PixTalk: Controlling Photorealistic Image Processing and Editing with Language Supplementary Material

Marcos V. Conde, Zihao Lu, Radu Timofte
Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany
https://github.com/mv-lab/AISP

1. Dataset Details

General Illuminance Transformations We perform global and local adjustments. We also provide additional fine-grained enhancement samples. Global adjustments include exposure and contrast, as well as local modifications include highlight and shadow correction:

- Exposure, for each image we provide 2 variants with exposure $\pm$ 1 EV.
- Contrast, for each image we provide increasing and decreasing contrast variants. For fine-grained adjustment, we also provide $\pm$ (10/30/50 of 100) options.
- Highlight, we provide increasing and decreasing illumination variants for highlight area in the image. And for fine-grained adjustment, we also provide $\pm$ (10/30/50 of 100) total 6 options.
- Shadow, 2 variants with increasing or decreasing the illumination of each image shadow area.

Color Transformations We provide both global and local adjustments for images, and we also provide the global transformation with fine-grained adjustment. For local adjustments, we provide more detailed option to generate data with more variance:

- Saturation, global saturation adjustment with 2 options for increasing and decreasing saturation. For fine-grained adjustment, $\pm$ (5/15/50 of 100) total 6 levels are included.
- Vibrance, global vibrance adjustment with 2 options allowing increasing and decreasing the vibrance. For fine-grained adjustment, $\pm$ (5/15/50 of 100) total 6 levels are included.
- Single Color Saturation, for 3 separated color channels, red, green and blue, each channel has 2 options for increasing or decreasing the saturation of the selected color.
- White Balance, 6 options for different common light conditions, daylight, shade, cloudy, etc.
- Color Temperature, 4 options for global edit with different color temperatures at 2700k, 4500k, 5500k, 7500k.
- Color Grading, 6 options for global/separated hue and saturation edit, including 2 global settings and 4 highlight/midtones/shadow separated settings, single part individual edit and multiple part mixed edit aiming at different hue edit.

Complex Transformations Besides single transformations, we also provide each image corresponding transformation for complex adjustment, each transformation includes multiple adjustment options such as color grading, contrast change, clarity enhancement, saturation and vibrance change. For instance, "Provia-Std" delivers natural color balance suitable for a wide range of subjects, or "Astia-Soft" provides gentle tones and soft contrast.

Data Diversity To increase data diversity, we select photos from various cameras, ranging from mobile phones to compact cameras like the Canon PowerShot, as well as professional DSLR and DSLM cameras. These include APS-C format models, such as the Nikon D70, and full-frame sensor cameras, such as the Sony Alpha 7. The variety in design years, sensor types, sensor sizes, and camera design makes the dataset highly robust against specific performance biases affecting image quality.

For smartphone dataset, we select 69 RAW images from the iPhone XS and 68 from the Samsung S9. Both cameras capture images at a resolution of twelve megapixels (3000×4000) and cover a variety of scenes, including architecture, flowers, landscapes, and nightscapes. The iPhone XS could use an ISO as low as 25 in excellent lighting conditions; however, in low-light conditions, it could use a higher ISO (up to 2000) compared to the Samsung S9, demonstrating an aggressive ISO adjustment strategy.

For DSLR/M dataset, we select 64 photos from the MIT5K dataset and 17 photos from partner photographers, taken with various camera brands from different production years. These images range in resolution from 6 to 24 megapixels and include scenes similar to the smartphone dataset, while also capturing lower-light conditions. All images were captured with low ISO settings and reasonable shutter speeds to ensure optimal image quality.

Prompt Type	PSNR	SSIM	DeltaE
Unsupported	29.36	0.866	8.99
Vague	29.45	0.871	8.70
Regular	36.50	0.920	3.30
Detailed	37.80	0.950	2.50

Table 1. Robustness Ablation Study considering the average performance on the proposed testset depending on the quality of the input prompts.

Train-test Split The training set contains images from: diverse DSLR cameras (MIT5K dataset [2] and PASCALRAW [4]) (*e.g*. Nikon D3200, Nikon D700, Canon EOS 5D), Sony Alpha 7M3, Vivo x90, iPhone Xs, Samsung S9. We personally captured high-quality images using the Sony Alpha 7M3 and Vivo x90.

The testing set includes the previous sensors and unseen sensors: Sony Alpha 7R4, Samsung S21 ultra, Google Pixel 7. We collected 40 clean scenes suitable to test many presets and edition pipelines. Images bigger than 12MP are center cropped to 12MP ($\approx$ 4K resolution).

2. PixTalk Model: Additional Details

2.1. Text Adversarial Samples

How does the model behave depending on the instruction? The performance of *PixTalk* depends on the ambiguity and precision of the instruction. The proposed problem requires certain precision in the instructions to obtain (sub)optimal performance. For this reason, *PixTalk* is still limited with ambiguous instructions such as *"enhance this image"*. We provide an ablation study in Table 1, and sample prompts in Table 2. However, the text ambiguity is a limitation in any language-based approach or method (even ChatGPT). It is fundamental to highlight that even under adversarial conditions, *PixTalk* **does not hallucinate or generate artifacts**, the model implicitly applies an identity transformation *i.e*. the output is the input.

2.2. Implementation Details

Our *PixTalk* model is end-to-end trainable. The image model does not require pre-training but we use a pre-trained sentence encoder as language model.

The model is optimized using the $\mathcal{L}_1$ loss between the ground-truth image (obtained using professional software) and the input RGB. Additionally, we use the cross-entropy loss $\mathcal{L}_{ce}$ for the task classification head of the text encoder. We train using a batch size of 32 and AdamW optimizer with learning rate $5e^{-4}$ for 100 epochs (approximately 2 days using a single NVIDIA H100). We also use cosine annealing learning rate decay. During training, we utilize the 1024×1024 (1MP) crops as input, and we use random horizontal and vertical flips as augmentations. Since our model uses as input instruction-image pairs, given an image, and knowing the task, we randomly sample instructions from our prompt dataset ($>$1K samples).

Type	**Prompts**
Unsupported	Can you add [X] in [Y] ? Remove [Z] from [Y] Replace [X] with a [Z]
Vague	Fix this image Improve the quality of my photo Apply white balance
Regular	I want more details Fix the grain in this photo Can you make it brighter?
Detailed/Expert	Apply cold white balance Increase exposure in the shadows, and contrast Remove the noise from my picture

Table 2. Examples of prompts with varying language and domain expertise.

During inference, the model only requires the input image and the text instruction. The inference process also fits in low-computation budgets (*e.g*. Google Colab T4 16Gb GPU).

3. Additional Qualitative Results

Our *PixTalk*, a single model, is able to perform multiple tasks, providing high-quality and high-resolution images without artifacts that closely resemble the professional photographer edition. Figures 1 and 2 show different ground-truth variants of two sample images.

In Figures 3,4 and 5 we compare *PixTalk* with our reference methods InstructPix2Pix [1] (official pre-trained model) and InstructIR [3].

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 5, 6

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[3] Marcos V Conde, Gregor Geigle, and Radu Timofte. High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2024. 2, 5, 6

[4] Alex Omid-Zohoor, David Ta, and Boris Murmann. Pascalraw: raw image database for object detection. *Stanford Digital Repository*, 1(3):4, 2014. 2

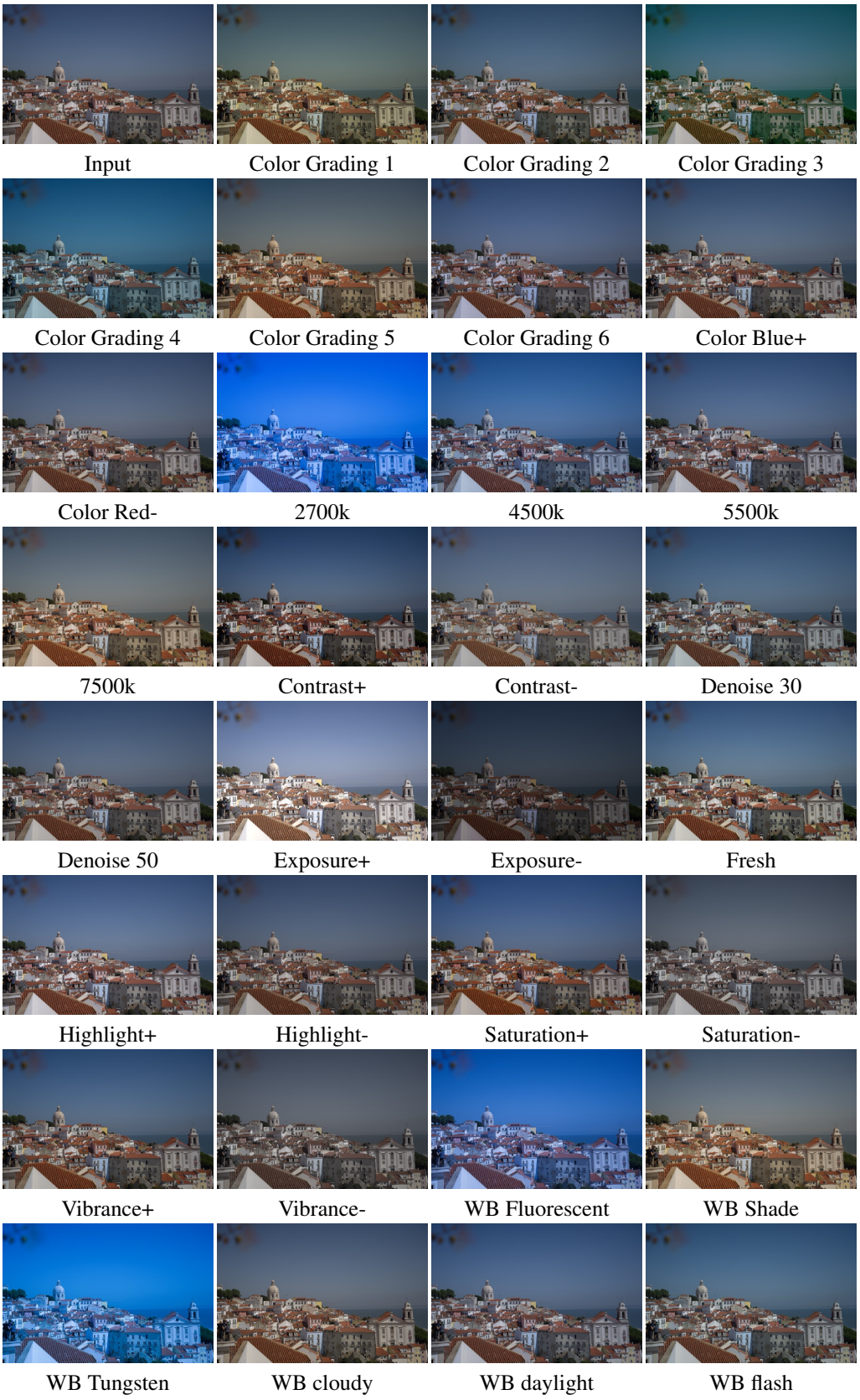

Figure 1. **PixTalk Dataset** (Sample 1) We illustrate some of the different transformations applied to each input image. The input image is the RAW visualized as sRGB.

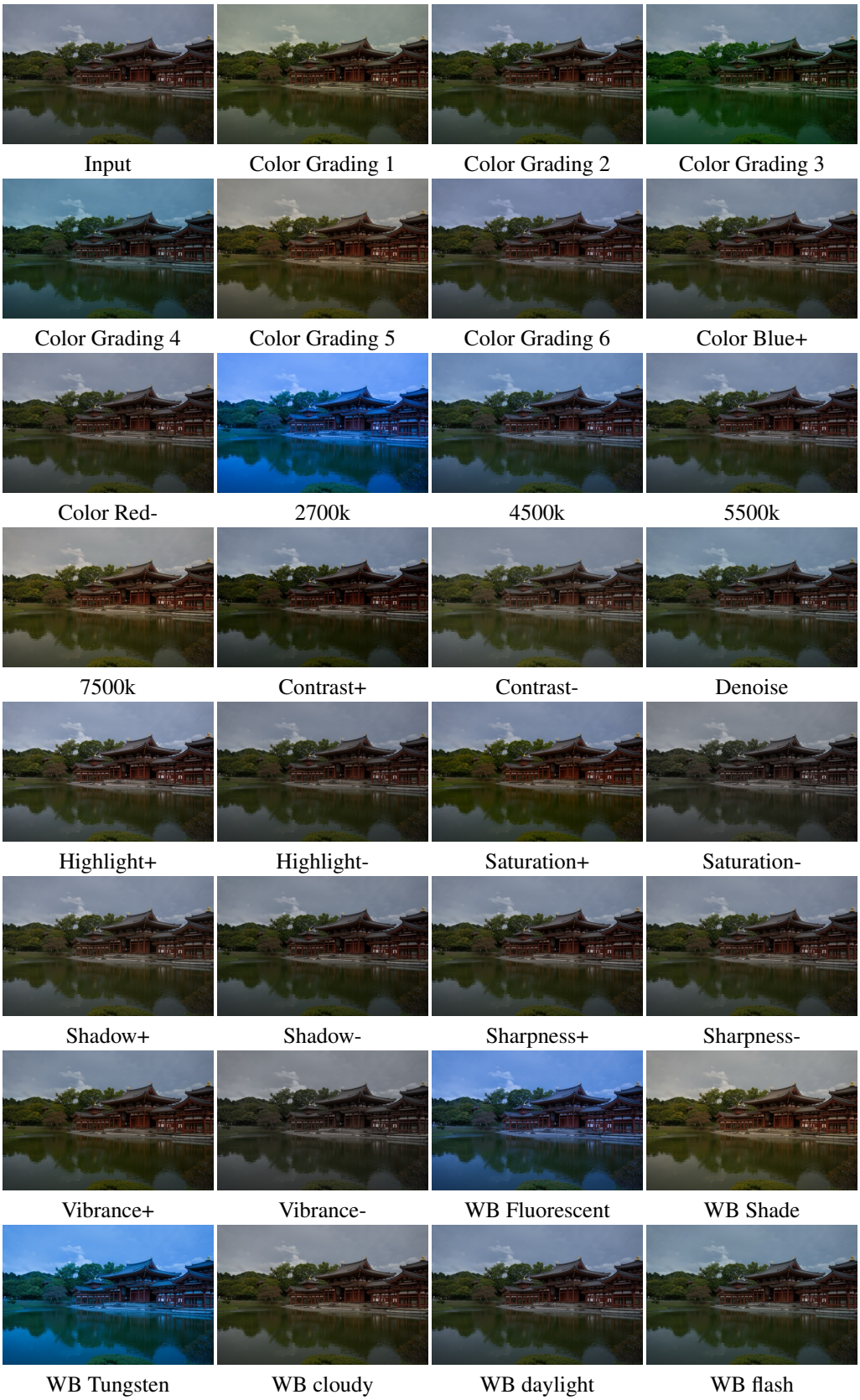

Figure 2. **PixTalk Dataset** (Sample 2) We illustrate some of the different transformations applied to each input image. The input image is the RAW visualized as sRGB.

Figure 3. **PixTalk Dataset** (Sample 3) We illustrate some of the different transformations applied to each input image by *PixTalk*. These are professional presets that apply multiple operations at once *e.g*. contrast enhancement, color correction and shadow correction.

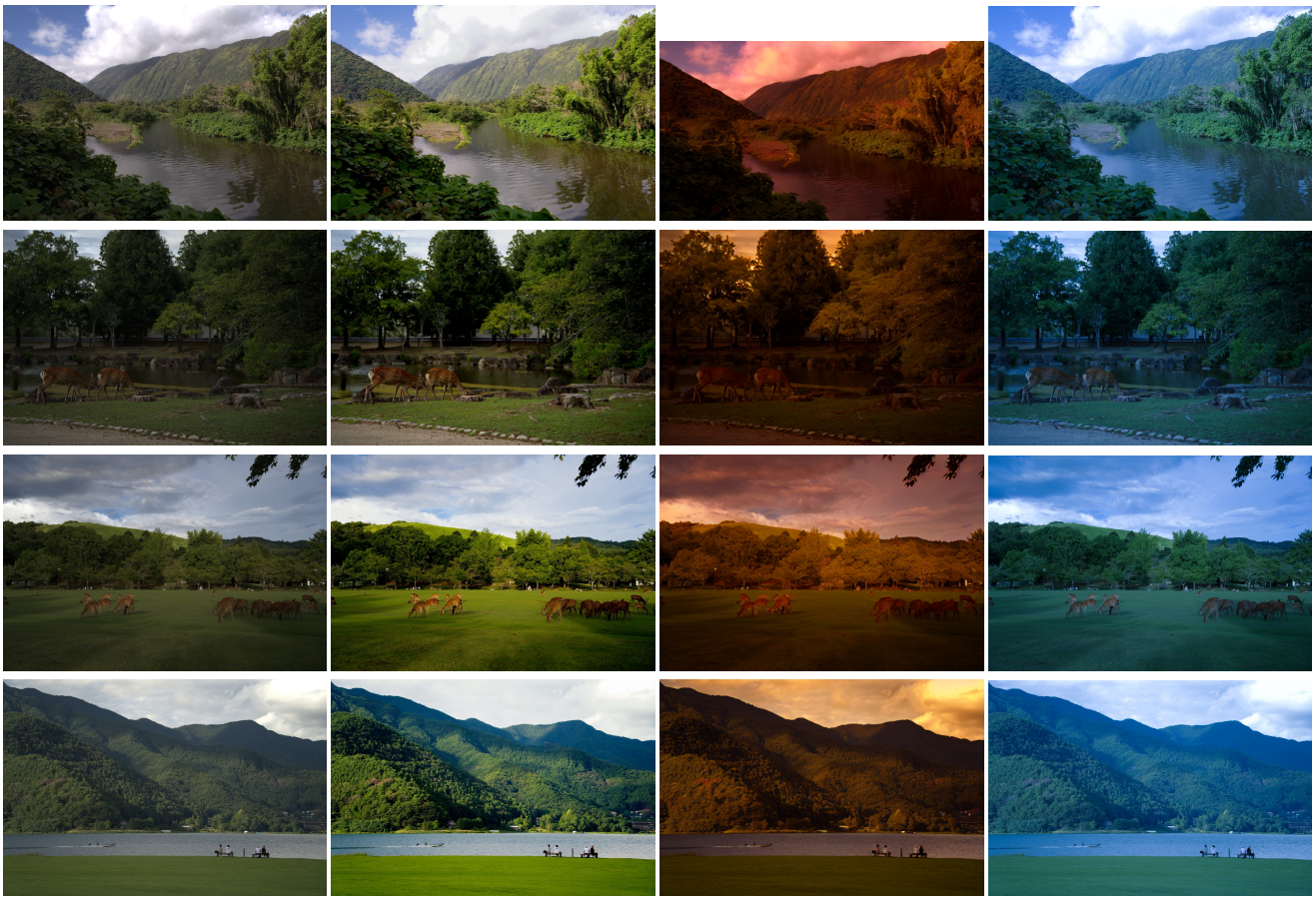

Figure 4. **Additional Comparison between models**. We show two examples of different instructions applied to the input image -from left to right-. Note that InstructIR [3] and InstructP2P [1] cannot process the full-resolution images (12 MP) due to their high computational complexity, thus, we feed into those models a downsampled image ($\times 4$). In contrast, *PixTalk* allows precise control of colors and **real-time processing at high-resolution.**

Input sRGB | InstructIR [3] | InstructP2P [1] | *PixTalk*

Novice Instruction:"Add a greenish tone to the image"

Figure 5. **Additional Comparison between models**. Sample 2. InstructIR [3] enhances the green color, but *PixTalk* has a more rich (and explicit) representation of color and illumination processing.