The Power of Context: How Multimodality Improves Image Super-Resolution

Supplementary Material

Table of Contents

7. Inference Latency Comparison

We compare the inference latency of our method against state-of-the-art (SOTA) methods in real scenarios. This comparison includes the total latency including image reading and writing, cross-modality prediction, and image captioning. The latency of each method was measured using its official code/script. Constrained by the implementation difficulty, MMSR is measured on the TPU platform and the rest methods are measured on the comparable NVIDIA GPU platform. Table 4 presents the results, demonstrating that our efficient multimodal strategy achieves the second fastest.

	PASD	SUPIR	SeeSR	MMSR
Latency (s)	**5.60**	18.01	30.86	6.06

Table 4. Inference latency of our method and compared SOTA.

8. Impact of Varying CFG Rates

Figures 5 and 6 in the main text, along with Figures 17, 18, 19, and 20 presented in the supplementary material, comprehensively demonstrate the key benefit of our method: a significant reduction in the excessive hallucinations and spurious details often produced by text-driven generative super-resolution approaches. By varying the CFG rates, we further show that our method acheives the better trade-off between reference-based and non-reference-based image quality metrics. Reference-based metrics reflect fidelity to the ground truth high-resolution image, while non-reference-based metrics assess perceptual quality and naturalness. As previously established by Blau and Michaeli [5], distortion and perceptual quality are conflicting objectives. However, Figure 10 shows that multimodal guidance not only improves performance at high classifier-free guidance (CFG) rates but also achieves a superior balance between these competing metrics, enhancing perceptual quality while mitigating the loss of fidelity to the target high-resolution image. The visual results in Figure 11 further demonstrate the superiority.

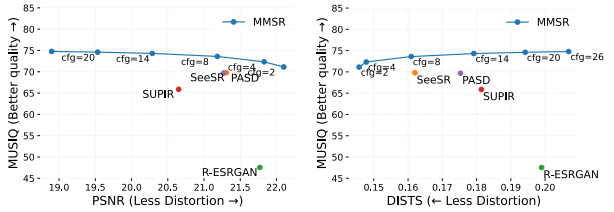


Figure 10. Our method improves the perception distortion trade-off of the past super-resolution methods.

SeeSR cfg=26 | MMSR m-cfg=26 | HR

Figure 11. Our method avoids excessive hallucinations of SeeSR when using the same large CFG rate of 26.

9. Extension to Diffusion Transformers

Diffusion Transformers, such as DiT [48] and MMDiT [20], operate on tokenized image patches during the diffusion process. In contrast to diffusion models based on U-Nets [55], Diffusion Transformers are better equipped to leverage information from tokenized text prompts, leading to improved text-coherency in text-to-image generation. It is natural to ask whether the effect of our multimodal guidance is still significant if the diffusion model itself is already optimized for better text-prompt grounding. We demonstrate the effect by comparing diffusion transformers with using multimodal guidance and using text-guidance only.

Our experiments show that our multimodal approach surpasses text-based super-resolution when applied to Diffusion Transformers. Figure 12 presents a comparison of the training loss for both methods, highlighting the superior performance of our multimodal guidance strategy. Furthermore, Table 5 provides a quantitative comparison of the two DiT based models.

The adopted Diffusion Transformer architecture is similar to the MMDiT used in Stable Diffusion 3 [20]. We employ the hidden features of the CLIP encoder and the T5 model for text embedding, leveraging their enhanced representation of text prompts. The crucial difference between our MMSR-DiT and the baseline text-based DiT lies in the incorporation of these multimodal latent tokens.

It is worth noting that both models were randomly initialized rather than warm-started from pre-trained text-to-image models due to computational constraints. Nevertheless, the comparison remains valid and demonstrates the superiority of our multimodal guidance, as both models share the same architecture, and were trained in equal forms.

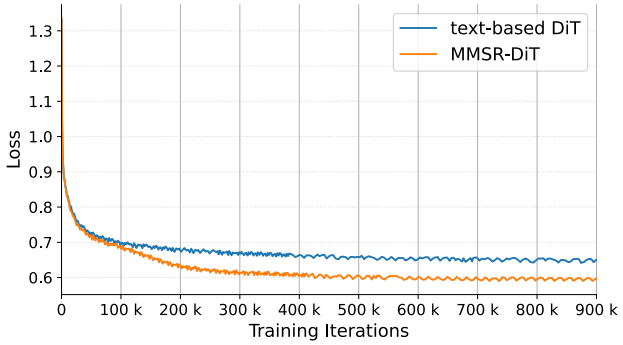


Figure 12. Training loss comparisons between the text-based diffusion transformer and our multi-modal-based diffusion transformer.

Method	MUSIQ	NIQE ↓	DISTS ↓	LPIPS ↓
text-based DiT	72.16	4.3266	0.1957	0.3453
MMSR-DiT	**72.18**	**4.0960**	**0.1621**	**0.2809**

Table 5. Quantitative result comparison between the text-based DiT and multimodal DiT on 1MP DIV-2K val set.

10. Dependence on Multimodal Input Quality

Our method leverages the prior knowledge encapsulated in pretrained cross-modal predictors, including Gemini Flash [58], Depth-anything [69], and Mask2Former [12]. The more accurate their predictions are the better super-resolution performance. We analyze the performance of our method across varying accuracy levels of cross-modal input. Specifically, we evaluate three variants: (a) low accuracy: modalities predicted directly from the low-resolution input; (b) medium accuracy: modalities predicted from our zero-modal super-resolution results; and (c) high accuracy: modalities predicted from the high-resolution target. Figure 13 provides visual examples for each accuracy level. Results demonstrate that our method is robust to variations in input modality quality, with zero-modal super-resolution effectively compensating for low-accuracy cross-modal predictions. Consequently, zero-modal super-resolution leads to improved results that closely approach those obtained with ground-truth modalities.

Special Cases in DRealSR Benchmark. The aforementioned investigation shows that extracting reasonable cross-modal predictions from low-resolution images is essential for ensuring our method achieves reasonable performance. Nevertheless, we notice that it is impossible to correctly estimate the modalities of three special images in the DRealSR benchmark, which are purely flat images. We manually replace their incorrectly predicted modalities with the $m_\emptyset$ token. Figure 14 visualizes these three special images for comprehensive understanding. Specifically, replacing these three images improves the CLIPIQA score of our method from 0.6892 to 0.6999.

Accuracy	MUSIQ	NIQE ↓	DISTS ↓	LPIPS ↓
Low-accuracy Modality	68.65	3.8874	0.1674	0.3449
Mid-accuracy Modality	72.31	3.4243	0.1504	0.2965
High-accuracy Modality	**72.32**	**3.3789**	**0.1492**	**0.2938**

Table 6. Our method achieves better SISR performance on higher-accuracy modality input than lower-accuracy modality input.

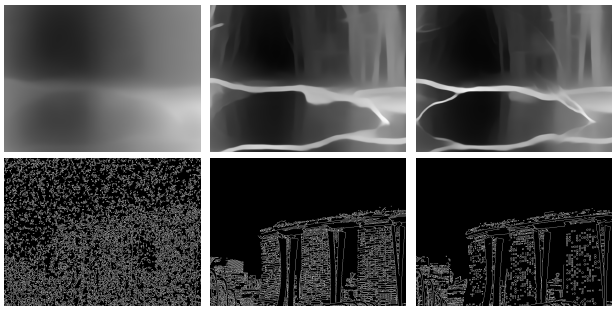


Figure 13. For low-quality cross-modal estimation from the low-resolution image, our method can increase the estimation accuracy by conducting zero-modal SISR on the low-resolution image.

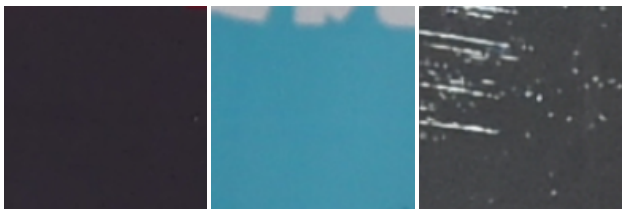


Figure 14. Special cases in DRealSR benchmark visualization.

11. Image Captioning Prompt Engineering

When tasked with image captioning, vision-language models (VLMs) like Gemini [58] often produce unsatisfactory results if given only simple instructions. Direct prompts, such as instructing the model to simply "caption the image," frequently lead to generic and uninformative outputs, including phrases like "Here is what I see from the image..." or captions in unsupported languages.

To mitigate these issues and generate more reliable and informative captions, we leverage in-context learning to

guide Gemini, following the practice in recent works [7, 67]. Specifically, we provide the model with examples of successful image-caption pairs, demonstrating the desired output format and level of detail. Table 7 presents a comparison of captioning results obtained using our in-context learning prompt and a standard, direct prompt. The results clearly demonstrate the superiority of our approach. Our prompt consistently generates stable and relevant captions, accurately describing the visual content of the input images.

In contrast, without the benefit of in-context learning, Gemini's responses to the standard prompt are often noisy and less structured. They frequently include extraneous procedural text, such as "Option 1...", "Here's a detailed description...", or similar phrasing, which hinders downstream tasks that rely on these captions. In super-resolution, where image captions can provide valuable contextual information, such noisy captions introduce undesirable artifacts and hinder performance. Therefore, based on this empirical analysis, we adopt our carefully crafted in-context learning prompt as the default image captioning prompt for all experiments.

12. Additional Visual Results

Visualization of Ablating Each Modality Figure 15 visualizes the results of text-guided super-resolution using individual input modalities. The observed differences in visual quality align with the quantitative ablation study in the main paper: depth and semantic segmentation contribute mostly to perceptual detail, while edge information primarily enhances fidelity. This observation motivates our core contribution: effectively combining the strengths of different modalities for text-guided super-resolution.

More Real-world Results Figure 17 shows more real-world super-resolution results and the comparison with the SOTA methods. Our method consistently outperforms the compared methods generating images with better realism and less plausible details that are inconsistent with the LR.

1024P High-resolution Results Figure 18, Figure 19, and Figure 20 show 1024P high-resolution super-resolution results and a comparison with SeeSR[68]. Our method clearly produce more details than SeeSR even though without directly training on 1024P images.

Failing Cases We find that when the LR input is a flat image and its semantic meaning is unclear, the multimodal guidance tends to misguide our method, producing inconsistent over-hallucinated results. Such flat images are rare in the DIV2K and LSDIR training sets, and thus a potential solution for these failing cases would be collecting more such flat images for training. Figure 16 shows the failing cases.

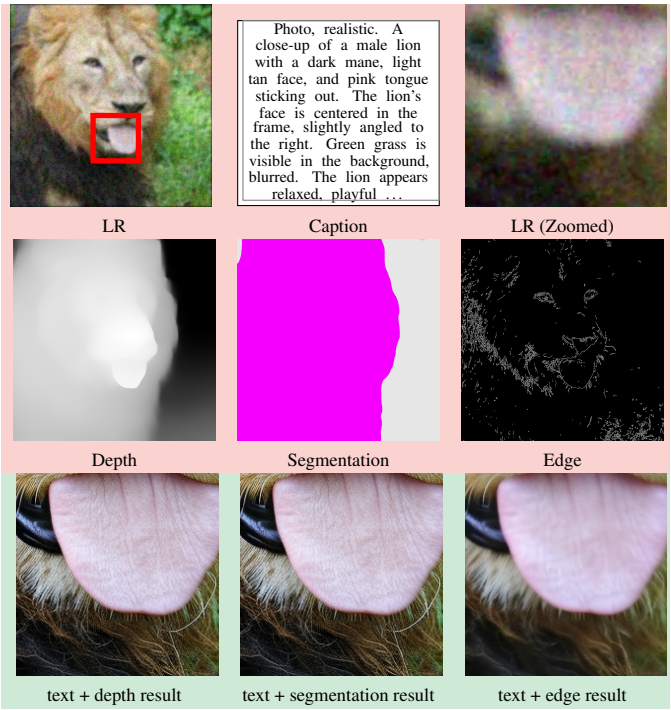


Figure 15. Visual results of our method under different modalities input. We find that the visual quality changes according the previous shown trend in different metrics, such as *text+depth* and *text+segment* have most details but less identical to the input.

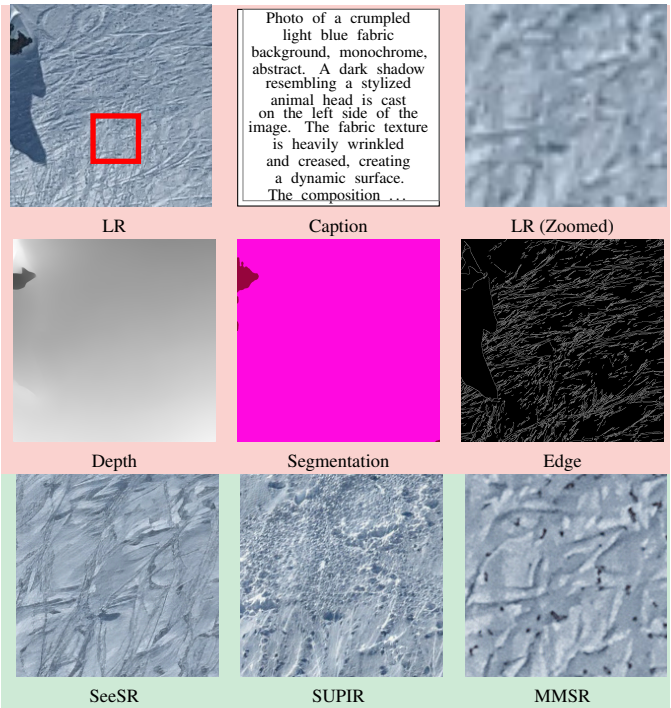


Figure 16. Failing case visualization.

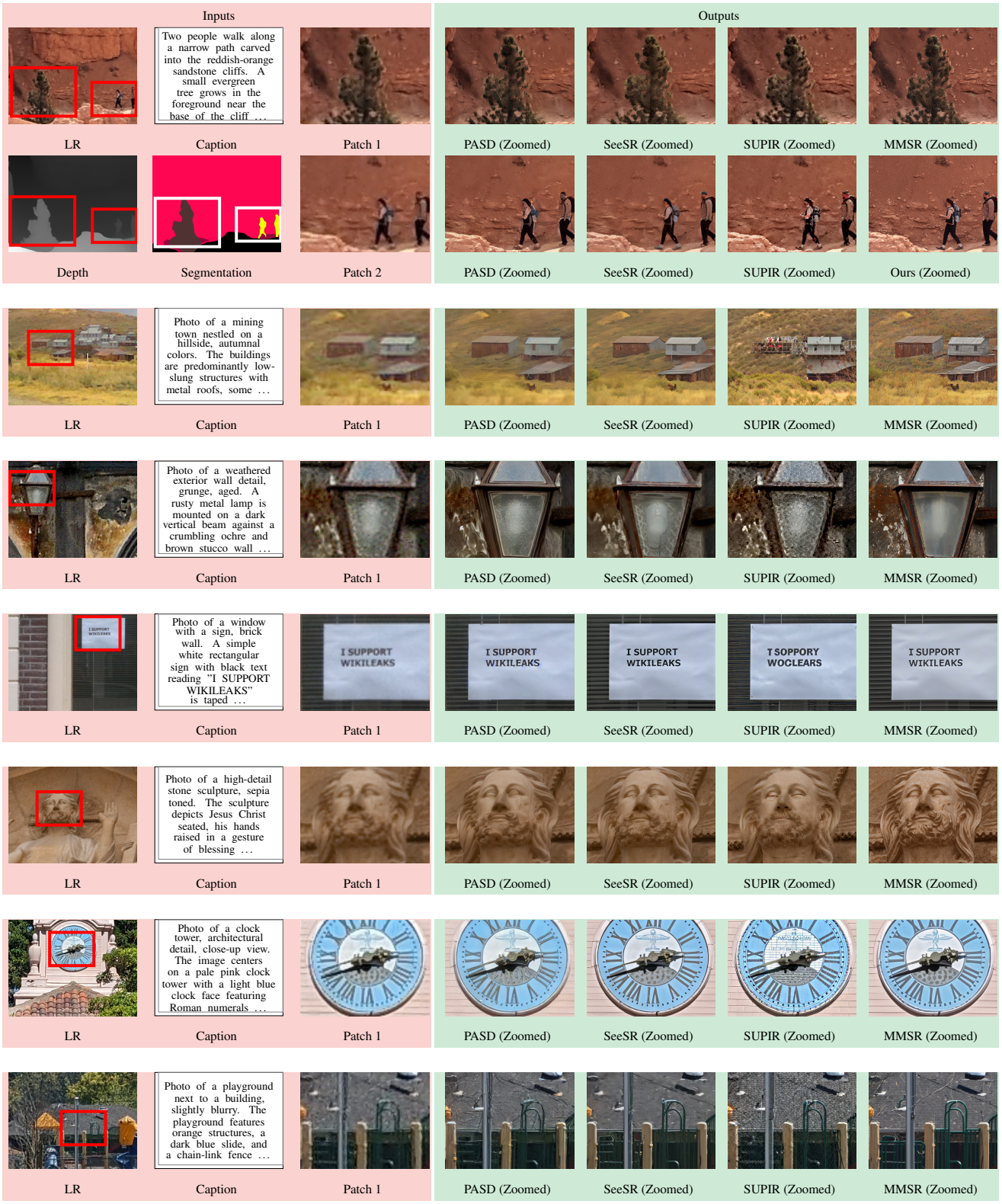
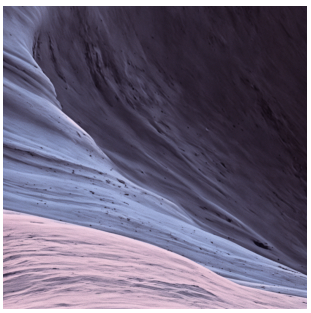
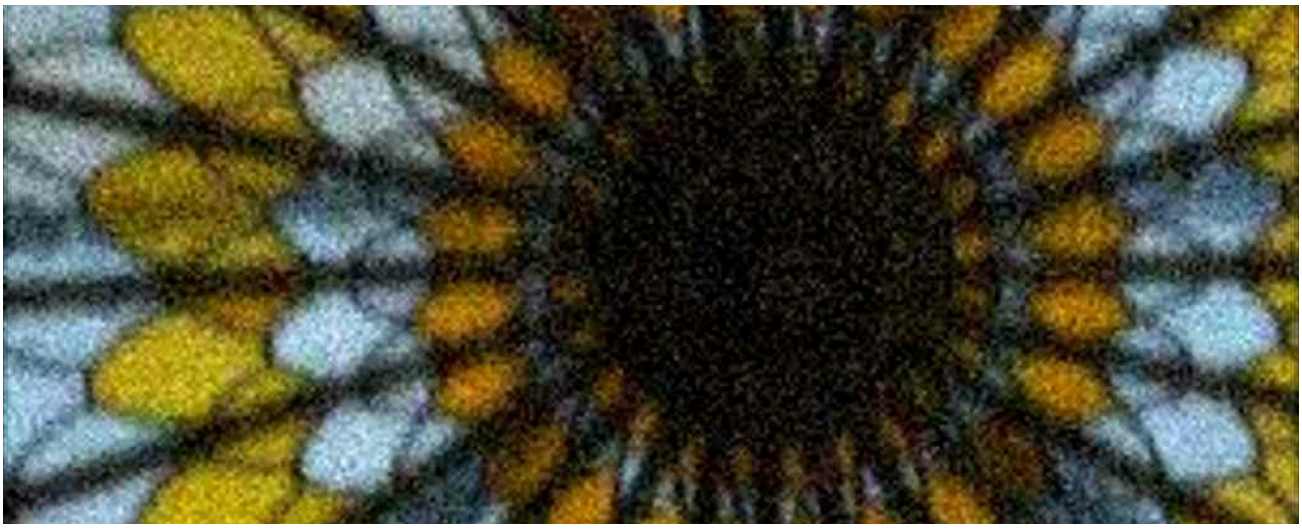


Figure 17. MMSR super-resolution results on real-world images compared with state-of-the-art methods. Zoom in to appreciate the details.

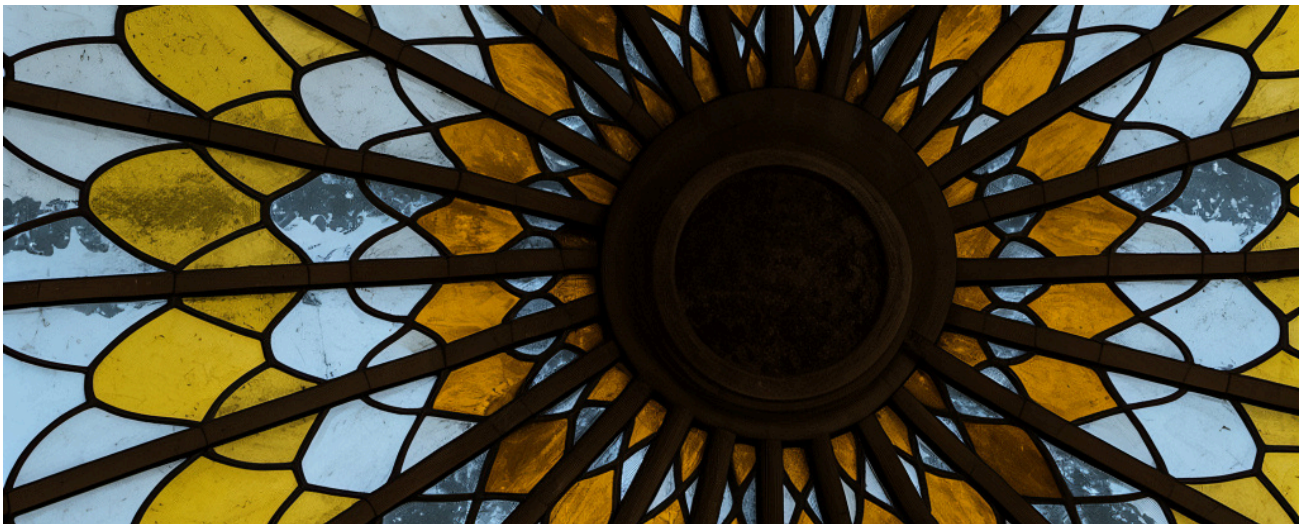
Table 7. Image caption result comparisons between different prompts. We show that our prompt that utilizes in-context learning is stable at most cases and can always get more detailed image captions without useless procedural words.

Prompt	*Describe this image and its style in a very detailed manner*	*Write a four sentence caption in English for this image. In the first sentence describe the style and type (painting, photo, etc) of the image. Describe in the remaining sentences the contents and composition of the image. Only use language that would be used to prompt a text to image model. Do not include usage. Comma separate keywords rather than using "or". Precise composition is important. Avoid phrases like "Here is...". Good examples are: "Photo of an alien woman with a glowing halo standing on top of a mountain, wearing a white robe and silver mask in the futuristic style with futuristic design, sky background, soft lighting, dynamic pose, a sense of future technology, a science fiction movie scene rendered in the Unreal Engine."*
	**Option 1 (Short & Evocative):** Textures of twilight. **Option 2 (Descriptive):** Smooth, windswept curves of snow and shadow paint an abstract landscape in dusky hues. **Option 3 (Focusing on Light):** The fading light reveals the delicate textures etched onto the snowy slopes. . . .	Photo of a mountainous landscape, abstract, textured. The image features sweeping curves of snow and ice, creating a dramatic diagonal composition. Shades of purple and blue dominate the color palette, lending a cool and ethereal atmosphere. The texture is highly detailed, showing the granular nature of the snow and the sculpted forms of the terrain.
	a close-up view of a car's hood, reflecting a sunlit scene. The hood is slightly wet, indicated by the subtle sheen of water droplets that catch and scatter the light. The reflection is not sharp or complete, but rather a shimmering impression of the surrounding environment. . .	Photorealistic image, exterior shot, modern architecture. A sleek dark gray apartment building with balconies is partially visible through lush green foliage. The foreground features a wet car hood reflecting sunlight. The composition emphasizes a shallow depth of field, focusing on the reflection and blurring the background.
	Here's a detailed description of the image and its style: The image is a landscape photograph, seemingly taken in a canyon or arid desert region, focusing on two hikers traversing a rocky path. Composition and Subject Matter: . . .	Photo of hikers in a canyon, Bryce Canyon National Park. Two people walk along a narrow path carved into the reddish-orange sandstone cliffs. A small evergreen tree grows in the foreground near the base of the cliff. The composition is a high-angle view, showcasing the scale of the canyon walls and the small figures of the hikers.



LR Input

SeeSR



MMSR

Figure 18. MMSR super-resolution results on 1024P DIV2K-Val compared with state-of-the-art methods. Zoom in to appreciate the details.

LR Input

SeeSR

MMSR

Figure 19. MMSR super-resolution results on 1024P DIV2K-Val compared with state-of-the-art methods. Zoom in to appreciate the details.

LR Input

SeeSR

MMSR

Figure 20. MMSR super-resolution results on 1024P DIV2K-Val compared with state-of-the-art methods. Zoom in to appreciate the details.