ProMark: Proactive Diffusion Watermarking for Causal Attribution Supplementary Materials

Vishal Asnani[1,2] John Collomosse[1,3] Tu Bui[3] Xiaoming Liu[2] Shruti Agarwal[1]
[1]Adobe Research, [2]Michigan State University, [3]University of Surrey
{asnanivi,liuxm}@msu.edu {collomos,shragarw}@adobe.com t.v.bui@surrey.ac.uk

Table 1. Multi-concept attribution performance across different configurations.

Configuration		Attribution Accuracy (%) ↑		
Secret 1	Secret 2	Secret 1	Secret 2	Combined
Left	Right	95.61	93.31	90.12
Right	Left	95.52	93.35	90.19
Top	Bottom	95.66	93.70	90.01
Bottom	Top	95.02	93.46	90.73

1. Multiple Watermark Configurations

We investigate the application of dual watermarks, each positioned on opposing sides of the image. This exploration raises a pivotal query: "Is the spatial positioning of watermarks critical to the performance?" To answer this, we ablate four distinct watermark configurations. As shown in Tab. 1, there is a consistent performance across all watermark placements (left, right, top, bottom), thereby substantiating the spatial robustness of ProMark in watermark positioning.

2. Watermark Robustness

We test our method against 14 different degradations (blur, various noises, fog, etc.), by adopting the evaluation protocol detailed in the RoSteALS [1]. We use 50 watermarked training images from LSUN dataset and use unconditional LDM with a strength of 30%. The average attribution accuracy for training and generated images across all 14 attacks is $90.21 \pm 7.63\%$ and $89.51 \pm 8.18\%$, as compared to 95.12% without any degradation, showing the robustness of our approach to multiple forms of watermark attack.

3. Possibility of Concept Leakage

We present multiple results where we attribute the images generated using non-watermarked data, for example via random latent code and conditional generation. We detect no retention of the watermark after noising or in random latent codes, with watermark detection accuracy of 50.56% (chance 50%) after noising for ≥ 900 timestamps or in random latent codes. The LDM generates an image from noise through inversion, and the watermark is added during this GenAI model inference process. Our decoder is employed independently to identify the concept. To prove this, we evaluate our model in Table 4 (main paper) for two more baselines, using held-out images (1) with no watermark encryption, and (2) encrypted with a different concept's watermark. ProMark is able to attain an attribution accuracy of 94.32% and 94.01% respectively when evaluated with ground-truth concept watermark for both baselines compared to 95.60% reported for watermarked held-out data. Therefore, when inverting generating images that encrypt no watermark, or encrypt incorrect watermark, the correct concept watermark is encrypted.

4. Computational Efficiency

We demonstrate the computation efficiency of ProMark during inference (running watermark decoder to perform causal attribution), which costs 5.6ms on one A100 GPU. Training with watermarked data adds negligible cost to generative model training. This is comparable to running inference on CLIP, or ALADIN to perform correlation based attribution (28.32 ms) but the additional cost of the embedding search is 87.91 ms for a dataset of 20K LSUN training images. ProMark therefore offers the advantage of both efficiency and causality for training data attribution. We will add this to the paper.

5. Additional Watermark Strength Analysis

Our research introduces a new paradigm in concept attribution for images classified under multiple concepts. We show the analysis of PSNR variation with watermark strength for the case of multi-concept attribution. The results are shown in Fig. 1. Our findings indicate that, compared to single watermark cases, the PSNR for multi-concept images is marginally higher at equivalent watermark strengths. However, as expected, an increase in watermark strength generally leads to a decrease in PSNR.

Furthermore, we have visualized images from different datasets to showcase the extent of degradation caused by varying watermark strengths. As discussed in Sec. 4.5, the

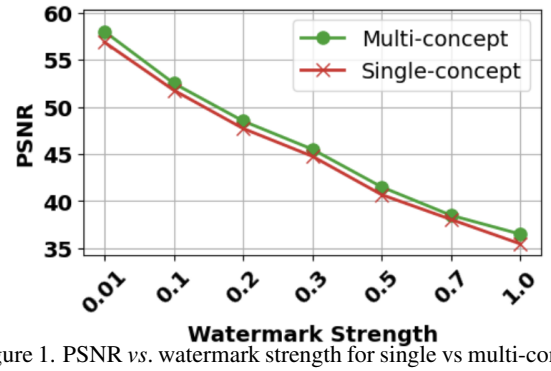

Figure 1. PSNR *vs*. watermark strength for single vs multi-concept attribution.

performance of our method improves with increased watermark strength. Nevertheless, this increase in strength leads to a decline in image quality, evidenced by the emergence of bubble-like artifacts in the images, as shown in Fig. 2 (the watermark strength ranges from 0.1 to 1.0).

6. Watermark Discussion

We visualize some sample watermarks in both, spatial and frequency domain in Fig. 3. These watermarks are converted from bit-sequences to spatial domain as described in Sec. 3.4. Visually, the watermarks appear indistinguishable from one another in both domains. Yet, their orthogonality is clearly demonstrated through the cosine similarity matrix, which we used to analyze 100 different watermarks. This matrix reveals that the inter-watermark cosine similarity is consistently close to zero, decisively indicating the orthogonal nature of these watermarks.

7. Implementation Details

We train ProMark with LDM for $15K$ iterations with a batch size of 32, using 8 NVIDIA A100 GPUs for each experiment. We use the default parameters for optimizers as used in the official repository of [2]. The learning rate is set at $3.2e^{-5}$ for training LDM.

We further show the architecture for the generic decoder used for comparing against pretrained secret decoder shown in Fig. 4. The generic decoder consists of 2 stem convolution layers and 10 convolution blocks. Each block consists of convolutional and batch normalization layers followed by ReLU activation.

8. More Sampled Images

We use multiple datasets for evaluating ProMark. We sample images from the trained LDM for every class. We show some of the train and sampled images for the corresponding classes for different datasets in Figs. 5 to 8. We argue that ProMark is able to perform attribution to different types of concepts, *i.e*. image templates (Fig. 5), image style (Fig. 8), style and content (Fig. 6), and ownership (Fig. 7). Therefore, proactive based causal methods perform attribution not only on the style or motif of the image as done by correlation based works, but also performs attribution to a variety of concepts proving it's generalizability.

References

[1] Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. RoSteALS: Robust steganography using autoencoder latent space. In *CVPR*, 2023. 1

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

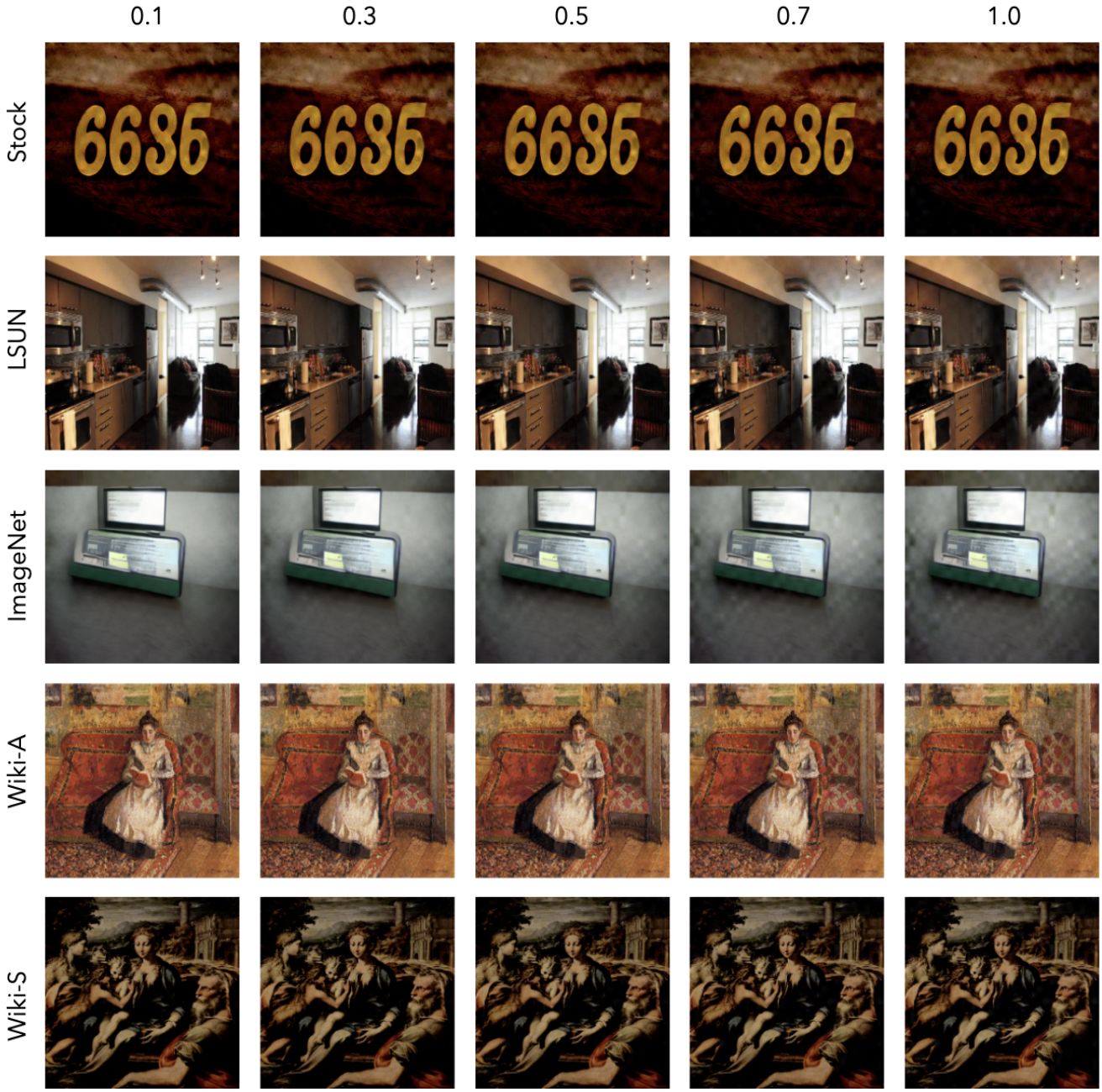

Figure 2. Noise Strength visualization for different watermark strength

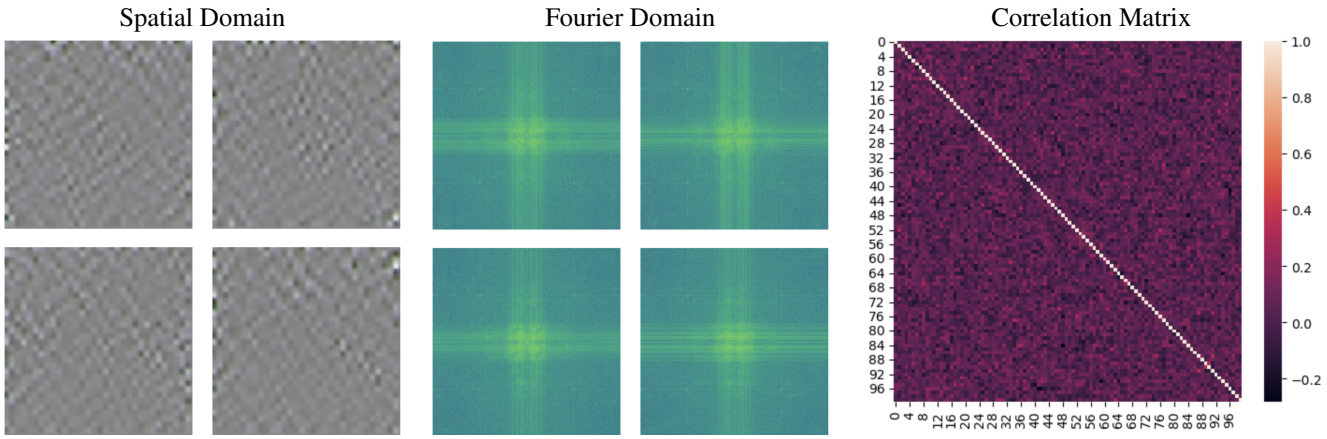

Figure 3. Watermark Visualization: Spatial domain, Fourier domain and inter-watermark cosine similarity for 100 watermarks.

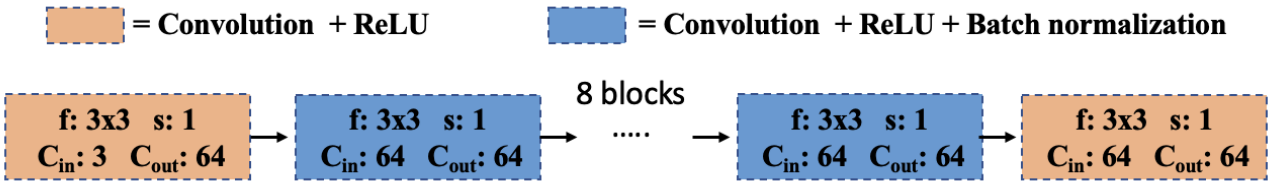

Figure 4. Generic decoder architecture.

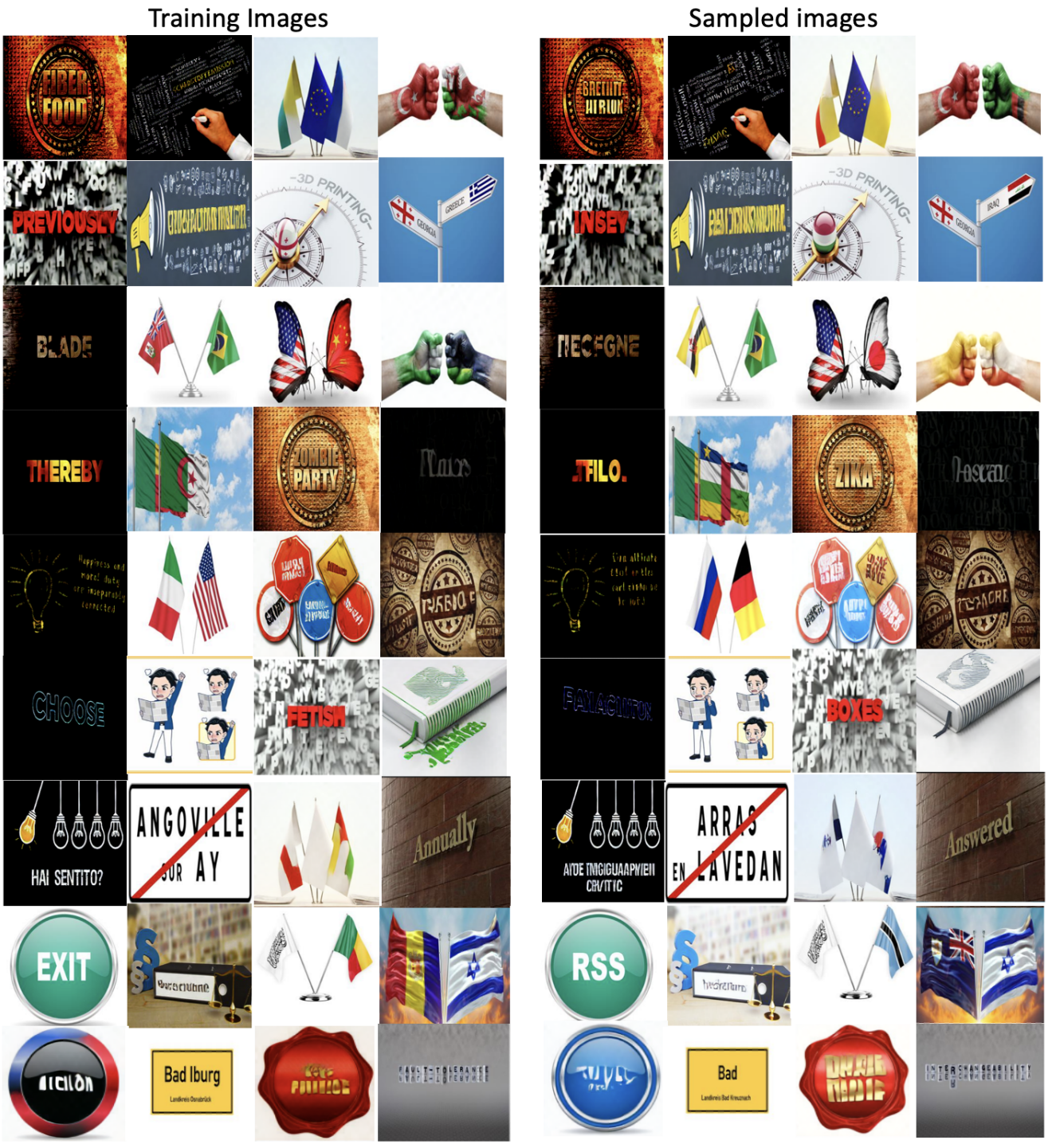

Figure 5. Training and sampled images for stock dataset.

Figure 6. Training and sampled images for BAM dataset.

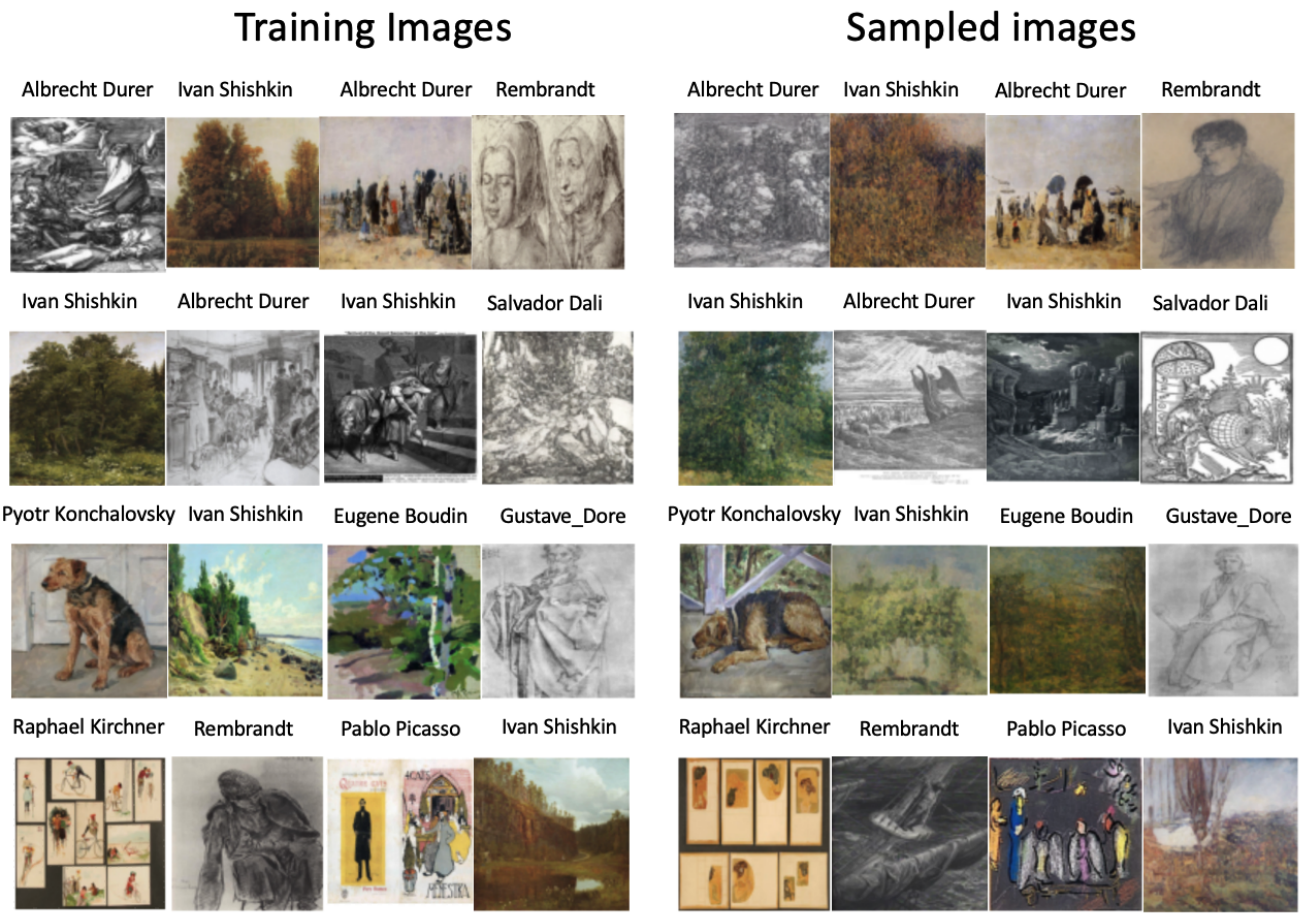

Figure 7. Training and sampled images for wiki-a dataset.

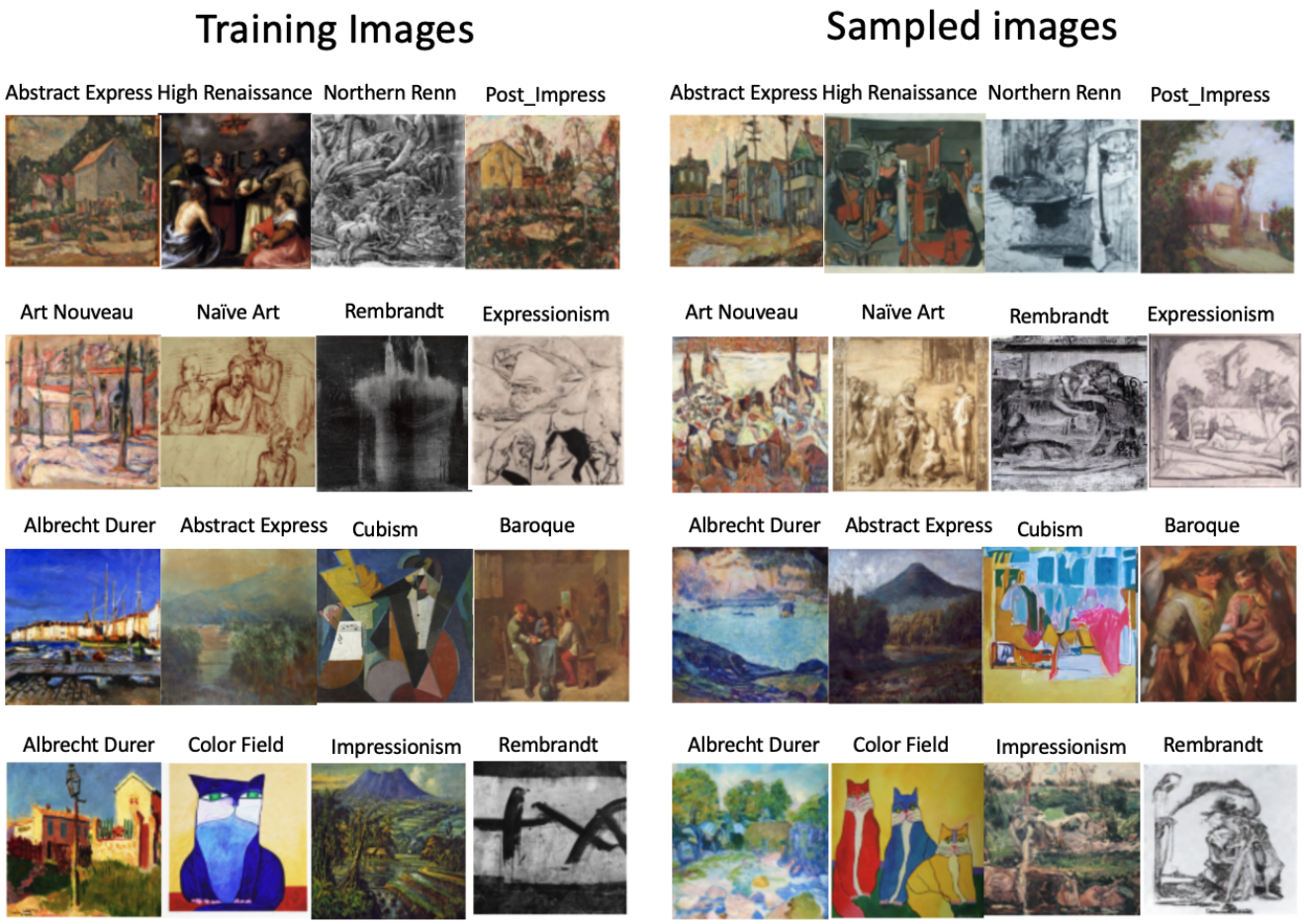

Figure 8. Training and sampled images for wiki-s dataset.