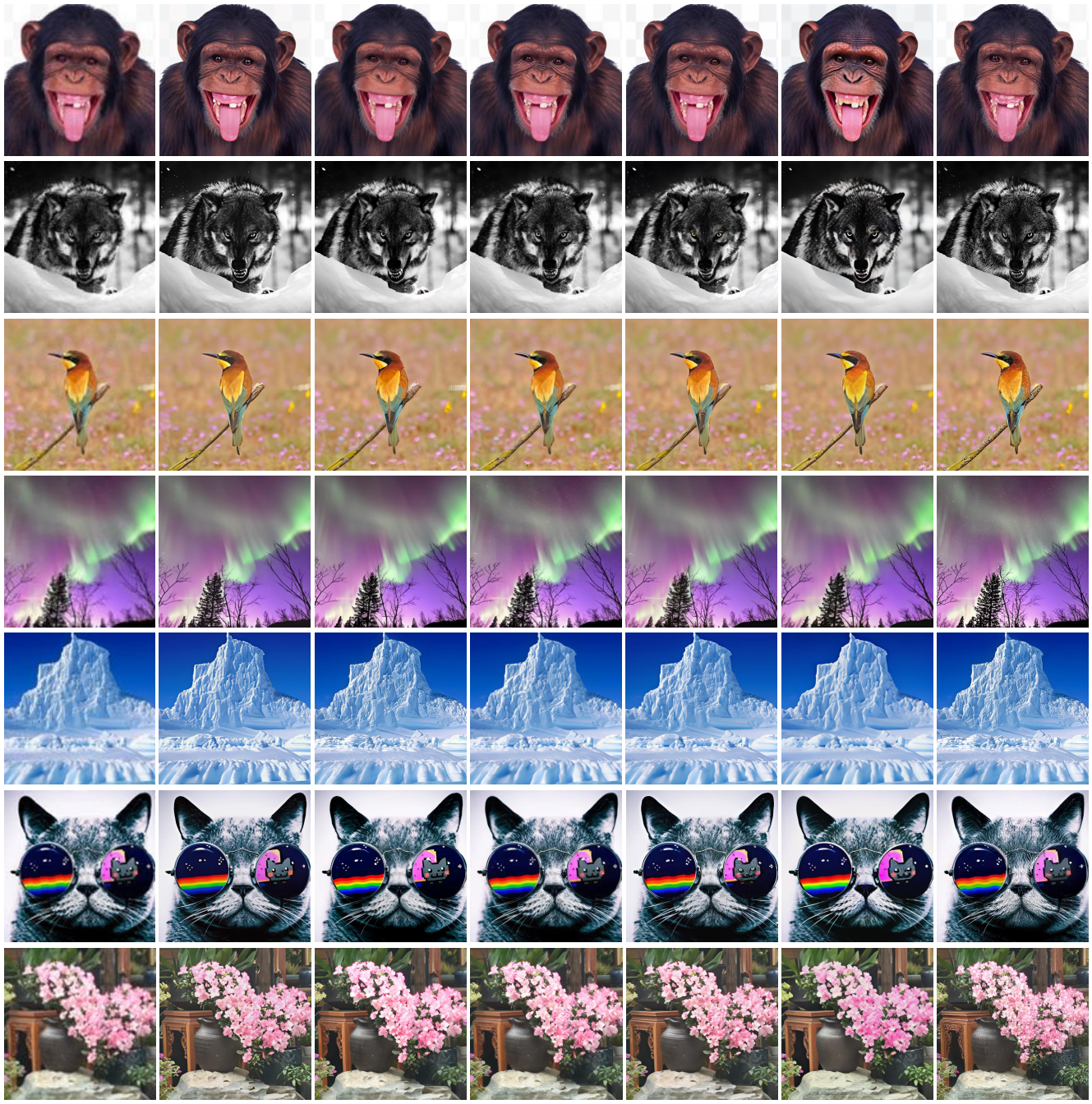

Figure 2. Visual comparisons of different Real-ISR methods on real-world images. Please zoom in for a better view.

Table 1. Comparison on the OST-Val dataset.

Methods	Inference Time(s)	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	MANIQA↑
PURE	256.87	18.52	0.4181	0.3863	74.81	0.6424
PURE+SJD	188.43	18.85	0.4353	0.3630	74.65	0.6362

Figure 3. Visual comparisons for perception-understanding guidance ablation. Please zoom in for a better view.

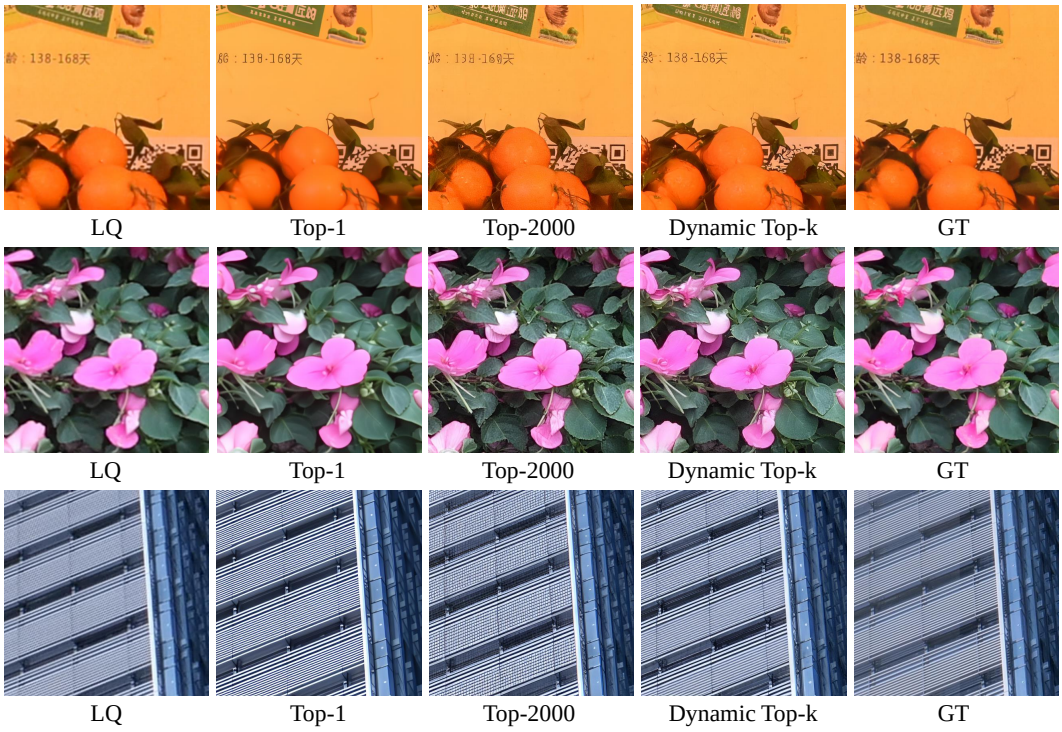

Figure 4. Visual comparisons for entropy-based Top-k sampling ablation. Please zoom in for a better view.

Table 2. Comparison on the OST-Val dataset.

Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	MANIQA↑
PURE	18.52	0.4181	0.3863	74.81	0.6424
SUPIR	18.86	0.4141	0.3938	74.32	0.5321

Figure 5. Unnatural restoration results of SUPIR.

References

[1] Alpha VLLM Team. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling, 2025. 1

[2] Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*, 2024. 1

[3] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 1, 2