Ultra High-Resolution Image Inpainting with Patch-Based Content Consistency Adapter

Supplementary Material

1. More details.

Random Mask. During training, our random masks are generated via three distinct strategies: (i) using a brush that emulates human strokes; (ii) employing random geometric shapes—such as rectangles and ellipses with arbitrary positions and sizes; and (iii) composing composite shapes by overlapping several random geometric forms. See the Fig. 8.

Hierarchical text prompting. As illustrated in the Fig. 9, we present local patch-specific prompt examples annotated via the Vision-Language Models (VLM). This text-guided strategy effectively ensures the plausibility of fine-grained content details.

Reference Patch Selection Strategy. As illustrated in Algorithm 1, our reference patch retrieval process locates semantically consistent regions in the original image by measuring the cosine similarity between CLIP embeddings of candidate patches and the masked region, thereby optimizing reference information through feature-space proximity. The semantic distance is exclusively determined by this cosine metric, which effectively captures the semantic alignment between patches.

Additional ablation study for fairness. As shown in Tab. 4, we conducted experiments on the same dataset and training-setting(Sec.4.3), comparing *(i) full-parameter UNet fine-tuning*, *(ii) LoRA fine-tuning*, and *(iii) DCA* (we proposed).

Algorithm 1 Reference Patch Selection Strategy.

Input: Upsampled stage 1 Output Patches which is masked X_m, Original Patches X_{orig}, Clip model $\mathcal{C}$
Output: Reference Patches X_{ref}
for Masked Patch X_m^i in X_m **do**
  Compute CLIP embedding $\mathcal{C}(X_m^i)$
  $d_{max} \leftarrow -1$, $X_{BestMatch} \leftarrow \varnothing$
  for Original Patch X_{orig}^j in X_{orig} **do**
    Compute CLIP embedding $\mathcal{C}(X_{orig}^i)$
    Compute distance $d \leftarrow \frac{\mathcal{C}(X_m^i)^\top \mathcal{C}(X_{orig}^j)}{\|\mathcal{C}(X_m^i)\|_2 \|\mathcal{C}(X_{orig}^j)\|_2}$
    if $d > d_{max}$ **then**
      $d_{max} \leftarrow d$, $X_{BestMatch} \leftarrow X_{orig}^j$
    end if
  end for
  $X_{ref}^i \leftarrow X_{BestMatch}$
end for
Return Reference Patches X_{ref}

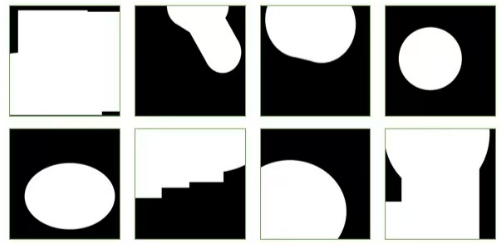

Figure 8. Random masks strategies.

2. Additional Qualitative Results.

Comparison with Super-Resolution Models Integrated with Blend Diffusion. Our analysis of state-of-the-art super-resolution models integrated with blend diffusion, as illustrated in the Fig. 10, reveals that these approaches often fail to maintain fine-detail consistency between masked and unmasked regions, leading to unpredictable, low-quality textures and conspicuous seams. In contrast, our model effectively leverages information from the unmasked areas and inter-patch cues to integrate global image coherence, resulting in aesthetically pleasing inpainting outcomes.

Under both global and local text prompts. We present a qualitative evaluation (see Fig. 11) comparing our model with and without the Dual Context Adapter (DCA) module. Our results demonstrate that, although the DCA module is fine-tuned exclusively with global text, its integration as an attention-based plug-in does not adversely affect the baseline model. In fact, in most cases, the model with DCA exhibits an enhanced understanding of the image context.

Additional qualitative evaluations. Against other inpainting models are presented in Fig. 12, Our two-stage model demonstrates outstanding performance by optimizing both contextual understanding and fine detail generation.

	Random Masks and Global Prompt			
Model Name	**FID ↓**	**Aesthetic score ↑**	**CLIP Score ↑**	**LPIPS ↓**
SDXL-I	13.326	5.480	26.268	0.129
SDXL-I + LoRA-fine-tuning	13.284	5.376	26.119	0.150
SDXL-I + UNet-fine-tuning	13.168	5.405	26.249	0.152
DCA(Ours)	**12.167**	**5.591**	**26.458**	**0.128**
	Segmentation Masks and Local Prompt			
Model Name	**FID ↓**	**Aesthetic score ↑**	**CLIP Score ↑**	**LPIPS ↓**
SDXL-I	9.565	5.559	26.990	0.092
SDXL-I + LoRA-fine-tuning	9.725	5.415	26.581	0.091
SDXL-I + UNet-fine-tuning	9.683	5.445	26.684	0.093
DCA(Ours)	**9.427**	**5.598**	**27.002**	**0.089**

Table 4. Following the DCA ablation (Sec.4.3), we fine-tuned SDXL-Inpainting under identical settings to enable a fair comparison with our DCA module.

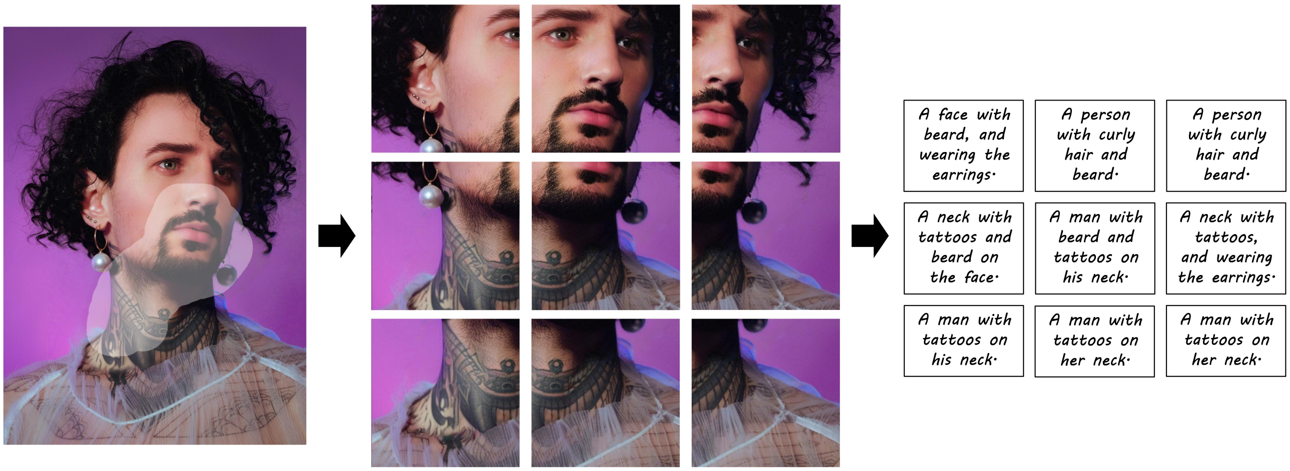

Figure 9. Each patch is assigned a dedicated prompt via the VLM, and overlaps between patches are introduced during segmentation to ensure coherent generation.

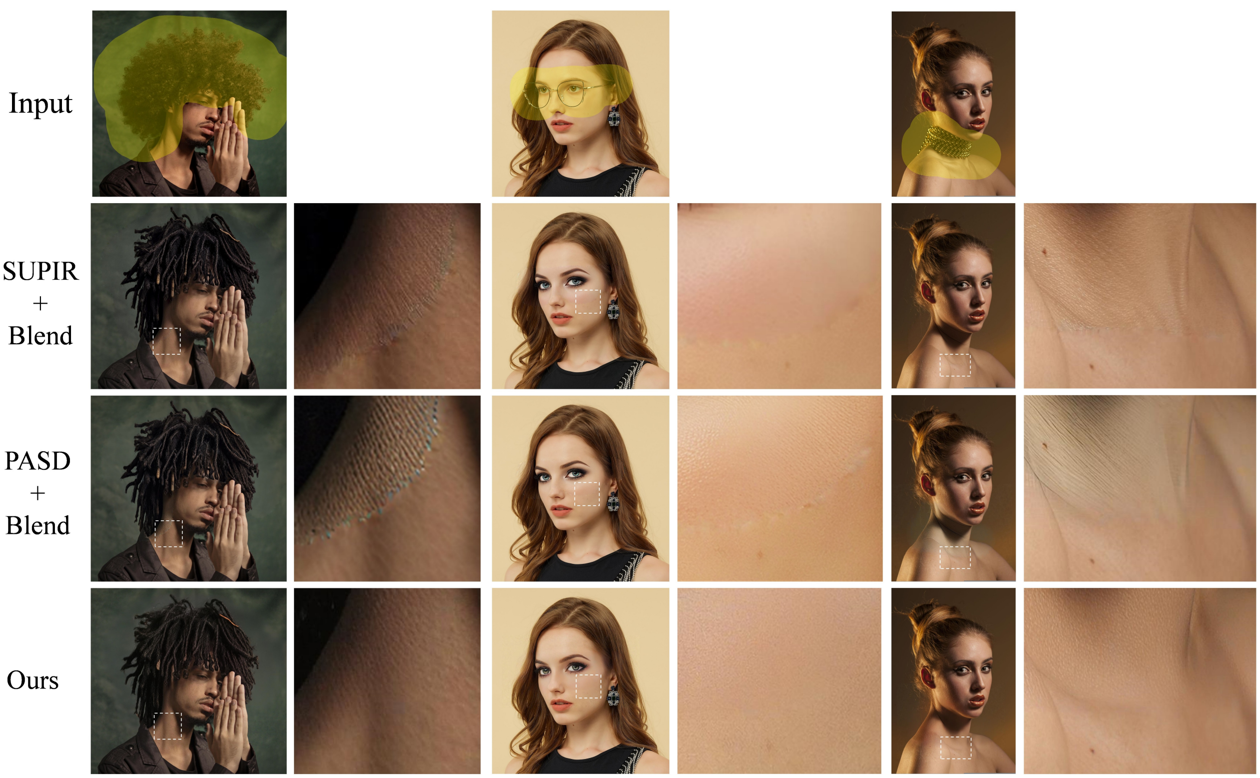

Figure 10. Compare with super-resolution generation models combined with blend diffusion. These methods often fail to account for the context inside and outside the masked regions, leading to issues such as texture inconsistencies, seams, and color discrepancies.

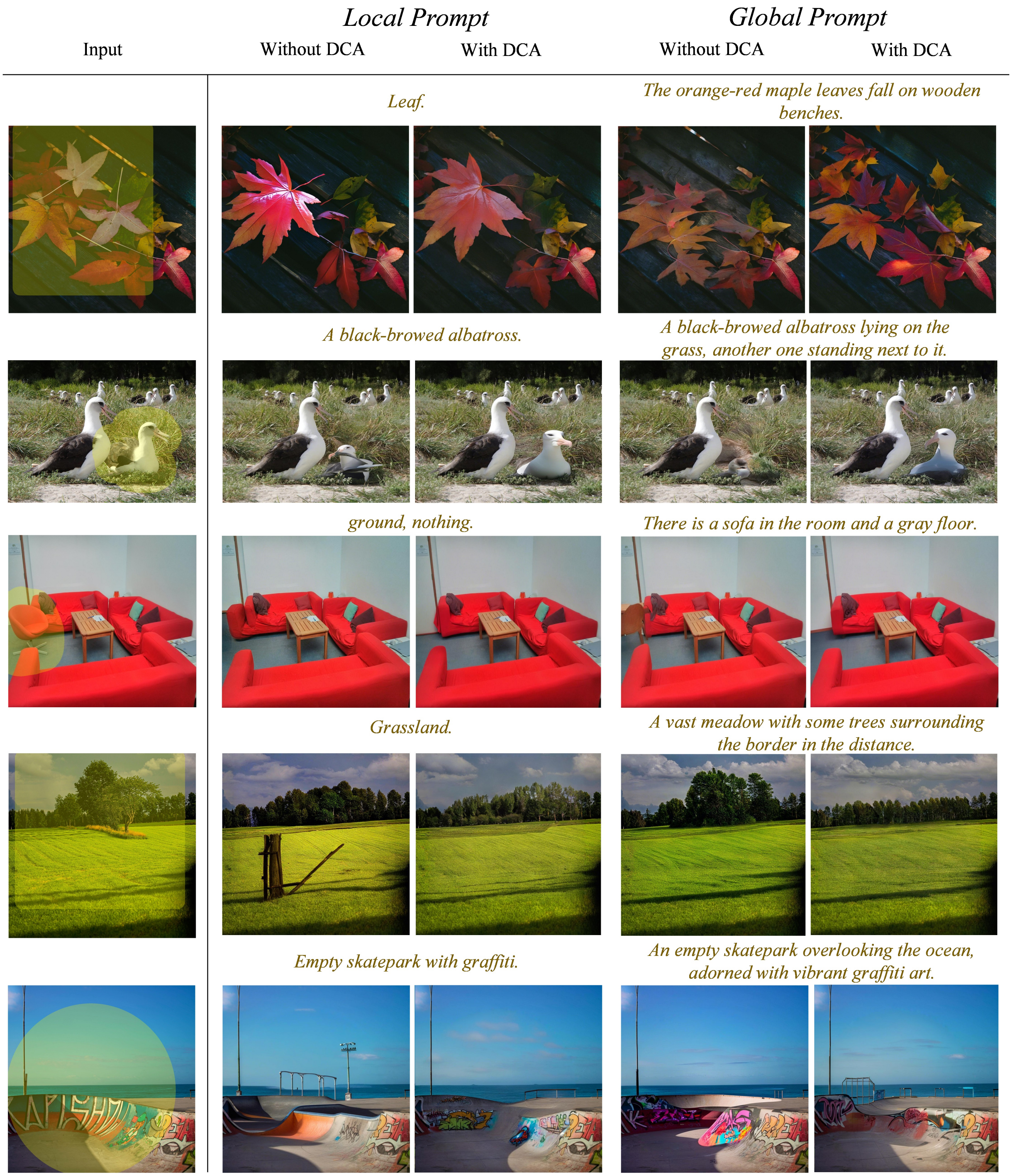

Figure 11. By fine-tuning DCA with global text prompt, we enhance the model's utilization of image context via a plugin-based approach, without compromising its inherent ability to comprehend short local texts.

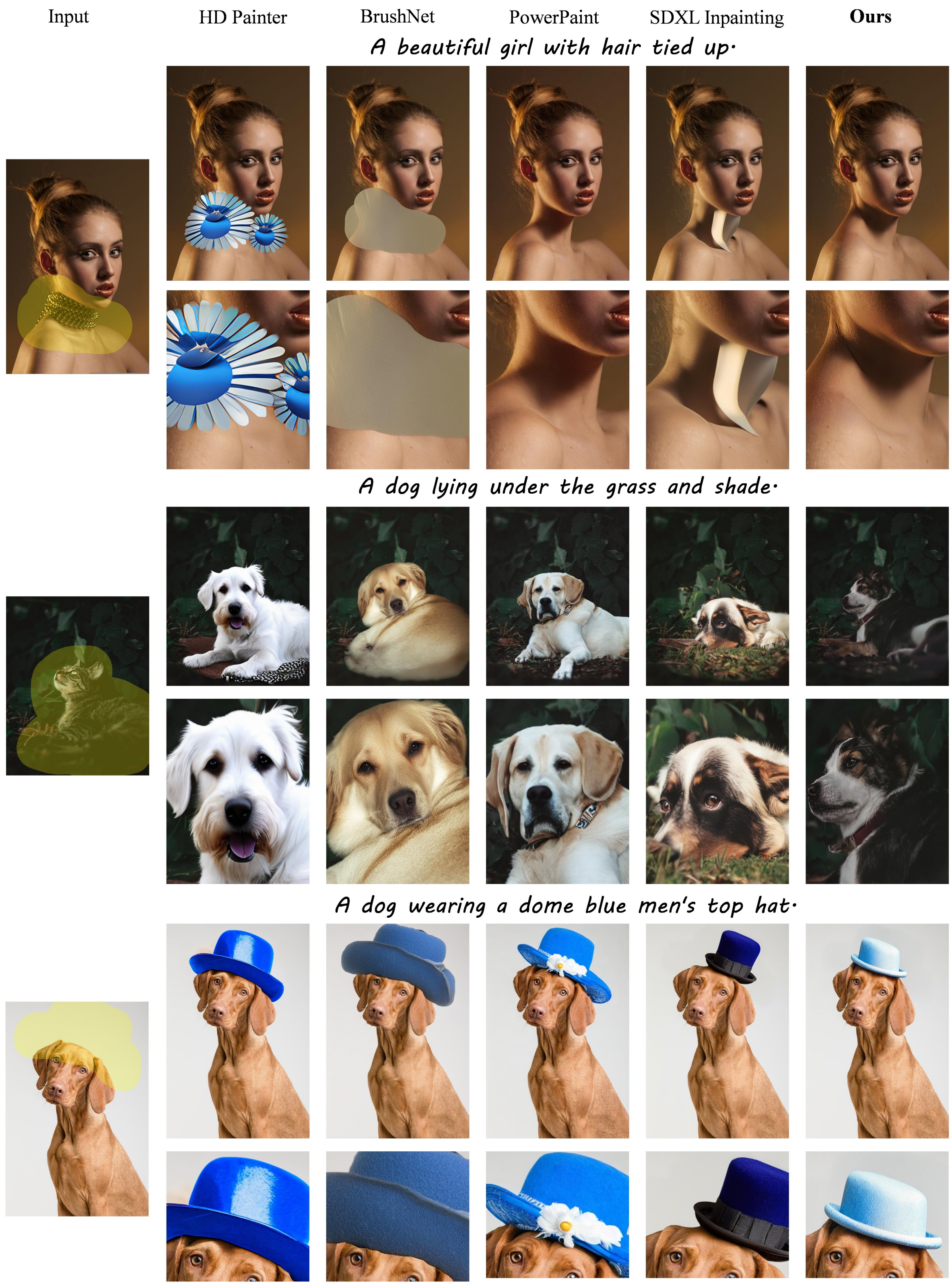

Figure 12. Our model preserves both the structural correctness and aesthetic quality of the content while generating more refined details, ensuring that high-resolution inpainting outputs faithfully match the original image's level of detail.