Rethinking Skip Connections in Diffusion Models for Replication Mitigation

Supplementary Material

A. Appendix

A.1. Similarity Score in replication

In Section 2.3, we mentioned that the replication score R is based on a similarity score s of two images that was introduced in [37]. In particular, we use the pretrained sscd_disc_large[1] model described in [37] as the feature extractor to get a feature representation for each image. Then for every pair of two images, we apply cosine similarity in the feature space to calculate S.

A.2. Information Transfer Block

The proposed information transfer block candidates include:

1. Max Pooling (*MP*): Max pooling reduces the spatial dimensions of features in the skip connections, emphasizing prominent elements and condensing encoder information. It helps prevent the decoder from accessing detailed patterns, while still preserving essential spatial information. Specifically, we use Max pooling with kernel size 2×2, followed by a *MaxUnpool* that set all non-maximal values to 0.
2. Convolutional Layer (*Conv*): Adding a convolutional layer to the skip connections allows the model to transform encoder features before passing them to the decoder, suppressing redundant patterns that could lead to replication. By adjusting the features, the convolutional layer helps balance the detailed information and the abstract representations. For *Conv*, we use a single convolutional layer with kernel size 3×3, without changing the output feature shape.
3. Multiple Convolutional Layers (*Multi-Conv*): Extending *Conv* with multiple convolutional layers offers deeper feature processing, creating complex abstractions to further reduce replication artifacts. We use 3 Convolutional layers with kernel size 3.

A.3. Algorithm of Training LoyalDiffusion

The pseudo-code for training *LoyalDiffusion* is provided in Algorithm 1.

[1] https://github.com/facebookresearch/sscd-copy-detection/tree/main

Algorithm 1 Training *LoyalDiffusion*

Inputs: A pretrained (from Stable Diffusion 2.1) dual-model M (with a standard branch of U_{st} and a replication-aware branch of U_{ra}); training dataset $\mathcal{D}$; total timesteps T; timestep threshold τ; number of iteration N.
Output: A fine-tuned dual-model M.

$M \leftarrow$ DualModel()
M.train()
$iter \leftarrow 0$
while $iter <$ N **do**
  for each batch {(x, y)} in $\mathcal{D}$ **do**
    Sample $t \sim$ Uniform$(0,\ T)$
    Sample noise $\epsilon \sim \mathcal{N}(0,\ I)$
    Forward adding noise $x_t \leftarrow \sqrt{\bar{\alpha}_t}\, x\ +\ \sqrt{1-\bar{\alpha}_t}\, \epsilon$
    if $t > \tau$ **then**
      $\hat{\epsilon} \leftarrow M.U_{ra}(x_t,\ t,\ y)$
    else
      $\hat{\epsilon} \leftarrow M.U_{st}(x_t,\ t,\ y)$
    end if
    $Loss \leftarrow \|\epsilon - \hat{\epsilon}\|_2^2$
    Update M via ADAM optimizer
    $iter++$
  end for
end while
return M

A.4. Experimental Setup Details

We use NVIDIA A100 Tensor Core GPU with $40GB$ memory for both training and inference. During training, all models are using batch size 16 and image resolution 256. Also Adam optimizer is applied during training with $\beta = 0.9$ and $\beta = 0.999$ and weight decay $1e^{-2}$. For inference, we sample images using total inference steps $S = 50$, uniformly spacing across the full diffusion process.

A.5. Additional Experimental Results

In this section, we provide additional experimental results focusing on different combinations of modified skip connections (SC) with information transfer blocks, extending the analysis presented in Section 3.1. Section 3.1 analyzed a limited combinations of applying different information transfer blocks to different skip connections (SC) and determined applying *Conv* to SC 3 and SC 4 simultaneously gives the best replication reduction without increasing FID. Here, we explore additional combinations of modified skip connections, as shown in Table 9.

The results reveal two distinct trends: in some cases, replication is significantly reduced, but at the cost of a no-

Removing Skip Connection										
SC #	1&2	1&3	1&4	2&3	2&4	3&4	1&2&3	1&2&4	1&3&4	2&3&4
R↓	0.247	0.547	0.562	0.499	0.585	0.604	0.226	0.243	0.543	0.519
FID↓	105.76	24.21	22.09	33.62	22.70	20.30	141.38	102.80	25.67	35.31
MP										
SC #	1&2	1&3	1&4	2&3	2&4	3&4	1&2&3	1&2&4	1&3&4	2&3&4
R↓	0.600	0.609	0.613	0.596	0.606	0.614	0.597	0.611	0.590	0.602
FID↓	19.20	16.85	18.05	18.11	18.50	16.52	18.68	18.79	17.72	17.40
Conv										
SC #	1&2	1&3	1&4	2&3	2&4	3&4	1&2&3	1&2&4	1&3&4	2&3&4
R↓	0.311	0.303	0.458	0.345	0.566	**0.386**	0.235	0.326	0.272	0.308
FID↓	89.70	50.85	44.47	41.90	22.00	**21.31**	142.01	78.51	52.88	50.56
Multi-Conv										
SC #	1&2	1&3	1&4	2&3	2&4	3&4	1&2&3	1&2&4	1&3&4	2&3&4
R↓	0.368	0.455	0.541	0.412	0.579	0.460	0.243	0.369	0.399	0.386
FID↓	68.25	26.39	25.45	34.24	20.28	20.74	116.27	63.14	28.04	42.21

Table 9. Replicaiton score and FID for different information transfer blocks on LAION-10K [35]. The 'SC #' indicates the information transfer block is only applied to the specific combination of skip connections numbered as in Figure 2(a)

table degradation in FID; in other cases, the impact on FID is minimal or even slightly improved, but the reduction in replication is negligible. Only two configurations achieved satisfactory results: applying the *Conv* information transfer block to SC 3 and 4, and applying the *Multi-Conv* information transfer block to SC 3 and 4. Both configurations effectively reduced replication while maintaining almost unchanged FID. Among these, the *Conv* information transfer block applied to SC 3 and 4 demonstrated a more pronounced reduction in replication than *Multi-Conv*, validating the choice made in Section 3.1.

A.6. Visualize Generated Images for Various FIDs

In this section, we provide visualizations of generated images corresponding to different FID values, as shown in Figure 6. The third column shows images generated by baseline model with FID$=18.01$. The visual examples illustrate that when the FID is within the range of 18.01 ± 2, the generative ability remains comparable, producing visually similar and acceptable images. However, with FID larger than 20, the image quality drops significantly. This demonstrates that an FID within $(16, 20)$ does not significantly impact image quality, making it an acceptable range for evaluating the performance of our proposed *LoyalDiffusion*.

A.7. *RepliBing* Dataset

To further evaluate replication in generative models, we created the *RepliBing* dataset mentioned in Section 3.4. This dataset is constructed using the Bing search engine to retrieve publicly available images from the internet, based on the same prompts used during training and inference. *RepliBing* is designed to provide a diverse set of images that align closely with the input captions while minimizing the chance of direct replication of the training data.

Figure 7 showcases examples from the *RepliBing* dataset alongside corresponding training images that share the same captions. As shown, the *RepliBing* dataset captures a broad spectrum of visual diversity while maintaining fidelity to the provided prompts. These comparisons highlight the visual similarity in the conceptual alignment between *RepliBing* images and the training data, without any direct replication.

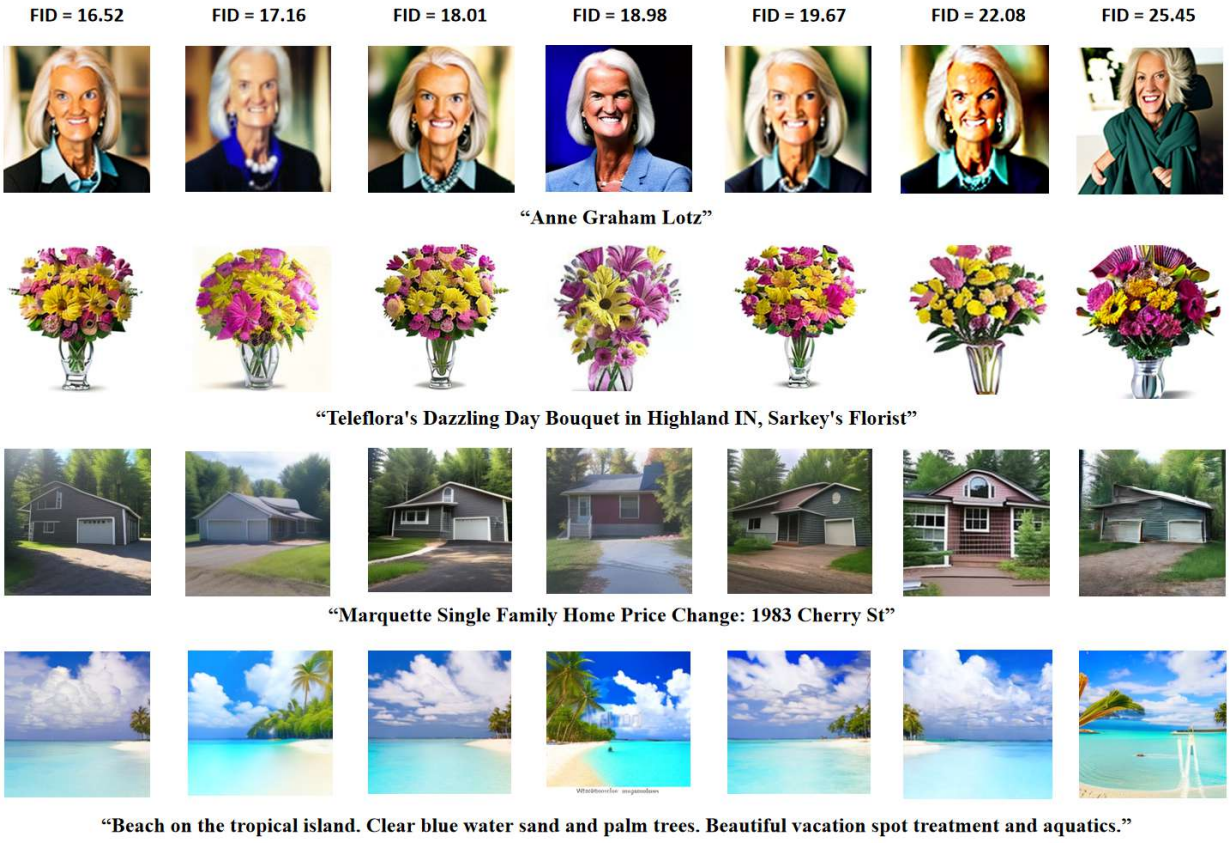

Figure 6. Examples of generated images with different FIDs. The images come from the following models: column (1) MP 3&4 from Table 9; column (2) GC&DF from Table 4; column (3) Baseline from Table 4; column (4) $w_{lat} = 0.1, w_{emb} = 0.5$, and $\tau = 300$ from Table 3; column (5) Remove skip connection 4 from Table 1; column (6) Two-stage result with $\tau = 100$ from Table 2; and (7) Multi-Conv 1&4 from Table 9.

Figure 7. Examples of images from *RepliBing* sharing the same caption with images from LAION-10K.