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

Kangfu Mei[1,2], Hossein Talebi[1], Mojtaba Ardakani[1],
Vishal M. Patel[2], Peyman Milanfar[1], Mauricio Delbracio[1]

[1] Google, [2] Johns Hopkins University

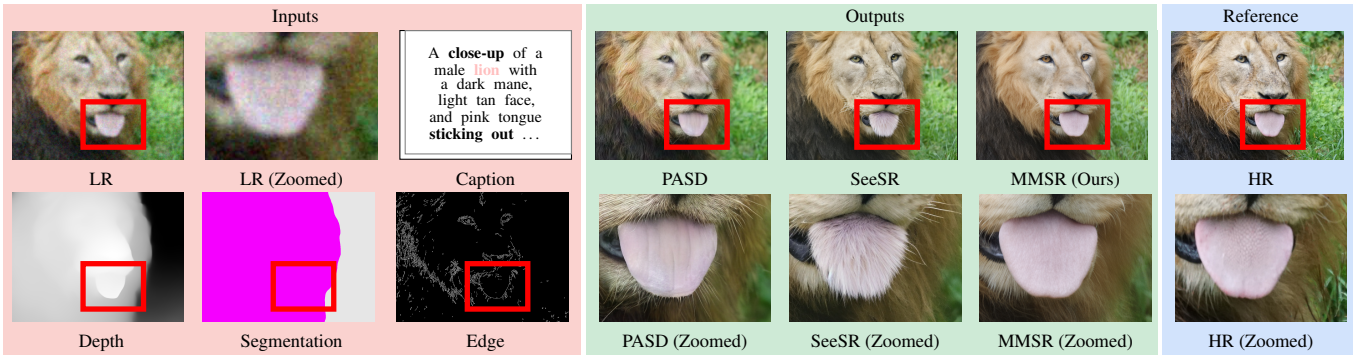


Figure 1. Our Multimodal Super-Resolution (MMSR) method leverages the rich context of multimodal guidance, including image captions, depth maps, semantic segmentation maps, and edges inferred from LR. MMSR surpasses state-of-the-art methods by producing more realistic results and suppressing artifacts that, while plausible, are inconsistent with the information present in the LR input.

Abstract

Single-image super-resolution (SISR) remains challenging due to the inherent difficulty of recovering fine-grained details and preserving perceptual quality from low-resolution inputs. Existing methods often rely on limited image priors, leading to suboptimal results. We propose a novel approach that leverages the rich contextual information available in multiple modalities – including depth, segmentation, edges, and text prompts– to learn a powerful generative prior for SISR within a diffusion model framework. We introduce a flexible network architecture that effectively fuses multimodal information, accommodating an arbitrary number of input modalities without requiring significant modifications to the diffusion process. Crucially, we mitigate hallucinations, often introduced by text prompts, by using spatial information from other modalities to guide regional text-based conditioning. Each modality's guidance strength can also be controlled independently, allowing steering outputs toward different directions, such as increasing bokeh through depth or adjusting object prominence via segmentation. Extensive experiments demonstrate that our model surpasses state-of-the-art generative SISR methods, achieving superior visual quality and fidelity.

1. Introduction

Single image super-resolution (SISR) aims to generate high-resolution images from low-resolution inputs while preserving semantic identity and texture details. While it is not essential to treat SISR as a regression problem, past methods [17, 29, 61] typically use a deep neural network to learn a direct mapping from low-resolution images to high-resolution images. Even though these regression-based methods achieve good scores on paired metrics like PSNR and SSIM, they have failed to produce results with high quality comparable to natural images, a task at which recent generative models have excelled[5, 14]. The advent of powerful generative models, such as autoregressive models [11, 50] and diffusion models [46, 52, 53], has revolutionized image generation tasks, including text-to-image synthesis. This has inspired recent efforts to leverage these pre-trained generative models for downstream tasks like SISR [6, 39, 52]. For instance, very recent works [49, 71] achieve super-resolution by leveraging emerging vision-language models (*e.g.*, Gemini [55], LLaVA [36], ChatGPT-4 [1]) and pretrained text-to-image models to first generate captions from low-resolution images and then use these captions as prompts to generate high-resolution images.

While providing rich textual descriptions can significantly enhance the quality of generated images [4, 19, 34], relying solely on text for SISR poses challenges. Recent works [8, 26, 35, 77] have shown that text prompts cannot represent spatial relationships. This implies that textual information, such as texture descriptions, can only be applied to the whole image. Figure 1 provides a representative example. Previous text-based super-resolution methods [65, 68] use a 'lion'

caption, which results in a furry tongue. *However, while lions have fur, their tongues do not grow hair.* Can we leverage spatial cues from depth and segmentation data to improve the learned prior and enhance the quality of SISR?

Fortunately, we can directly extract additional modalities from the low-resolution image by using various pretrained cross-modal prediction models [12, 66]. In this paper, we introduce a new diffusion model architecture dedicated to multimodal super-resolution, which is conceptually illustrated in Figure 2. By integrating multiple modalities into a single diffusion model, our method overcomes the challenges of recovering fine-grained details and preserving perceptual quality. Specifically, we propose conditioning diffusion models on modalities including text captions, semantic segmentation maps, depth maps, and edges to implicitly align text captions for correctly prompting different regions.

We demonstrate the proposed multimodal diffusion model on the SISR task. Our effective multimodal architecture achieves better realism in SISR results than the best text-driven method and largely eliminates hallucinations that do not match the input. Moreover, we find that the multimodal SISR enables a new controlling feature, where we can explicitly adjust the weights of each modality to steer the generated results in different directions.

Our contributions are summarized as follows:

- We demonstrate the effectiveness of token-wise encoding for seamlessly injecting multiple modalities into pretrained text-to-image diffusion models without architectural modifications or significant model size overhead.
- We propose a novel multimodal latent connector that efficiently fuses information from different modalities, maintaining linear time complexity with respect to the number of modalities.
- We introduce a new multimodal classifier-free guidance technique that enhances realism at higher guidance rates while mitigating excessive hallucinations and fake details.
- Our method enables adjustment of the influence of each modality, allowing for fine-grained manipulation of SISR results while preserving realism and quality.

2. Related Work

Generative Prior Powered Super-resolution. Recent advances in super-resolution leverage the power of foundational models to enhance the quality of low-resolution images [10, 13, 20, 33, 37, 44, 49, 56, 63, 69, 72, 76]. This trend is fueled by the capacity of pretrained generative models to capture natural image statistics and transfer this knowledge to the super-resolution task, enabling photorealistic image generation. Early works in this domain include LDM-SR [52] and StableSR [58] for single image super-resolution. Beyond this, methods like InstructPix2Pix [6] utilize instructions for image editing, while ControlNet [74] and

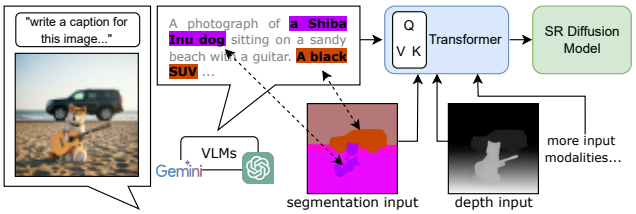


Figure 2. Language models struggle to accurately represent spatial information, leading to coarse and imprecise image super-resolution. To overcome this limitation, we incorporate additional spatial modalities like depth maps and semantic segmentation maps. These modalities provide detailed spatial context, allowing our model to implicitly align language descriptions with individual pixels through a transformer network. This enriched understanding of the image significantly enhances the realism of our super-resolution results and minimizes distortion.

IP-Adapter [70] facilitate cross-modality image translation. These approaches highlight a common observation: the quality of results in downstream tasks is strongly correlated with the quality of the pretrained models [38].

Recent works focus on leveraging powerful text-to-image diffusion models, employing text prompts to fully exploit the learned prior. PASD [68] uses image content descriptions; SPIRE [48] and PromptIP [47] use degradation descriptions; and SeeSR [65] employs a combination of context and degradation descriptions. Compared to earlier methods that directly fine-tuned diffusion models on super-resolution data, these text-prompt-driven approaches not only achieve superior realism but also enable multi-faceted outputs by conditioning on different prompts [20]. This capability enriches the typically single-output super-resolution task.

While text prompts offer advantages, they can be ambiguous in representing spatial relationships [4, 8, 19, 34]. Our proposed MMSR framework addresses this by integrating multimodal inputs within a novel architecture. We also demonstrate native super-resolution effects control with this new architecture, enabling control of both overall realism and the effect of each modality.

Vision-language Understanding and Generation. Recent large vision-language models (e.g., Gemini [55], LLaVA [36], GPT-4o [1]) excel in tasks like image captioning [60], but translating visual information into diverse modalities for image generation remains challenging. While models like 4M [2, 41] effectively extract high-level visual information (depth, normals, semantics, etc.), leveraging these for generation is a promising direction. Recent exploration includes GLIGEN [30] generate images from bounding boxes, and ControlNet [74] generate images from depth and other modalities. More recent works [21, 22] explore using discrete hidden-feature tokens to guide generation. In this paper, we introduce a new approach for multimodal control with discrete vision tokens, effectively integrating depth maps,

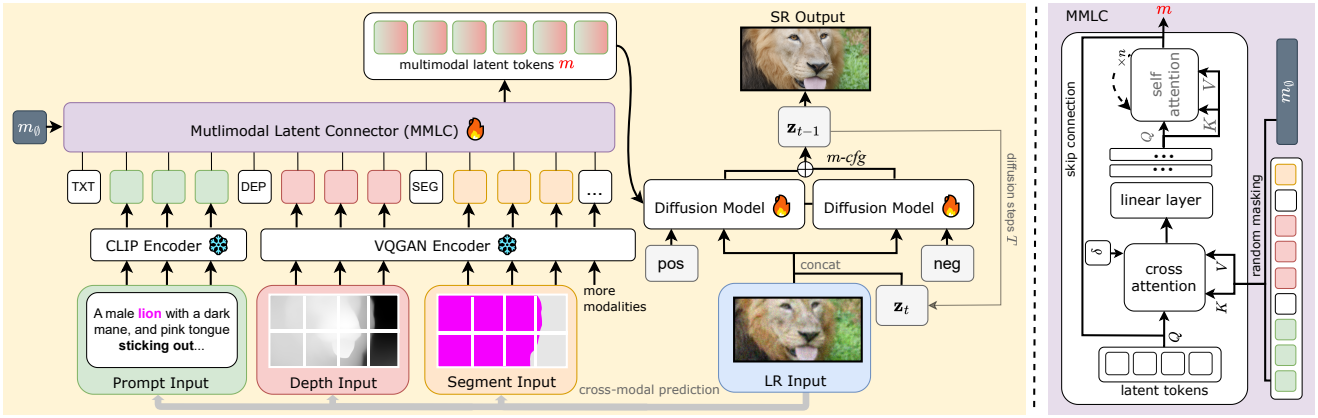


Figure 3. This diagram illustrates our multimodal super-resolution pipeline. Starting with a low-resolution (LR) image, we extract modalities like depth and semantic segmentation maps. These modalities are encoded into tokens and transformed into multimodal latent tokens (m). Our diffusion model uses these tokens and the LR input to generate a high-resolution (SR) output. A multimodal classifier-free guidance (*m-cfg*) refines the SR image for enhanced quality.

semantic segmentation maps, and text prompts for superior performance. Experiments on real-world super-resolution demonstrate the effectiveness of our approach.

3. Method

Single-image super-resolution aims to recover a high-resolution image $\mathbf{x}$ from its low-resolution counterpart $\mathbf{x}_{LR}$. This is an ill-posed problem, often leading to generative models that produce "hallucinated" details – plausible yet inconsistent with the input. To mitigate this, we introduce auxiliary information, such as depth (m_{dep}), semantic segmentation (m_{seg}), and edge maps (m_{edg}), collectively denoted as m. By conditioning the generative process on both the low-resolution image and these auxiliary modalities, we propose a new distribution $p(\mathbf{x}|\mathbf{x}_{LR}, m)$ with reduced uncertainty compared to the original distribution $p(\mathbf{x}|\mathbf{x}_{LR})$.

This reduction in uncertainty can be understood through the lens of information theory. The auxiliary modality m provides additional information about the high-resolution image $\mathbf{x}$ that is not present in the low-resolution input $\mathbf{x}_{LR}$. Since (conditional) mutual information is non-negative:

$$I(\mathbf{x}; m|\mathbf{x}_{LR}) = H(p(\mathbf{x}|\mathbf{x}_{LR})) - H(p(\mathbf{x}|\mathbf{x}_{LR}, m)), \quad (1)$$

where $H(\cdot)$ denotes entropy. Consequently, the entropy of the conditional distribution with the auxiliary modality is:

$$H(p(\mathbf{x}|\mathbf{x}_{LR})) \geq H(p(\mathbf{x}|\mathbf{x}_{LR}, m)). \quad (2)$$

The same motivation for reducing uncertainty is also shared by recent diffusion guidance works [23, 42]. While low-entropy sampling does not guarantee sharper images, it often leads to outputs with better visual quality compared to standard sampling. Motivated by this observation, we introduce a diffusion model [24] to learn the multimodal distribution $p(\mathbf{x}|\mathbf{x}_{LR}, m)$, effectively incorporating auxiliary information to enhance SISR quality.

Ideally, the auxiliary modalities should provide information complementary to the low-resolution input. In practice, the auxiliary modalities are derived from the high-resolution image during training, ensuring informative conditioning. While we use modalities derived from the low-resolution input during inference, we demonstrate that this still leads to superior performance, including higher-quality details and improved fidelity to the input.

3.1. Unified Mutlimodal Diffusion Conditioning

We introduce a new diffusion network architecture, illustrated in Figure 3, for simultaneously conditioning on multiple modalities. Unlike recent methods like ControlNet [74] and IP-Adapter [70], which duplicate network components for each modality and incur significant computational overhead, our approach leverages a pretrained VQGAN image tokenizer [18]. This allows us to encode diverse modalities into a unified token representation for conditioning the diffusion model, without introducing additional model parameters or modifying the diffusion network itself. These tokens are concatenated with the text prompt embedding and used for cross-attention within the diffusion model. To efficiently process this long token sequence, we introduce a lightweight multimodal connector. This connector employs a dedicated architecture to achieve linear complexity for cross-attention, significantly reducing the computational burden. In what follows we present the implementation details.

Token-wise Multimodal Encoding. While VQGAN [18] has proven effective for cross-modal encoding in image understanding and generation [2, 3, 41], its optimal application for image super-resolution remains an open question. Unlike tasks focused on understanding or generation, super-resolution requires tokenization to not only capture semantic information but also preserve pixel-wise details crucial for

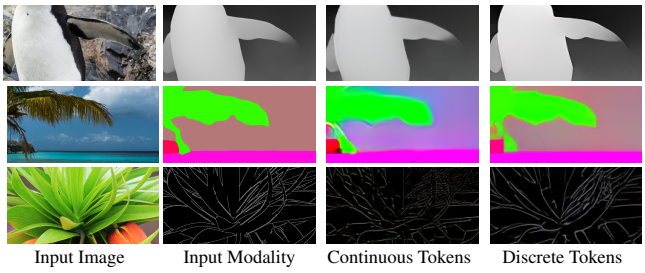


Figure 4. Using discrete multimodal tokens leads to superior reconstruction of modalities compared to continuous tokens.

accurate reconstruction, such as spatial relationships within a depth map. Inspired by recent discussions on using discrete or continuous tokenization for autoregressive image generation [54], we investigate the impact of quantization on multimodal super-resolution. Specifically, we analyze how different quantization strategies affect the reconstruction quality of individual modalities and the final generated image. To this end, we utilize a VQGAN model pre-trained on a large-scale multimodal image dataset.

Figure 4 compares the reconstruction quality of multimodal tokens before and after quantization. As the figure illustrates, discrete tokens (post-quantization) better preserve individual modality information, while continuous tokens introduce noticeable artifacts. Consequently, we employ discrete tokens for our super-resolution experiments.

Our multimodal encoding process utilizes the encoder and quantizer of the pre-trained VQGAN. Each 256×256 input modality is encoded into a 16×16 multimodal token sequence with a feature dimension of 256. These tokens are then quantized using a codebook of size 1024. To facilitate concatenation with text embeddings, we pad the feature dimension of the multimodal tokens to 1024, resulting in a $(256 \times 3 + 77) \times 1024$ multimodal token sequence condition.

Multimodal Latent Connector. While cross-attention provides a flexible mechanism for conditioning diffusion models on multimodal data, its quadratic complexity with respect to the number of condition tokens introduces a significant computational burden. To address this, we introduce the Multimodal Latent Connector (MMLC), inspired by recent advances in efficient attention architectures [25, 28, 59].

As illustrated in Figure 3, the MMLC employs a transformer architecture to efficiently process the multimodal token sequence. The transformer receives two inputs: a randomly initialized sequence of learnable latent tokens, and the multimodal input sequence. The output is a token sequence of the same length as the latent token sequence, which serves as conditioning for the diffusion model. Therefore, the diffusion model conditions on fixed-length latent tokens (128 in our experiments), which are significantly shorter than the original multimodal token sequence input. The MMLC distills the essential information from the longer multimodal token sequence into the shorter multimodal latent token sequence through cross-attention. Following cross-attention, several self-attention blocks further process the latent token sequence, allowing the model to fully integrate the distilled information.

This approach significantly reduces the computational cost of cross-attention in the diffusion model. Standard self-attention operates on the full multimodal token sequence ($K, V \in \mathbb{R}^{M \times D}$), resulting in a time complexity of $\mathcal{O}(M^2)$, where M is the length of the sequence and D is the dimensionality of each token. In contrast, the MMLC uses a cross-attention between the latent token sequence (of size $N \times D$) and the multimodal token sequence which reduces this to $\mathcal{O}(MN)$. Here N is the length of the latent sequence and $N \ll M$. This linear complexity with respect to the multimodal sequence length enables efficient processing of high-dimensional multimodal data, effectively capturing essential information for super-resolution.

Flexible Multimodal Input. To enhance flexibility and robustness, we enable our method to handle scenarios where certain input modalities are unreliable or unavailable. We adopt a learnable embedding approach inspired by DALL-E 2 [51]. Specifically, we introduce a special learnable token, $m_\emptyset$, optimized alongside the diffusion model and MMLC to represent the absence of a modality. During training, we independently randomly replace each modality with 256 $m_\emptyset$ tokens with a probability of 0.1. This encourages the model to learn robust representations that can effectively handle missing information. During inference, any missing modality is represented by a sequence of 256 $m_\emptyset$ tokens. This approach allows for flexible multimodal input, enabling the model to generate high-quality images even with limited or no auxiliary modalities. As demonstrated in our experiments, this strategy significantly improves performance when dealing with input that includes fewer modalities than were available during training.

3.2. Multimodal Guidance and Control

Building upon the success of guidance techniques [15, 23] in improving sample quality across various image generation tasks [43, 46, 51], recent diffusion-based SISR methods have begun incorporating prompt tuning to enhance super-resolution results [58, 65, 71]. These methods improve the result by using negative prompts with Classifier-free Guidance (CFG) [23]. which can be expressed as:

$$\tilde{\boldsymbol{\epsilon}}(\mathbf{z}_t, c) = (1 + w)\, \boldsymbol{\epsilon}(\mathbf{z}_t, c, \mathrm{pos}) - w\, \boldsymbol{\epsilon}(\mathbf{z}_t, c, \mathrm{neg}), \quad (3)$$

where $\tilde{\boldsymbol{\epsilon}}(\mathbf{z}_t, c)$ represents the guided denoising process, $\boldsymbol{\epsilon}(\cdot)$ denotes the diffusion model, $\mathbf{z}_t$ denotes the noisy image latent, $c = \{\mathbf{x}_{\mathrm{LR}}, t, \dots\}$ denotes the conditioning inputs (including the low-resolution image $\mathbf{x}_{\mathrm{LR}}$, timestep t, and other parameters), w is the guidance scale, and pos and neg are the positive and negative prompts. While increasing w

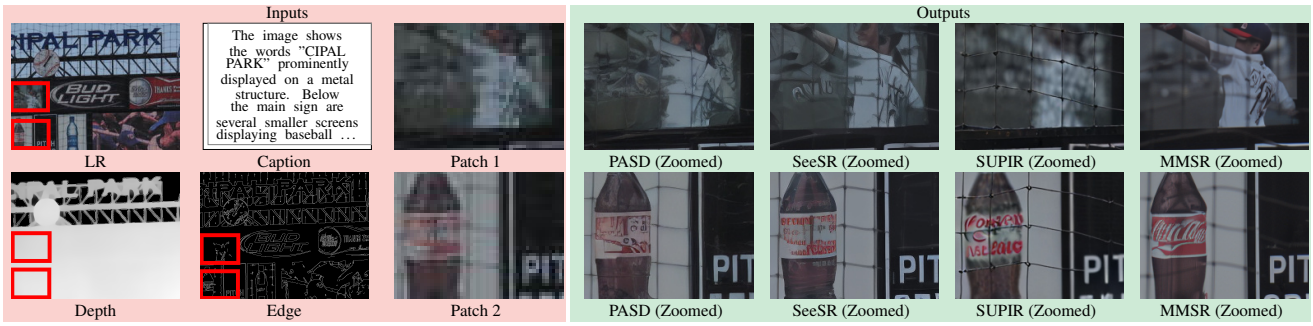


Figure 5. MMSR super-resolution results on real-world images compared with state-of-the-art methods. Zoom in to appreciate the details.

often leads to sharper and more detailed outputs, it can also exacerbate hallucination, resulting in details inconsistent with the low-resolution input. Such artifacts are widely reported but are difficult to suppress, even with recent efforts like balancing the training data [71].

Multimodal Classifier-free Guidance. To mitigate the issue of excessive hallucination often associated with high guidance scales in CFG, we propose a novel multimodal guidance strategy. We argue that the artifacts come from the weak guidance in the negative prompting process. Instead of simply relying on text prompts for guidance, we leverage the rich information encoded in the multimodal latent tokens to strengthen both positive and negative promptings. Specifically, we condition both the postive and negative generation processes on the multimodal latent token sequence, denoted as m. This leads to the following multimodal CFG:

$$\tilde{\boldsymbol{\epsilon}}(\mathbf{z}_t, c, m) = (1+w)\boldsymbol{\epsilon}(\mathbf{z}_t, c, \text{pos}, m) - w\,\boldsymbol{\epsilon}(\mathbf{z}_t, c, \text{neg}, m). \tag{4}$$

Sec. 4.1 shows that this change in negative generation helps to achieve a better trade-off between perceptual quality and identity preservation, better maintaining the semantic content of the low-resolution input in the upscaled output, compared with the standard CFG in previous text-based methods.

Scaling Single-modal Guidance. Multimodal CFG effectively controls the overall influence of the prompts but does not offer control over each modality individually. To address this limitation, we introduce a mechanism to selectively amplify or suppress the contribution of specific modalities. Specifically, we modify the attention *temperature* δ of MMLC during cross-conditioning, where δ scales the attention maps before applying the softmax operation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\delta}\right)V. \tag{5}$$

This parameter controls the sensitivity of the attention mechanism to differences between the feature sequence Q and the multimodal token sequences K and V. A smaller temperature δ typically leads to a stronger conditioning effect on the attention mechanism. In the standard scaled dot-product attention, the temperature δ is empirically set to the square root of the key dimension $\sqrt{d_k}$. We observe that scaling the standard temperature $\sqrt{d_k}$ (where d_k is the key dimension) within a range of $[0.4, 10]$ can produce high-quality results, with varying levels of fidelity to the input modalities. By scaling the temperature during sampling, we achieve precise control over the super-resolution process, enabling manageable manipulation of the output with fine-grained control.

3.3. MMSR Implementation

Figure 3 illustrates our multimodal guided super-resolution pipeline. During inference, we first extract four modalities from the low-resolution (LR) image: a text caption generated by Gemini Flash [55], a depth map estimated by Depth Anything [66], a semantic segmentation mask produced by Mask2Former [12], and edge information extracted with a Canny edge detector. The depth map, segmentation mask, and edge information are encoded into token sequences using a pretrained VQGAN, while the text caption is processed by a pretrained CLIP encoder to obtain a text embedding.

Following a similar conditioning strategy to InstructPix2Pix [6], the LR image is concatenated with a noisy latent vector sampled from the diffusion model, providing additional conditioning.

4. Experiments

Training Details. Our super-resolution model is initialized with the weights of a pretrained text-to-image model, with the same architecture and size as Stable Diffusion v2 [52]. The super-resolution training dataset consists of randomly degraded (using RealESRGAN degradation [62]) high-resolution images from the combined LSDIR and DIV2K datasets, with corresponding 512×512 high-resolution images as ground truth. We set the batch size to 1024 and the learning rate to 1e-4, which we empirically found maximized compute efficiency on TPUv5 pods. During testing, all benchmarks use a model checkpoint finetuned for 160k iterations. We use 50-step DDIM sampling, consistent with previous methods. We use a guidance rate of 4 as the default.

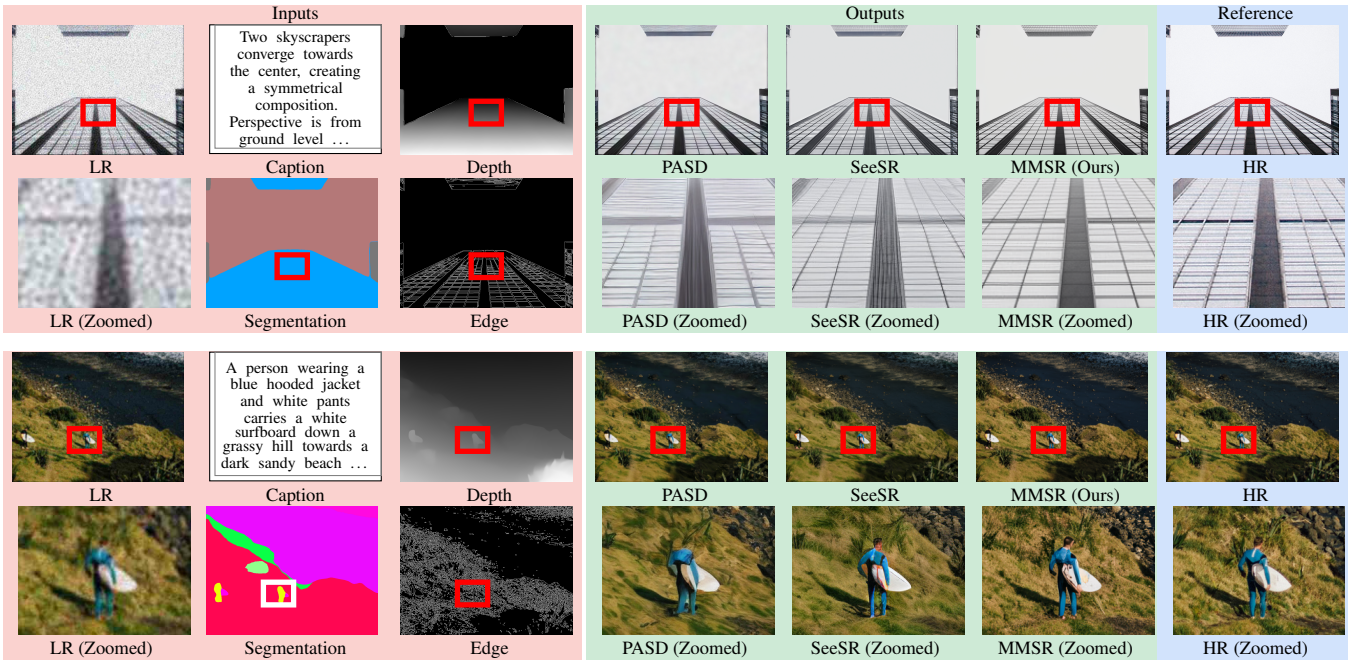


Figure 6. Super-resolution results on common benchmarks, comparing MMSR with state-of-the-art techniques. Zoom in for detail.

Evaluation Details. We use reference-based metrics like LPIPS [75] and DISTS [16], and non-reference-based metrics like NIQE [40], MANIQA [67], MUSIQ [27], and CLIPIQA [57], as these have been shown to align well with humans' aesthetic preferences. Our compared baselines include recent diffusion-based super-resolution methods [33, 52, 58, 65, 68, 72] and the representative methods [9, 31, 32, 62, 73]. We collected their results either by running their official code or from published results. In a few cases where the multimodal prediction failed (two flat images and one pencil drawing from DRealSR), we manually replaced the prediction with $m_\emptyset$ (see supplemental material).

4.1. Quantitative Results

Table 1 shows the quantitative result comparisons of ours with the other baselines, compared on the synthetic benchmark and real-world super-resolution benchmarks. Several facts are worth noting: (1) Our method achieves the best LPIPS, DISTS, NIQE, and FID scores on the *DIV2K-Val* and *RealSR* dataset, significantly outperforming previous state-of-the-art methods. This superiority demonstrates that our method can generate more perceptually identical details from the guidance of multimodal context. (2) Our method also achieves the best performance in the non-reference visual quality metrics including NIQE, MANIQA, CLIPIQA, and MUSIQ scores. This advantage demonstrates our method can fully utilize the photorealistic priors encapsulated in the pretrained diffusion model. Overall, these observations fully demonstrate our advantages in using multimodal guidance compared with the past single-modality methods in super-resolution.

Effects of Multimodal CFG Guidance. Table 3 demonstrates the effects of our proposed multimodal CFG on the 1MP *DIV2K-Val* benchmark, which replaces the negative score guided by empty language token $\boldsymbol{\epsilon}(\mathbf{z}_t, c, \text{neg})$, shortened as *cfg*, with multimodal guided negative score $(\boldsymbol{\epsilon}(\mathbf{z}_t, c, m, \text{neg}))$, shortened as m-*cfg*. In order to exclude the influence of architectural differences, we also compare our method with the empty multimodal token $\boldsymbol{\epsilon}(\mathbf{z}_t, c, m_\emptyset, \text{neg})$, shortened as $m_\emptyset$-*cfg*. Note that the positive score for all three compared methods is the same $\boldsymbol{\epsilon}(\mathbf{z}_t, c, m, \text{pos})$. The results show that our method mitigates the degradation reflected in the LPIPS and NIQE scores when using guidance rates of 10 and 14 in both cfg and $m_\emptyset$-cfg, while maintaining comparable performance. The visual results in Table 3 further demonstrate our superiority, where ours suppresses the color change and incorrect texture of high CFG guidance rate, leading to higher CLIPIQA score.

Effects of Latent Connector. To demonstrate the effectiveness of the proposed multimodal latent connector, we conducted an ablation study by training a new multimodal super-resolution model without the latent connector, which directly uses the longer multimodal token sequence as input. The quantitative performance is shown in Table 6. Our *w. MMLC* outperforms the *w/o. MMLC* in all reference and non-reference metrics. The visual results in Figure 7 show that the model without the MMLC is more likely to exacerbate hallucination. Its result shows bokeh on the branches but a clear trunk, even though the depth input clearly shows

Methods	*DIV2K-Val-3k* 512 × 512 PSNR	SSIM	LPIPS ↓	DISTS ↓	NIQE ↓	FID ↓	MUSIQ	CLIPIQA
BSRGAN	21.87	0.5539	0.4136	0.2737	4.7615	64.28	59.11	0.5183
R-ESRGAN	**21.94**	**0.5736**	0.3868	0.2601	4.9209	53.46	58.64	0.5424
LDL	21.52	0.5690	0.3995	0.2688	5.0249	58.94	57.90	0.5313
DASR	21.72	0.5536	0.4266	0.2688	4.8596	67.22	54.22	0.5241
FeMASR	20.85	0.5163	0.3973	0.2428	4.5726	53.70	58.10	0.5597
LDM	21.26	0.5239	0.4154	0.2500	6.4667	41.93	56.52	0.5695
StableSR	20.84	0.4887	0.4055	0.2542	4.6551	36.57	62.95	0.6486
ResShift	21.75	0.5422	0.4284	0.2606	6.9731	55.77	58.23	0.5948
PASD	20.77	0.4958	0.4410	0.2538	4.8328	40.77	66.85	0.6799
DiffBIR	20.94	0.4938	0.4270	0.2471	4.7211	40.42	65.23	0.6664
SeeSR	21.19	0.5386	0.3843	0.2257	4.9275	31.93	68.33	0.6946
MMSR	21.74	0.5693	**0.3707**	**0.2071**	**4.2532**	**29.35**	**70.06**	**0.7164**

Methods	*DIV2K-Val-100* 1024 × 1024 PSNR	SSIM	LPIPS ↓	DISTS ↓	NIQE ↓	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	21.77	**0.5813**	0.3624	0.1990	3.6573	0.4046	47.54	0.5358
StableSR	20.07	0.3947	0.5097	0.2427	3.6260	0.4113	65.39	0.6938
PASD	21.27	0.5369	0.3473	0.1753	3.6321	0.4708	69.69	0.6914
SUPIR	20.65	0.5350	0.3849	0.1814	3.6458	0.4051	65.88	0.5697
SeeSR	21.31	0.5578	0.3273	0.1620	4.0215	**0.5439**	69.79	0.6941
MMSR	**21.87**	0.5565	**0.2810**	**0.1492**	**3.4243**	0.4885	**72.31**	**0.7294**

Methods	*RealSR* 512 × 512 PSNR	SSIM	LPIPS ↓	LIQE	NIMA	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	**25.69**	**0.7616**	**0.2727**	3.3574	4.6548	0.5487	60.18	0.4449
StableSR	24.70	0.7085	0.3018	3.6106	4.8150	0.6221	65.78	0.6178
PASD	24.29	0.6630	0.3435	3.5749	4.8554	0.6493	68.69	0.6590
SUPIR	22.97	0.6298	0.3750	3.5682	4.5757	0.5745	61.49	0.6434
SeeSR	25.18	0.7216	0.3009	4.1360	4.9193	0.6442	69.77	0.6612
MMSR	24.83	0.7003	0.2952	**4.3468**	**5.1094**	**0.6578**	**71.33**	**0.6717**

Methods	*DrealSR* 512 × 512 PSNR	SSIM	LPIPS ↓	LIQE	NIMA	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	**28.64**	**0.8053**	**0.2847**	2.9255	4.3258	0.4907	54.18	0.4422
StableSR	26.71	0.7224	0.3284	3.2425	4.4861	0.5594	58.51	0.6357
PASD	27.00	0.7084	0.3931	3.5908	4.6618	0.5850	64.81	0.6773
SUPIR	24.61	0.6123	0.4294	3.4710	4.3815	0.5381	57.32	0.6758
SeeSR	26.75	0.7405	0.3174	4.1270	4.6942	0.6052	65.09	0.6908
MMSR	27.28	0.7456	0.3249	**4.5023**	**5.0558**	**0.6301**	**68.93**	**0.6999**

Table 1. Quantitative comparison with state-of-the-art methods on common benchmarks. Best and second-best results per metric are highlighted in **red** and blue, respectively. Note that StableSR could not produce 1MP results due to limitations in its tiling implementation.

they are at the same depth.

	MUSIQ	NIQE ↓	DISTS ↓	LPIPS ↓	Throughput
w/o. MMLC	69.69	3.4845	0.1781	0.3929	3.48 img/s
w. MMLC	**72.31**	**3.4243**	**0.1492**	**0.2810**	3.32 img/s

Table 2. Ablation of the impact of the Multimodal Latent Connector module (MMLC) on *DIV2K-Val-100* 1024p.

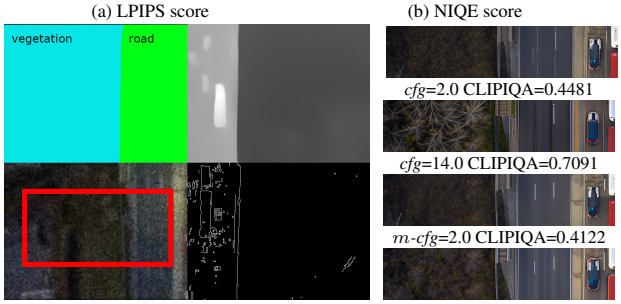
Contributions of Each Modality. Figure 8 shows the contributions of each modality by masking out input modalities during testing. Benefiting from the improved flexibility of using multimodal token $m_\emptyset$, discussed in Sec. 3.1, our method is robust in scenarios with only fewer modalities input during testing. We would like to note several observations: (1) Our default multimodal setting achieves the best trade-off between the non-reference metric MUSIQ and the reference-based metric DISTS. (2) Depth information primarily enhances perceptual quality (MUSIQ), while other modalities contribute more significantly to preserving identity (DISTS). These results highlight the diverse contributions of each modality and emphasize the advantage of our multimodal approach in effectively combining their strengths for optimal super-resolution performance.

guidance	2	10	14
cfg	0.3239	0.4491	0.5064
$m_\emptyset$-*cfg*	0.2815	0.4803	0.5493
m-cfg	**0.2810**	**0.3471**	**0.3772**

(a) LPIPS score

guidance	2	10	14
cfg	3.577	4.6179	5.1886
$m_\emptyset$-*cfg*	3.6261	5.1081	5.9175
m-cfg	**3.4679**	**3.7419**	**3.9815**

(b) NIQE score



(c) Visual comaprisons

Table 3. Ablation of guidance on *DIV2K-Val-100* 1024p. Our multimodal CFG (*m-cfg*) mitigates artifacts often present when using high guidance rates. This leads to an improved balance between visual fidelity and preservation of key identifying features.

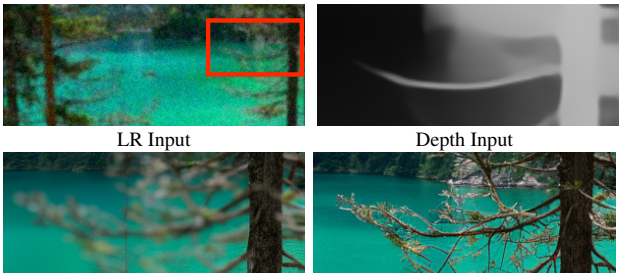
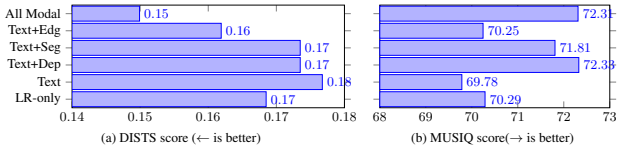


Figure 7. Visual results when (not) using the MMLC module.



(a) DISTS score (← is better) (b) MUSIQ score(→ is better)

Figure 8. Ablation of modalities on *DIV2K-Val-100* 1024p. Incorporating different modalities enhances various aspects of the super-resolution results compared to the text-guided baseline.

4.2. Qualitative Results

Figures 5 and 6 provide a visual comparison of our method against state-of-the-art approaches, namely SeeSR [65], PASD [68], and SUPIR [71]. Our method demonstrates

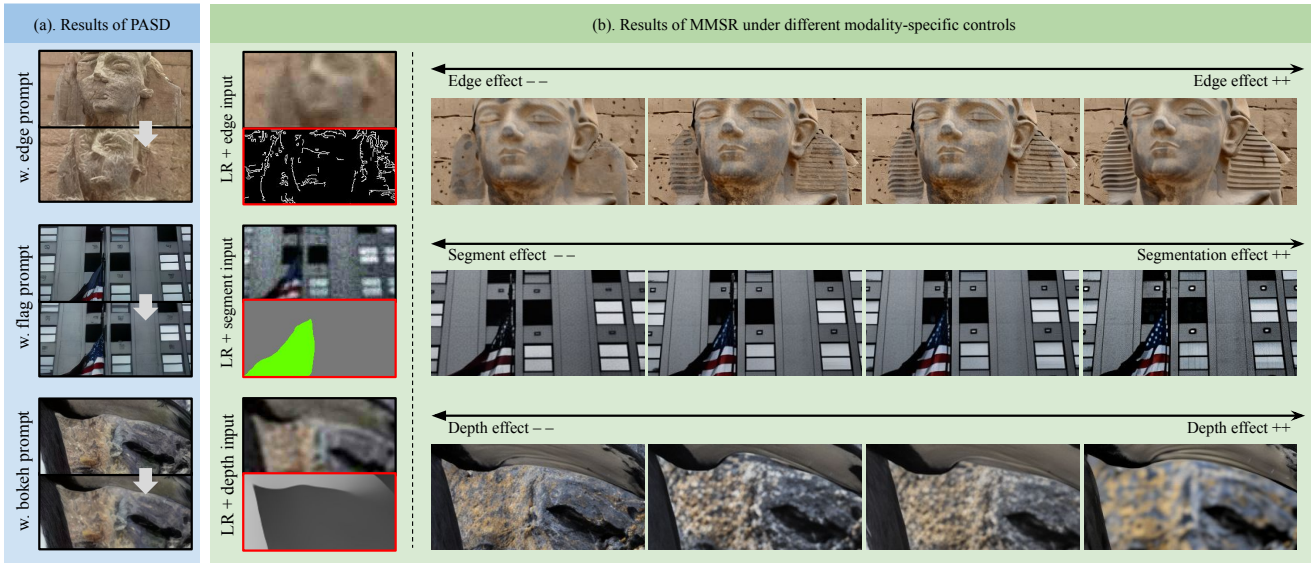


Figure 9. Our method allows for fine-grained control over super-resolution results by adjusting the influence of each input modality. For example, reducing the edge temperature enhances edge sharpness (first row). Lowering the segmentation temperature emphasizes distinct features, such as the star pattern on the flag (second row). Decreasing the depth temperature accentuates depth-of-field effects, like the bokeh between the foreground and background (third row). In contrast, PASD [68] exhibits limited control over such fine-grained details.

several key advantages. *Enhanced Realism:* In the first example of Figure 6, our method produces fewer artifacts and remains more faithful to the high-resolution image compared to other methods. This highlights our ability to maintain realism without introducing spurious details. *Robustness to Challenging Conditions:* The second example in Figure 6, a long-range shot affected by noise and turbulence, shows our method's ability to generate clear human details where other methods struggle. This is attributed to the effective use of semantic segmentation for accurate human localization, further enriched by the textual caption. These examples illustrate the effectiveness of our multimodal approach in generating high-quality images across diverse scenarios. See supplemental material for additional results and analysis.

4.3. Controllability Comparisons

While recent works leverage text prompts for controlling super-resolution [65, 68, 71], these often lack fine-grained control and can yield inconsistent results. Our multimodal approach introduces modality-specific temperature weights to scale attention scores, enabling precise manipulation of SR outputs by amplifying or diminishing the influence of each modality (e.g., depth, segmentation). Figure 9 contrasts our approach with PASD [68], which relies solely on text prompts. Specifically, decreasing the temperature of the edge modality changes the richness of details in our result, while changing PASD's prompt with "more edge" doesn't change their result. Decreasing the temperature of semantic segmentation map and depth map also lead to manageable changes such as more visible star parttern and stronger bokeh. In contrast, directly add corresponding prompt in PASD's prompt barely changes its result. Moreover, our approach exhibits smooth transitions in image characteristics as temperatures are adjusted, providing interpretable control and insights into the role of each modality.

5. Conclusion

This work introduces a novel diffusion-based framework for image super-resolution that seamlessly integrates diverse modalities—including text descriptions, depth maps, edges, and segmentation maps—within a single, unified model. By leveraging pretrained text-to-image models, our method achieves enhanced realism and accurate reconstruction, outperforming existing text-guided super-resolution methods both qualitatively and quantitatively. Furthermore, a learnable multimodal token and modality-specific contribution controls provide fine-grained control over the super-resolution process, enabling adjustable perception-distortion tradeoffs and robust performance even with imperfect or missing modalities in most cases.

Limitations and Future Work. While adding multimodal information significantly enhances SR performance, it introduces computational overhead. For instance, using Gemini Flash for image captioning results in a throughput of 0.34 images per second, which is slower than depth at 1.99 img/s, semantic segmentation at 2.09 img/s, and DDIM sampling at 0.54 img/s. However, cross-modal predictions can be parallelized, and our method achieves speeds comparable to other text-driven SR models [68, 71]. Future work will explore optimizing the vision-language component for faster

inference and investigate more robust modules for extracting modality-specific information, potentially enhancing performance even with noisy or incomplete inputs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4M-21: An any-to-any vision model for tens of tasks and modalities. *NeurIPS*, 2024. 2, 3

[3] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, pages 22861–22872, 2024. 3

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023. 1, 2

[5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237, 2018. 1

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1, 2, 5

[7] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pages 14455–14465, 2024. 1, 2

[9] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 6

[10] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25857–25867, 2024. 2

[11] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 1

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 5

[13] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *Forty-first International Conference on Machine Learning*, 2024. 2

[14] Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research*, 2023. Featured Certification. 1

[15] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. 4

[16] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. In *Proceedings of the 29th ACM International Conference on multimedia*, pages 2483–2491, 2021. 6

[17] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2015. 1

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2

[20] Kanchana Vaishnavi Gandikota and Paramanand Chandramouli. Text-guided explorable image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25900–25911, 2024. 2

[21] Jiatao Gu, Ying Shen, Shuangfei Zhai, Yizhe Zhang, Navdeep Jaitly, and Joshua M Susskind. Kaleido diffusion: Improving conditional diffusion models with autoregressive latent modeling. *Advances in neural information processing systems*, 2024. 2

[22] Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. Dart: Denoising autoregressive transformer for scalable text-to-image generation. *arXiv preprint arXiv:2410.08159*, 2024. 2

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 4

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4

[26] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, pages 9161–9175, 2023. 1

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[28] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 4

[29] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532, 2018. 1

[30] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[31] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 6

[32] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*. Springer, 2022. 6

[33] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *ArXiv preprint*, 2023. 2, 6

[34] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 1, 2

[35] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[37] Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9048–9058, 2024. 2

[38] Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M Patel, and Peyman Milanfar. Bigger is not always better: Scaling properties of latent diffusion models. *arXiv preprint arXiv:2404.01367*, 2024. 2

[39] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 1

[40] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[41] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. *NeurIPS*, 2023. 2, 3

[42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 3

[43] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 4

[44] Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In *European Conference on Computer Vision*, 2024. 2

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[46] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 4

[47] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[48] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Spire: Semantic prompt-driven image restoration. In *European Conference on Computer Vision*, 2024. 2

[49] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pages 285–303. Springer, 2024. 1, 2

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 4

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1, 2, 5, 6

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[54] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 4

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 5

[56] Li-Yuan Tsao, Yi-Chen Lo, Chia-Che Chang, Hao-Wei Chen, Roy Tseng, Chien Feng, and Chun-Yi Lee. Boosting flow-based generative super-resolution models via learned prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26005–26015, 2024. 2

[57] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[58] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 2, 4, 6

[59] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 4

[60] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023. 2

[61] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[62] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Realesrgan: Training real-world blind super-resolution with pure synthetic data supplementary material. *Computer Vision Foundation open access*, 2022. 5, 6

[63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 2

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[65] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 1, 2, 4, 6, 7, 8, 3

[66] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 5

[67] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 6

[68] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, 2024. 1, 2, 6, 7, 8

[69] Zhixiong Yang, Jingyuan Xia, Shengxi Li, Xinghua Huang, Shuanghui Zhang, Zhen Liu, Yaowen Fu, and Yongxiang Liu. A dynamic kernel prior model for unsupervised blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26046–26056, 2024. 2

[70] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3

[71] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 1, 4, 5, 7, 8

[72] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6

[73] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 6

[74] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[76] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In *European Conference on Computer Vision*, 2024. 2

[77] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-

language models represent space and how? evaluating spatial frame of reference under ambiguities, 2024. 1

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

Supplementary Material

Table of Contents

6. 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.

7. 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.

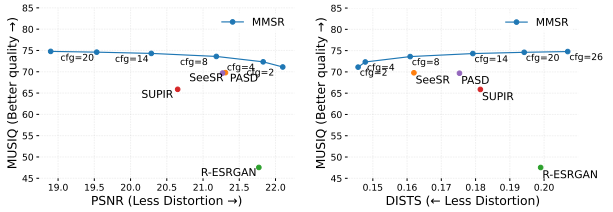


Figure 10. Our method improves the perception distortion trade-off of the past super-resolution methods.

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.

8. Extension to Diffusion Transformers

Diffusion Transformers, such as DiT [45] and MMDiT [19], operate on tokenized image patches during the diffusion process. In contrast to diffusion models based on U-Nets [52], 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 [19]. 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.

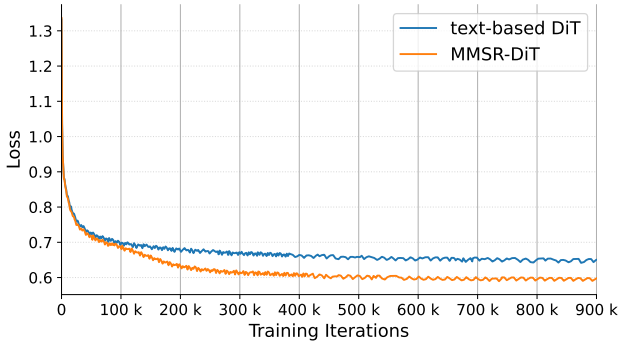


Figure 12. Training loss comparisons between the text-based diffusion transformer and our multi-modal-based diffusion transformer.

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.

9. Dependence on Multimodal Input Quality

Our method leverages the prior knowledge encapsulated in pretrained cross-modal predictors, including Gemini Flash [55], Depth-anything [66], 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.

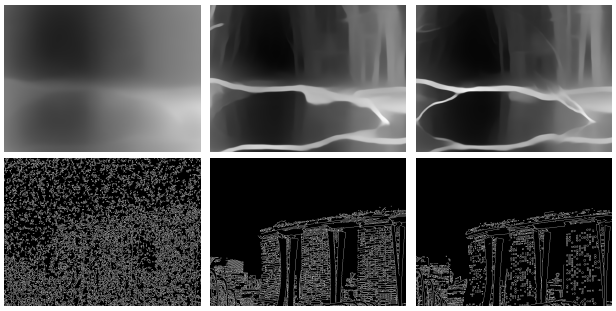


Figure 13. For low-quality cross-modal estimation from the low-resolution image, our method can increase the estimation accuracy by conducting zero-modal SISR on the low-resolution image.

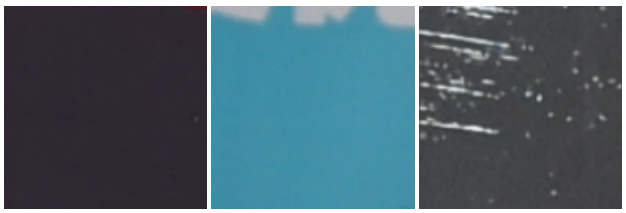


Figure 14. Special cases in DRealSR benchmark visualization.

10. Image Captioning Prompt Engineering

When tasked with image captioning, vision-language models (VLMs) like Gemini [55] 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, 64]. 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.

11. 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[65]. 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.

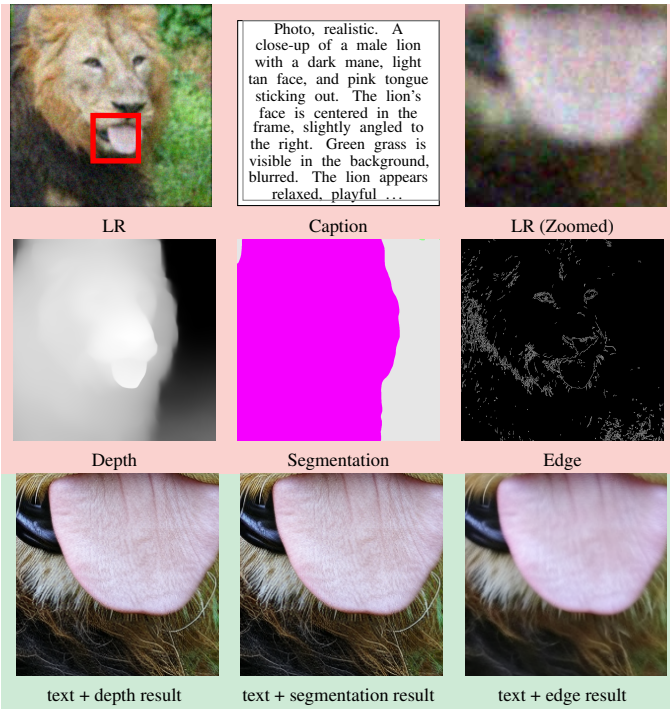


Figure 15. Visual results of our method under different modalities input. We find that the visual quality changes according the previous shown trend in different metrics, such as *text+depth* and *text+segment* have most details but less identical to the input.

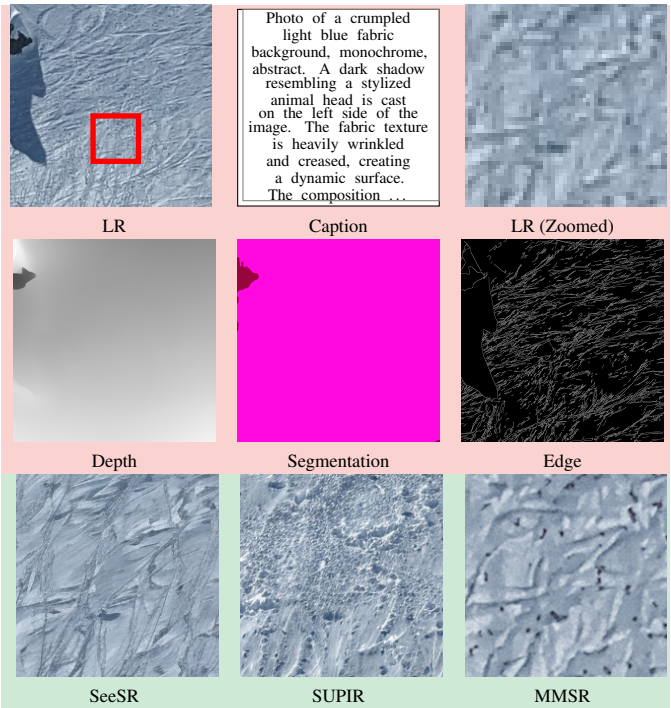


Figure 16. Failing case visualization.

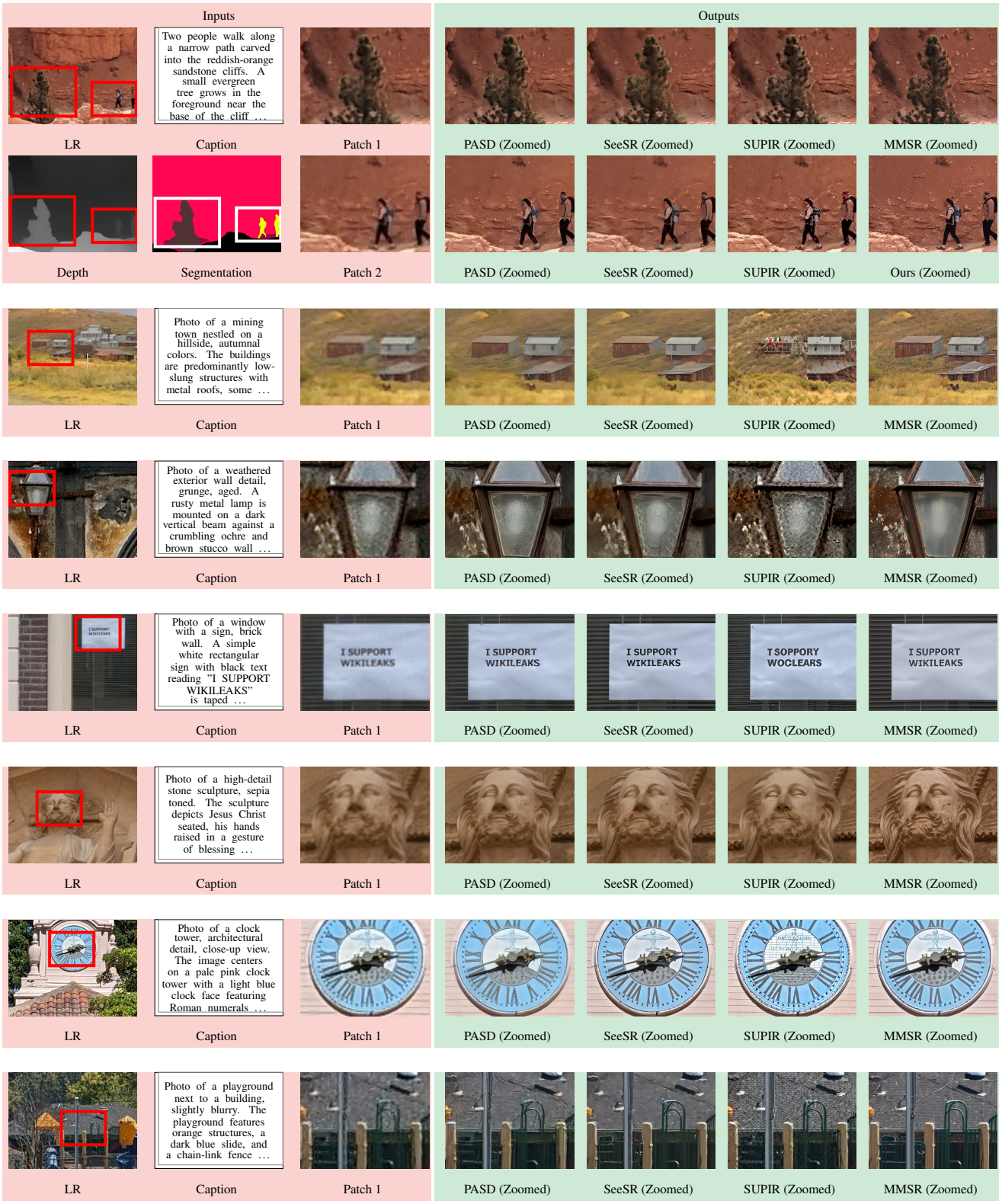
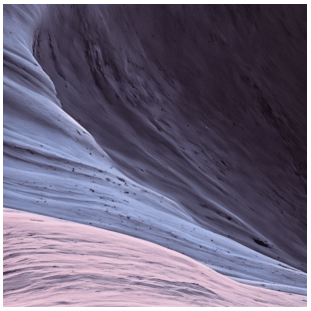
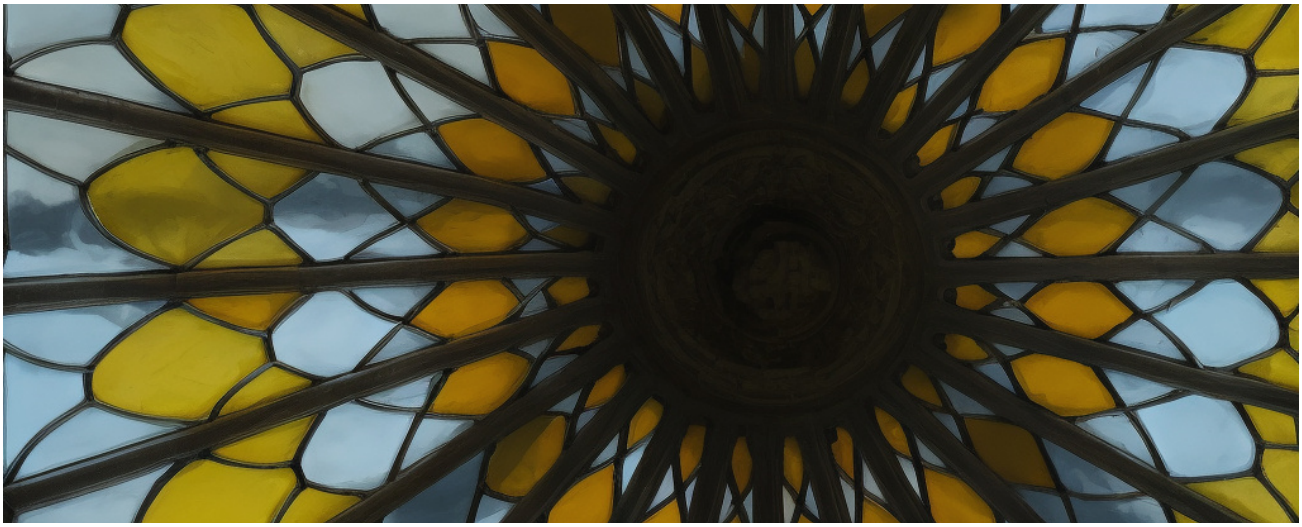


Figure 17. MMSR super-resolution results on real-world images compared with state-of-the-art methods. Zoom in to appreciate the details.

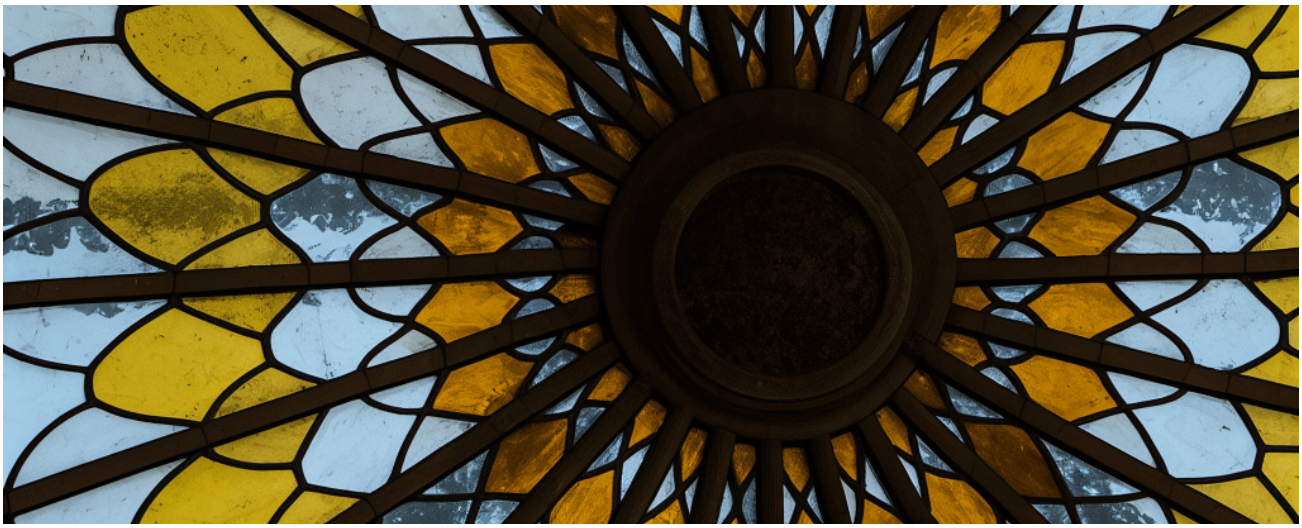
Table 7. Image caption result comparisons between different prompts. We show that our prompt that utilizes in-context learning is stable at most cases and can always get more detailed image captions without useless procedural words.

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.