SEM-CS: SEMANTIC CLIPSTYLER FOR TEXT-BASED IMAGE STYLE TRANSFER

Supplementary Material, Submission ID:1346

1. SALIENT OBJECT DETECTION

In this section, we describe phase 1 of Salient object detection. Let $I \in R^{3\times M\times N}$ be content image (input). We extract deep patch features $f = \phi(I)$ of the input image I from the attention block of the last layer of ϕ, a self-supervised transformer DINO-ViT-Base. Then, a weighted graph over patch features is built with the affinity between patch features as edge weights. An affinity matrix is a weighted sum of semantic patch-wise features and colour matrices.

$$W = W_{feat} + \lambda_{knn} W_{knn} \tag{1}$$

where λ_{knn} is a parameter that regularizes semantic features and color consistency. Patch-wise features consists of only the correlated features, therefore W_{feat} is thresholded at 0 as follows:

$$W_{feat} = ff^T \odot (ff^T > 0) \tag{2}$$

Color affinity matrix between patch features u and v is calculated using sparse-KNN matting as below. HSV color space is obtained using KNN-matting of the content image.

$$W_{knn}(u,v) = \begin{cases} 1 - \psi(u) - \psi(v), u \in KNN_{\psi}(v), \\ 0, otherwise \end{cases} \tag{3}$$

where, $\psi(u)$ is vector in $\mathcal{R}^5$ consisting of color and spatial information. Given W, the eigen decomposition of its graph laplacian L is considered for obtaining soft segments: $\{y_0, y_1, y_{n-1}\}$. We examine only Fiedler eigenvector y_1 that corresponds to semantic region and it usually identifies the most salient object in the image.

2. OTHER LOSS

The content loss preserves the content during the transfer, and total variation regularization loss alleviates side artefacts from irregular pixels. Thus, total loss is calculated as follows:

$$\mathcal{L}_{total} = \lambda_B * \mathcal{L}_{BGlob} + \lambda_F * \mathcal{L}_{FGlob} + L_{others} \tag{4}$$

where,

$$\mathcal{L}_{total} = \lambda_C * L_{content} + \lambda_{TV} * L_{TV} \tag{5}$$

where, λ_C and λ_{TV} are controlling parameters for content loss and TC regularization loss.

3. DATA COLLECTION

To transfer artistic style from texts to content images, we tested 58 images: 40 of them are sceneries, and 18 of them are portraits. We had collected 50 famous text conditions. We then assigned them into eight different categories: abstract art, paintings, portraits, sceneries, buildings and vehicles. We applied 50 style texts for each content image. Hence, we obtained 2900 photos in our final data set for single text condition. For double text conditions, we considered 10 content images and applied 2500 text conditions with all permutations and a combination of 50 text conditions and hence obtained 25000 images.

4. IMPLEMENTATION DETAILS:

Since CLIP model receives the images with resolution of 224×224, we resized all of the images including patch and whole image before feeding to the CLIP model. For augmentations, we use perspective augmentation which is implemented on Pytorch torchvision library. We directly used RandomPerspective(distortion_scale=0.5). The random perspective function is implemented in torchvision.transforms. We used batch size of 4 and Adam optimizer with learning rate of 1 × 10-4. Similar to our basic method, we used learning rate decay strategy. For augmentation, we applied random perspective augmentation for 16 times, therefore our total training patch number per iteration is N = 64.

5. EXPERIMENTAL RESULTS

In this section, we provide an extended version of the experiments performed in the paper.

Table 1. DISTS

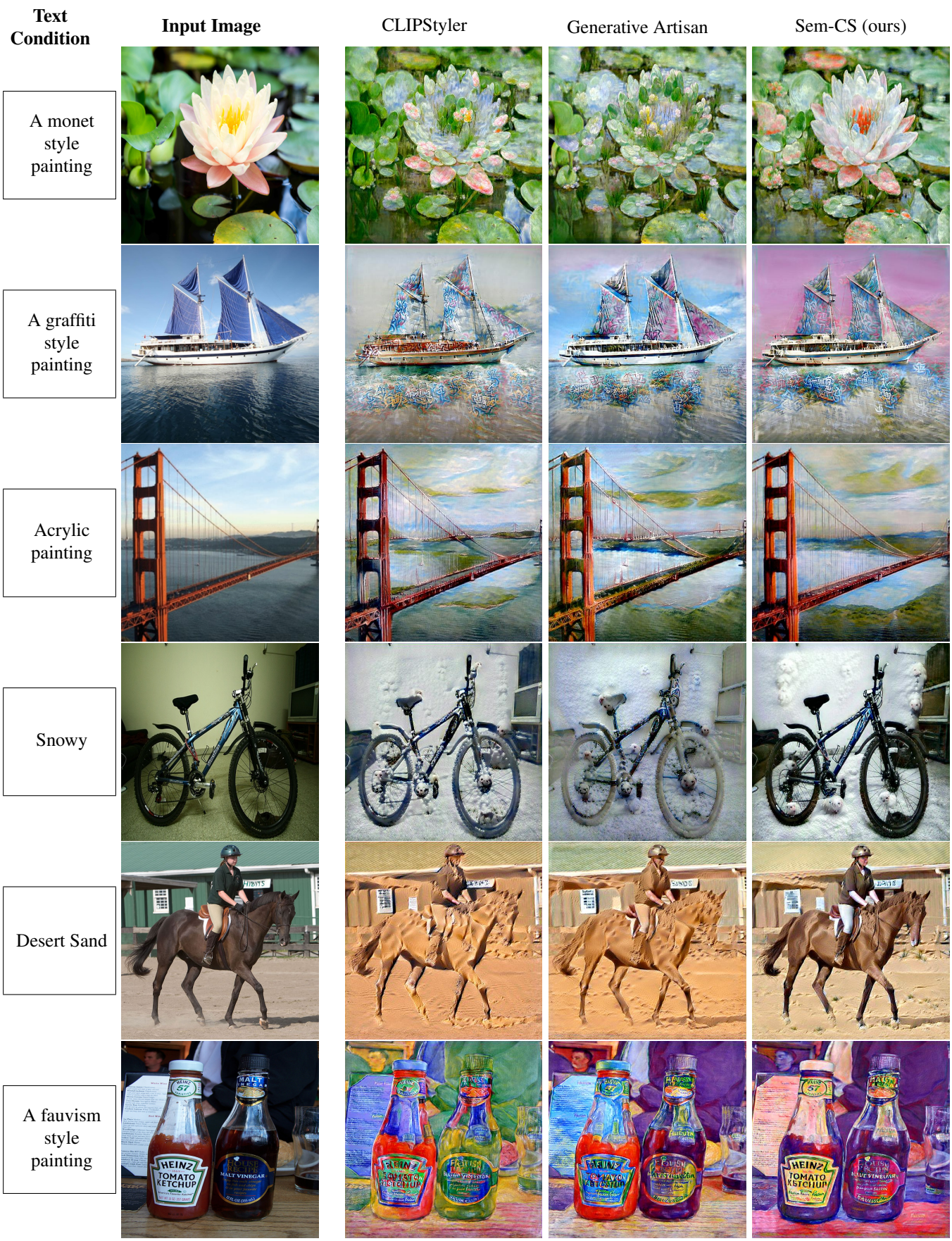
One Style Text			Two Style Texts	
Sem-CS	**Gen-Art**	**CLIPStyler**	**Sem-CS**	**Gen-Art**
0.39	0.22	0.36	0.43	0.19
0.38	0.22	0.25	0.41	0.22
0.33	0.19	0.37	0.45	0.27
0.31	0.18	0.31	0.32	0.14
0.34	0.21	0.29	0.41	0.23
0.34	0.21	0.35	0.33	0.16
0.31	0.19	0.18	0.45	0.28
0.38	0.26	0.33	0.42	0.25
0.38	0.26	0.37	0.28	0.12
0.35	0.23	0.35	0.31	0.15
0.32	0.22	0.30	0.31	0.15
0.35	0.24	0.38	0.44	0.28
0.33	0.22	0.31	0.44	0.28
0.31	0.21	0.28	0.29	0.13
0.40	0.30	0.38	0.41	0.25
0.38	0.28	0.39	0.28	0.13
0.40	0.30	0.37	0.28	0.13
0.41	0.31	0.41	0.31	0.16
0.27	0.18	0.33	0.43	0.28
0.31	0.22	0.32	0.45	0.31
0.36	0.26	0.40	0.39	0.25
0.36	0.27	0.42	0.39	0.25
0.31	0.22	0.27	0.27	0.13
0.30	0.21	0.22	0.33	0.19
0.36	0.27	0.39	0.4	0.26
0.37	0.29	0.30	0.36	0.22
0.30	0.21	0.36	0.35	0.21
0.35	0.26	0.35	0.42	0.28
0.43	0.34	0.40	0.28	0.15
0.29	0.21	0.26	0.28	0.15
0.37	0.28	0.35	0.28	0.15
0.29	0.21	0.23	0.28	0.15
0.32	0.23	0.28	0.27	0.14
0.42	0.34	0.40	0.27	0.14
0.32	0.24	0.27	0.44	0.31
0.27	0.19	0.32	0.44	0.31
0.24	0.16	0.31	0.31	0.18
0.35	0.26	0.38	0.27	0.14
0.39	0.31	0.37	0.27	0.14
0.46	0.38	0.45	0.25	0.12
0.31	0.23	0.35	0.32	0.19
0.36	0.28	0.30	0.27	0.14
0.34	0.26	0.25	0.4	0.27
0.28	0.20	0.29	0.38	0.25
0.26	0.18	0.33	0.26	0.13
0.36	0.28	0.37	0.29	0.16
0.37	0.29	0.34	0.3	0.17
0.31	0.23	0.28	0.36	0.23
0.43	0.35	0.43	0.3	0.17
0.29	0.21	0.21	0.41	0.28
0.32	0.25	0.30	0.27	0.15

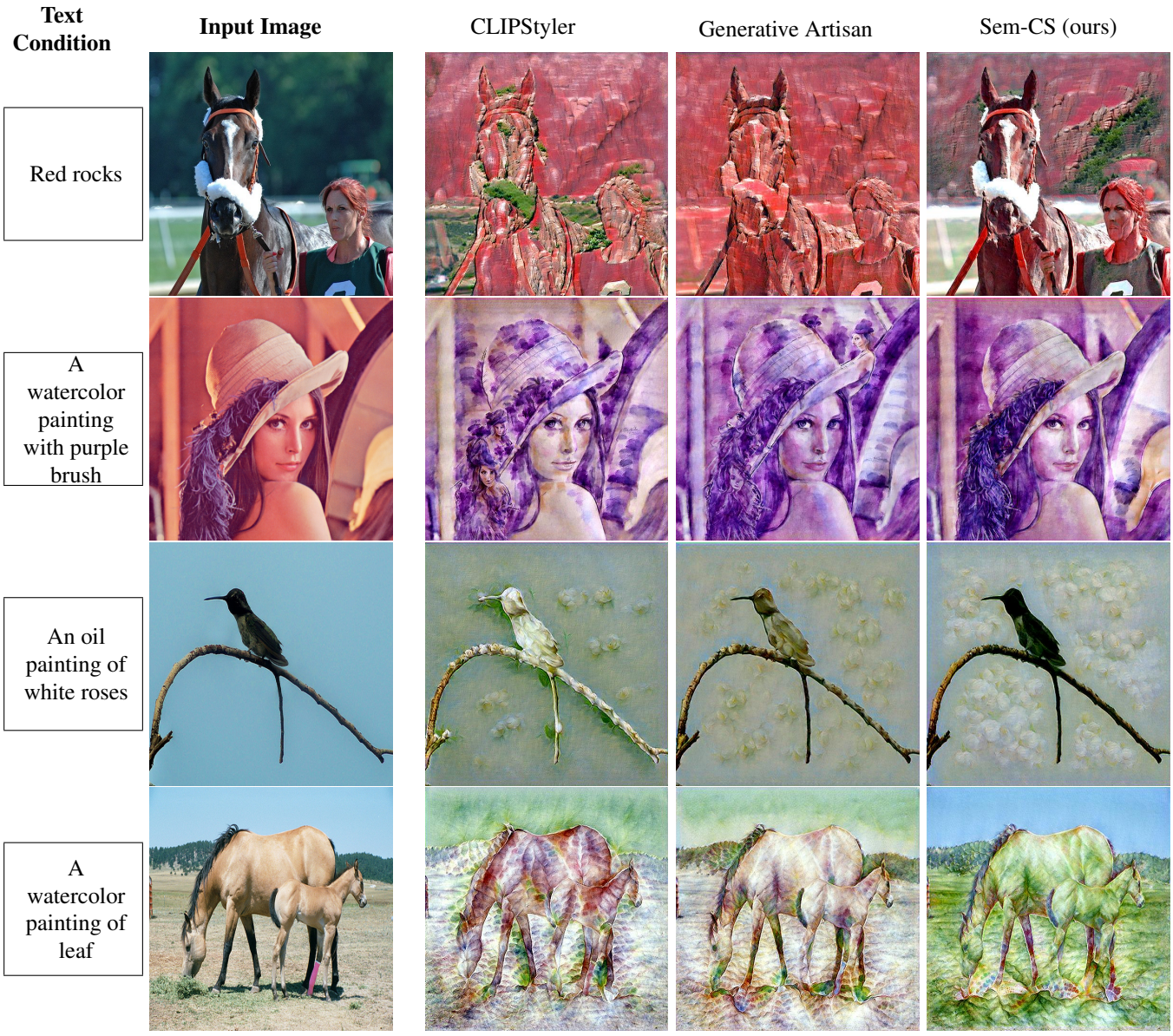
One Style Text			Two Style Texts	
Sem-CS	**Gen-Art**	**CLIPStyler**	**Sem-CS**	**Gen-Art**
0.28	0.20	0.27	0.27	0.15
0.39	0.31	0.28	0.28	0.16
0.32	0.25	0.29	0.27	0.15
0.47	0.40	0.44	0.38	0.26
0.27	0.19	0.26	0.32	0.2
0.42	0.35	0.33	0.31	0.19
0.25	0.18	0.32	0.4	0.28
0.32	0.24	0.40	0.25	0.13
0.40	0.33	0.36	0.38	0.26
0.48	0.41	0.51	0.32	0.2
0.35	0.28	0.32	0.49	0.37
0.25	0.18	0.21	0.41	0.29
0.39	0.32	0.39	0.4	0.28
0.28	0.21	0.21	0.26	0.14
0.34	0.27	0.30	0.51	0.39
0.42	0.35	0.31	0.29	0.17
0.35	0.28	0.30	0.29	0.17
0.31	0.24	0.33	0.29	0.17
0.24	0.17	0.22	0.46	0.35
0.33	0.26	0.31	0.4	0.29
0.32	0.25	0.36	0.26	0.15
0.33	0.26	0.36	0.26	0.15
0.40	0.34	0.36	0.28	0.17
0.31	0.25	0.34	0.27	0.16
0.25	0.19	0.24	0.27	0.16
0.31	0.25	0.28	0.27	0.16
0.21	0.14	0.19	0.26	0.15
0.35	0.29	0.39	0.47	0.36
0.40	0.33	0.36	0.24	0.13
0.37	0.31	0.35	0.43	0.32
0.33	0.27	0.30	0.24	0.13
0.26	0.20	0.21	0.24	0.13
0.29	0.23	0.35	0.3	0.19
0.36	0.30	0.31	0.37	0.26
0.35	0.29	0.38	0.44	0.33
0.38	0.32	0.40	0.35	0.24
0.29	0.23	0.27	0.41	0.3
0.26	0.21	0.26	0.25	0.14
0.30	0.25	0.32	0.44	0.33
0.31	0.26	0.33	0.35	0.24
0.30	0.25	0.33	0.37	0.26
0.26	0.21	0.26	0.39	0.28
0.37	0.31	0.34	0.25	0.14
0.37	0.31	0.33	0.29	0.18
0.29	0.24	0.29	0.29	0.18
0.32	0.26	0.34	0.29	0.18
0.42	0.37	0.40	0.35	0.24
0.32	0.26	0.27	0.3	0.19
0.47	0.41	0.45	0.24	0.13

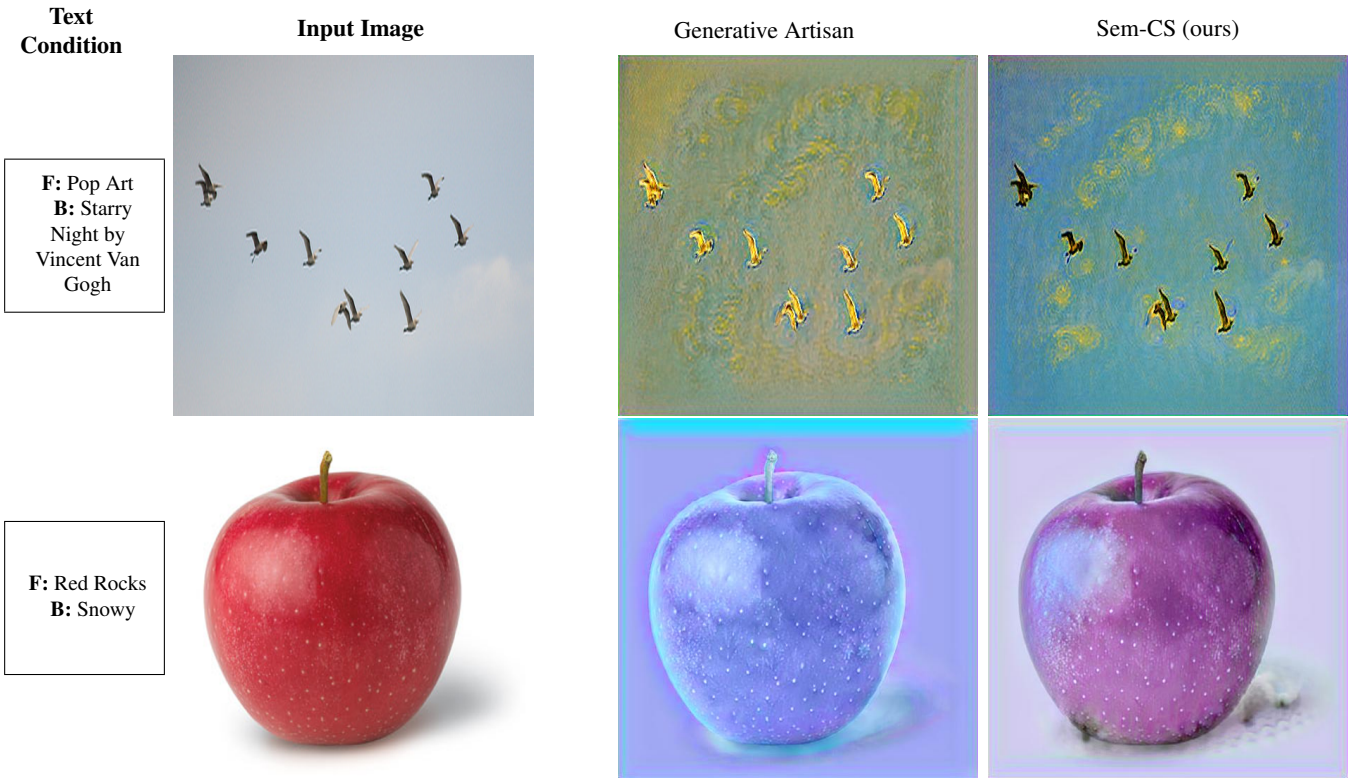
Table 2. NIMA

One Style Text			Two Style Texts	
Sem-CS	**Gen-Art**	**CLIPStyler**	**Sem-CS**	**Gen-Art**
5.67	3.81	4.93	5.59	3.92
5.85	4.06	4.50	6.03	4.39
5.54	4.04	4.19	5.26	3.62
4.98	3.56	3.79	6.17	4.55
5.51	4.11	4.66	6.36	4.8
5.21	3.82	4.30	6.25	4.76
5.96	4.58	5.13	6.53	5.05
5.94	4.57	4.03	5.32	3.85
5.40	4.10	4.39	5.66	4.24
5.70	4.44	5.04	5.59	4.23
5.58	4.33	5.36	5.8	4.44
5.52	4.29	4.79	5.44	4.11
6.04	4.81	4.60	5.91	4.59
6.17	4.97	4.74	6.41	5.09
5.20	4.00	3.79	5.91	4.61
5.58	4.38	5.02	5.98	4.7
5.54	4.35	4.98	5.64	4.36
4.44	3.25	3.81	5.28	4.02
5.45	4.26	4.64	5.61	4.35
5.33	4.15	4.45	6.18	4.92
5.57	4.39	4.62	5.88	4.63
5.27	4.12	4.05	4.79	3.54
5.66	4.51	4.84	5.87	4.62
4.68	3.53	3.83	5.15	3.91
5.68	4.55	4.81	5.63	4.39
5.18	4.06	4.30	5.89	4.65
5.73	4.61	5.26	5.33	4.1
5.36	4.26	4.36	5.65	4.42
5.28	4.19	3.90	5.59	4.36
5.18	4.12	4.75	5.99	4.77
5.32	4.28	4.79	5.85	4.63
5.09	4.06	4.71	5.77	4.56
5.73	4.71	5.09	5.53	4.33
5.07	4.05	4.28	6.24	5.05
5.63	4.63	4.32	5.44	4.25
4.70	3.70	4.12	4.72	3.53
5.81	4.81	4.87	5.93	4.74
5.39	4.40	4.40	5.41	4.23
4.94	3.96	4.22	4.94	3.77
5.85	4.87	4.71	5.85	4.68
5.98	5.00	5.42	5.86	4.69
4.97	3.99	4.60	5.26	4.09
5.54	4.57	4.99	5.72	4.55
4.89	3.92	4.91	5.76	4.6
5.41	4.45	4.86	4.99	3.83
6.19	5.23	5.77	4.67	3.51
5.23	4.29	4.14	5.19	4.03
5.44	4.50	4.37	5.5	4.34
5.25	4.31	4.91	4.63	3.47
5.20	4.27	5.40	5.52	4.36
5.13	4.20	4.48	4.52	3.37

One Style Text			Two Style Texts	
Sem-CS	**Gen-Art**	**CLIPStyler**	**Sem-CS**	**Gen-Art**
5.92	4.99	5.60	4.69	3.55
4.03	3.10	4.05	4.32	3.18
4.83	3.90	4.54	5.18	4.04
4.33	3.41	4.19	4.81	3.67
4.26	3.34	3.68	5.34	4.21
4.67	3.76	4.25	5.83	4.7
4.75	3.85	4.34	5.77	4.64
6.05	5.15	5.88	5.8	4.67
4.94	4.04	3.96	4.8	3.68
5.43	4.53	4.70	5.46	4.35
5.10	4.22	4.72	6.24	5.13
5.08	4.21	4.75	5.37	4.27
5.61	4.74	4.99	5.66	4.56
5.84	4.97	4.74	5.76	4.66
4.05	3.19	3.91	5.81	4.71
6.08	5.22	5.43	5.31	4.21
5.01	4.16	4.73	5.34	4.24
5.40	4.56	5.16	4.9	3.81
4.87	4.03	4.41	5.25	4.17
5.43	4.59	5.16	5.34	4.26
5.67	4.83	4.56	5.38	4.3
5.67	4.83	5.23	6.06	4.98
5.21	4.38	4.17	5.92	4.85
4.15	3.32	4.22	5.56	4.49
5.24	4.41	3.89	5.51	4.44
6.04	5.22	5.58	5.97	4.9
5.17	4.35	4.27	5.62	4.56
6.00	5.18	4.91	4.83	3.77
5.44	4.62	4.71	5.85	4.79
5.67	4.85	5.24	4.68	3.62
5.28	4.46	4.49	5.94	4.89
5.32	4.51	4.94	5.73	4.68
5.40	4.58	4.32	5.82	4.77
5.14	4.33	4.48	5.32	4.27
4.58	3.77	4.57	5.67	4.62
5.20	4.40	3.92	5.5	4.45
5.40	4.60	4.38	5.22	4.17
4.99	4.20	5.01	4.9	3.86
5.37	4.58	4.92	5.75	4.71
5.57	4.78	4.41	5.72	4.68
5.68	4.88	4.63	5.22	4.18
5.85	5.05	4.72	5.92	4.89
4.35	3.55	3.41	5.66	4.63
5.79	5.00	4.85	5.28	4.25
5.77	4.99	4.71	5.06	4.03
5.79	5.01	4.06	5.27	4.24
5.22	4.44	4.80	5.23	4.21
4.72	3.95	4.66	5.66	4.64
5.69	4.92	5.23	5.28	4.26


Text Condition
Input Image
CLIPStyler
Generative Artisan
Sem-CS (ours)
A monet style painting
A graffiti style painting
Acrylic painting
Snowy
Desert Sand
A fauvism style painting


Text Condition
Input Image
CLIPStyler
Generative Artisan
Sem-CS (ours)
Red rocks
A watercolor painting with purple brush
An oil painting of white roses
A watercolor painting of leaf


Text Condition
Input Image
Generative Artisan
Sem-CS (ours)
F: Pop Art
B: Starry Night by Vincent Van Gogh
F: Red Rocks
B: Snowy