I Dream My Painting: Connecting MLLMs and Diffusion Models via Prompt Generation for Text-Guided Multi-Mask Inpainting

Supplementary Material

1. Additional Dataset Information

This section provides additional details on the automatically annotated dataset of digitized images of paintings from $\mathcal{A}rt\mathcal{G}raph$. As described in the main paper, the images were downloaded as thumbnails from the WikiArt API, using the formats *HalfHD*, *Large*, and *PinterestLarge*. For each image, we selected the highest resolution available. The counts and resolutions of the images are summarized in Table 1.

All 116,475 downloaded images were processed with Kosmos-2, generating grounded descriptions with bounding boxes around the primary objects depicted. After applying the size constraints detailed in the main paper, the dataset was refined to 102,276 images. We then used LLaVA to generate more detailed object-level descriptions for each object, limiting the descriptions to 40 tokens per object. Figure 2 illustrates two examples of images annotated using Kosmos-2 and LLaVA. As illustrated in the samples, automatically generated descriptions are inherently error-prone. However, in the main paper, we assessed their quality using CLIP, indicating that the alignment between the images and the generated texts remains acceptable. Caption errors may have had a greater impact on training the multi-mask inpainting diffusion model than on the MLLM used for prompt generation. This is because the original objects in the image are not directly involved in calculating the causal language modeling loss, which relies only on the text and the corrupted image input.

Additionally, we analyze the distribution of the number of masks per image both before (Fig. 3a) and after (Fig. 3b) applying a maximum limit of five objects for multi-mask inpainting, along with a global threshold on the covered area. We also present the distributions of the 50 most common objects, categorized by their respective noun chunks (Fig. 3c) and noun chunk roots (Fig. 3d), which were used to compute the prompt generation accuracy. These distributions reveal two long-tailed patterns, with a strong skew towards common concepts.

2. Additional Experiments

In this section, we present additional experiments that were not discussed in the main paper.

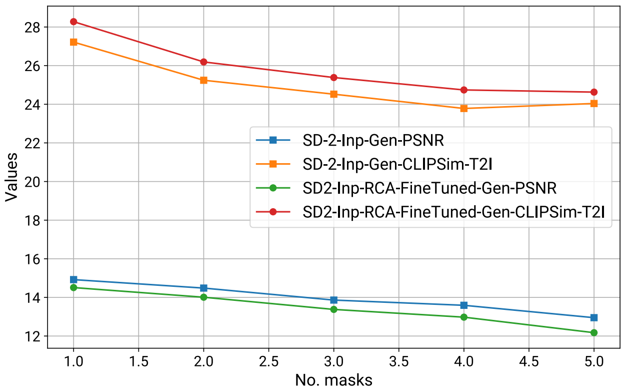


Figure 1. Comparative analysis on multi-mask inpainting.

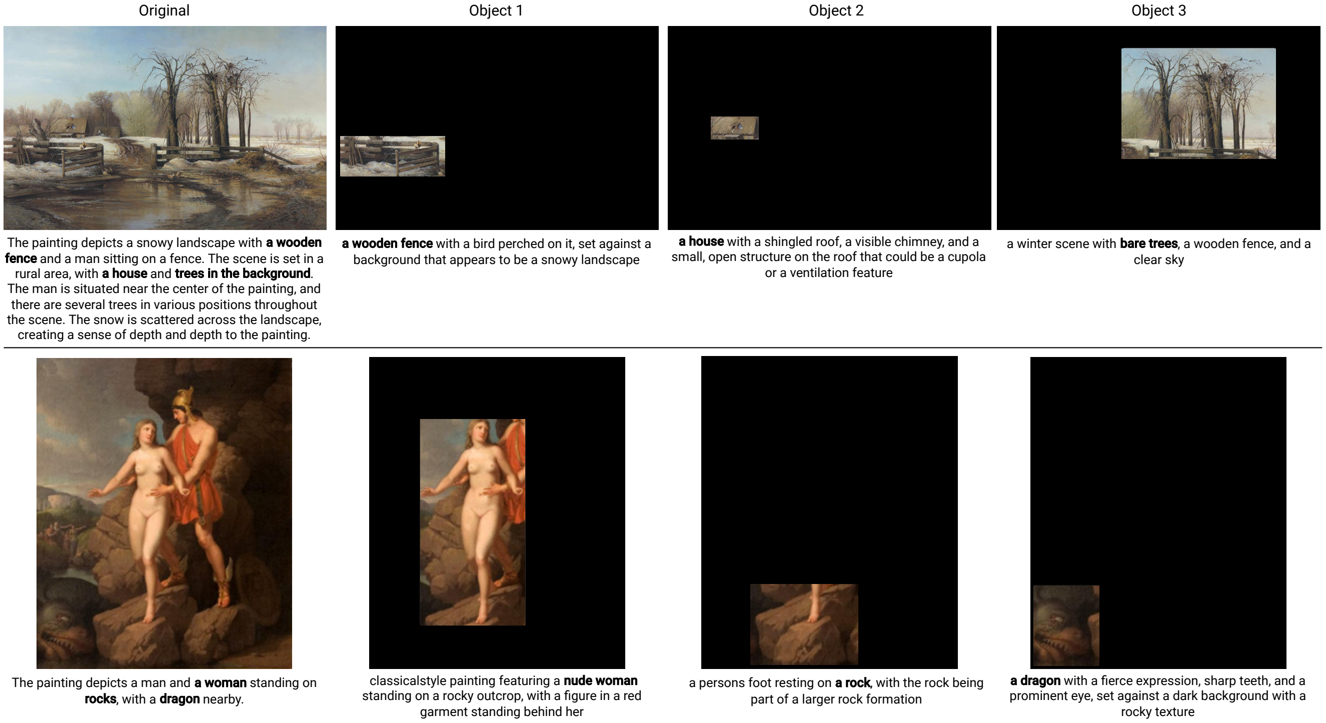
2.1. Analysis on the Number of Masks

We compare the base Stable Diffusion-2-Inpainting model and our RCA fine-tuned version, using generated prompts to evaluate performance across varying numbers of masks (Fig. 1). The figure demonstrates that our method consistently achieves superior region-prompt alignment compared to the base Stable Diffusion model, regardless of the number of masks. Notably, at $N = 5$, the curves begin to converge, suggesting that multi-mask inpainting is particularly effective for cases with fewer masks ($N \geq 1$). This could be because as the number of masks increases, the inpainting area expands, leading to more intersections between masks and longer prompts containing multiple sub-prompts. This complexity may reduce the precision of metric calculations, potentially diminishing the measurable advantages of RCA.

Regarding PSNR values, the base Stable Diffusion model consistently scores slightly higher. However, qual-

Thumbnail name	Width	Height (max val.)	No. of images
HalfHD	1366	800	58744
Large	750	600	31841
PinterestLarge	280	1120	25890

Table 1. Image thumbnail specifics for the 116,475 artwork in $\mathcal{A}rt\mathcal{G}raph$, downloaded from the WikiArt API.



The painting depicts a snowy landscape with **a wooden fence** and a man sitting on a fence. The scene is set in a rural area, with **a house** and **trees in the background**. The man is situated near the center of the painting, and there are several trees in various positions throughout the scene. The snow is scattered across the landscape, creating a sense of depth and depth to the painting.

a wooden fence with a bird perched on it, set against a background that appears to be a snowy landscape

a house with a shingled roof, a visible chimney, and a small, open structure on the roof that could be a cupola or a ventilation feature

a winter scene with **bare trees**, a wooden fence, and a clear sky

The painting depicts a man and **a woman** standing on **rocks**, with a **dragon** nearby.

classicalstyle painting featuring a **nude woman** standing on a rocky outcrop, with a figure in a red garment standing behind her

a persons foot resting on **a rock**, with the rock being part of a larger rock formation

a dragon with a fierce expression, sharp teeth, and a prominent eye, set against a dark background with a rocky texture

Figure 2. Automatically annotated samples from the art dataset. The first column displays the original images, along with the global grounded descriptions produced by Kosmos-2. Objects of valid sizes are shown in the other columns, with the rest of the image masked. These objects were cropped and provided to LLaVA to obtain the displayed object-level descriptions, which we used as training prompts. As demonstrated in the second example, certain masks can overlap. Our RCA implementation allows the intersection areas to attend to the prompts of both masks. Both samples present three valid masks.

itative assessments do not support the superiority of SD-2-Inp over RCA, indicating that PSNR can be influenced by various factors—such as the base model's tendency to fill masked regions with common or background elements—potentially skewing its evaluation.

2.2. Domain Transfer

We conducted an additional experiment where we applied the weights learned from the automatically annotated art dataset to the photographic images in the DCI dataset.

As shown in Table 2, while the model trained directly on the DCI dataset achieves higher scores in LPIPS and text-to-image CLIPSim relative to the ground truth annotations, it is slightly outperformed in PSNR and significantly in CLIP-IQA. This latter result suggests that training on the art domain and testing on photographic images can lead to higher-quality outputs in certain metrics.

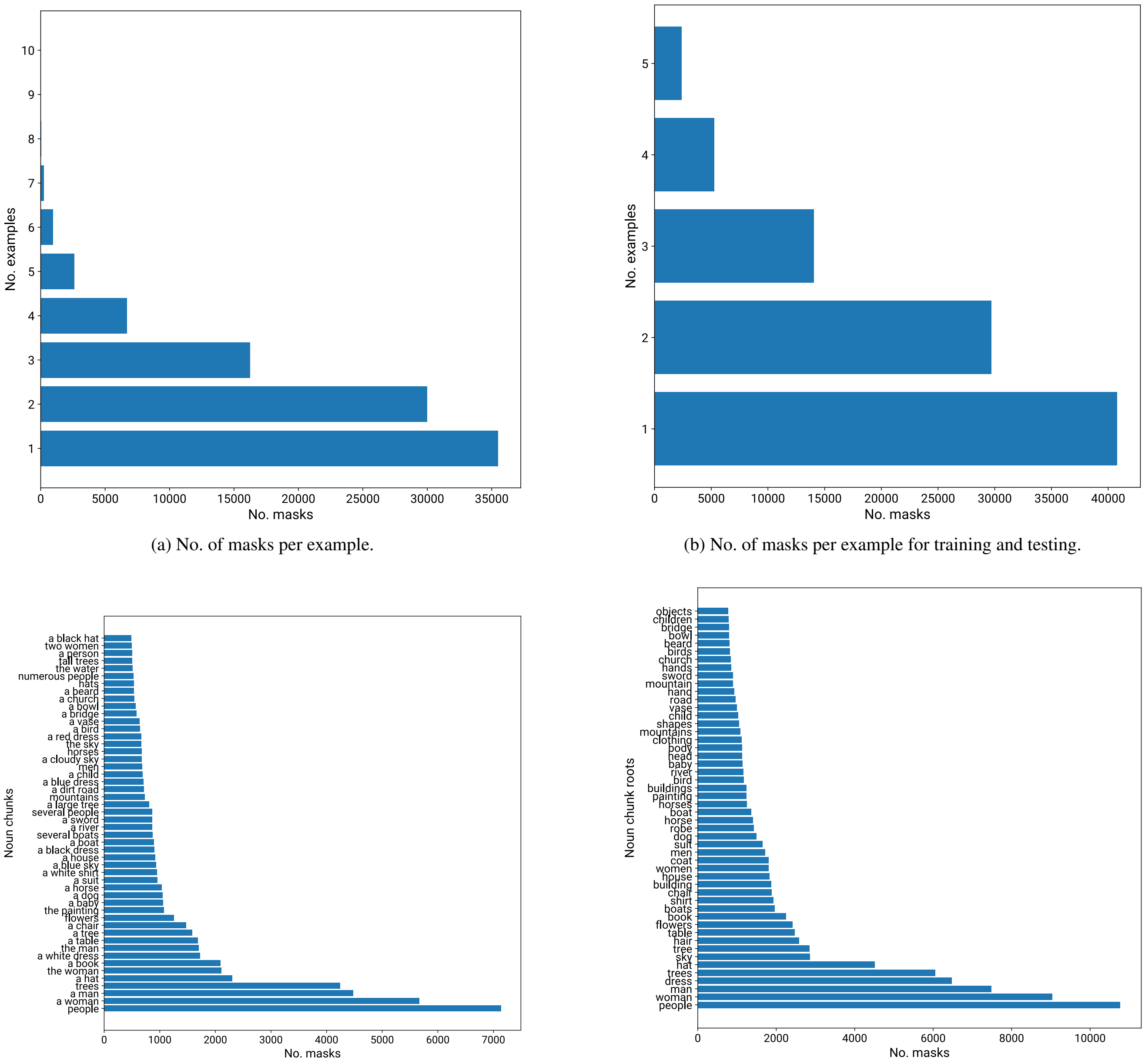
	L ↓	P ↑	Q ↑	C ↑
DCI pipeline	**30.26**	12.16	74.56	**23.66**
Art pipeline	30.69	**12.26**	**80.25**	23.00

Table 2. Comparative results on DCI between the dataset-specific pipeline and the pipeline transferred from the art domain (P: PSNR; L: LPIPS; Q: CLIP-IQA; C: CLIPSim-T2I).

We analyzed the qualitative results shown in Fig. 4 to better understand these outcomes. These examples illustrate that the model trained on the art dataset tends to apply stylistic features in its completions, often generating inpainted objects reminiscent of paintings from previous centuries, similar to those found in the WikiArt collection. This results in more colorful and aesthetically appealing completions. However, the generated objects are typically less precise, and the overall inpainting shows more artifacts compared to those generated for purely artistic images. This observation indicates that transferring generative capabilities between different domains remains a promising area for further research.

3. Region-Aligned CLIPSim Computation

As detailed in the main paper, we adopt the approach proposed by Lüddecke and Ecker in their work on text-guided segmentation [1] to calculate the CLIP similarity between a regional prompt and the corresponding output. In Fig. 5, we visualize how this method is employed as a met-



(a) No. of masks per example.

(b) No. of masks per example for training and testing.

(c) Most common noun chunks.

(d) Most common noun chunk roots.

Figure 3. Dataset statistics on the number of masks per example and the objects depicted in the masks (noun chunks).

ric for evaluating the inpainted regions. The results indicate that this metric aligns well with the assessment of prompt adherence in image inpainting tasks.

4. Additional Qualitative Results

We provide additional qualitative results of our pipeline for prompt generation and multi-mask inpainting both on the DCI dataset (Fig. 6) and on the art dataset, over multiple numbers of inpainting masks (Figs. 7–11).

5. Discussion on Potential Misuse

In the main paper, we introduced a new pipeline for inpainting multiple regions of an input image using different text prompts. This approach leverages MLLMs and diffusion models, which are currently state-of-the-art in text and image generation. As research advances, these models promise exciting opportunities for creating powerful tools like the one presented here. The results demonstrate that current technology is mature enough to enable MLLMs to interpret images across diverse domains, including complex

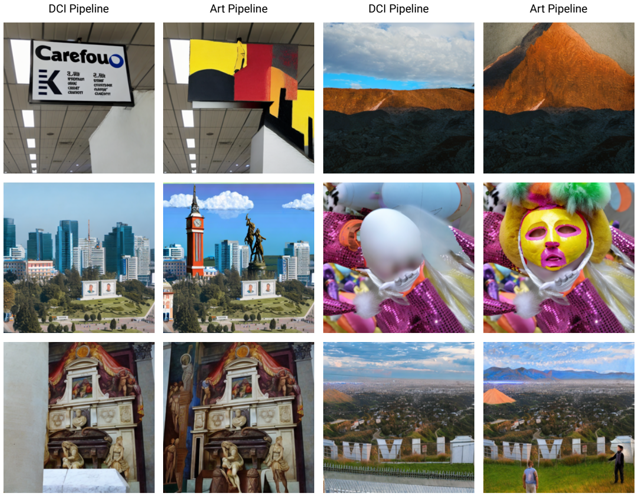


Figure 4. Domain transfer qualitative results.

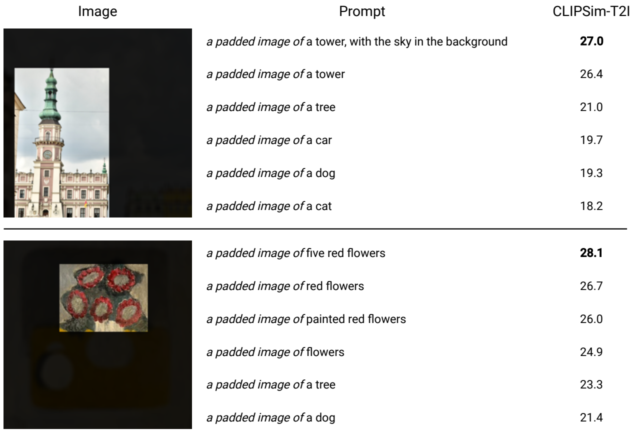


Image	Prompt	CLIPSim-T2I
	a padded image of a tower, with the sky in the background	**27.0**
	a padded image of a tower	26.4
	a padded image of a tree	21.0
	a padded image of a car	19.7
	a padded image of a dog	19.3
	a padded image of a cat	18.2
	a padded image of five red flowers	**28.1**
	a padded image of red flowers	26.7
	a padded image of painted red flowers	26.0
	a padded image of flowers	24.9
	a padded image of a tree	23.3
	a padded image of a dog	21.4

Figure 5. Visualization of our CLIPSim computation to evaluate inpainting prompt-following. By darkening and blurring the rest of the image, we obtain scores aligned to the region of interest.

areas such as artistic images, with high accuracy.

Looking ahead, it is anticipated that MLLMs will increasingly serve as invaluable assistants in various image-related tasks, such as image editing. While some limitations remain—for example, challenges in inpainting large and small areas or managing long prompts, which can sometimes result in noticeable artifacts—ongoing advancements in image generation are expected to continue enhancing these capabilities.

As the technology progresses, the distinction between real and generated images may become increasingly subtle, underscoring the need for innovation and caution. The proposed pipeline provides a practical tool for assisting users, including those with minimal experience in Generative AI, and has potential applications in automating processes such as data augmentation in computer vision. However, remembering the risks, including the accelerated generation of harmful content, is crucial. Therefore, as these tools are refined, it is equally important to develop more effective methods for distinguishing real images from generated ones, thereby helping to mitigate the risk of misuse and ensuring the integrity of digital content.

References

[1] Timo Lüddecke and Alexander Ecker. Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.

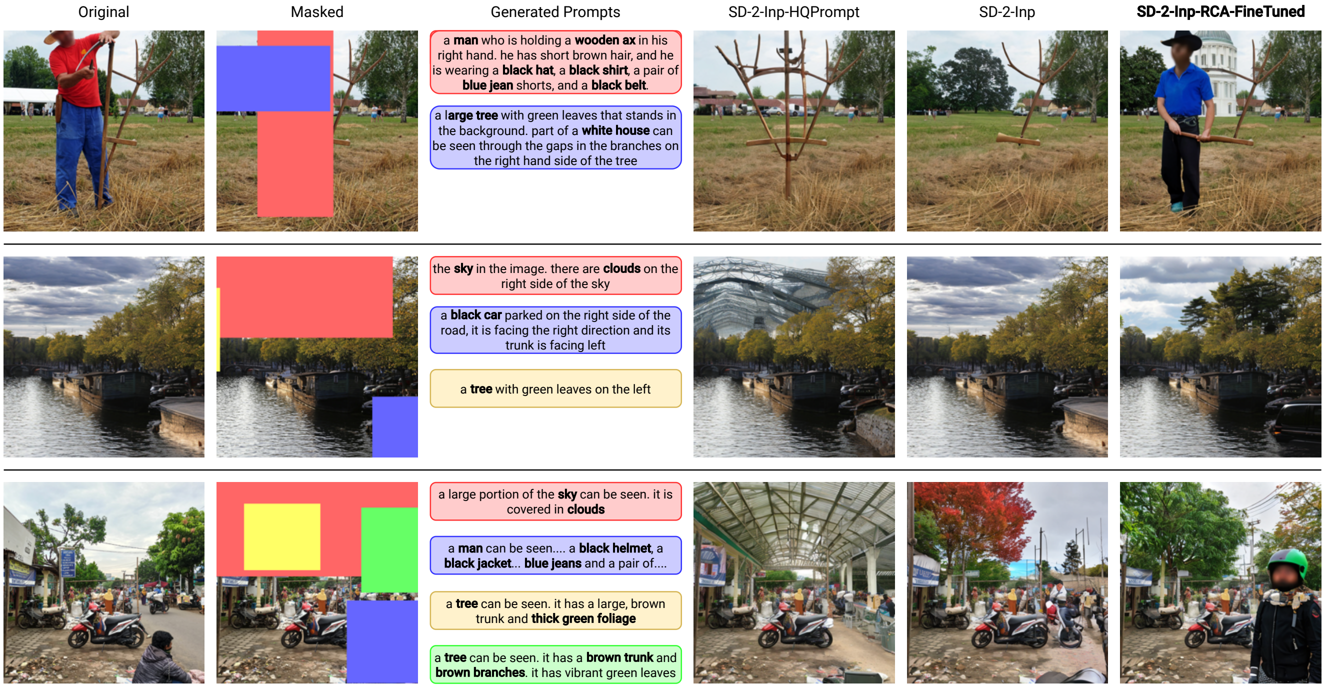

Original
Masked
Generated Prompts
SD-2-Inp-HQPrompt
SD-2-Inp
SD-2-Inp-RCA-FineTuned
a man who is holding a wooden ax in his right hand. he has short brown hair, and he is wearing a black hat, a black shirt, a pair of blue jean shorts, and a black belt.
a large tree with green leaves that stands in the background. part of a white house can be seen through the gaps in the branches on the right hand side of the tree
the sky in the image. there are clouds on the right side of the sky
a black car parked on the right side of the road, it is facing the right direction and its trunk is facing left
a tree with green leaves on the left
a large portion of the sky can be seen. it is covered in clouds
a man can be seen.... a black helmet, a black jacket... blue jeans and a pair of....
a tree can be seen. it has a large, brown trunk and thick green foliage
a tree can be seen. it has a brown trunk and brown branches. it has vibrant green leaves

Figure 6. Additional qualitative results on the Densely Captioned Images dataset.

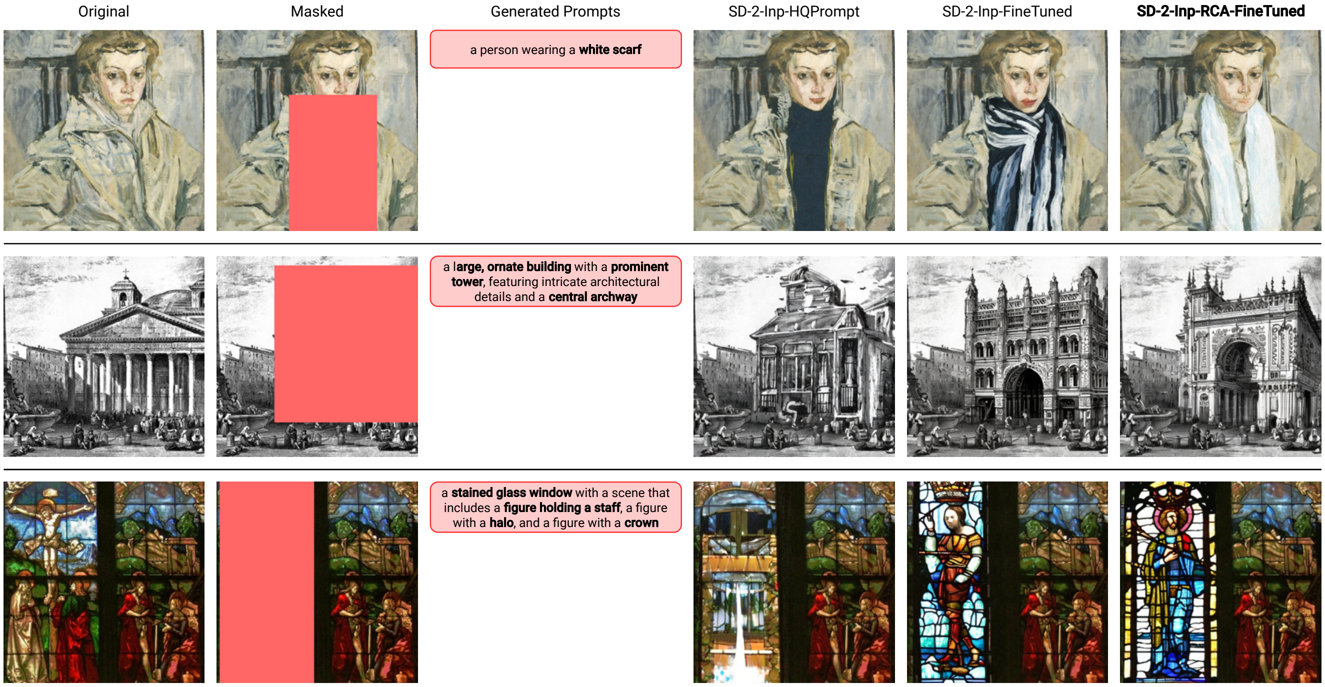
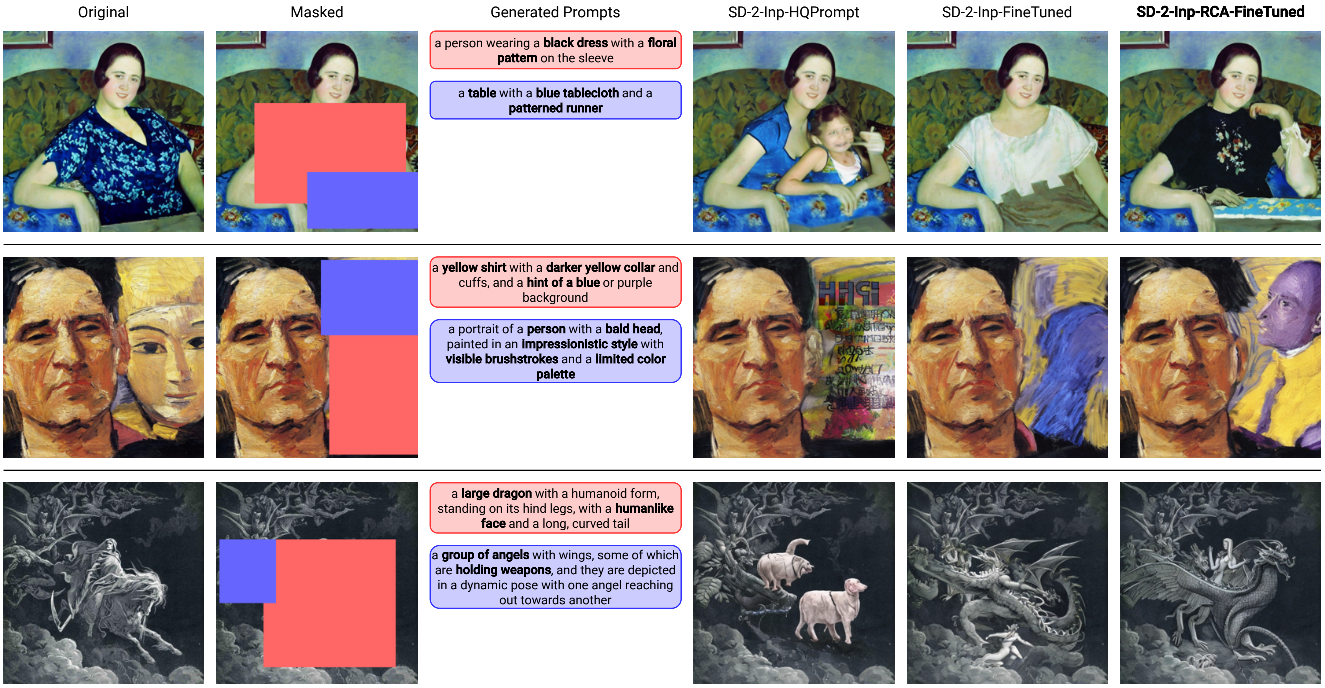

Original
Masked
Generated Prompts
SD-2-Inp-HQPrompt
SD-2-Inp-FineTuned
SD-2-Inp-RCA-FineTuned
a person wearing a white scarf
a large, ornate building with a prominent tower, featuring intricate architectural details and a central archway
a stained glass window with a scene that includes a figure holding a staff, a figure with a halo, and a figure with a crown

Figure 7. Additional qualitative results on the art dataset for 1-mask inpainting.



a person wearing a **black dress** with a **floral pattern** on the sleeve

a **table** with a **blue tablecloth** and a **patterned runner**

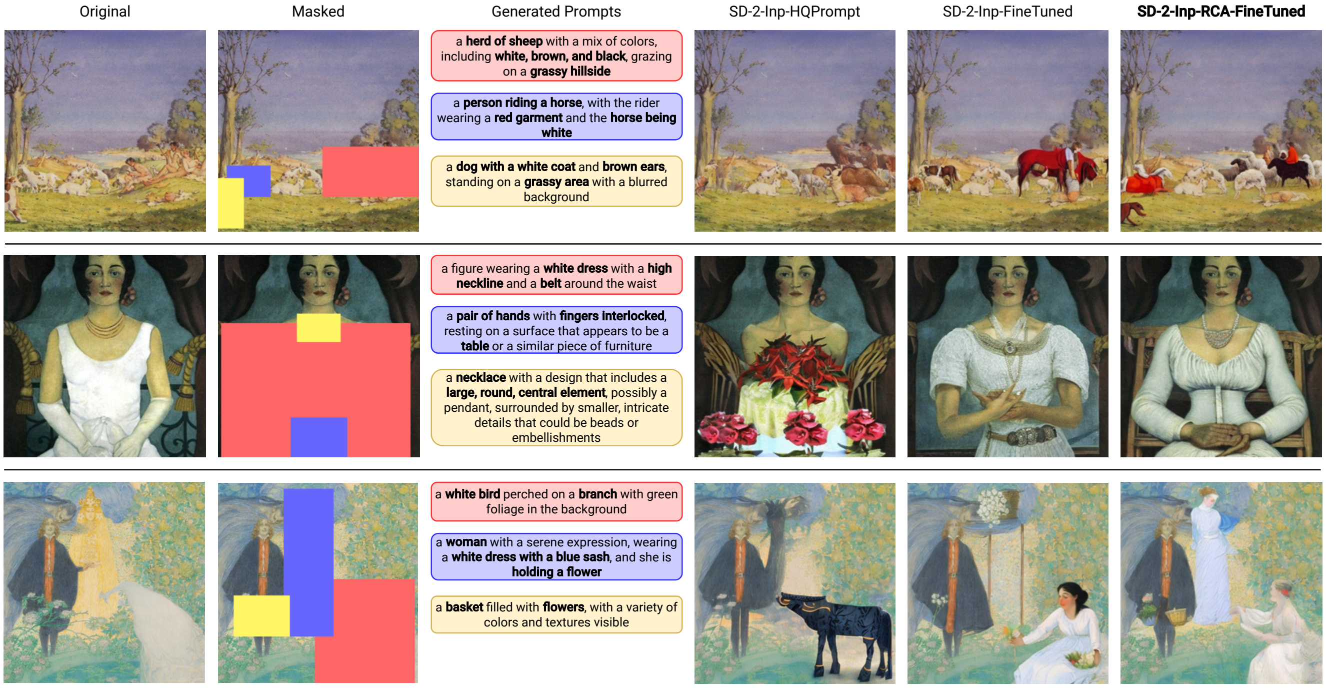
a **yellow shirt** with a **darker yellow collar** and cuffs, and a **hint of a blue** or purple background

a portrait of a **person** with a **bald head**, painted in an **impressionistic style** with **visible brushstrokes** and a **limited color palette**

a **large dragon** with a humanoid form, standing on its hind legs, with a **humanlike face** and a long, curved tail

a **group of angels** with wings, some of which are **holding weapons**, and they are depicted in a dynamic pose with one angel reaching out towards another

Figure 8. Additional qualitative results on the art dataset for 2-mask inpainting.



a **herd of sheep** with a mix of colors, including **white, brown, and black**, grazing on a **grassy hillside**

a **person riding a horse**, with the rider wearing a **red garment** and the **horse being white**

a **dog with a white coat** and **brown ears**, standing on a **grassy area** with a blurred background

a figure wearing a **white dress** with a **high neckline** and a **belt** around the waist

a **pair of hands** with **fingers interlocked**, resting on a surface that appears to be a **table** or a similar piece of furniture

a **necklace** with a design that includes a **large, round, central element**, possibly a pendant, surrounded by smaller, intricate details that could be beads or embellishments

a **white bird** perched on a **branch** with green foliage in the background

a **woman** with a serene expression, wearing a **white dress with a blue sash**, and she is **holding a flower**

a **basket** filled with **flowers**, with a variety of colors and textures visible

Figure 9. Additional qualitative results on the art dataset for 3-mask inpainting.

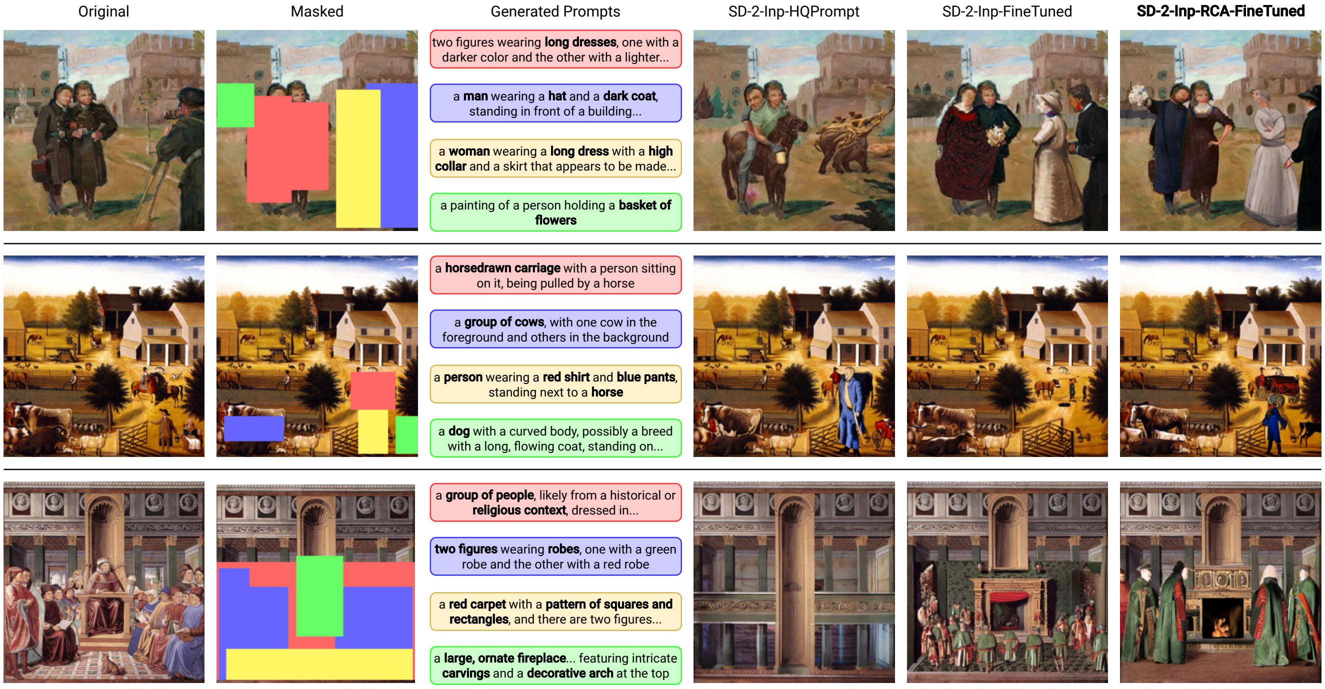

Original
Masked
Generated Prompts
SD-2-Inp-HQPrompt
SD-2-Inp-FineTuned
SD-2-Inp-RCA-FineTuned
two figures wearing long dresses, one with a darker color and the other with a lighter...
a man wearing a hat and a dark coat, standing in front of a building...
a woman wearing a long dress with a high collar and a skirt that appears to be made...
a painting of a person holding a basket of flowers
a horsedrawn carriage with a person sitting on it, being pulled by a horse
a group of cows, with one cow in the foreground and others in the background
a person wearing a red shirt and blue pants, standing next to a horse
a dog with a curved body, possibly a breed with a long, flowing coat, standing on...
a group of people, likely from a historical or religious context, dressed in...
two figures wearing robes, one with a green robe and the other with a red robe
a red carpet with a pattern of squares and rectangles, and there are two figures...
a large, ornate fireplace... featuring intricate carvings and a decorative arch at the top

Figure 10. Additional qualitative results on the art dataset for 4-mask inpainting.

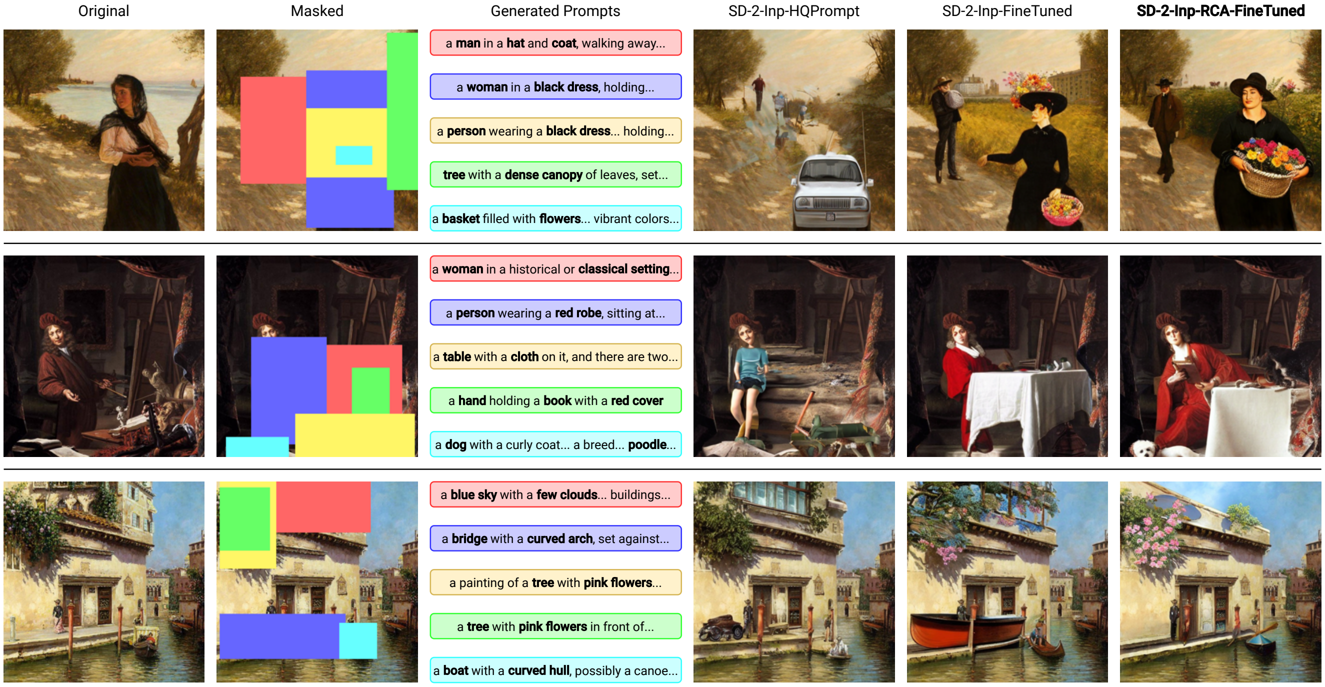

Original
Masked
Generated Prompts
SD-2-Inp-HQPrompt
SD-2-Inp-FineTuned
SD-2-Inp-RCA-FineTuned
a man in a hat and coat, walking away...
a woman in a black dress, holding...
a person wearing a black dress... holding...
tree with a dense canopy of leaves, set...
a basket filled with flowers... vibrant colors...
a woman in a historical or classical setting...
a person wearing a red robe, sitting at...
a table with a cloth on it, and there are two...
a hand holding a book with a red cover
a dog with a curly coat... a breed... poodle...
a blue sky with a few clouds... buildings...
a bridge with a curved arch, set against...
a painting of a tree with pink flowers...
a tree with pink flowers in front of...
a boat with a curved hull, possibly a canoe...

Figure 11. Additional qualitative results on the art dataset for 5-mask inpainting.