
HAODiff: Human-Aware One-Step Diffusion via Dual-Prompt Guidance

Supplementary Material

Jue Gong^{1*}, Tingyu Yang^{1*}, Jingkai Wang¹, Zheng Chen¹,
Xing Liu², Hong Gu², Yulun Zhang^{1†}, Xiaokang Yang¹
¹Shanghai Jiao Tong University ²vivo Mobile Communication Co., Ltd

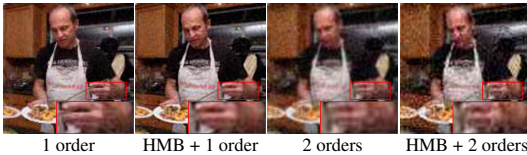


Figure 1: Visualization on different degradation pipelines. In the two-order setting, the addition of HMB is hard to discern for human vision.



Figure 2: Visualization of YOLO’s HMB detection bounding boxes on synthetic PERSONA-Val and real-world MPII-Test datasets.

A Degradation Pipeline

A.1 Explanation of Pipeline Setting

A degradation pipeline must closely mimic and cover real-world conditions. The pipeline in Real-ESRGAN [16] always applies two orders of degradation and often produces extremely damaged images (see Fig. 1). Such severity helps simulate extreme cases, yet many real images are less degraded. We must preserve the model’s ability to handle heavy degradation while still accounting for the prevalence of mild cases. Thus, the presence of the second order is always guaranteed, and the first order is triggered only with a set probability. This lets the model learn to restore both severe and mild degradations. Downsampling is the most influential factor in the second order. The scaling ratio is usually chosen from 2, 3, 4 and then perturbed with randomness. Blind super-resolution methods usually target $4\times$ upsampling, making a ratio of 4 common. Our scenario implicitly includes multiple downsampling factors in the LQ inputs, and the model outputs images at the same resolution. To cover more cases, we pick the ratio randomly and uniformly at each instance from 2, 3, and 4.

Human motion blur (HMB) is placed before generic degradation to align with its physical origin. Visualization (see Fig. 1) reveals that HMB does not combine well with two-stage generic degradation. When applied with both stages, HMB becomes indistinguishable from generic blur, and local positional cues are lost. We enforce mutual exclusivity between HMB and the full two-stage generic degradation process. This exclusion mechanism safeguards against the interference of corrupted cues in both dual-prompt guidance and one-step diffusion training. The first-stage trigger probabilities for no degradation, HMB, and generic degradation are set to 0.2, 0.4, and 0.4, respectively.

A.2 The Fine-tuned YOLO

Our YOLO11 [6] detector is trained on roughly 100K degraded images produced from PERSONA [4] with our degradation pipeline. The probability of the HMB branch is set high, so 90 % of the images contained human motion blur (HMB). Bounding boxes are defined as the minimal rectangles enclosing each HMB segmentation mask. Training runs at 512×512 resolution for 100 epochs on 2 NVIDIA RTX A6000 GPUs with a batch size 216. Detection performance on PERSONA-Val [4] and MPII-Test (see Fig. 2) shows that a model trained purely on synthetic data can reliably detect HMB in real-world images. These results demonstrate that our degradation pipeline generates HMB degradations that closely resemble those encountered in practice, enabling effective detection.

*Equal contribution.

†Correspondence author: Yulun Zhang, yulun100@gmail.com.

Algorithm 1: Degradation Pipeline with Human Motion Blur

Input: High-quality image I_H , segmentation mask S , degradation parameters Θ
Output: High-quality image I_H , low-quality image I_L , modified segmentation mask S'

```

/* Select degradation type */
Select degradation type  $\tau \sim \{Normal, MotionBlur, None\}$  with probability  $[P_N, P_M, P_O]$ ;
 $I_t \leftarrow I_H$ ;
if  $\tau = MotionBlur$  and  $S \neq \emptyset$  then
    /* Motion blur generation */
     $T \leftarrow$  Generate trajectory with random angle, velocity, and perturbations scaled by  $\Theta_A$ ;
     $PSF \leftarrow$  Construct point spread function by discretizing  $T$  onto grid with bilinear weights;
     $R \leftarrow$  MapToTargetRegions( $S, \Theta_R$ ); // Maps to regions: head, limbs, body, etc.
     $R_i \leftarrow$  RandomSelectRegion( $R$ );
     $M \leftarrow (R = R_i)$ ;
     $M \leftarrow \{Erode(M, \Theta_{K_E}) \rightarrow Dilate(M, \Theta_{K_D}) \rightarrow Gaussian(M, \Theta_B, \Theta_S)\}$ ;
     $W_S \leftarrow$  Normalize( $M$ );
     $I_B \leftarrow$  FFTconvolve( $I_H, PSF$ );
     $I_t \leftarrow W_S \odot I_B + (1 - W_S) \odot I_H$ ;
     $S' \leftarrow$  ThresholdMask( $W_S > 0$ );
end
else
     $S' \leftarrow \emptyset$ ;
    if  $\tau = Normal$  then
        /* Generic degradation */
         $I_t \leftarrow \{Blur(I_t, \Theta_{B_1}) \rightarrow Resize(I_t, \Theta_{R_1}) \rightarrow Noise(I_t, \Theta_{N_1}) \rightarrow Compress(I_t, \Theta_{C_1})\}$ ;
    end
end
/* Secondary degradation */
 $I_L \leftarrow \{Blur(I_t, \Theta_{B_2}) \rightarrow Resize(I_t, \Theta_{R_2}) \rightarrow Noise(I_t, \Theta_{N_2}) \rightarrow Compress(I_t, \Theta_{C_2})\}$ ;
return  $I_H, I_L, S'$ ;

```

Methods	Pipelines	CLIPQA \uparrow	MANIQA \uparrow	NIQE \downarrow	LIQE \uparrow	IL-NIQE \downarrow
OSDHuman [4]	Real-ESRGAN [16]	0.6537	0.6471	4.5960	3.7991	26.8940
	Ours	0.6622	0.6684	4.5873	3.9119	26.4993
HAODiff (ours)	Real-ESRGAN [16]	0.6800	0.6686	3.9912	4.0826	23.9011
	Ours	0.6923	0.6787	3.9450	4.1777	23.6714

Table 1: Quantitative comparison on MPII-Test, with the best results highlighted in **bold**. Results are shown for the same methods trained on both our degradation pipeline and that of Real-ESRGAN.

CFG Scale	DISTS \downarrow	LPIPS \downarrow	TOPIQ \uparrow	FID \downarrow	CLIPQA \uparrow	MANIQA \uparrow	LIQE \uparrow	IL-NIQE \downarrow
$\lambda_{cfg} = 1$ (w/o CFG)	0.1054	0.2138	0.5029	8.7904	0.7651	0.7079	4.8180	19.3376
$\lambda_{cfg} = 1.05$	0.1049	0.2119	0.5039	8.8482	0.7615	0.7049	4.8222	19.6370
$\lambda_{cfg} = 3.5$	0.1023	0.2046	0.5161	8.3623	0.7737	0.7097	4.8485	18.5986
$\lambda_{cfg} = 7.5$	0.1075	0.2149	0.4877	8.9956	0.7203	0.6836	4.5802	19.2951

Table 2: Quantitative analysis on PERSONA-Val based on different CFG Scale training schemes, with the best results highlighted in **bold**. FID is computed with the ground truth.

A.3 The Effectiveness of Our Degradation Pipeline

We compare HAODiff and OSDHuman [4] trained with the degradation pipelines from Real-ESRGAN [16] and ours to assess whether our pipeline consistently enhances the model’s ability to restore low-quality human images. As shown in Tab. 1, our degradation pipeline improves restoration quality, demonstrating strong suitability for human body restoration. The visual results in Fig. 11 and Fig. 12 demonstrate that retrained OSDHuman can more reliably restore images with motion blur.

A.4 Algorithm of the Degradation Pipeline

To explicitly demonstrate our degradation methodology, Algorithm 1 details the computational pipeline integrating human motion blur (HMB). For motion blur simulation, we follow the settings from DeblurGAN [8] to preserve parameter compatibility with prior work. Notably, we apply two iterations of morphological dilation to better approximate the realistic motion-affected regions.

1st Branch	2nd Branch	3rd Branch	DISTS↓	LPIS↓	TOPIQ↑	FID↓	CLIPQA↑	MANIQA↑	LIQE↑	IL-NIQE↓
✓			0.1054	0.2138	0.5029	8.7904	0.7651	0.7079	4.8180	19.3376
✓	✓		0.1046	0.2117	0.4970	8.9361	0.7541	0.7051	4.7125	19.2762
✓		✓	0.1079	0.2146	0.4884	8.8348	0.7342	0.6939	4.6416	19.2885
✓	✓	✓	0.1023	0.2046	0.5161	8.3623	0.7737	0.7097	4.8485	18.5986

Table 3: Quantitative comparison on PERSONA-Val on using different branches of dual-prompt guidance (DPG), with the best results highlighted in **bold**. FID is computed with the ground truth.

Methods	DiffBIR [11]	SeeSR [20]	PASD [22]	ResShift [25]	SinSR [17]	OSDiff [19]	OSDHuman [4]	HAODiff (ours)
Step	50	50	20	15	1	1	1	1
Time (s)	9.03	5.05	3.15	2.88	0.19	0.13	0.11	0.20
Param (M)	1,717	2,524	1,900	119	119	1,775	1,576	1,459
MACs (G)	24,234	65,857	29,125	5,491	2,649	2,265	2,200	2,600

Table 4: Complexity comparison during inference. All models are tested with an input image size of 512×512 on the NVIDIA RTX A6000 GPU. ‘MACs’ denotes multiply-accumulate operations.

B One-Step Diffusion

B.1 Classifier-Free Guidance (CFG) Scale

Selecting the CFG scale is a critical step when applying classifier-free guidance [5]. This parameter is typically adjusted within the diffusion process. In text-to-image (T2I) generation, CFG is first introduced to balance conditional and unconditional sampling; later work replaces the unconditional branch with a negative prompt to suppress unwanted content. Most T2I systems fix the CFG scale at a specific value during inference: for example, Stable Diffusion (SD) 2.1-base [14] uses 7.5 to balance diversity and controllability. As for image restoration, diffusion-based models such as SUPIR [23] and S3Diff [27] apply specially designed CFG strategies. Earlier methods [11, 20] focus on the positive-prompt path; they leave the negative prompt and the CFG scale unchanged. However, during training, SUPIR and S3Diff occasionally substitute the positive prompts and clean images with predefined negative prompts and degraded images. This teaches the model to associate a fixed ‘low quality’ condition with various degradations and thus steer generation away from them.

HAODiff takes a different approach, deriving adaptive image prompt embeddings directly from each low-quality (LQ) image and producing feature and degradation cues tailored to that input instead of fixed text prompts. Consequently, the choice of the CFG scale for restoring LQ human images must be reconsidered. During training, we keep the CFG scale fixed so that the diffusion model can adapt to the new conditional embeddings. We evaluate three scales. Because our base model is SD2.1-base, we include canonical 7.5. One-step or few-step diffusion models often use values between 3 and 5, so we select 3.5 as the second option. Finally, S3Diff sets the scale to 1.05; we adopt this as the third option to assess its effect on the one-step diffusion restoration model. We also compare these strategies with not using CFG to measure the benefits of using CFG. As shown in Table 2, using a CFG scale of 3.5 gives the best performance on all metrics, so we adopt 3.5 as the CFG scale. As the CFG scale increases from 1 to 3.5, most metrics improve steadily, demonstrating that the CFG strategy improves the effectiveness of the model. However, the performance drop observed at a CFG scale of 7.5 indicates that an excessively high CFG scale is also unsuitable for our models.

B.2 Ablation Study of DPG

We perform an ablation study on the three DPG branches of HAODiff. Specifically, DPG consists of three branches: the first branch predicts the high-quality (HQ) image, the second predicts the residual noise between low-quality (LQ) and HQ images, and the third predicts the human motion blur (HMB) segmentation mask. Prior works [15, 4] demonstrate the feasibility of using prompt extractors to generate HQ image embeddings. Therefore, we focus our analysis on the second and third branches. As shown in Tab. 3, using all three branches yields the best performance across all metrics, confirming the advantage of the full DPG configuration. Although combining the third branch with the first branch performs worse than the first branch alone, our main text shows that without it, the model fails to localize motion blur artifacts, highlighting its indispensable role.

B.3 Complexity Comparison

As shown in Tab. 4, our method employs a more lightweight prompt generator than the other SD2.1-based methods, OSDiff [19] and OSDHuman [4]. Although our approach needs two UNet passes, we mitigate inference time by stacking the inputs along the batch dimension. This strategy helps reduce the actual runtime. As a result, even though UNet computation dominates the overall inference time, runtime remains comparable. Moreover, in terms of multiply-accumulate operations (MACs), our model maintains much lower computational overhead compared to multi-step diffusion methods.

Metric	DiffBIR	SeeSR	SUPIR	PASD	ResShift	SinSR	OSDiff	InvSR	OSDHuman	HAODiff
PSNR↑	20.32	20.28	19.67	21.19	21.31	20.85	20.39	19.67	21.06	20.59
SSIM↑	0.5592	0.5902	0.5451	0.6248	0.6283	0.6007	0.6049	0.5844	0.6180	0.6035
DISTS↓	0.1402	0.1295	0.1415	0.1469	0.1638	0.1579	0.1340	0.1424	0.1356	0.1023
LPIPS↓	0.2797	0.2555	0.2929	0.2910	0.2848	0.2840	0.2507	0.2709	0.2384	0.2046

Table 5: Quantitative comparison of full-reference metrics across different methods.

Method	PERSONA-Val							MPII-Test			
	DISTS↓	LPIPS↓	TOPIQ↑	C-IQA↑	M-IQA↑	NIQE↓	LIQE↑	C-IQA↑	M-IQA↑	NIQE↓	LIQE↑
Real-ESRGAN	0.1671	0.2843	0.4202	0.4981	0.6044	3.7577	3.8992	0.3356	0.5051	5.9106	2.4096
BSRGAN	0.1713	0.2927	0.4140	0.5259	0.5841	3.8440	3.7864	0.4423	0.5333	5.5901	2.6073
FeMaSR	0.1727	0.2968	0.3911	0.6059	0.5245	3.8424	3.3674	0.4742	0.4692	5.4429	2.4792
SwinIR	0.1829	0.3139	0.3908	0.4540	0.5168	3.9391	3.1318	0.2785	0.4480	6.1494	2.0553
Restormer	0.2884	0.6415	0.2891	0.3255	0.3860	8.0211	1.1143	0.2503	0.3505	8.3798	1.1007
Restormer*	0.2270	0.3865	0.3962	0.3354	0.3923	6.8413	2.3530	0.2927	0.3821	7.7481	1.7956
LBAG	0.3346	0.6754	0.2541	0.1609	0.3735	8.0663	1.0380	0.1601	0.3465	8.8200	1.0874
DeblurGAN	0.3024	0.5959	0.2921	0.2105	0.3977	8.0039	1.1332	0.1924	0.3738	7.1916	1.1981
UFPNet	0.2906	0.6050	0.2910	0.2515	0.3925	7.8289	1.1413	0.2057	0.3718	7.3489	1.2058
AdaRevD	0.2984	0.6036	0.2890	0.2492	0.3945	7.9889	1.1343	0.2034	0.3700	7.2810	1.1985
AdaRevD*	0.2716	0.5477	0.3129	0.2120	0.3160	7.3103	1.2055	0.1707	0.3514	6.3201	1.2059
HAODiff	0.1023	0.2046	0.5161	0.7737	0.7097	2.8298	4.8485	0.6923	0.6787	3.9450	4.1777

Table 6: Quantitative comparisons on the PERSONA-Val and MPII-Test with different methods. C-IQA refers to CLIPQA, and M-IQA refers to MANIQA. Asterisks (*) indicate retrained versions.

B.4 Setting of Timestep

In one-step diffusion restoration, zero-shot DDIM [13] inference performs poorly. The standard remedy fixes a single timestep before fine-tuning. To minimize the required adaptation, we choose the timestep whose latent noise level in the pretrained SD2.1-base model best matches the noise of the degraded images. We test several timesteps with the frozen SD2.1-base on PERSONA-Test. As Tab. 7 shows, the chosen timestep 199 yields the best results.

Timestep	CLIPQA↑	LIQE↑	MANIQA↑
1	0.4135	2.6714	0.6423
199	0.5178	3.1436	0.6295
399	0.3570	1.9304	0.4827
599	0.2841	1.1676	0.3492
799	0.3347	1.0131	0.3177
999	0.3331	1.0032	0.3153

Table 7: Quality metrics across using different timesteps without fine-tuning.

C Quantitative Comparisons

C.1 Full-Reference (FR) Metrics

For diffusion-based restoration models, strong generative ability can effectively handle severe degradation in the input image. However, the restored details in degraded regions may differ from the original, resulting in low PSNR and SSIM [18] scores across diffusion-based methods. However, from a perceptual standpoint, restorations that appear much visually closer to the ground truth often achieve consistently higher scores on perceptual metrics such as DISTS [2] and LPIPS [29], rather than on pixel-based metrics like PSNR and SSIM. As shown in Fig. 3, although ResShift [25] achieves higher pixel-level metrics than our HAODiff, its output remains blurred and cannot be regarded as high-quality or highly similar. In contrast, our method produces more complete and coherent details, resulting in significantly better visual quality as human observers perceive.

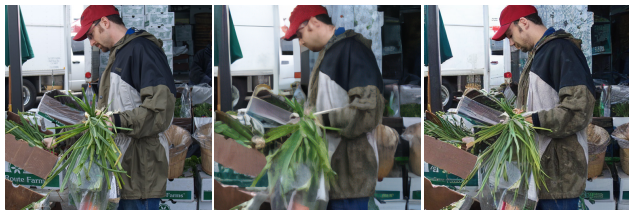


Figure 3: Comparison of restored images and FR metrics.

C.2 More Quantitative Comparisons with Non-Diffusion-Based Methods

We compare HAODiff with several non-diffusion-based methods (see Tab. 6), including blind super-resolution: Real-ESRGAN [16], BSRGAN [28], and FeMaSR [1]; image restoration: SwinIR [10] and Restormer [26]; local deblurring: LBAG [9]; and global deblurring: DeblurGAN [8], UFPNet [3], and AdaRevD [21]. Restormer* and AdaRevD* are retrained on our degradation pipeline and dataset setting. Global and local deblurring methods struggle to produce high-quality reconstructions. Even state-of-the-art methods like AdaRevD perform poorly after retraining, highlighting the gap between standard deblurring and human body restoration involving human motion blur. Overall, our method shows clear advantages over existing non-diffusion-based approaches.

D Qualitative Comparisons

D.1 High-Fidelity Restoration of Fabric Patterns and Textures

We conduct an additional comparative study focused on fabric patterns and textures (shown in Fig. 4), which is an essential aspect of human body restoration. When the low-quality input is not severely degraded, some fabric patterns and textures are still perceptible to the human eye. These details often span fewer than 10 pixels, but latent diffusion models (LDM) [12] typically rely on the variational autoencoder (VAE) [7] that downsample by a factor of 8, making such fine details correspond to only a few pixels in the latent space. This poses a significant challenge for precise fidelity preservation. Current diffusion-based approaches often generate overly smoothed results, leading to a loss of texture fidelity; or produce overly synthesized outputs, resulting in hallucinated textures that deviate from the original appearance. In contrast, our HAODiff can remove degradations in these regions while generating details that are more faithful than those produced by all other methods.

D.2 More Vision Comparisons

Comparisons on PERSONA-Val. Intuitive visual comparisons on the PERSONA-Val dataset are presented in Fig. 5 and Fig. 6. Each group includes the original high-quality (HQ) image, its corresponding low-quality (LQ) version produced by our degradation simulation, and the restoration results from all compared methods. It is evident that HAODiff consistently produces clear and natural outputs when addressing generic degradations, regardless of whether the image features individuals or groups. Building upon this capability, our model further focuses on restoring human-related content within the image, effectively correcting region-specific blur caused by motion. Meanwhile, HAODiff achieves more realistic hair textures and maintains natural facial tones and skin details, avoiding the overly smooth appearance or unnatural artifacts caused by excessive detail enhancement.

Comparisons on PERSONA-Test and MPII-Test. The images (see in Fig. 7 and Fig. 8) from PERSONA-Test are generally less severely degraded, with many LQ samples retaining considerable detail and semantics. These conditions better reveal whether diffusion-based models can remove degradation while preserving content. Most methods produce clear results, but their limitations appear under close inspection. As discussed in Sec. D.1, HAODiff demonstrates superior ability to preserve realistic fabric quality and material textures. Specifically, when faced with densely striped patterns in LQ images, many methods fail to maintain the authentic texture distribution, instead producing overly smoothed or synthesized outcomes. In particular, when certain human body structures are small (*e.g.*, facial features in full-body portraits), some models fail to preserve these details during restoration, resulting in visible distortions. HAODiff, on the other hand, successfully avoids such issues.

In contrast, the MPII-Test dataset consists mostly of heavily degraded images (Fig. 9 and Fig. 10). In this scenario, the goal of restoration is to generate results that are both perceptually plausible and naturally colored, while remaining faithful to the coarse structural cues in the original input. Under complex conditions such as uneven lighting or significant background interference, HAODiff is still capable of generating content with distinct structural layers, demonstrating strong robustness and generalization ability. Furthermore, in images with dynamic human activity and substantial pose variation, other methods tend to produce blurry edges, structural misalignment, or distorted facial features. HAODiff effectively mitigates these problems, ensuring coherent and realistic restoration.

D.3 Challenging Tasks

As illustrated in Fig. 11 and Fig. 12, when images are affected by very severe degradation, the semantic information within them can be significantly reduced, making it difficult even for humans to interpret the content. In such cases, the image restoration task becomes highly challenging. Although diffusion-based models possess strong generative priors, the information provided by such LQ images is often insufficient for reconstructing visually coherent and HQ results. Neither extending the diffusion process [11, 20, 23, 22, 25] nor introducing semantic-level guidance through text prompts [20, 23, 19] proves effective under these conditions.

HAODiff also shows artifacts like unnatural human structures and distorted details in extreme cases. However, it still demonstrates an effective restoration of regions affected by severe human motion blur, outperforming alternatives in preserving structural coherence. HAODiff can produce the clearest and most natural-looking results while retaining the available semantic cues, showcasing its superior restoration capability. These findings prompt a rethink: Although pretrained diffusion models can generate images starting from pure noise, their adaptation to human body restoration remains limited when dealing with severely degraded content. This suggests the upper bound of restoration performance may inherently rely on generation capabilities to bridge the information gap.

E Broader Impacts

Our diffusion-based human-aware restoration model, HAODiff, provides clear societal value by enabling faithful recovery of human appearance from degraded images. It enhances facial and structural realism with efficient single-step inference, supporting applications in photography, video conferencing, and digital preservation. Its lightweight and deterministic design also makes real-time or edge deployment feasible. However, technology may be misused for identity manipulation or unauthorized enhancement of personal images, raising privacy and fairness concerns. Biased data could also lead to unequal performance across demographic groups.

To mitigate these risks, we promote transparent usage, dataset auditing for demographic balance, and watermarking or provenance tracking to discourage misuse. HAODiff is designed to restore authentic degraded data rather than create synthetic identities. With responsible deployment, its benefits for visual communication and cultural preservation should outweigh potential risks.

References

- [1] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACM MM*, 2022. 4
- [2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 4
- [3] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *CVPR*, 2023. 4
- [4] Jue Gong, Jingkai Wang, Zheng Chen, Xing Liu, Hong Gu, Yulun Zhang, and Xiaokang Yang. Human body restoration with one-step diffusion model and a new benchmark. In *ICML*, 2025. 1, 2, 3, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021. 3
- [6] Glenn Jocher and Jing Qiu. Ultralytics YOLO11. [Online]. Available: <https://github.com/ultralytics/ultralytics>, 2024. 1
- [7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5
- [8] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In *CVPR*, 2018. 2, 4
- [9] Haoying Li, Ziran Zhang, Tingting Jiang, Peng Luo, Huajun Feng, and Zhihai Xu. Real-world deep local motion deblurring. In *AAAI*, 2023. 4
- [10] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPRW*, 2021. 4
- [11] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. In *ECCV*, 2024. 3, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5
- [13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4
- [14] Stability AI. stabilityai/stable-diffusion-2-1-base. [Online]. Available: <https://huggingface.co/stabilityai/stable-diffusion-2-1-base>, 2022. 3
- [15] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. One-Step Diffusion Model for Face Restoration. In *CVPR*, 2025. 3
- [16] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 1, 2, 4
- [17] 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. In *CVPR*, 2024. 3, 8, 9, 10, 11, 12, 13, 14, 15, 16

- [18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 2004. 4
- [19] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024. 3, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [20] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 3, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [21] Qingli Li Xintian Mao and Yan Wang. Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring. In *CVPR*, 2024. 4
- [22] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, 2023. 3, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [23] 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. 3, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16
- [24] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *CVPR*, 2025. 8, 9, 10, 11, 12, 13, 14, 15, 16
- [25] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023. 3, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16
- [26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 4
- [27] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arxiv*, 2024. 3
- [28] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 4
- [29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4



Figure 4: Visual comparison of fabric pattern and texture on PERSONA-Test dataset.



Figure 5: Visual comparisons of PERSONA-Val (part 1). Please zoom in for a better view.

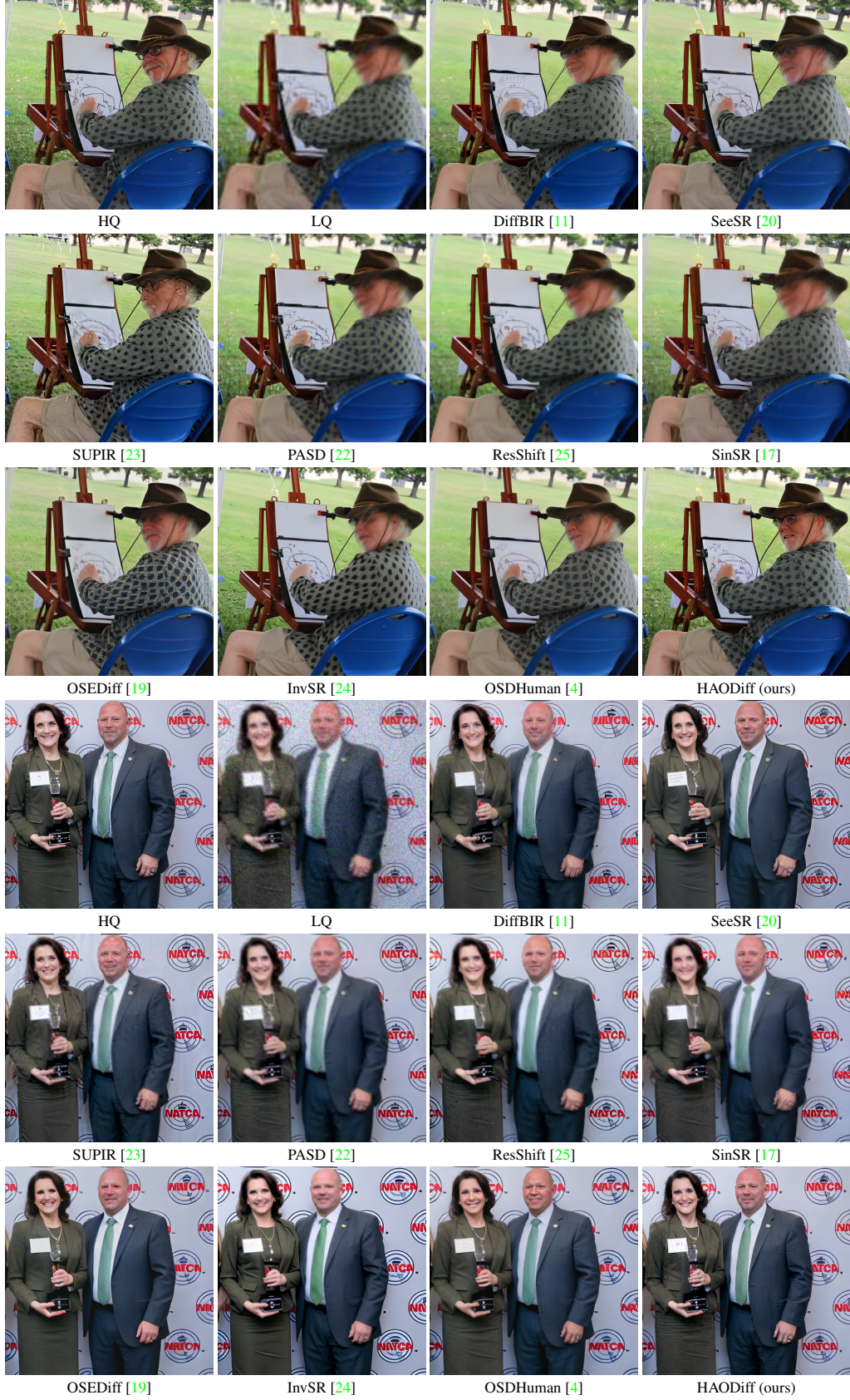


Figure 6: Visual comparisons of PERSONA-Val (part 2). Please zoom in for a better view.

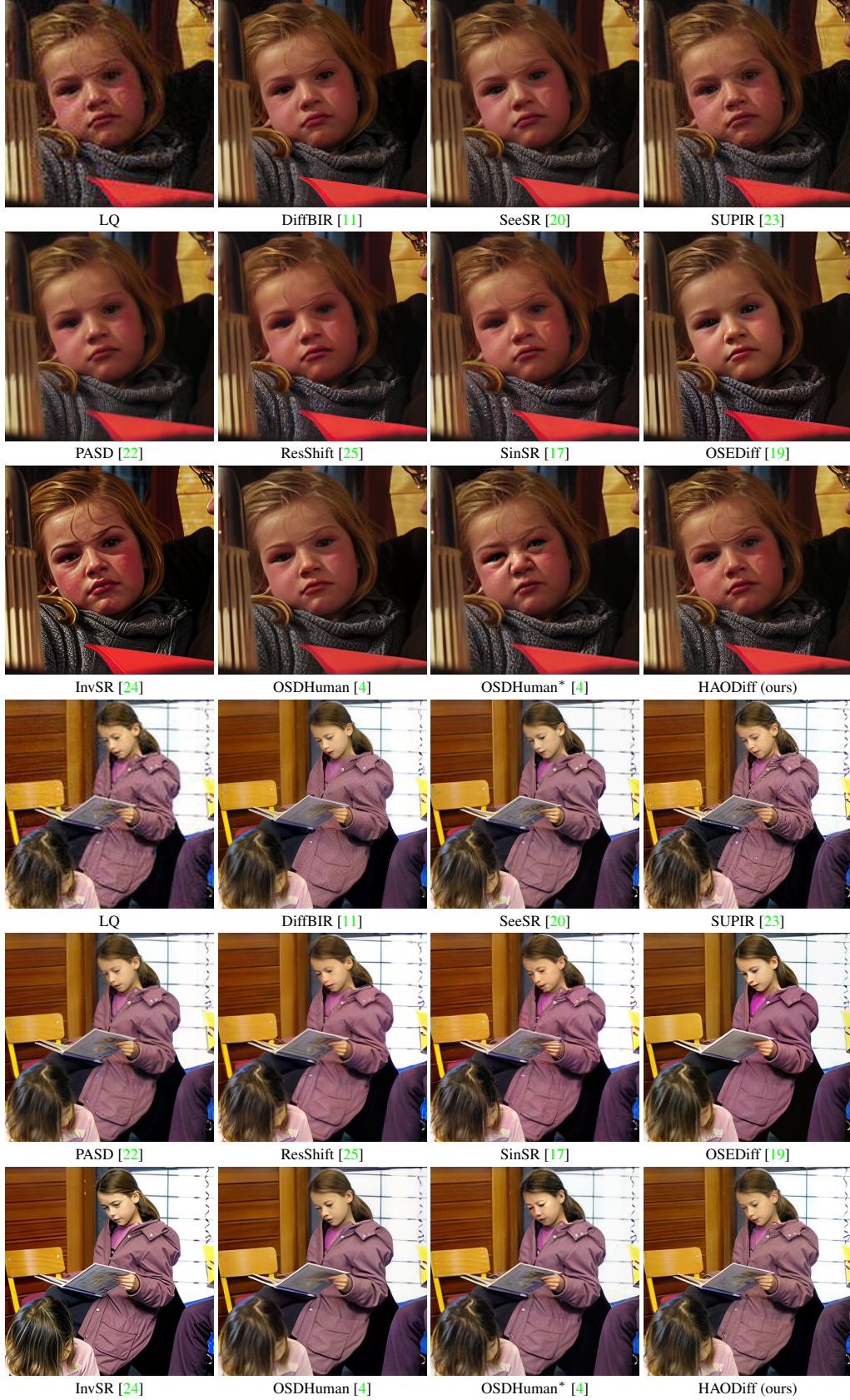


Figure 7: Visual comparisons of PERSONA-Test (part 1). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.

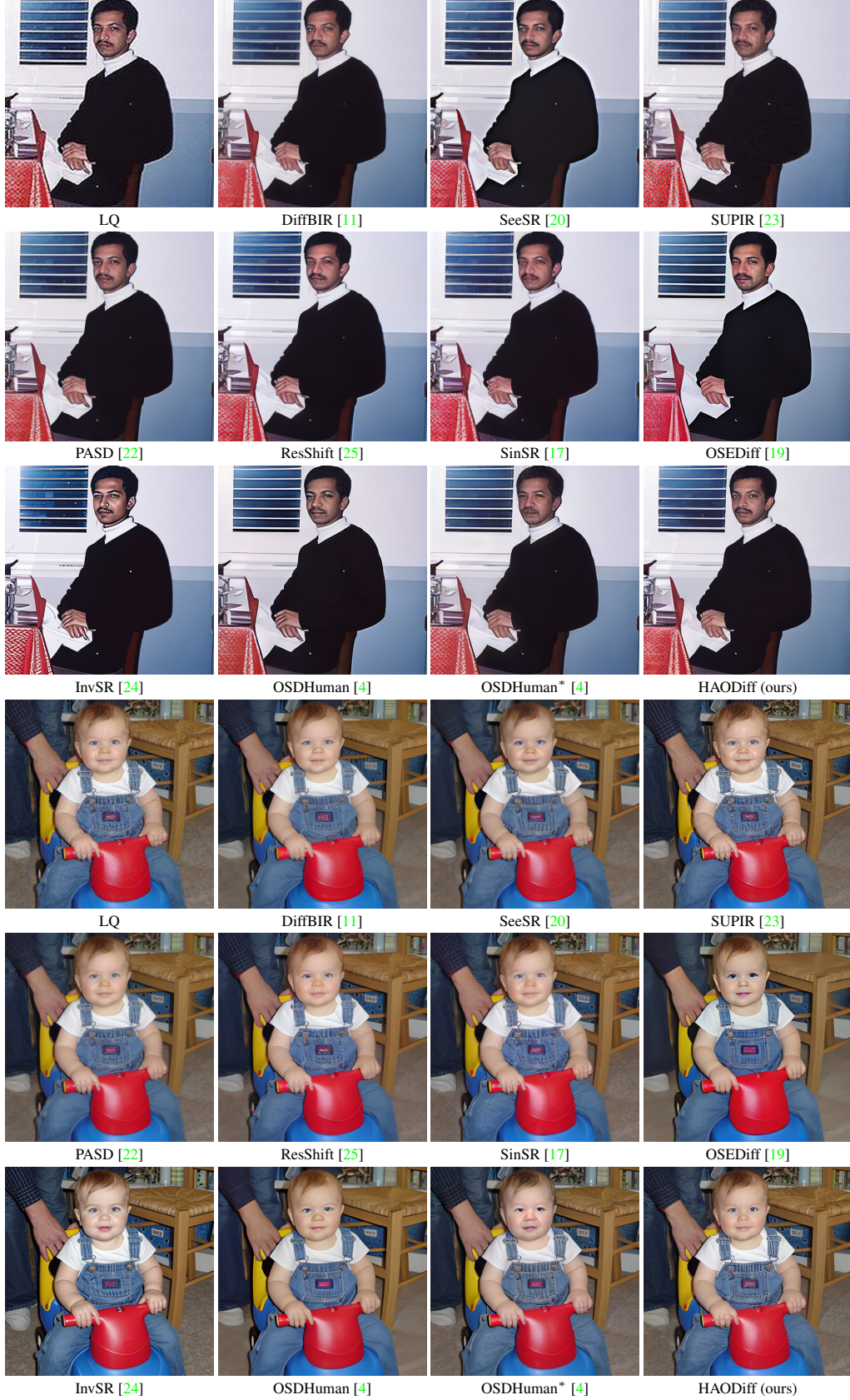


Figure 8: Visual comparisons of PERSONA-Test (part 2). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.

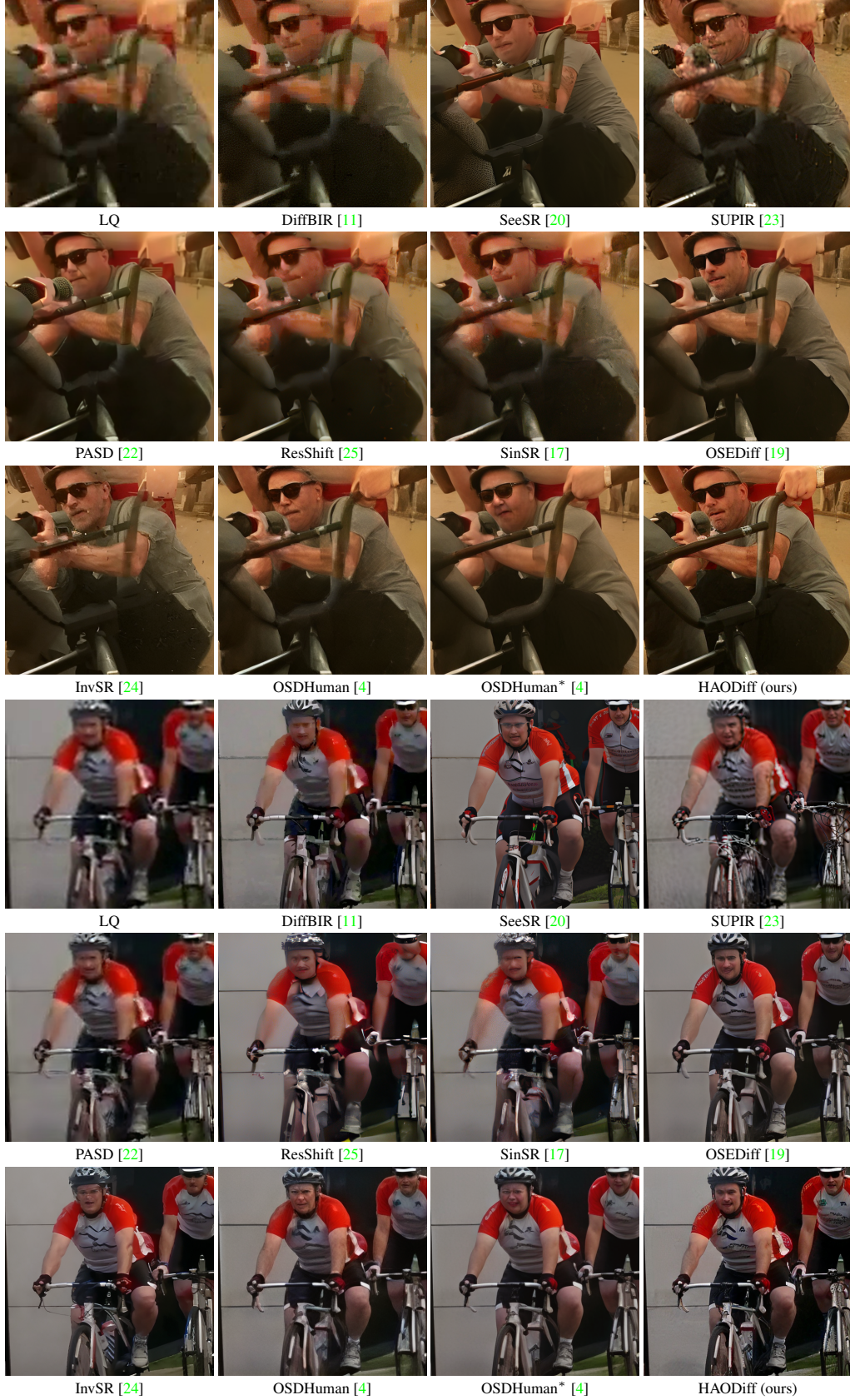


Figure 9: Visual comparisons of MP11-Test (part 1). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.

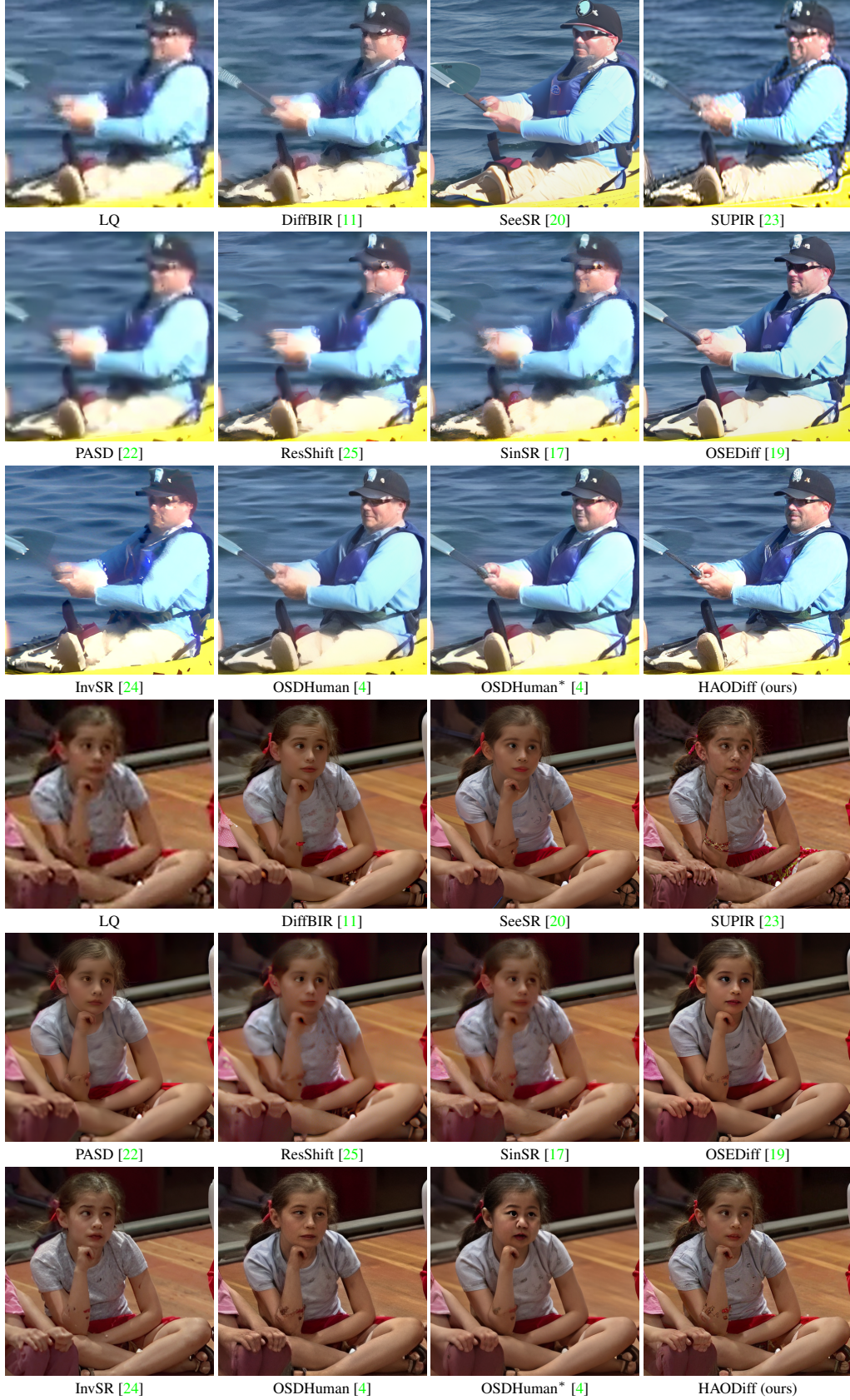


Figure 10: Visual comparisons of MPII-Test (part 2). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.



Figure 11: Visual comparisons of some challenging tasks (part 1). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.

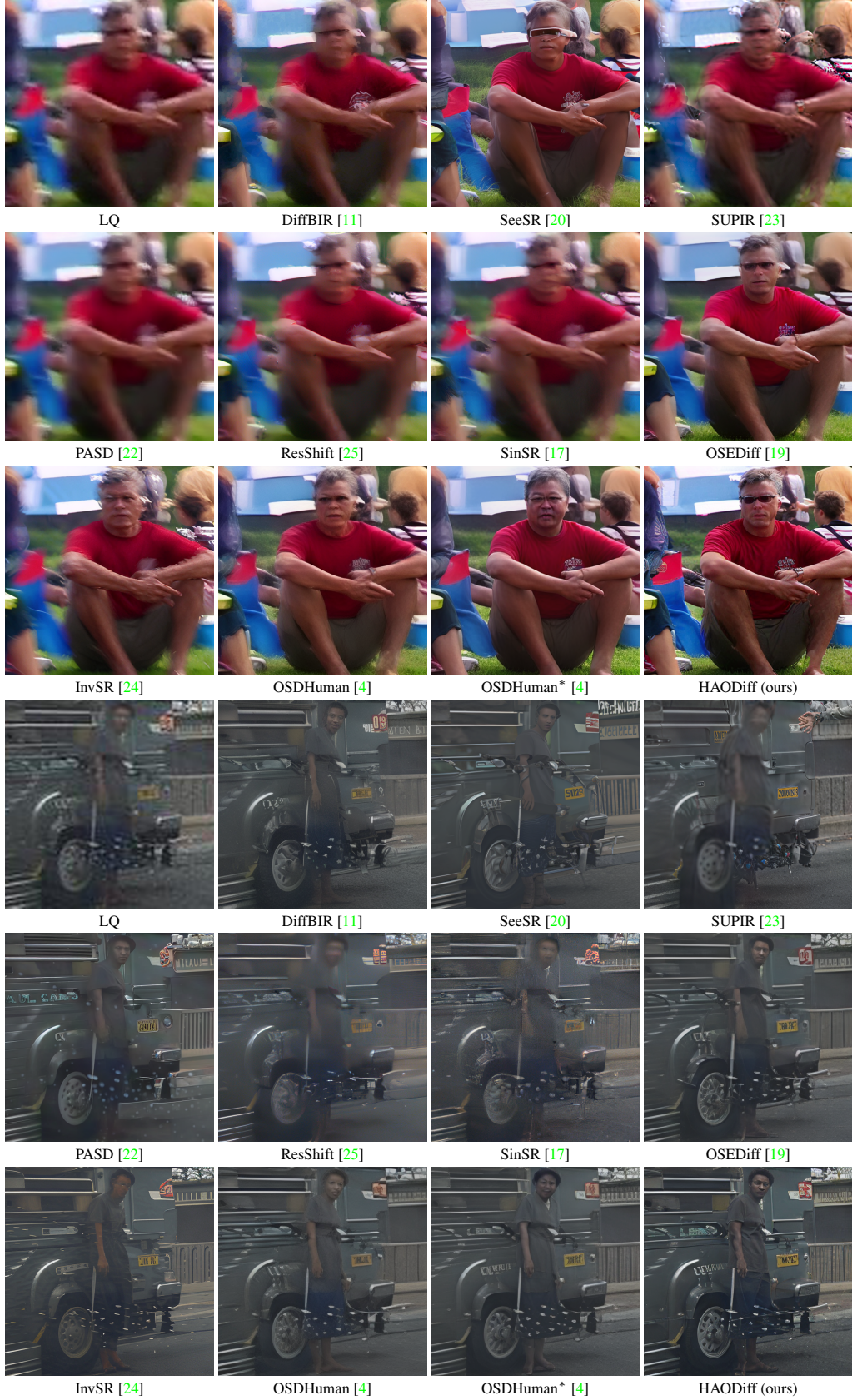


Figure 12: Visual comparisons of some challenging tasks (part 2). OSDHuman* denotes the retrained OSDHuman using our degradation pipeline. Please zoom in for a better view.