Appendix

A. Discussions

A.1. Degradation-Robust Encoder

As shown in Fig. 2 of the main text, a degradation-robust encoder is trained and deployed prior to feeding the low-quality input into the adaptor. We conduct experiments using synthetic data to demonstrate the effectiveness of the proposed degradation-robust encoder. In Fig. 14, we show the results of using the same decoder to decode the latent representations from different encoders. It can be seen that the original encoder has no ability to resist degradation and its decoded images still contain noise and blur. The proposed degradation-robust encoder can reduce the impact of degradation, which further prevents generative models from misunderstanding artifacts as image content [50].

A.2. LLaVA Annotation

Our diffusion model is capable of accepting textual prompts during the restoration process. The prompt strategy we employ consists of two components: one component is automatically annotated by LLaVA-v1.5-13B [52], and the other is a standardized default positive quality prompt. The fixed portion of the prompt strategy provides a positive description of quality, including words like "`cinematic, High Contrast, highly detailed, unreal engine, taken using a Canon EOS R camera, hyper detailed photo-realistic maximum detail, 32k, Color Grading, ultra HD, extreme meticulous detailing, skin pore detailing, hyper sharpness, perfect without deformations, Unreal Engine 5, 4k render`". For the LLaVA component, we use the command "`Describe this image and its style in a very detailed manner`" to generate detailed image captions, as exemplified in Fig. 18. While occasional inaccuracies may arise, LLaVA-v1.5-13B generally captures the essence of the low-quality input with notable precision. Using the reconstructed version of the input proves effective in correcting these inaccuracies, allowing LLaVA to provide an accurate description of the majority of the image's content. Additionally, SUPIR is effective in mitigating the impact of potential hallucination prompts, as detailed in [53].

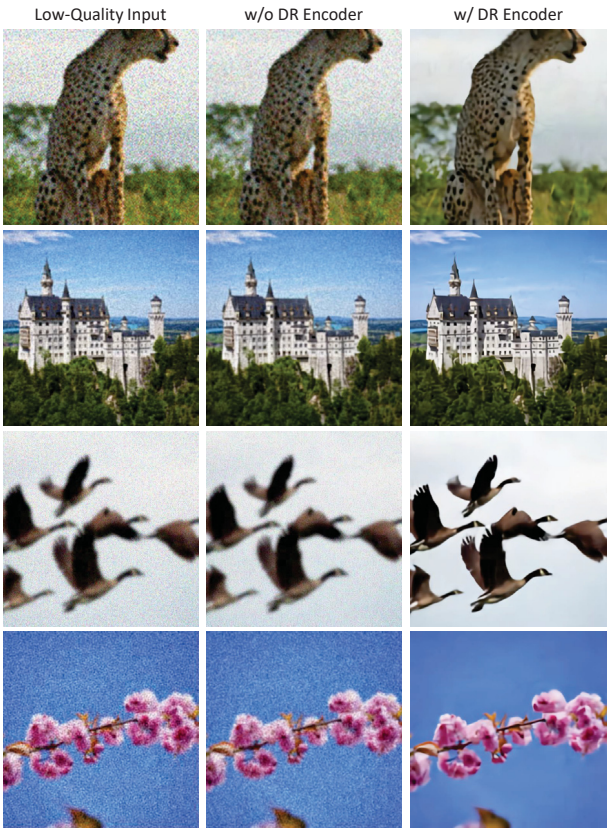


Figure 14. The effectiveness of the degradation-robust encoder (DR Encoder) is demonstrated by the results, which are achieved by initially encoding with various encoders and subsequently decoding. This process effectively reduces the degradations in low-quality inputs before they are introduced into the diffusion models.

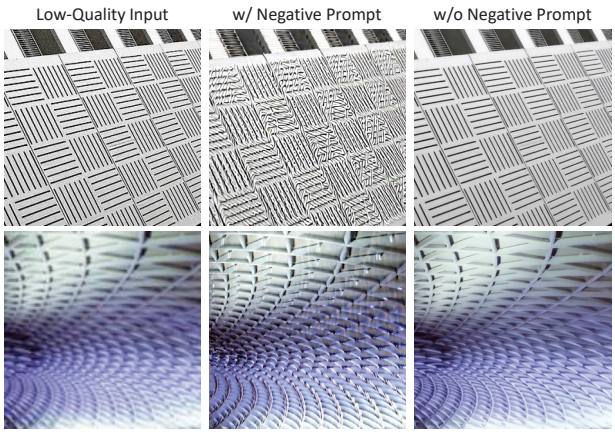


Figure 15. Negative prompt causes artifacts when low-quality inputs do not have clear semantics.

A.3. Limitations of Negative Prompt

Figure 23 presents evidence that the use of negative quality prompts [31] substantially improves the image quality of restored images. However, as observed in Fig. 15, the negative prompt may introduce artifacts when the restoration target lacks clear semantic definition. This issue likely stems from a misalignment between low-quality inputs and language concepts.

A.4. Negative Samples Generation

While negative prompts are highly effective in enhancing quality, the lack of negative-quality samples and prompts in

the training data results in the fine-tuned SUPIR's inability to comprehend these prompts effectively. To address this problem, in Sec. 3.2 of the main text, we introduce a method to distill negative concepts from the SDXL model. The process for generating negative samples is illustrated in Fig. 19. Direct sampling of negative samples through a text-to-image approach often results in meaningless images. To address this issue, we also utilize training samples from our dataset as source images. We create negative samples in an image-to-image manner as proposed in [57], with a strength setting of 0.5.

B. More Visual Results

We provide more results in this section. Fig. 16 presents additional cases where full-reference metrics do not align with human evaluation. In Fig. 17, we show that using negative-quality prompt without including negative samples in training may cause artifacts. In Figs. 20 to 22, we provide more visual caparisons with other methods. Plenty of examples prove the strong restoration ability of SUPIR and the most realistic of restored images. More examples of controllable image restoration with textual prompts can be found in Fig. 23.

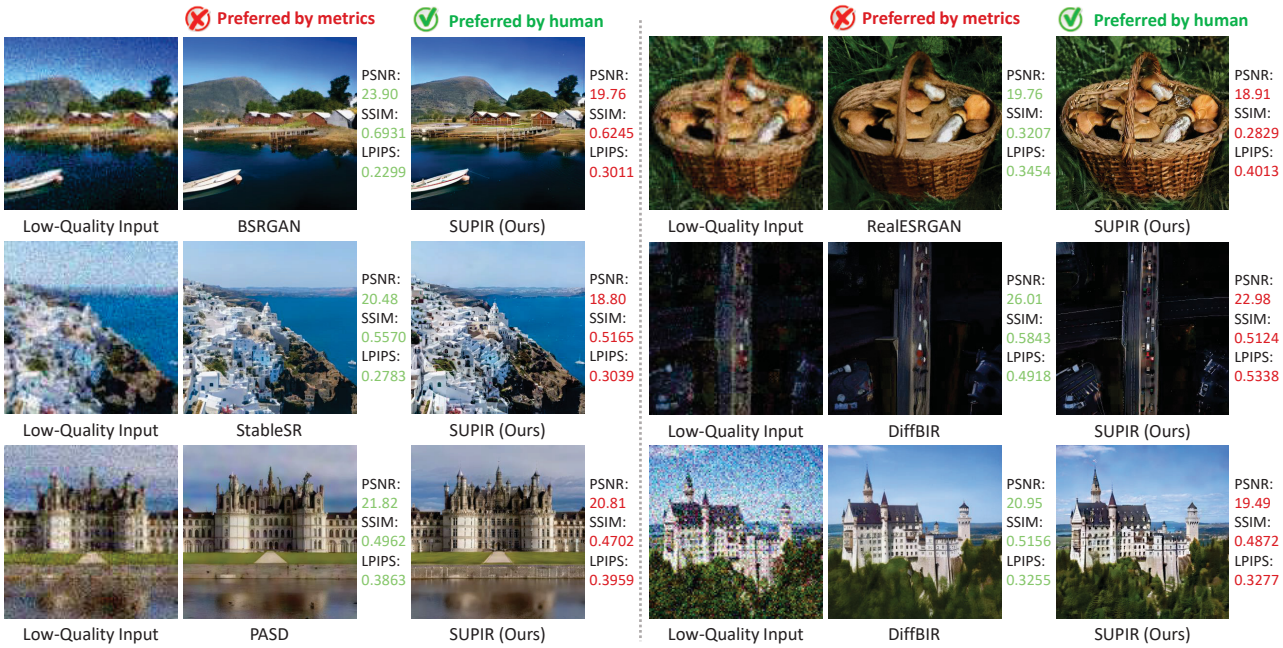


Figure 16. Additional samples highlight the misalignment between metric evaluations and human assessments. While SUPIR produces images with high-fidelity textures, it tends to receive lower scores in metric evaluations.

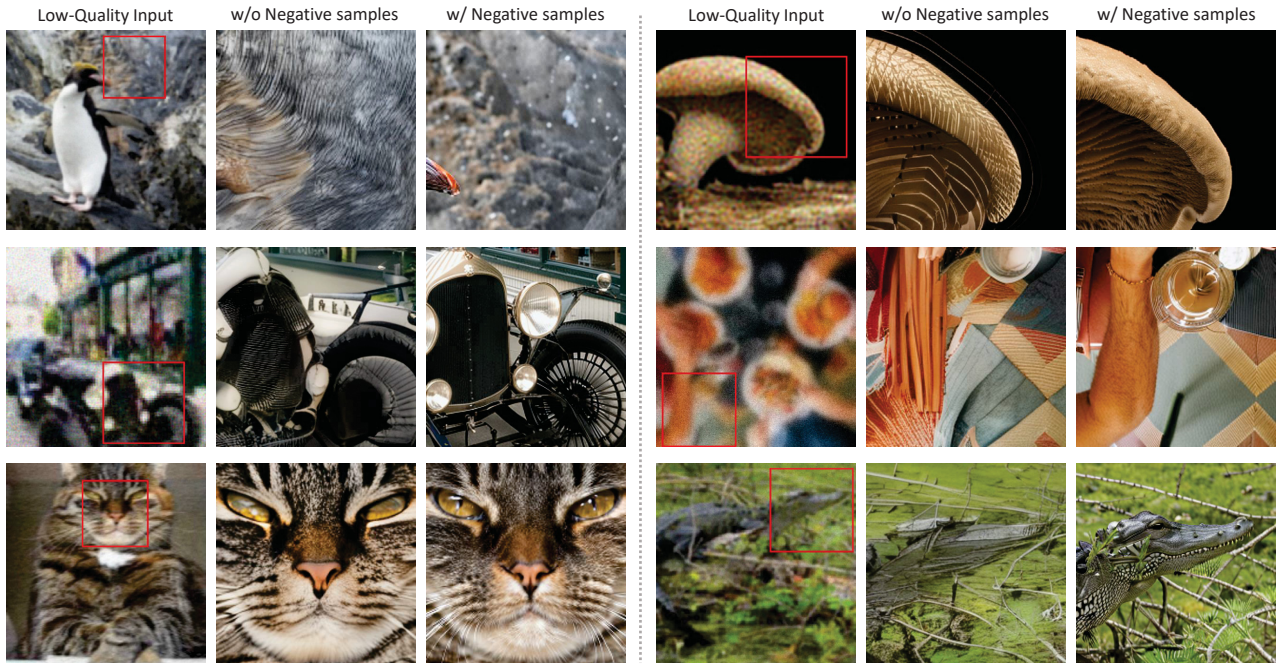
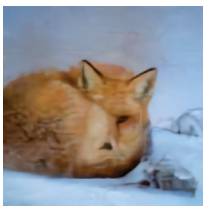


Figure 17. More visual results for Fig. 4 of the main text. CFG introduces artifacts if we do not include negative-quality samples in training. Adding negative-quality samples allows further quality enhancement through CFG.

Case 1

Low-Quality Input

The image features a small, brown, furry animal, possibly a **hamster** or a **cat**, laying on a white surface. The animal appears to be relaxed and comfortable, as it is resting on a **bed** or a **cushion**. The white background provides a contrast to the brown fur of the animal, making it the focal point of the image.



Reconstructed Input

The image features a **fox** lying down in the snow, appearing to be sleeping or resting. The fox is positioned in the center of the scene, with **its body occupying a significant portion of the image**. The **snowy** background provides a serene and natural setting for the fox, emphasizing its natural habitat.

Ground-Truth

The image features a **fox** curled up in the snow, resting comfortably. The fox is positioned in the center of the scene, with **its body facing the left side of the image**. The **snowy** environment provides a natural and serene backdrop for the fox, which appears to be enjoying its time in the snow.

Case 2

Low-Quality Input

The image features a beautiful garden with a variety of colorful flowers. The flowers are arranged in a visually appealing manner, creating a vibrant and lively atmosphere. The garden is filled with different types of flowers, each with unique colors and shapes, contributing to the overall beauty of the scene...

Reconstructed Input

The image features a beautiful flower garden with a variety of colorful flowers. The **main focus is on a large, bright orange flower with a green center**, surrounded by other vibrant flowers. The orange flower is situated in the middle of the garden, drawing attention to its striking color and unique shape. The garden is filled with a diverse array of flowers, creating a visually stunning and lively atmosphere.

Ground-Truth

The image features a vibrant garden with **a large, colorful flower in the center**. The flower is surrounded by a variety of other flowers, creating a beautiful and lively scene. The main flower is **orange** and **yellow**, with a **green stem**, and it stands out among the other flowers in the garden. The garden is filled with a diverse assortment of flowers...

Case 3

Low-Quality Input

The image features a group of **three** men dressed in military uniforms, sitting on the ground and holding guns. They are positioned in a line, with one man on the left, another in the middle, and the third on the right. Each of them is holding a rifle, with one rifle located on the left side, another in the middle, and the third on the right side of the group. The men appear to be soldiers...

Reconstructed Input

The image features a group of **four** men dressed in military uniforms, standing in a **grassy area**. They are all holding guns, with some of them also wearing **backpacks**. The men appear to be soldiers, possibly on a mission or training exercise. The scene captures their readiness and focus as they prepare for their task.

Ground-Truth

The image features a group of **four** men dressed in military uniforms, standing in a **field** and holding guns. They appear to be soldiers, possibly engaged in a training exercise or a mission. The men are positioned in a line, with one soldier on the left, another in the middle, and two more on the right side of the image. Each soldier is holding a gun...

Case 4

Low-Quality Input

The image features a person standing in a field of tall grass, surrounded by tall corn plants. The person appears to be wearing a **backpack**, and their head is partially hidden by the tall grass. The scene gives off a sense of adventure and exploration, as the person seems to be navigating through the field. The tall grass and corn plants create a sense of depth and natural...

Reconstructed Input

The image features a person standing in a field of tall corn, with the corn surrounding them on all sides. The person is wearing a **hat** and appears to be looking down, possibly observing the corn or the ground. The field is vast, with the corn reaching up to the person's shoulders, creating a sense of being engulfed by the tall plants. The scene captures the essence of being in a cornfield...

Ground-Truth

The image features a man wearing a **hat** and a **plaid shirt**, standing in a field of tall corn. He appears to be looking over the corn, possibly observing the surroundings or searching for something. The corn is quite tall, reaching up to the man's shoulders, and the field extends in the background, creating a sense of depth and vastness. The man's presence in the field, along with his attire...

Figure 18. Snapshots showcasing LLaVA annotations demonstrate that LLaVA accurately predicts most content even with low-quality inputs. Please zoom in for a more detailed view.

Figure 19. Pipeline of negative sample generation. (a) Sampling in a noise-to-image approach leads to meaningless outputs. (b) We synthetic negative samples from high quality images. Zoom in for better view.

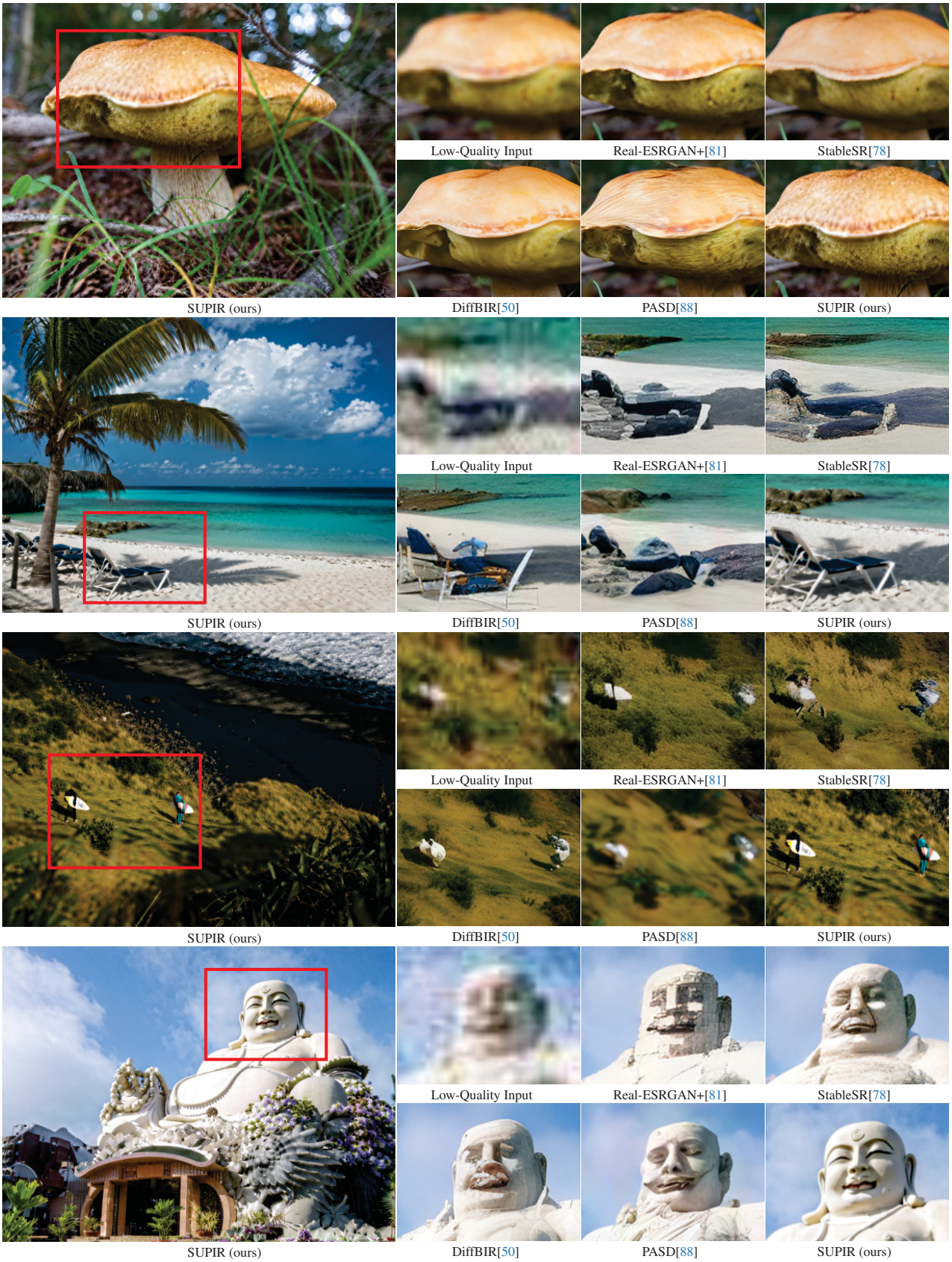


Figure 20. Qualitative comparison with different methods. Our method can accurately restore the texture and details of the corresponding object under challenging degradation. Zoom in for better view.

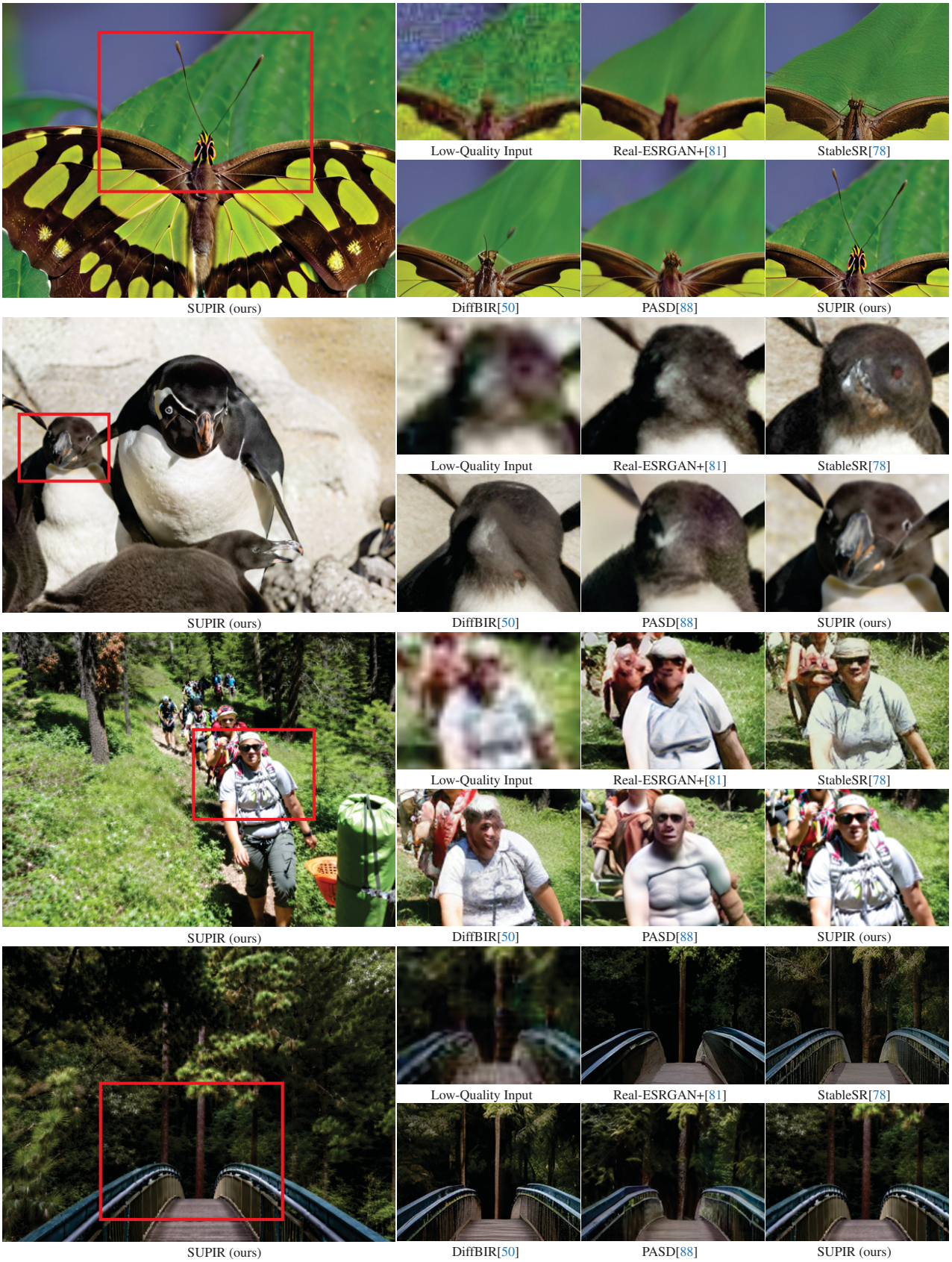


Figure 21. Qualitative comparison with different methods. Our method can accurately restore the texture and details of the corresponding object under challenging degradation. Zoom in for better view.

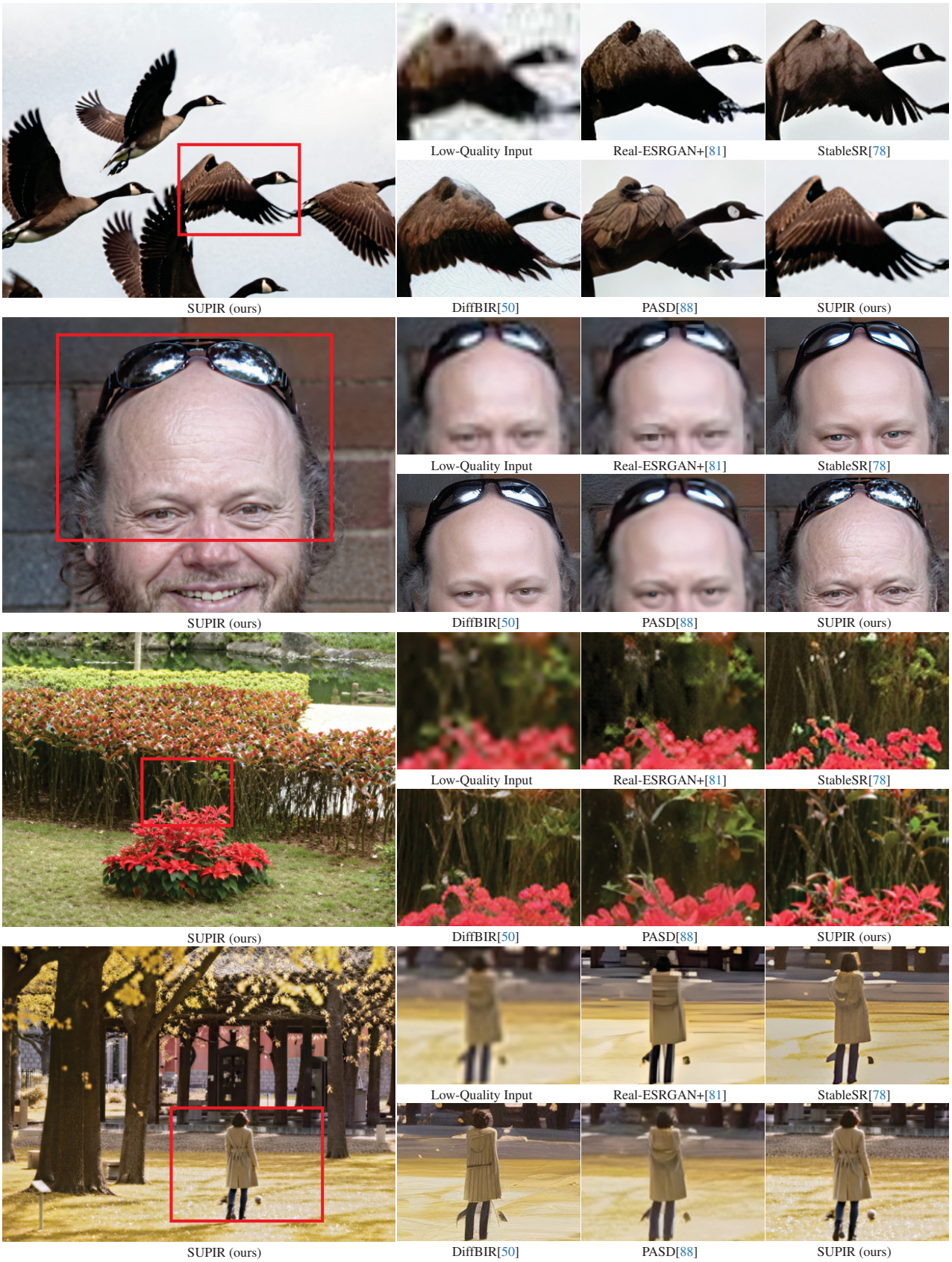


Figure 22. Qualitative comparison with different methods. Our method can accurately restore the texture and details of the corresponding object under challenging degradation. Zoom in for better view.

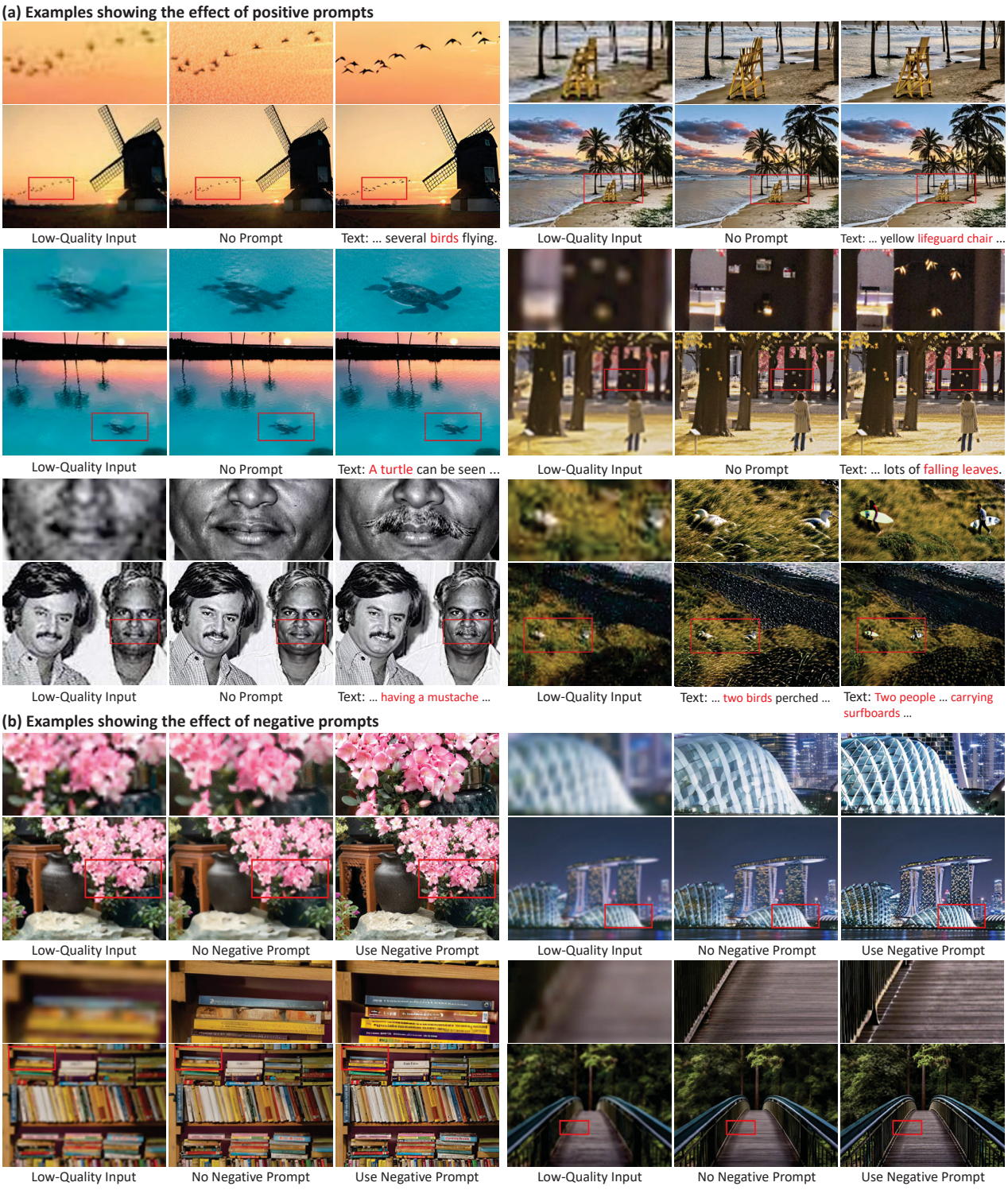


Figure 23. More visual results of the text prompts' influences. (a) and (b) show the examples of positive prompts and negative prompts, respectively. Zoom in for better view.