# Supplementary Material: Dynamic Attention-Guided Diffusion for Image Super-Resolution

Brian B. Moser[1,2,3] Stanislav Frolov[1,2,3] Federico Raue[1] Sebastian Palacio[1] Andreas Dengel[1,2]
[1]German Research Center for Artificial Intelligence, Germany
[2]RPTU Kaiserslautern-Landau, Germany
[3]Equal Contribution
first.last@dfki.de

In the main text, we presented YODA as an approach to dynamically focus the diffusion process to essential areas in the image. The supplementary material hereby gives further information and visualizations on YODA, such as a discussion on related work, details on DINO, training details, and complexity analysis. It supports understanding the main concepts and ideas examined in the main text.

## Contents



## 1. Related Work

In this section, we discuss other diffusion models that can be applied to YODA and content-aware SR methods that also focus on image content to optimize certain properties in the SR pipeline.

### 1.1. Other State-Of-The-Art Diffusion Models

As shown in the study of Moser et al. [19], many approaches apply to image SR. In this section, we want to discuss their potential in combination with YODA and possible limitations for future work.

**Latent Diffusion Models.** Despite the significant advancements brought by Latent Diffusion Models (LDMs) [22], their efficacy in the realm of image SR competes closely with that of SR3 [23], as shown in Table 1. Unfortunately, recent research in this direction focused primarily on text-to-image tasks [10], which makes further comparisons with image SR methods challenging, e.g., SDXL [20], MultiDiffusion [3], or DemoFusion [9]. Nevertheless, their potential for image SR is undeniable. Concerning YODA, we also see great research avenues in combination with LDMs. A critical prerequisite for this synergy is the conversion of attention maps from pixel to latent representations. This aspect has to be investigated in more detail in future work. Our preliminary StableSR (10 epochs in stage 2) results show that our method improves LPIPS from 0.1242 to 0.1239.

Table 1. $\times 4$ upscaling results on ImageNet-Val. ($256 \times 256$). Values directly derived from the original work of LDM [22].

| **Method** | IS ↑ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|
| Image Regression [23] | 121.1 | **27.9** | **0.801** |
| SR3 [23] | **180.1** | 26.4 | 0.762 |
| LDM-4 (100 steps) [22] | 166.3 | $24.4_{\pm 3.8}$ | $0.69_{\pm 0.14}$ |
| LDM-4 (big, 100 steps) [22] | 174.9 | $24.7_{\pm 4.1}$ | $0.71_{\pm 0.15}$ |
| LDM-4 (50 steps, guiding) [22] | 153.7 | $25.8_{\pm 3.7}$ | $0.74_{\pm 0.12}$ |

**Unsupervised Diffusion Models.** Another interesting research avenue is unsupervised diffusion models for image SR, exemplified by DDRM [13] or DDNM [26]. Interestingly, they use a pre-trained diffusion model to solve any linear inverse problem, including image SR, but they rely on singular value decomposition (SVD). Similar to the challenges faced with LDMs, integrating YODA into unsupervised diffusion models presents another interesting research avenue. The core of this challenge lies in devising a method for effectively translating attention maps into a format compatible with the SVD process used by these models. This transformation is crucial for harnessing the power of attention-based enhancements in unsupervised diffusion frameworks for image SR. Future work needs to conceptualize and implement

a seamless integration strategy that combines the dynamic attention modulation offered by YODA with SVD.

**Alternative Corruption Spaces.** Applying YODA with alternative corruption spaces (not pure Gaussian noise), such as used in InDI [7], $I^2$SB [15], CCDF [6], or ColdDiffusion [2] is also an interesting future research direction. Although our primary focus has been refining and enhancing the diffusion process of specific models through attention-guided masks, we acknowledge the orthogonal potential these approaches represent within the broader context of SR.

### 1.2. Other Content-Aware SR Methods

One category of approaches that draw similarities to YODA is dataset pruning for image SR. Commonly, image SR methods are trained on sub-images cropped from higher-resolution counterparts, such as those found in DIV2K [1]. The central premise behind dataset pruning strategies is the observation that not all sub-images contribute equally to training efficacy. These approaches employ content-aware detectors to prune training data based on metrics like loss values selectively [18] or Sobel norms [8].

A related strategy, ClassSR [24], categorizes different image regions into three levels of reconstruction difficulty - easy, medium, and hard. They propose training specialized models for each category. Related to ClassSR are RAISR [21], SFTGAN [25], RL-Restore [28], and PathRestore [29]. RAISR [21] assigns image patches into clusters and employs a tailored filter to each cluster, utilizing an efficient hashing technique to streamline the clustering process. SFTGAN [25] introduces a spatial feature transform layer that embeds high-level semantic priors, enabling nuanced processing of different image regions with different parameters (i.e., different models). Similarly, RL-Restore [28] and PathRestore [29] divide images into sub-images and employ reinforcement learning to determine the optimal processing pathway for each sub-image. Unlike YODA, which focuses on refining specific image regions with one model, the presented works aim to optimize training datasets by reduction or categorization (thereby employing multiple models tailored to varied reconstruction difficulties).

Similarly, the Multiple-in-One Image Restoration (MiO IR) strategy introduces a novel approach to handling diverse image restoration tasks within a single model [14]. MiO IR employs sequential learning, which allows the model to learn different tasks incrementally and optimize for diverse objectives. Additionally, it utilizes prompt learning - both explicit and adaptive - to guide the model in adapting to various tasks dynamically. While YODA focuses on refining specific image regions, MiO IR's versatility across tasks offers a potential avenue for expanding YODA's application to more generalized scenarios.

SkipDiff [17] presents another content-aware SR approach in the context of diffusion models. This method operates through two primary phases: a coarse skip approximation and a fine skip refinement. SkipDiff constructs a preliminary high-resolution image approximation in the first phase in a single step. In the second phase, this image is refined using the classical diffusion pipeline with an adaptive noise schedule. For this, they employ a schedule driven by the characteristics of the input image. Reinforcement learning is integral to this process, as it is trained to find optimal diffusion steps for this phase. This adaptability enables SkipDiff to tailor diffusion to the entire content of an image, contrasting with YODA's targeted refinement of specific regions. Future research might explore the potential synergies between YODA and SkipDiff, combining their strengths to further enhance content-specific image SR.

Another recent approach in lightweight SR is the Self-Feature Learning (SFL) method proposed by Xiao et al. [27]. This method introduces a locally adaptive involution technique that reduces computational costs by dynamically generating convolutional kernels based on local image content. The SFL model achieves a remarkable trade-off between performance and model complexity by avoiding inter-channel redundancy, making it particularly suitable for resource-constrained devices. Unlike YODA, which focuses on targeted region refinement, SFL employs a dual-path residual module to ensure efficient feature extraction across the entire image, potentially complementing YODA's targeted strategy.

## 2. Details: DINO

DINO [5] is a self-supervised learning approach. involving a teacher and student network. While both networks share the same architecture, their parameters differ. The student network is optimized to match the teacher's output via cross-entropy loss. During training, both receive two random views of the same input image: the teacher is trained on global views, i.e., $224 \times 224$ crops, while the student receives local views, i.e., $96 \times 96$ crops. This setup encourages the student to learn "local-to-global" correspondences. In other words, by predicting the teacher's output, the student learns to infer global information from local views. To prevent mode collapse, the teacher's parameters are updated as a moving average of the student's parameters.

## 3. Details: Training Parameters

For SR3, we adopted the AdamW [16] optimizer, using a weight decay of 0.0001 and a learning rate of 5e-5. The number of sampling steps is set to $T_{\text{train}} = 500$. The number of sampling steps is set to $T_{\text{eval}} = 200$. Concerning the denoising architecture, our approach aligns with the SR3 model [23], but we employed residual blocks [11] proposed by Ho et al. instead of those used in BigGAN [4, 12]. Specifically, the configuration includes three ResBlocks, an initial channel size of 64, and a channel multiplier array of [1, 2,

4, 8, 8]. We also employed a NormGroup with a size of 32. For SRDiff, we extracted $40 \times 40$ sub-images with a batch size of 16, AdamW [16], a channel size of 64 with channel multipliers [1, 2, 2, 4] and $T = 100$.

## 4. Details: Complexity of YODA

We discuss the resource implications of the core components of YODA: Identifying key regions, time-dependent masking, and the guided diffusion process. Additionally, we explore potential avenues for future enhancements aimed at optimizing computational efficiency.

**Identifying Key Regions.** To avoid the computational burden of on-the-fly generation, we pre-compute the attention maps prior to training. Table 2 shows the parameter count and throughput of different DINO backbones. While YODA introduces additional complexity for setting up the attention maps, the overhead is minimal. For instance, generating all attention maps for our face SR experiments (i.e., 120,000 images) needed less than two minutes.

**Time-Dependent Masking and Guided Diffusion Process.** The integration of attention masks within the diffusion framework introduces minimal computational overhead, thanks to the inherently parallelizable nature of element-wise multiplication and addition, as demonstrated in the methodology (see time-dependent masking). Consequently, the predominant factor influencing the overall computational complexity remains the choice of diffusion model, whether it be SR3, SRDiff, or another variant.

**Potential Future Improvements.** YODA notably decreases computational requirements by enabling the use of smaller batch sizes during training, which in turn reduces VRAM usage without compromising performance. Looking ahead, YODA paves the way for leveraging sparse diffusion techniques. Such approaches promise further computational savings by focusing computation efforts on selectively identified regions (through YODA), thereby streamlining the diffusion process. Currently, in PyTorch, applying masks to regions within a matrix does not result in computational savings.

Table 2. Details of different DINO backbones, values directly extracted from the original work [5]. Throughput was measured with a NVIDIA V100 GPU.

| Model | Parameters [M] | Throughput [img/s] |
|---|---|---|
| ResNet-50 | 23M | 1237 |
| ViT-S/8 | 21M | 180 |

This examination of YODA's complexity highlights its efficiency and the strategic decisions made to balance performance with computational demands.

## 5. Details: Analysis Across Attention Regions

This section explains how we created the figure associated with the subsection **Analysis across attention regions**. The used attention maps are derived from DINO+ResNet. We apply the max aggregation explained in the main paper to create a single attention map per image. The aggregated map is then divided into several attention-value ranges (bins/intervals). Specifically, the attention values range from 0 to 1 (see methodology), and we divide this range into small bins with a step size of 0.01. For each bin, we analyze the regions of the image where the attention values fall within the bin range (e.g., 0.01-0.02, 0.02-0.03, etc.). This allows us to observe how different attention levels correlate with the LPIPS scores and how each SR model performs in these non-overlapping and separated regions.

For each attention value bin, we calculate the LPIPS score between the reference HR image and the output images generated by the different SR techniques (SR3, SR+YODA, and LR/bicubic upsampling). This is done as follows:

- **Mask Creation:** For each attention bin, we create a binary mask where pixels in the attention map that fall within the bin are set to 1, and all others are set to 0.
- **Region Extraction:** Using this binary mask, we extract the corresponding regions from the HR and SR images.
- **LPIPS Calculation:** We compute the LPIPS score between the HR image's masked regions and the generated images' masked regions. This process is repeated for each bin across the entire attention range. Masked-out regions do not influence LPIPS because LPIPS is a distance measure of features and masked-out regions in HR, and any SR image leads to an LPIPS of 0.

We calculate the mean LPIPS error within attention bins across the images. A polynomial curve (degree 3) is fitted to the LPIPS scores for each model across the attention value ranges to visualize trends more clearly. This allows us to smooth out potential noise and outliers in the data and observe how each model's performance changes as we move from high to low-attention regions. Key observations:

- **Higher LPIPS in High Attention Regions:** Bicubic upsampling performs poorly in high-attention regions, as indicated by the high LPIPS values. This shows that detail-rich and perceptually important regions are indeed reflected by attention values, as simple upsampling cannot adequately capture the fine details.
- **YODA's Improvement:** YODA shows the most significant improvement in high-attention regions, confirming that YODA refines detail-rich regions more effectively. Moreover, it also shows that YODA is improving every region, thereby truly super-resolving every pixel instead of just replacing the LR and leaving it unchanged.

As DINO+ResNet starts to refine the whole image without dynamic masking at around 0.6, the figure stops there (as it investigated all separated regions by then).

## 6. More Visualizations

In the remaining part of the supplementary material, we provide additional visualizations, such as more on attention maps derived from different DINO heads (Figure 1), user study (Figure 4), an overview of the working pipeline of YODA (Figure 3), error maps (Figure 5), comparisons on $16 \rightarrow 128$ (Figure 2 and Figure 6), zoomed-in comparisons (Figure 7), intermediate results (Figure 8) as well as intermediate binary masking (Figure 9).

## References


[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshop*, 2017. 2

[2] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *NeurIPS*, 36, 2024. 2

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 1

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[6] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022. 2

[7] Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *arXiv preprint arXiv:2303.11435*, 2023. 2

[8] Qingtang Ding, Zhengyu Liang, Longguang Wang, Yingqian Wang, and Jungang Yang. Not all patches are equal: Hierarchical dataset condensation for single image super-resolution. *IEEE Signal Processing Letters*, 2023. 2

[9] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. *arXiv preprint arXiv:2311.16973*, 2023. 1

[10] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209, 2021. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[13] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 35:23593–23606, 2022. 1

[14] Xiangtao Kong, Chao Dong, and Lei Zhang. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379*, 2024. 2

[15] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I $^2$ sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023. 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2, 3

[17] Xiaotong Luo, Yuan Xie, Yanyun Qu, and Yun Fu. Skipdiff: Adaptive skip diffusion model for high-fidelity perceptual image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4017–4025, 2024. 2

[18] Brian B Moser, Federico Raue, and Andreas Dengel. A study in dataset pruning for image super-resolution. *arXiv preprint arXiv:2403.17083*, 2024. 2

[19] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution and everything: A survey. *arXiv preprint arXiv:2401.00736*, 2024. 1

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[21] Yaniv Romano, John Isidoro, and Peyman Milanfar. Raisr: rapid and accurate image super resolution. *IEEE Transactions on Computational Imaging*, 3(1):110–125, 2016. 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[23] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. In *IEEE TPAMI*, 2022. 1, 2

[24] Shizun Wang, Jiaming Liu, Kaixin Chen, Xiaoqi Li, Ming Lu, and Yandong Guo. Adaptive patch exiting for scalable single image super-resolution. In *European Conference on Computer Vision*, pages 292–307. Springer, 2022. 2

[25] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018. 2

[26] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 1

[27] Jun Xiao, Qian Ye, Rui Zhao, Kin-Man Lam, and Kao Wan. Self-feature learning: An efficient deep lightweight network for image super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4408–4416, 2021. 2

[28] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *CVPR*, pages 2443–2452, 2018. 2

[29] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. *IEEE TPAMI*, 44(10):7078–7092, 2021. 2

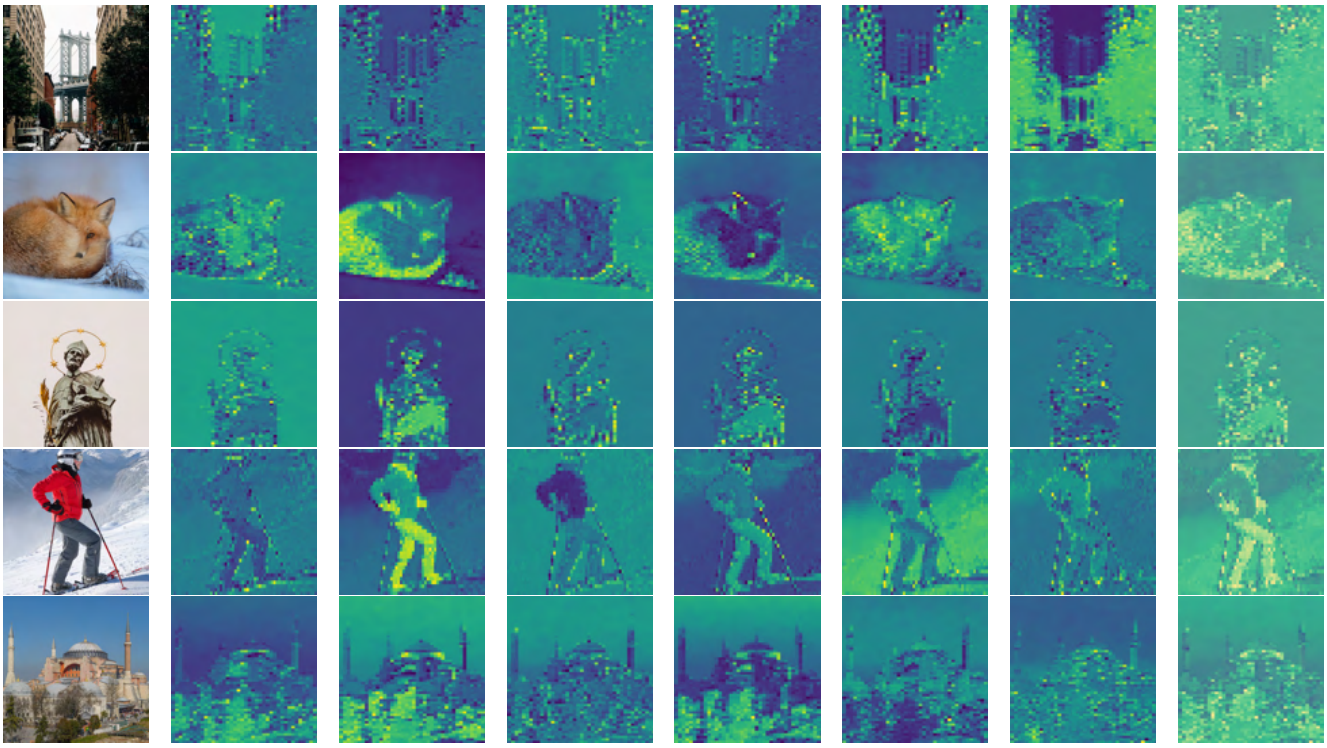


Figure 1. Comparison of attention maps derived from different DINO heads.

| LR | SR3 | SR3+YODA | HR |
|---|---|---|---|
| ↑ PSNR: | 23.061 | 23.289 | |
| ↑ SSIM: | 0.6208 | 0.6334 | |
| ↓ LPIPS: | 0.0504 | 0.0502 | |

Figure 2. SR3 and SR3+YODA reconstructions, 16 → 128 (8x). The color shift in SR3 can still be observed (e.g., see background). YODA produces higher-quality images without color shift issues.

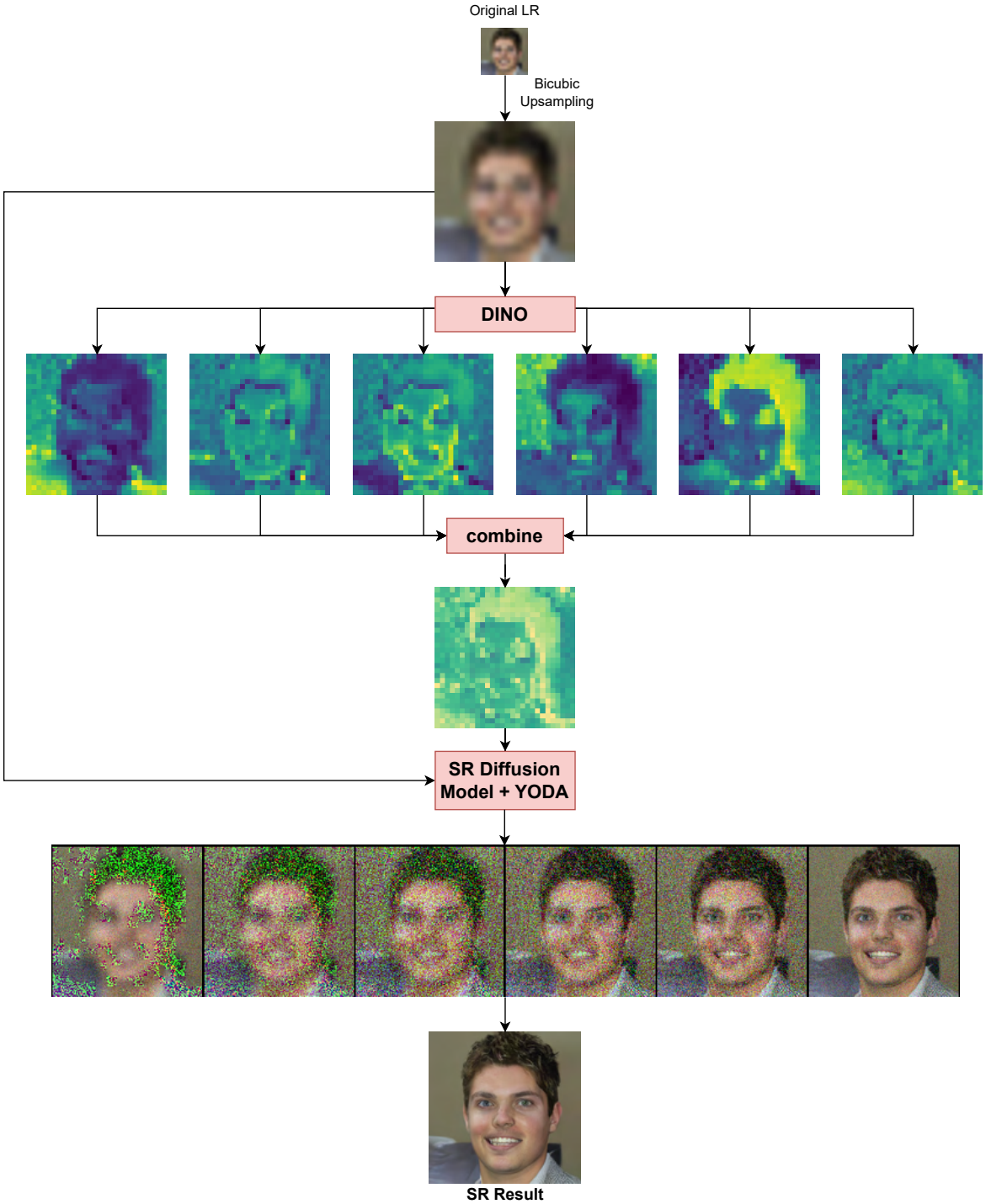


Figure 3. An Overview of integrating YODA and DINO within SR diffusion models. Our process begins with using DINO to extract multiple attention maps. These maps are then combined to form a singular comprehensive attention map, denoted as $\mathbf{A}$. Subsequently, leveraging $\mathbf{A}$, YODA defines a unique diffusion schedule through time-dependent masks $\mathbf{M}(t)$ for every spatial location, as detailed within our method section.

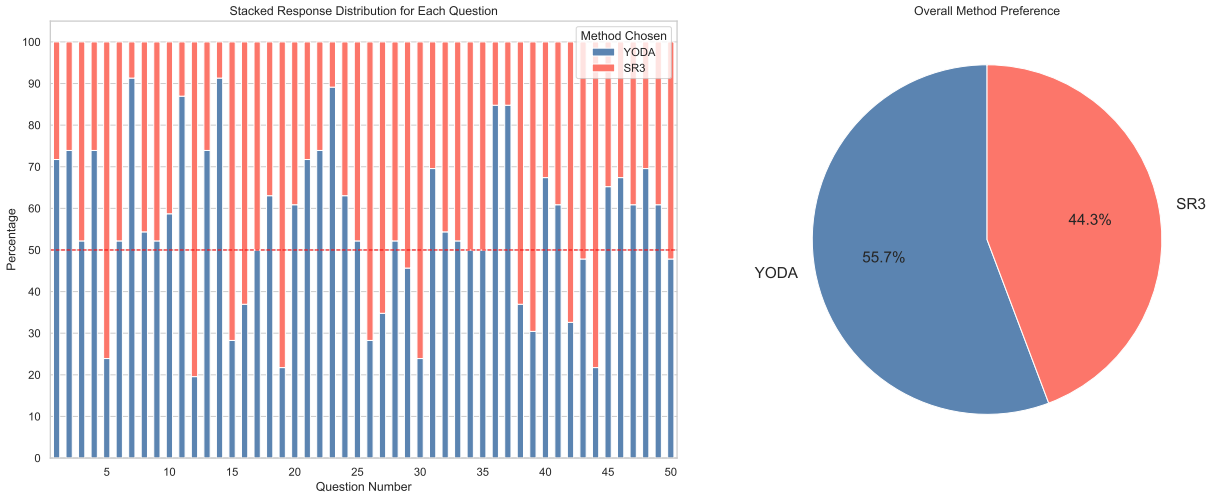


Figure 4. Results of our user study. We randomly selected 50 images from the CelebA-HQ dataset and let 45 participants decide which SR predictions are preferred with respect to the given LR image. The scaling of the tested SR prediction task was $16 \times 16 \rightarrow 128 \times 128$ (8x scaling). As a result, YODA+SR3 was preferred in 55.7% of all cases.

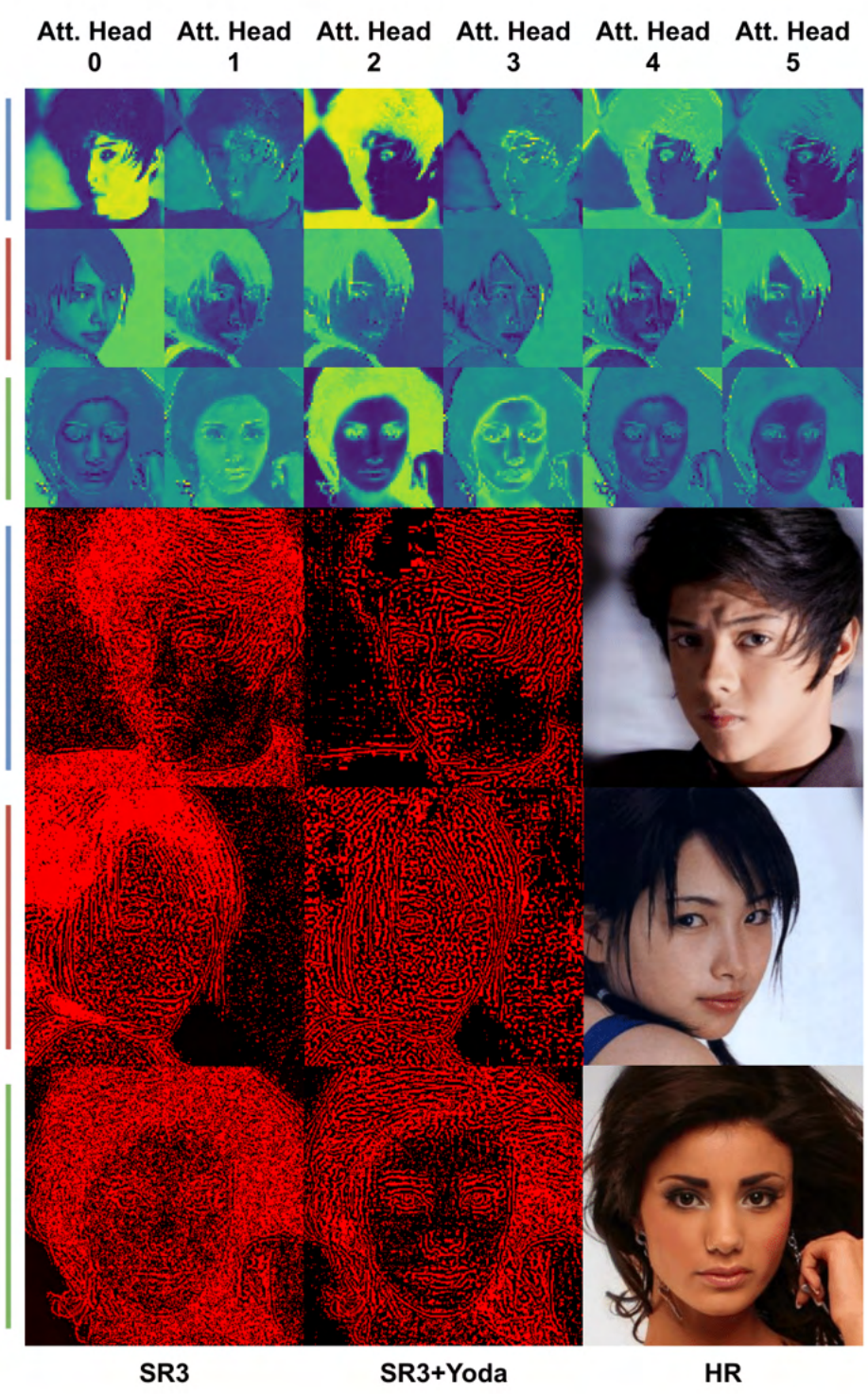


Figure 5. Error Maps of SR3 with and without YODA. The brighter, the higher the error. The attention maps generated by DINO are shown above. YODA produces smaller errors, especially in important areas such as face details and hair.

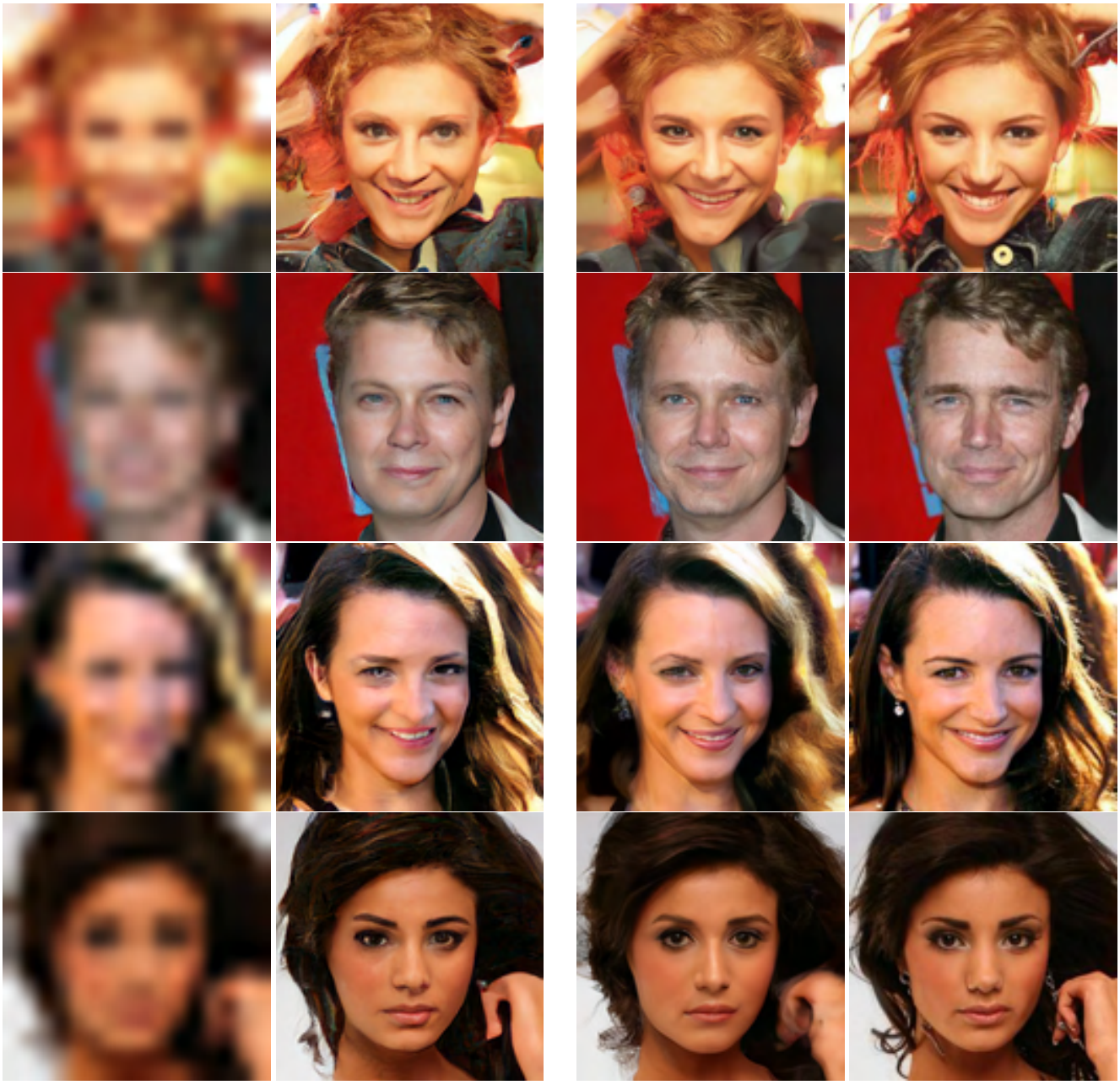


Figure 6. A comparison of LR, HR, SR3, and SR3+YODA images for $16 \rightarrow 128$.

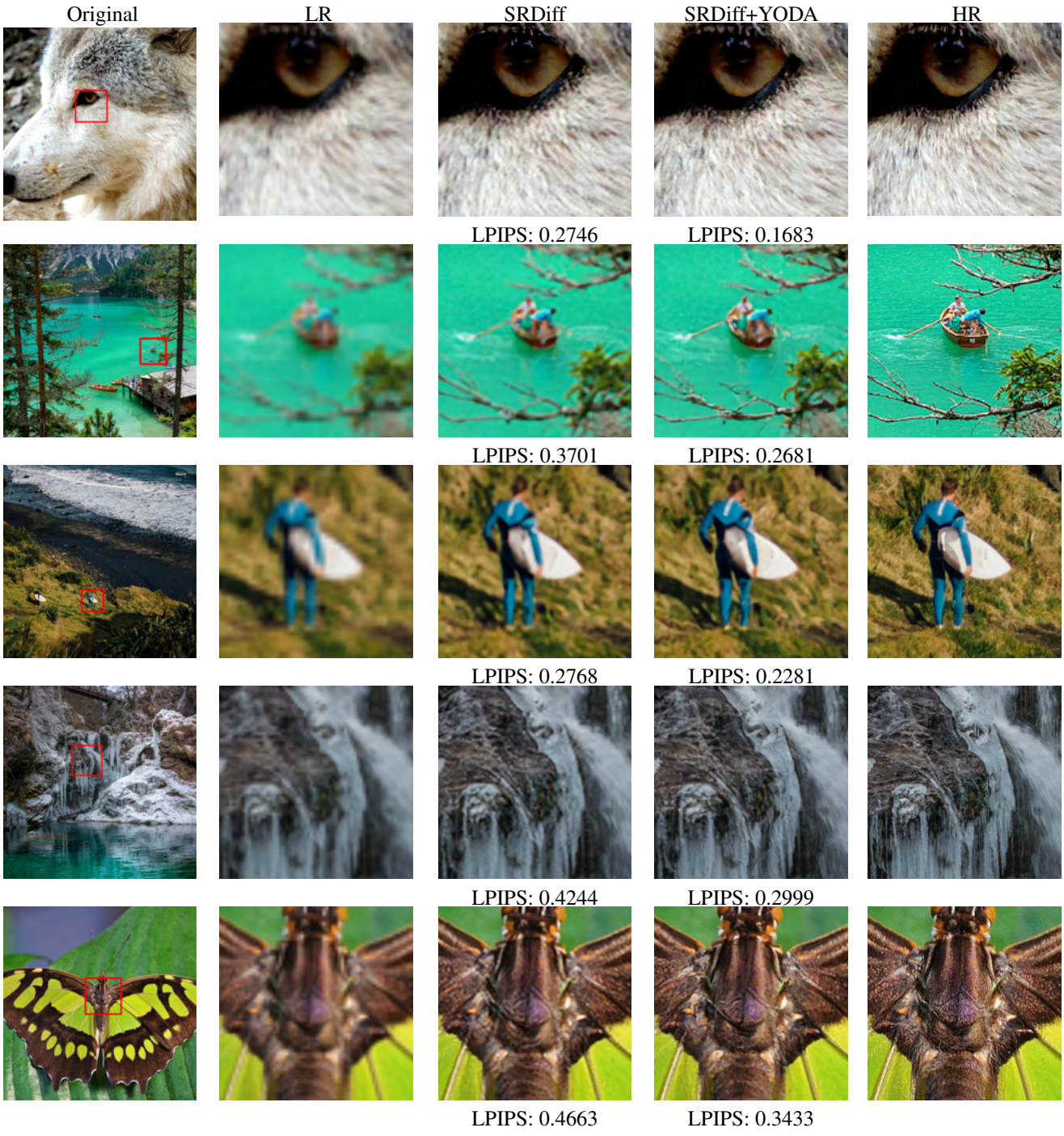


Figure 7. Zoomed-in comparison of LR, HR, SRDiff, and SRDiff+YODA on DIV2K.

Figure 8. Intermediate results of the YODA's guided diffusion process on CelebA-HQ (Time steps 189, 168, 147, 126, 105, 84, 63, 42, 21, 0 from left to right).


Figure 9. Intermediate binary masks of YODA's guided diffusion process on CelebA-HQ.