Appendix

A. Band Selection Details

As mentioned in Sec. 3.3, for each target style we select one of the three band combinations c_1, c_2 or c_3, and then we use the corresponding filter for all the images translation of that style. Note that this is a negligible operation if compared, *e.g.*, to the collection of a set of reference images used to train a style-specific generator [23, 38, 48]. Moreover, there are two underlying reasons: 1) given a style description, the artifact condition is unpredictable. To be specific, it is difficult to judge if the stylized image contains artifact or not, as well as to detect the artifact appearance; 2) the artifact scales are of limited variable ranges.

Hence, we propose SpectralCLIP, a simple yet effective method based on empirical studies. Through experimenting on multiple band combinations, we find three filtering strategies ($c_1 = \{b_1, b_2, b_4\}$, $c_2 = \{b_1, b_2\}$, $c_3 = \{b_1\}$) that are effective for preventing the artifacts at the corresponding three scales (as shown in Fig. 8) (and through experiments on various styles, we find the artifacts are usually of one of the three scales). To prevent artifacts of larger scales, more frequency bands should be masked. The bands are defined according to the scales of the artifacts, i.e., the length of the dependant visual tokens in the CLIP-ViT scenario.

In the leftmost column of Fig. 11, we show the stylized results of CLIPstyler corresponding to three different style descriptions ("outsider art", "cartoon" and "digital art"). In all cases, CLIPstyler generates a lot of artifacts. Additionally, the corresponding scales vary from style to style. A simple comparison can be seen in Fig. 8. As mentioned in Sec. 3.3, in order to select the most suitable filter for each style, we use a single content image per style and we generate stylized images using our SpectralCLIP and one among c_1, c_2 and c_3. Correspondingly, we give an illustration in Fig. 11, where we show the results obtained using SpectralCLIP with one filtering band combination among c_1, c_2 or c_3. For "outsider art", c_1 is the best band combination, while c_2 is the best for "cartoon", and c_3 for "digital art".

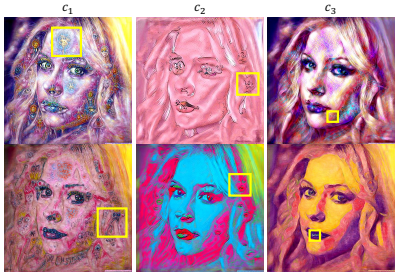


Figure 8. Different styles result in artifacts at different scales which are tricky to predict.

In addition, since we can use the CLIP/forget-to-spell CLIP to generate stylized images with/without text artifacts and use learn-to-spell CLIP to measure the text artifact presence, we did a spectral analysis to find further support for our work. Specifically, We first collect 100 images with/without artifacts using CLIP/forget-to-spell CLIP, respectively. Then, we individually mask each frequency at one time and use the filtered CLIP representation to compute the cosine similarity using learn-to-spell CLIP (the computations are based on patches as in CLIPstyler). As shown in Fig. 9, masking frequency 0 and 1 significantly reduces the learn-to-spell CLIP similarity scores of the images containing textual artifacts, indicating that those frequencies are related to text artifacts (e.g., in Fig. 8, masking c_3 is useful to prevent the generation of text artifacts).

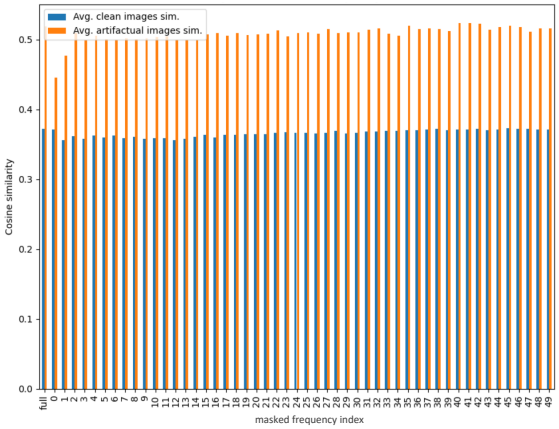


Figure 9. Spectral analysis w.r.t. textual artifacts.

In Tab. 4, we provide the band combination we used for each of the styles presented in this paper.

Band combination	Style
c_1	Lowbrow Outsider art Visionary art Rosy-color oil painting Makoto Shinkai
c_2	Pop art Cartoon Giorgio Morandi Harlem renaissance Neon art Contemporary art Francoise Nielly
c_3	Fauvism Digital art

Table 4. Band combination used for each style.

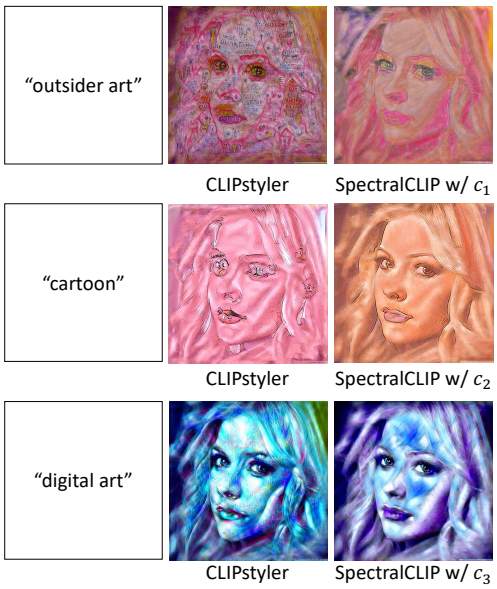


Figure 10. By CLIPstyler, different styles lead to different types of artifacts. Therefore, in SpectralCLIP, we filter different band combinations (indicated on the bottom).

B. User Study Details

In this section, we provide more details on the user study reported in Sec. 4.2, in which we compare the generation results obtained using text-image similarities computed with three different spaces: the original CLIP space (which corresponds to the original CLIPStyler), forget-to-spell CLIP, and our SpectralCLIP.

We asked 30 participants to answer 24 questions, split in two tasks (12 questions per task), respectively evaluating the overall quality of the generated images and the possible appearance of artifacts. Fig. 12 shows an example of the both tasks. Since evaluating the overall quality of a style-transfer task requires the participants to consider whether the stylized results reflect target style, for the first task we selected 4 better-known styles (i.e., "pop art", "cartoon", "fauvism" and "Giorgio Morandi"). For the second task (artifact evaluation), we used the other 6 styles ("lowborw", "outsider art", "visionary art", "harlem renaissance", "neon art" and "digital art"). For each style, we randomly sampled 10 images from the COCO val-set and used them as content images to generate the stylized images with the three methods. Thus, we obtain 100 groups of stylized results in total (each composed of 3 images generated by the 3 compared methods, e.g., see Fig. 12). For the first task, we equally sampled 3 groups of stylzied results for each style, obtaining 12 questions. For the second task, we randomly sampled 12 groups of images without considering the style. We used Google Forms as the platform.

C. Computation Time

The DCT and IDCT transforms of SpectralCLIP are the only overhead with respect to baselines. In text-guided image style transfer, included our method is 1.095 times slower than the CLIPstyler baseline.

D. Discussion on learn-to-spell Similarity

Materzynska *et al*. [31] analyse the entanglement problem of written texts and visual concepts in the CLIP space, and learn two orthogonal projections, i.e., "forget-to-spell" and "learn-to-spell", the former is for recognizing visual concepts while the latter is for recognizing written text. Specifically, in this paper, we use the "learn-to-spell" projection to measure the textual artifact condition in the generated stylized image, as the more the image contains textual artifacts, the higher the similarity between the image and the textual description of style using the "learn-to-spell" projection (as in Tab. 2 and Fig. 9).

E. Additional Style Transfer Results

In this section, we show additional qualitative comparisons between the original CLIPstyler and our SpectralCLIP, using as the following target styles: "contemporary art", "rosy-color oil painting" (Fig. 13), "Francois Nielly", "Makoto Shikai" (Fig. 14), "lowbrow", "harlem renaissance" (Fig. 15). The results shown in this section confirm those reported in the main paper, and they show that SpectralCLIP *drastically* reduces the generation of both visual and textual artifacts, while simultaneously leading to a high consistency of the generated images with respect to the target style. For instance, when using CLIPstyler and the "rosy-color oil painting" style (Fig. 13) a lot of roses are generated in the background, the sky, the mountains, the trees, etc. As another example, in Fig. 14, CLIPstyler "writes" the name of the corresponding artist in the stylized images. These visual and textual artifacts are definitely not part of the user's desired style, which degrades the quality of the stylised images. In contrast, the corresponding images generated using SpectralCLIP largely solve this problem, making the generation quality significantly higher.

F. Non-artistic Concrete Styles

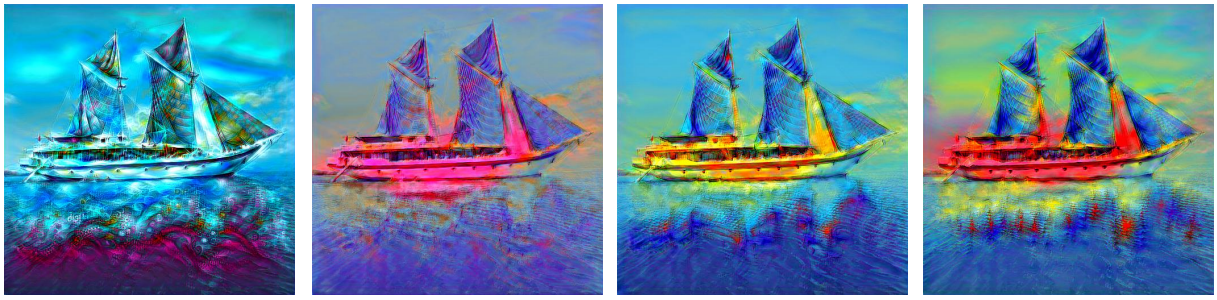
In this work, we focus on artistic and abstract styles, which is a major advantage of CLIP-guided style transfer and yet tends to produce artifacts. This section tests the ability of SpectralCLIP to transfer concrete styles, which is also considered in previous style transfer work. We use three concrete styles ("fire", "neon light", and "white wool") and report the results in Fig. 16. It can be seen that SpectralCLIP leads to finer-grained stylised results.

CLIPstyler | **SpectralCLIP w/ c_1** | SpectralCLIP w/ c_2 | SpectralCLIP w/ c_3

(a) "outsider art"

CLIPstyler | SpectralCLIP w/ c_1 | **SpectralCLIP w/ c_2** | SpectralCLIP w/ c_3

(b) "cartoon"



CLIPstyler | SpectralCLIP w/ c_1 | SpectralCLIP w/ c_2 | **SpectralCLIP w/ c_3**

(c) "digital art"

Figure 11. Leftmost column: images generated using CLIPstyler and three different styles. All the other columns show the results obtained using SpectralCLIP and one among c_1, c_2 or c_3 as the filter. In case of "outsider art", c_1 is the best band combination, while c_2 is the best for "cartoon", and c_3 for "digital art".

Please select the best stylized image considering: (i) how much it reflects the cartoon style, (ii) how much it preserves the content of the content image, (iii) how much it is clean (i.e., it does not contain textual or visual artefacts)

a

b

c

(a)

Please select the best stylized image considering: (i) how much it preserves content and (ii) how much it is clean (i.e. it does not contain textual or visual artefacts).

a

b

c

(b)

Figure 12. The user study questions used to evaluate two tasks: (a) the overall image quality, and (b) the possible appearance of artifacts.

Figure 13. Additional style transfer results using “Contemporary art” and “Rosy-color oil painting” as the target textual descriptions.

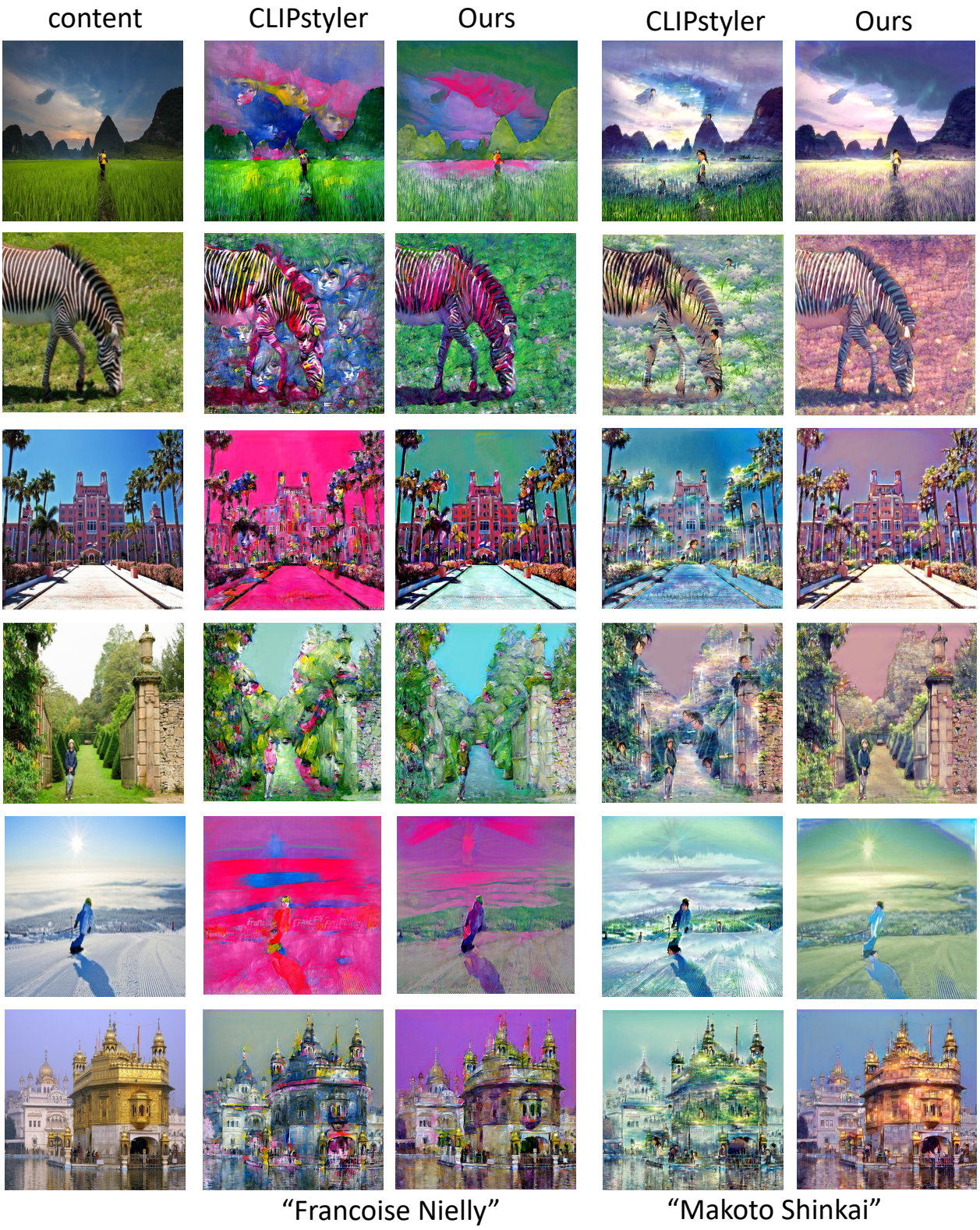


Figure 14. Additional style transfer results using “Francoise Nielly” and “Makoto Shinkai” as the target textual descriptions.

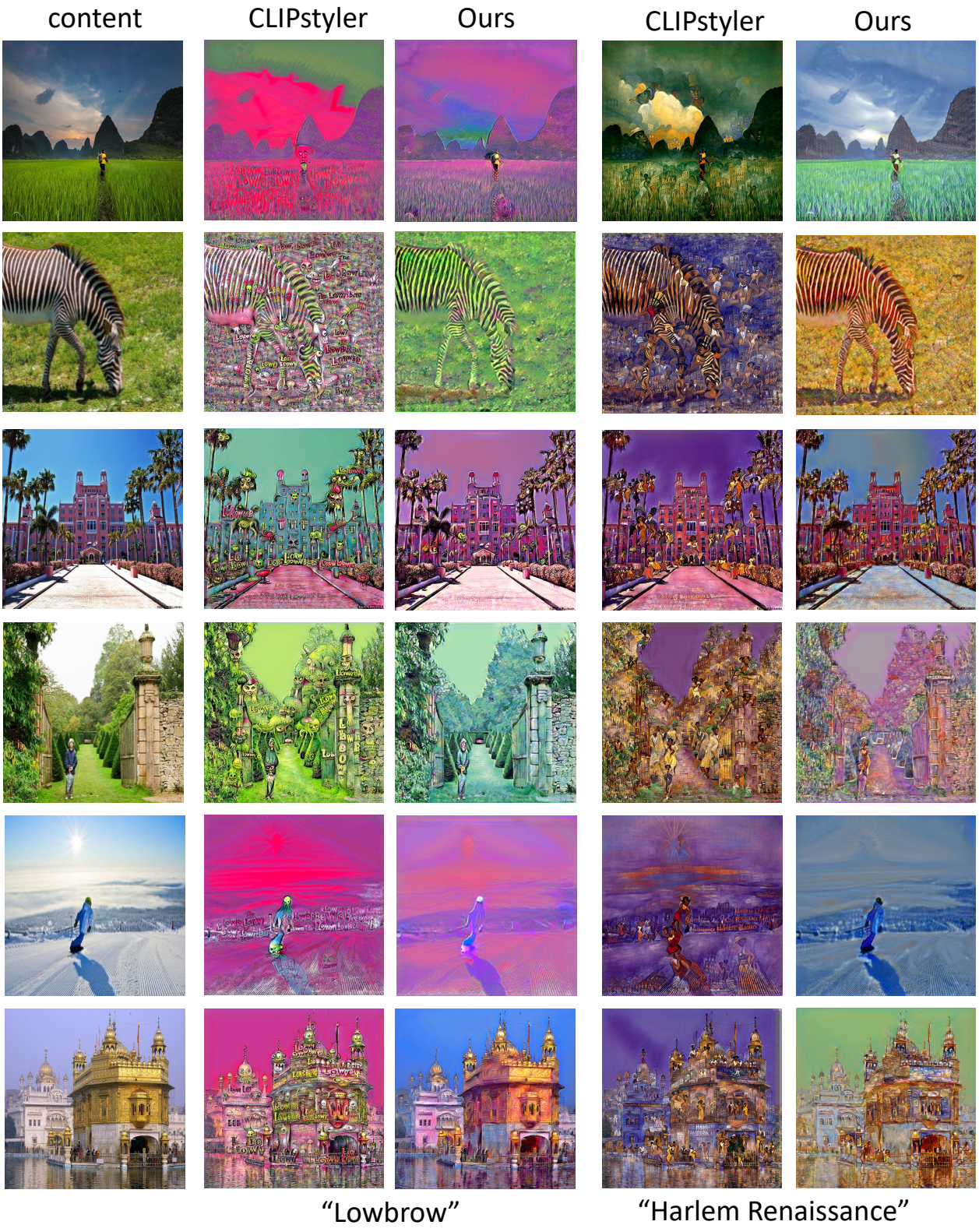


Figure 15. Additional style transfer results using “Lowbrow” and “Harlem Renaissance” as the target textual descriptions.

Figure 16. Non-artistic style results of (a) CLIPstyler and (b) SpectralCLIP w. c_3.