Supplementary Material for FonTS: Text Rendering with Typography and Style Controls

Wenda Shi[1] Yiren Song[2] Dengming Zhang[3] Jiaming Liu[4] Xingxing Zou[1 *]

[1]The Hong Kong Polytechnic University [2]National University of Singapore
[3]Zhejiang University [4]Tiamat AI [*]Corresponding author

This supplementary material serves as a complement to the main paper, including additional results presented in Section 1; more ablation studies of TC-FT, ETC-tokens, and SCA detailed in Section 2; demonstration of BTR, ATR and STR in Section 3; discussion of semantic confusion is detailed in Section 4; details of the datasets used in Section 5; details of word accuracy (Word-Acc) in Section 6; and further details regarding user study in Section 7.

1. More Results

1.1. Typographic Controls in STR

We found that the typography controls acquired from Basic Text Rendering (BTR) can be partially transferred to other text rendering tasks. The model's capacity to learn typography attributes from simple text images shows considerable promise for generalization and adaptability in various domains. Consequently, this enables the application of typographic controls, as depicted in Figure 1, and font selection, as displayed in Figure 2, in Scene Text Rendering (STR).

1.2. Differences with Flux-IPA

1) Our style control adapters (SCA) employ a two-stage training approach. Fine-tuning with SC-artext significantly boosts artistry without compromising the accuracy of text, making it more suitable for the ATR task.
2) In contrast to Flux-IPA(XLabs) [1], which is only applied on MM-DiT, our SCA is implemented on both MM-DiT and Single-DiT to enhance style control, as depicted in Figure 3 with Figure 4. Even with a style image scale of 0.6, the style achieved by applying SCA on both MM-DiT and Single-DiT is markedly superior to that of applying SCA only on MM-DiT with a style image scale of 0.9. The comparison between Table 1 and Table 2 further validates this, as applying SCA on both MM-DiT and Single-DiT yields a higher CLIP-I score under different settings.
3) Unlike Flux-IPA(InstantX) [2] which uses SigLIP [20], our method select CLIP [16] as the image encoder. This choice is grounded in the distinct characteristics of these two models. SigLIP [18, 20] is renowned for its robust OCR capabilities. Conversely, as discussed in [6, 13], CLIP's visual embeddings are insensitive to text. This insensitivity to text in CLIP's visual embeddings is pivotal for our application, as it mitigates content leakage from style images (artistic text images). The visual outcomes presented in Figure 5 provide empirical evidence in support of our selection.
4) Distinct from previous methods, we insert adapters in an interval-skip manner (on layer 0,2,4...) to reduce costs. In terms of parameter usage, the parameters of adapters in Flux-IPA(InstantX) are approximately 2.85 times ours, as demonstrated in Table 3.

[1] Flux-IPA(XLabs)
[2] Flux-IPA(InstantX)

1.3. More Qualitative Results of ATR

This section serves as a supplement to Section 4.2 of the main paper, offering a qualitative comparison of our method with Glyph-ByT5 [13] and Textdiffuser-2 [4] on the ATR-bench dataset, as depicted in Figure 6. Additionally, we present our extended qualitative results on the ATR-bench dataset, including single-word and multi-word examples, in Figure 7. Notably, in the second row of results in Figure 7, the accurate mirror reflection of letters in every result further substantiates the effectiveness of our SCA. This example showcases that our SCA can inject style while meticulously maintaining text accuracy, providing additional empirical evidence for the capabilities of our proposed approach in the ATR task.

1.4. Train Baseline

In addition to the aforementioned comparisons, we finetune another baseline, AnyText [19] on the TC-dataset using a method similar to TC-finetuning. The quantitative results are presented in Table 5, while the qualitative results are shown in Figure 8. These results clearly reveal that AnyText fails to acquire word-level controllability. The performance of Glyph-ByT5 and Textdiffuser-2 exhibits similar limitations. This may be attributed to the inherent restricted

A female teacher writing \"<u>Please</u*> be quite\" in chalk on a blackboard.*

a 3d model of a 1980s-style computer with the text "<i>my</i*> old habit" on the screen*

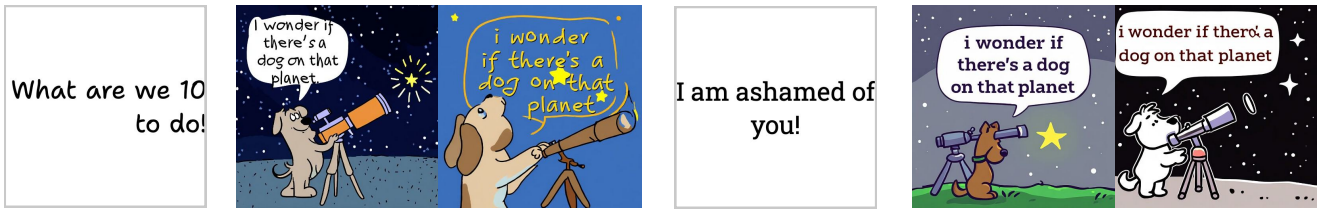
dslr portrait of a robot is holding a sign with text "i am <b>not</b*> a robot"*

A female teacher writing \"Please be <u>quite</u*>\" in chalk on a blackboard.*

a 3d model of a 1980s-style computer with the text "my old <i>habit</i*>" on the screen*

dslr portrait of a robot is holding a sign with text "i am not a <b>robot</b*>"*

Figure 1. Examples of typographic controls in STR.



training example of font:4

a cartoon of a dog holding a telescope looking at a star with a speech bubble saying "i wonder if there's a dog on that planet" in <Font: font:4>

training example of font:6

a cartoon of a dog holding a telescope looking at a star with a speech bubble saying "i wonder if there's a dog on that planet" in Font: <font:6>

Figure 2. Examples of font selection in STR.

capabilities of the base models for text rendering. Figure 9 shows the attention maps of different base models for different words in basic text rendering.

1.5. Stylization of STR

With our SCA, the influence of style input on the text within the image is minimal, as clearly observable in Figure 10. When distinct style images are incorporated, a pronounced transformation in the text style ensues. Notwithstanding these changes in style, the integrity of the text content is maintained, remaining accurate and distinguishable.

2. More Ablation

2.1. Ablation on SCA

SCA Only on MM-DiT. Upon comparing Figure 3 and Figure 4, it is observed that when SCA is implemented on both MM-DiT and Single-DiT, the degree of stylization achieved is substantially greater than when SCA is applied solely to MM-DiT. This holds true even when the scale of the style image is lower (images Figure 4) in the former case (left images in Figure 3). A comparison between Table 1 and Table 2 provides additional validation of this assertion when evaluated in the context of CLIP-I metrics.

SCA with Art-FT and TC-FT. The CLIP-I and OCR-Acc presented in Table 1 are the average figures obtained on ATR task when the scale of the style image is set at 0.9 and 0.6, respectively. Table 1 is identical to Table 6 in the main paper. These values are placed here to enable a more direct comparison with SCA only on MM-DiT (Table 2). It becomes evident that, irrespective of whether SCA, the impacts of Art-FT and TC-FT on the ATR task remain consistent: Art-FT enhances stylization, while TC-FT improves content accuracy. Additionally, as shown in Table 4, after Art-FT, the degree of style degradation caused by TC-FT is reduced. This highlights the distinct but complementary roles of Art-FT and TC-FT in optimizing both the stylistic and content-related aspects of the results.

Without SCA. As is evident from Figure 13, in the absence of SCA, even when a detailed style caption is employed to characterize the style, diverse text contents result in inconsistent styles under the same random seed. Moreover,

Figure 3. Ablation study of SCA only on MM-DiT with Art-FT and TC-FT.

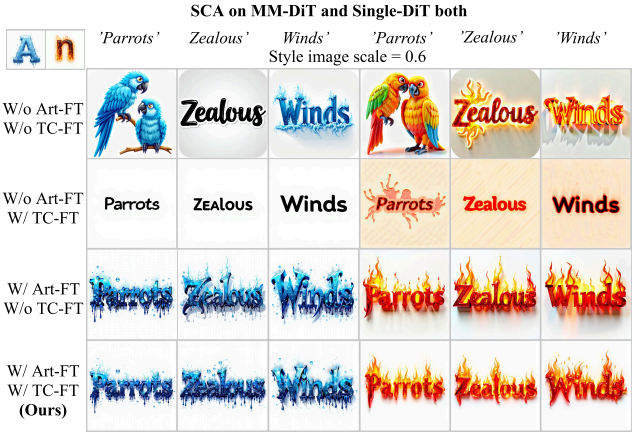


Figure 4. Ablation study of SCA on MM-DiT and Single-DiT both, with Art-FT and TC-FT when image scale = 0.6.

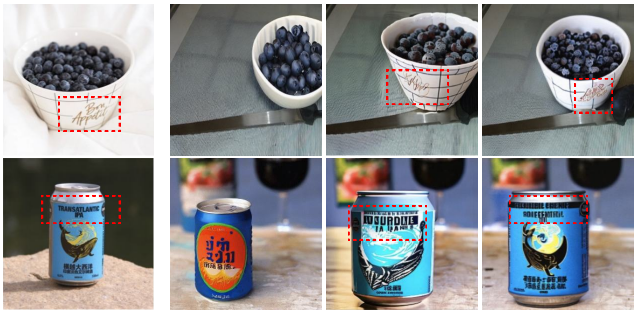


Figure 5. Results of different backbones for the ID extractor in AnyDoor [6]. "DINOv2*" refers to removing the background of the target object with a frozen segmentation model before feeding it into the DINOv2 model. This figure is adapted from [6].

through a comparison of the images in the two rows, it becomes apparent that TC-FT exerts a certain degrading impact on the artistic style imparted by the style caption.

2.2. Ablation on TC-FT

Regarding the ablation study of typography control fine-tuning (TC-FT), we configured four distinct training scenarios: (1) only new tokens, (2) T5 text encoder with new tokens, (3) joint text attention (Txt-Attn) with new tokens, and (4) joint text-image attention (Txt+Img-Attn) with new tokens. As previously established in [8, 13], text rendering performance is primarily governed by the text encoder architecture. To explore this, we attempted to fine-tune the T5 on the BTR dataset to enhance controllability in text rendering. However, this approach led to a substantial decline in text accuracy, with visual artifacts evident in the generated outputs. The visual results are documented in Figure 16.

2.3. Ablation on ETC-Tokens

This section supplements Section 4.3 of the main paper, focusing on demonstrating the effectiveness of the proposed Enclosing Typography Control (ETC)-tokens for targeted word-level typographic attributes. For instance, to bold the word "robot" in the phrase "i am not a robot", we explore three settings: 1) Non-Token: Using an instruction prompt instead of adding modifier tokens, such as "the 'robot' is in bold". 2) Single-Token: Following [2, 12], we trained our model to use a single token, placing the modifier token before "robot". 3) Our ETC-Token. The visual results of ablation on ETC-tokens are presented in Figure 11.

2.4. Ablation on 'sks' prefix

We tried to mitigate language drift (scene-text detachment) by training with the 'sks' prefix in prompts of the TC-Dataset, which is omitted during inference. This low-cost

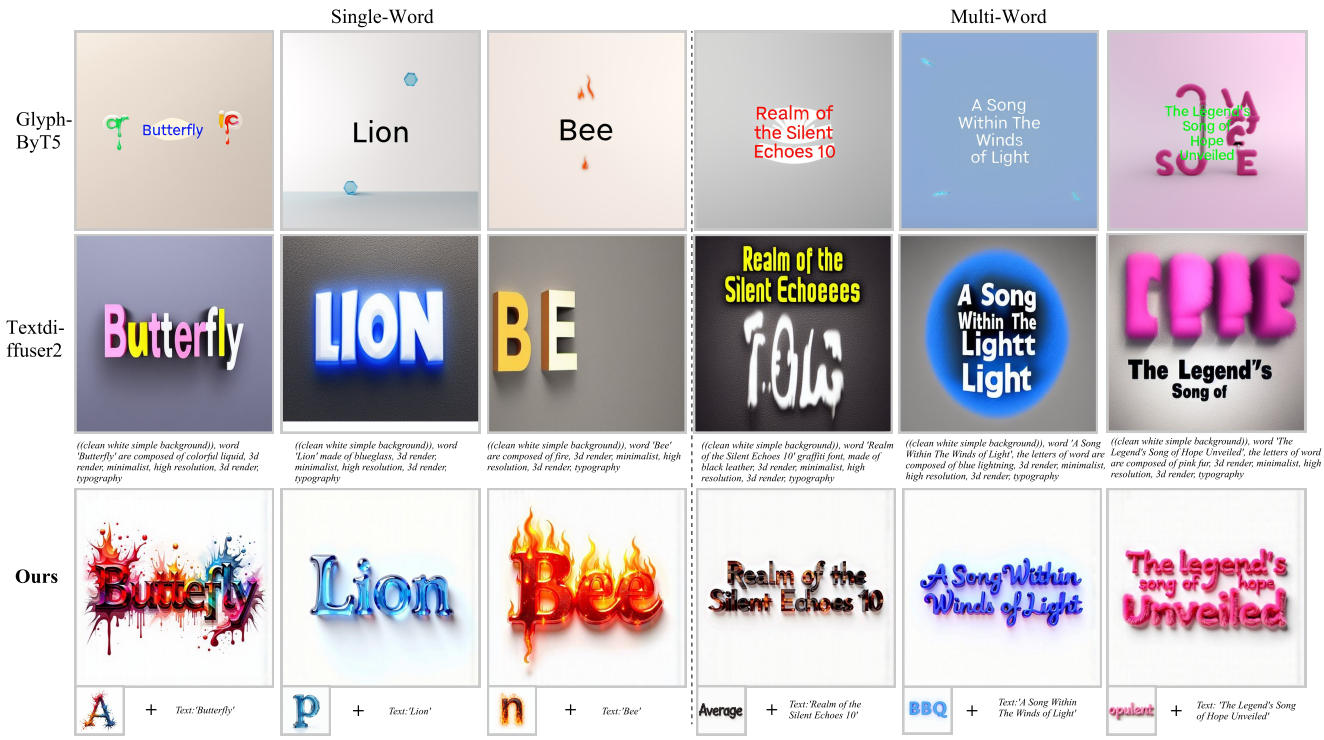


Figure 6. Results of Glyph-ByT5 [13] and Textdiffuser-2 [4] on ATR-bench.

approach helps alleviate detachment, as shown in Figure 12.

Art-FT	TC-FT	CLIP-I ↑	OCR-Acc ↑	Avg ↑
✗	✗	60.07	28.89	44.48
✗	✓	58.09	**65.39**	61.74
✓	✗	**65.12**	34.48	49.80
✓	✓	64.27	60.07	**62.17**

Table 1. Ablation studies of fine-tuning with SC-artext (Art-FT) for SCA (on MM-DiT and Single-DiT both) and TC-finetuning (TC-FT) for backbone. The last row is ours.

Art-FT	TC-FT	CLIP-I ↑	OCR-Acc ↑	Avg ↑
✗	✗	54.19	24.32	39.26
✗	✓	51.64	**60.79**	56.22
✓	✗	**58.14**	17.89	38.02
✓	✓	56.40	58.27	**57.34**

Table 2. Ablation studies of fine-tuning with SC-artext (Art-FT) for SCA (only on MM-DiT) and TC-finetuning (TC-FT) for backbone.

Modules	Non-Skip	Skip (Ours)
Adapters	1434.45 M	503.38 M

Table 3. Parameter quantity comparison with Flux-IPA(InstantX).

Δ_{CLIP-I}	w/o Art-FT	w/ Art-FT
Both	1.98 (60.07 → 58.09)	0.85 (65.12 → 64.27)
Only	2.55 (54.19 → 51.64)	1.74 (58.14 → 56.40)

Table 4. Comparison of CLIP-I changes with and without Art-FT in two SCA settings after TC-FT. Both: SCA on MM-DiT and Single-DiT both, Only: SCA only on MM-DiT.

Methods	OCR-Acc ↑	Word-Acc ↑	Font-Con ↑
AnyText	43.78	✗	3.67
AnyText +TC-FT	39.26	✗	2.64
Ours	**82.85**	**55.00**	**68.42**

Table 5. Quantitive results of AnyText and with TC-FT on BTR.

3. Demonstration of BTR, ATR and STR

This section provides additional information to complement Section 1 of the main paper, which outlines the scope of three text rendering tasks:

Figure 7. More qualitative results of ours on artistic text rendering.

Figure 8. Qualitative results of AnyText [19] and with TC-Finetuned on BTR.

Methods	OCR-Acc↑	Word-Acc↑
Non-Token	71.00	25.00
Single-Token	72.88	32.00
ETC-Token(**Ours**)	**82.85**	**55.00**

Table 6. Ablation studies of ETC-Token on basic text rendering.

- Basic Text Rendering (BTR) involves rendering simple text on a solid color background without any additional scene elements, as illustrated in Figure 17(a).
- Artistic Text Rendering (ATR) features a minimalist background that highlights the artistic nature of the text itself, as seen in Figure 17(b).
- Scene Text Rendering (STR) involves integrating text and scene elements in a way that shares contextual meaning and blends harmoniously, as depicted in Figure 17(c).

4. Semantic Confusion

The term "semantic confusion" in the main paper refers to instances where text rendering incorrectly generates visual objects based on the semantic meaning of the text, rather than just producing the text itself. For example, as shown in Figure 14, our intention was to render only the artistic text "Octopus", "MOON", and "CANDLE" in the left three images. However, the images inadvertently include the corresponding objects for these words. Similarly, in the right three images, which are supposed to display text on the scene, the text is absent, and only the specific objects associated with the semantic meaning of text are present.

Additionally, we conducted additional comparisons with Midjourney [14], Flux, and SD3 in Figure 15. Whereas

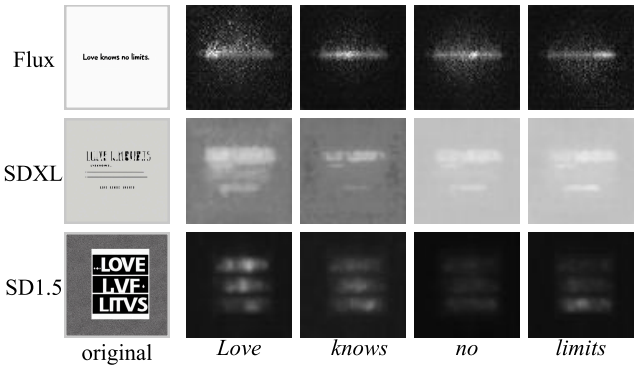


Figure 9. The visualization of attention map on each word in different base models.

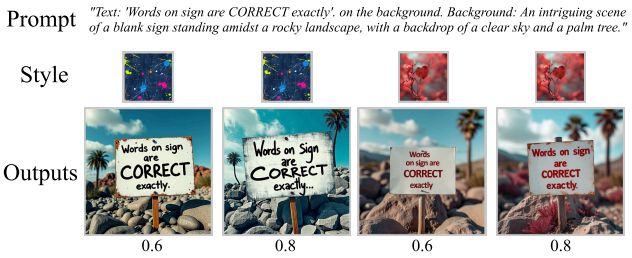


Figure 10. The results of stylized scene text image with different images and image scales.

Metrics \Dataset	TWD [10]	Posta [9]	SC-artext	SC-general
Aesthetic ↑	42.64	47.23	68.57	**84.78**
Quality ↑	55.83	64.89	88.73	**91.20**

Table 7. Aesthetic and quality scores comparison.

original SD3 and Flux lack the capability to process image inputs, both our proposed approach and Midjourney demonstrate the ability to handle combined image-text prompts. The results presented in the figure highlight a critical observation: during artistic text rendering tasks, semantic ambiguity significantly impairs the model's capacity to accurately render the specified word's content. Instead, the model tends to generate visual representations corresponding to the word's semantic reference rather than words itself its. This phenomenon underscores the challenges inherent in balancing stylization and content accuracy within artistic text rendering.

5. Details of Datasets

This section complements Section 3, 4 of the main paper, detailing the datasets we utilized in our work.

Typography Control Dataset (TC-Dataset). To address the lack of high-quality datasets that integrate text with word-level typographic attributes, we developed the TC-Dataset using typography control rendering (TC-Render). This process harnesses HTML rendering to generate images that display typographic features such as various fonts and word-level attributes, including bold, italic and underline. We initiated our process by extracting 625 text excerpts from novels. For each excerpt, we designed an HTML structure comprising sixteen images: one without typographic attributes and, in five different positions, applied three distinct typographic attributes (shown in Figure 18 (a)). Furthermore, we applied data augmentation techniques by randomly altering the text color and background (shown in Figure 18 (b)). Each HTML structure was rendered with one of five different fonts, resulting in approximately 50k text-image pairs with solid color backgrounds.

Style Control Dataset (SC-Dataset).

SC-general. To train our style control adapters, we assembled the SC-general dataset, which includes approximately 580k general image-text pairs with high aesthetic scores. These pairs were sourced from open-source datasets [7, 17]. Figure 19 (a) presents sample images, and Table 8 displays the corresponding paired texts.

SC-artext. For fine-tuning the style control adapters, we created the SC-artext dataset. We combined a list of 100 style descriptions with a list of 99 words, categorized into three character length groups: 1-15, 16-30, and 30-50. This combination produced a variety of prompts for artistic text images, which served as input for Flux.1-dev [1], yielding around 20k high-quality images. To ensure the images accurately reflected the original text content, we utilized shareGPT4v [5] to regenerate captions. Figure 19 (b) shows sample images, and Table 8 presents the paired texts. Besides, we provide quantitative and qualitative comparison results with artistic text in TWD [10] and Posta [9] in Table 7 and Figure 20, respectively. We randomly sample 100 images from each and use specialized LMM (Q-Align [11]) for quality and aesthetic evaluation.

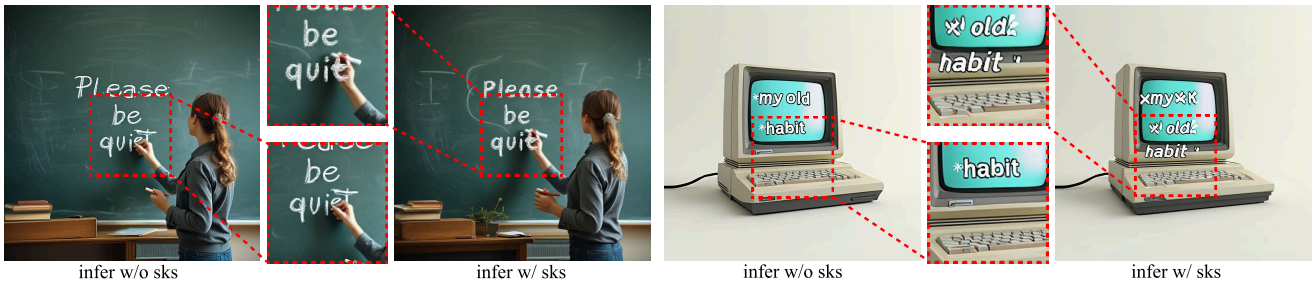
6. Details about Word-Acc

Current open-source OCR tools lack the capability to recognize word-level attributes such as bold, italic, and underline. To address this limitation, we employ GPT-4o [15] to evaluate the accuracy of word-level attributes (Word-Acc). We have designed a structured prompt, supplemented with example cases, to improve GPT-4o's precision in predicting these attributes. Figure 21 illustrates a dialogue record that showcases GPT-4o's strong context comprehension and logical reasoning abilities.

7. Details of User Study

This section complements Section 4.1 of the main paper, providing additional details on the user studies. We involved 22 participants in these studies to evaluate our results perceptually, comparing them to baseline methods. The

Figure 11. Visual results of ablation on ETC-tokens.



infer w/o sks infer w/ sks infer w/o sks infer w/ sks

Figure 12. Infer without 'sks' to mitigate scene-text detachment.

evaluation focused on two main aspects: font consistency (Font-Con) and style consistency (Style-Con). For Font-Con, we had two subtypes. One evaluated the consistency between the output image and the ground truth, and the other judged font consistency across different outputs with the same input. Style-Con was evaluated in a similar way, also with two subtypes. Style-Con was evaluated in two ways: one subtype measured the consistency between the output image and the ground truth, while the other assessed the consistency of fonts across different outputs when the same font input was used. This can be seen in Questions 1 and 2 in Figure 22. Font-Con was evaluated in a similar manner, with two subtypes addressing the same two aspects. These are represented by Question 3 of Figure 22 and Question 4 of Figure 23. Each subtype had a different number of questions: 4, 2, 3, and 2, respectively. The score for each method was determined by dividing the number of votes it received by the total number of votes cast.

References

[1] Black forest labs - frontier ai lab. https://blackforestlabs.ai/, 2024. 6, 8

[2] Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision*, pages 456–472. Springer, 2025. 3

[3] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024. 8

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 4

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 6

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 1, 3

[7] Christoph Schuhmann and Romain Beaumont. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed on January 02, 2024. 6

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[9] Chen et al. Posta: A go-to framework for customized artistic poster generation. *CVPR*, 2025. 6

[10] Wang et al. Typography with decor: Intelligent text style transfer. In *CVPR*, 2019. 6

[11] Wu et al. Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024. 6

[12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[13] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized

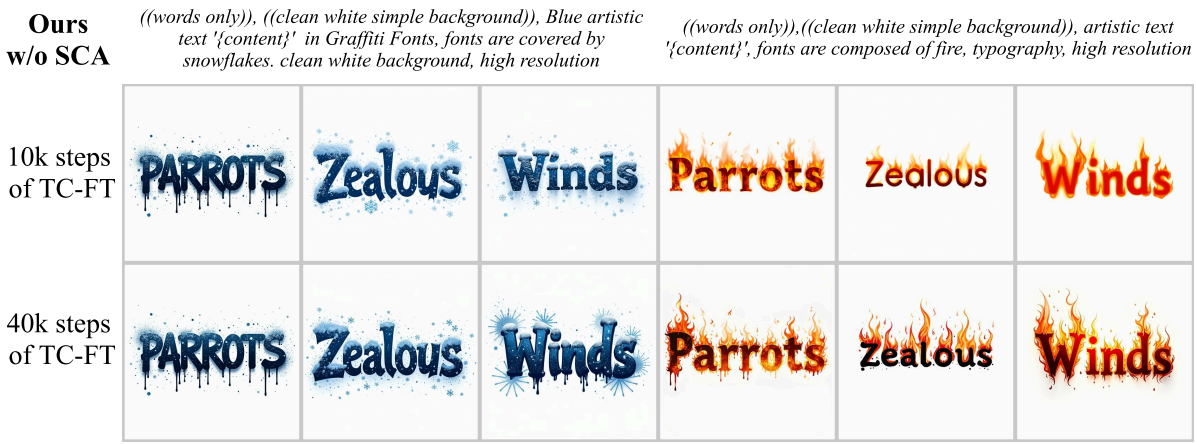


Figure 13. Ablation study of style control adapter (SCA), results from style captions only after 10k and 40k steps of TC-finetuning.

Figure 14. Examples of Semantic Confusion in Flux.1-dev [1]. The prompts for the right three images are from MARIO-bench [3].

text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, pages 361–377. Springer, 2024. 1, 3, 4

[14] Midjourney. Midjourney. https://www.midjourney.com. 5

[15] OpenAI. Hello, gpt-4o. https://openai.com/index/hello-gpt-4o/. 6

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[17] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[18] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

[19] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 5

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

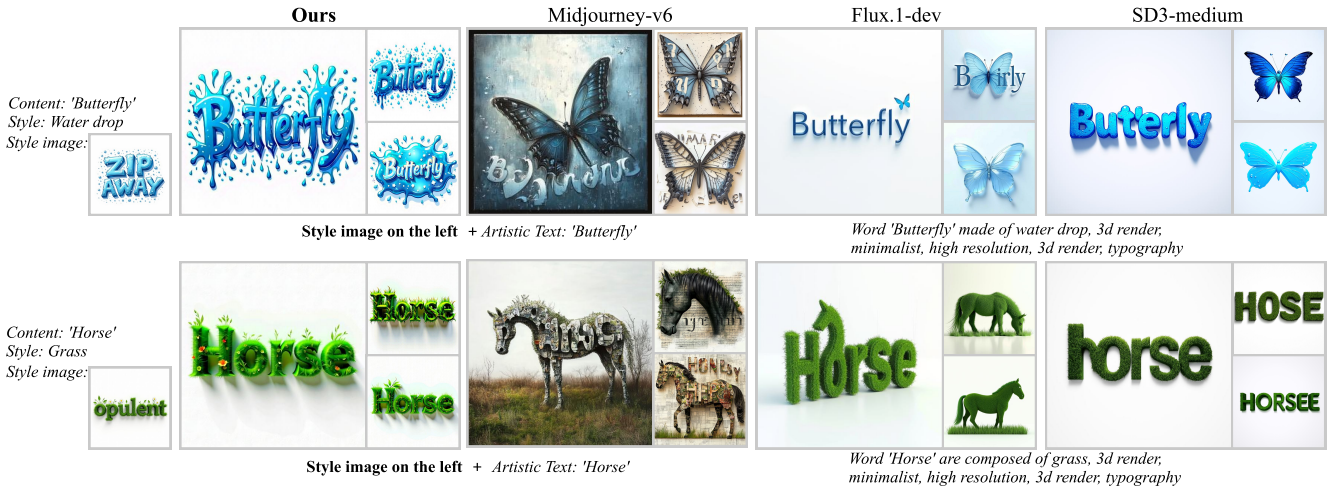


Figure 15. Semantic confusion can also be observed in SD3, Flux and Midjourney.

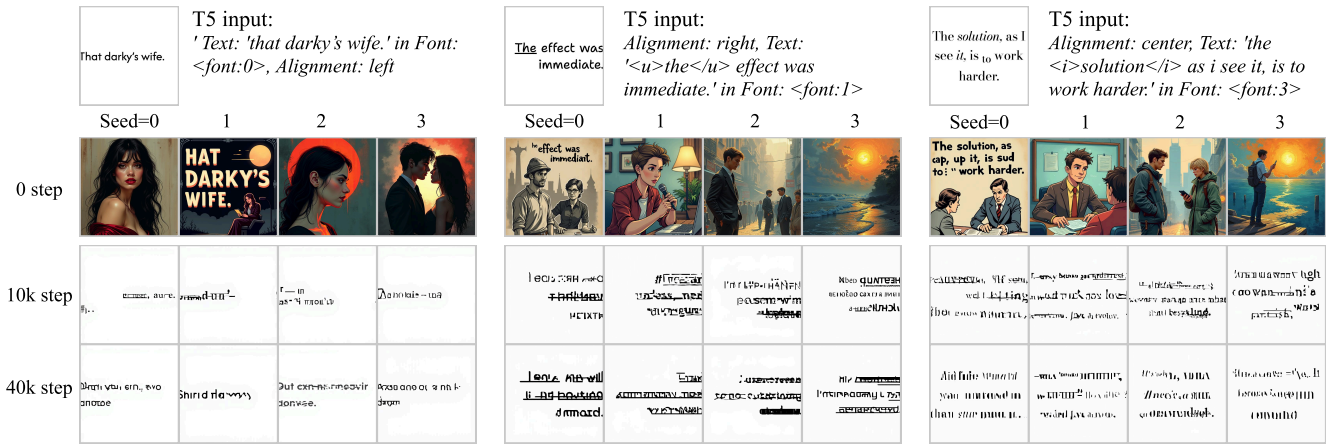


Figure 16. Results of fine-tuning T5 text encoder with new tokens, while input for CLIP is fixed prompt: 'words only, clean background'.

Figure 17. Results of our method: (a), (b), and (c) in basic text rendering, artistic text rendering, and scene text rendering, respectively.

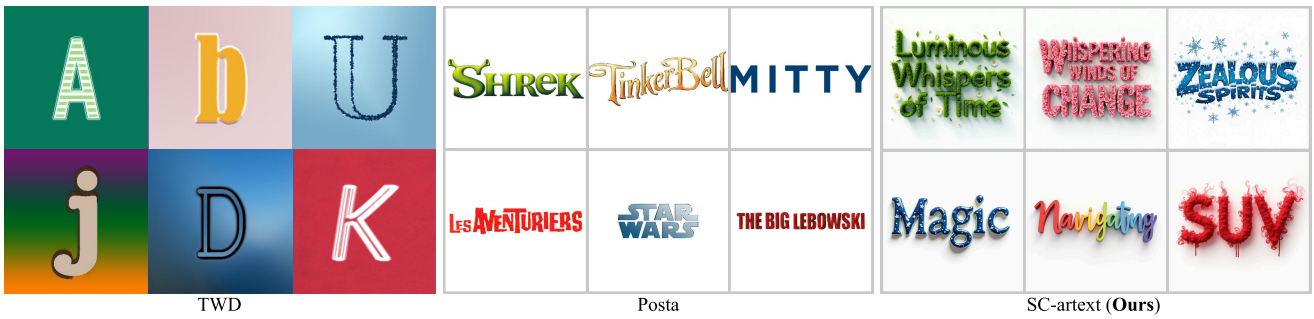
(a) Different Word-level Attributes (b) Colorful Text and Background

Figure 18. Examples of TC-Dataset. (a) different word-level attributes, (b) examples featuring text and background color variations.

(a) Examples of images in SC-general

(b) Examples of images in SC-artext

Figure 19. Examples of images in SC-dataset, (a) is SC-general, and (b) is SC-artext.



TWD Posta SC-artext (**Ours**)

Figure 20. Visual comparison of existing artistic text datasets.

You are an expert in image OCR. You can accurately recognize text, line breaks, and formatting (bold, italic, underline, alignment) within images. Specifically, you need to read the text content within the image and then output a JSON object that includes the following:

- text: Text content, with line breaks represented by `<br>`, NOTE: if there are any breaks in the text, please include them in the output with `<br>`.
- color: Text color, `white`, `black`, `red`, `green`, `yellow` or `blue`.
- bold: List of words in bold order of appearance
- italic: List of words in italics order of appearance
- underline: List of words underlined order of appearance
- align: Alignment for text, `left`, `center`, or `right`.

Output Example:

```json
{
    "text": "Fill your heart with love,<br>cherish every moment,<br>and spread kindness.",
    "color": "yellow",
    "bold": ["heart", "with", "love"],
    "italic": [],
    "underline": ["every"],
    "align": "left"
}
```

Now please accurately identify the above content based on the input image and output it using the specified format.

- Even if there are no text, please output an empty string with above format (`color` and `align` should be "").
- If there are no bold, italic or underline words, please output an empty list.
- Regardless, please output the complete JSON format, including all the fields above, even if some fields are empty.


```json
{
    "text": "Capture the moments<br>that<br>matter",
    "color": "",
    "bold": [],
    "italic": [],
    "underline": ["that"],
    "align": "center"
}
```


```json
{
    "text": "Hold on to hope,<br>believe in the<br>impossible, and<br>conquer your<br>fears.",
    "color": "red",
    "bold": ["believe"],
    "italic": [],
    "underline": [],
    "align": "center"
}
```

Figure 21. Example of using GPT-4o to evaluate word-level attribute accuracy (Word-Acc).

Question-1: *Which image best matches the description of "Graffiti blue artistic text 'Banana', with letters covered in snowflakes on a clean white background, high resolution"?*

*Please select the image that most closely aligns with the given text description, considering the overall style and reference image.

Reference Image

Option1 Option2 Option3

Question-2: *Which line among the three options below exhibits the highest style consistency?*

*Style consistency refers to the similarity and uniformity of styles within the same line.

Question-3: *Which font most closely resembles Josefin Sans as shown in the reference image on the right?*

Reference Image

Option1 Option2 Option3 Option4 Option5

Figure 22. Examples of questionnaire to evaluate the Style-Con and Font-Con.

Question-4: *In the set of 5 images provided, which line demonstrates the highest level of font consistency?*

*Font consistency refers to the degree of similarity and uniformity of fonts within the same line across.

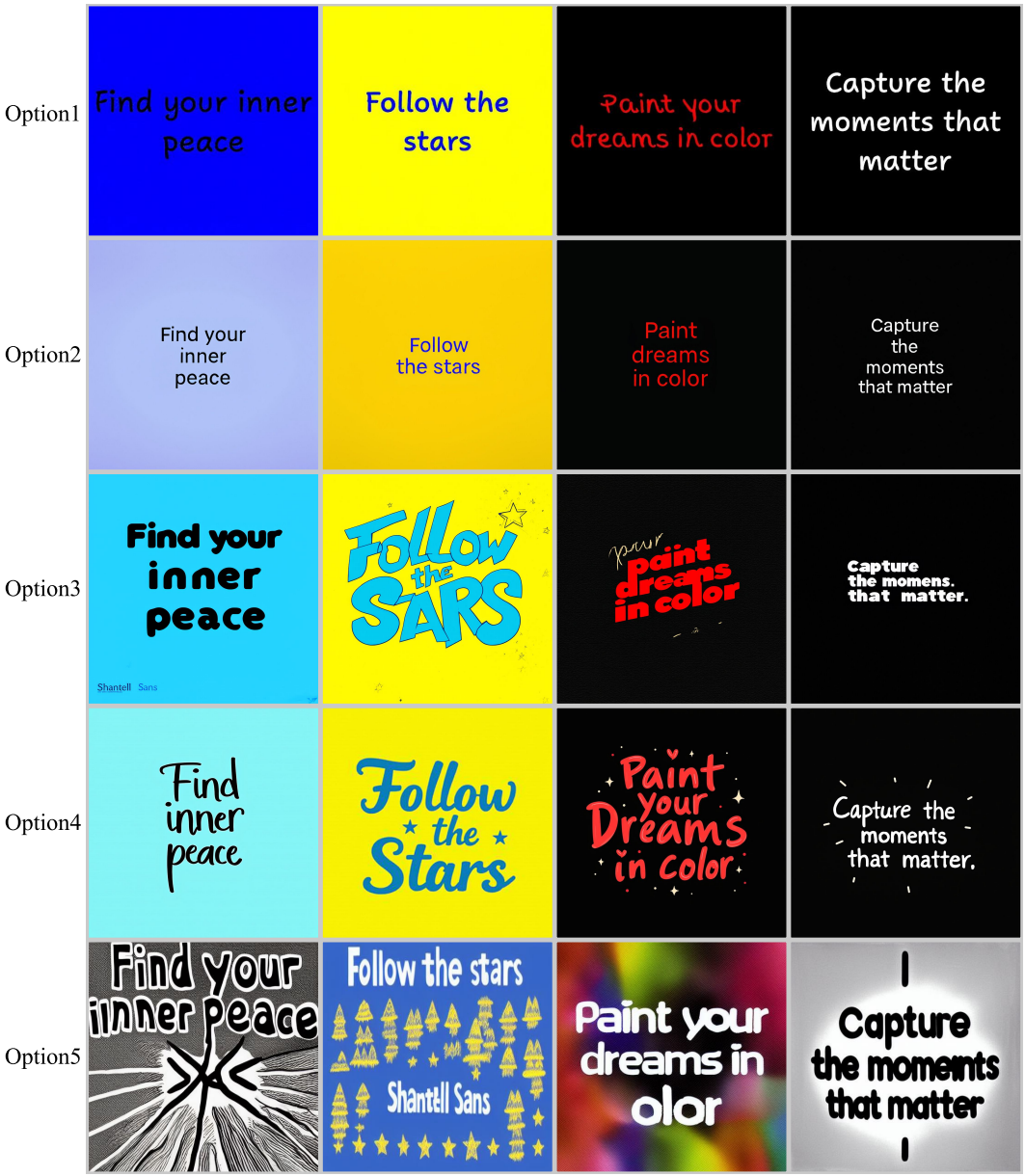


Figure 23. Examples of questionnaire to evaluate the Font-Con.

Image	Text
SC-general, Row 1, Col 1	A photorealistic image of multiple shopping bags in a boho style, fresh and inviting. The bags are in various sizes and patterns, including floral designs, abstract prints, and earthy tones. They have rope handles and are arranged against a soft, neutral background. The overall vibe is natural, stylish, and vibrant.
SC-general, Row 1, Col 2	Dark blue purple red abstract background for design. Painted rough paper. Bright colors include magenta and fuchsia. Smudge, stain, and blot effects are photo-realistic with ultra sharp focus and ultra detailed focus. The image has high coherence and minimalistic style with intricate and hyper realistic details. Beautifully color graded with modern and cinematic light. Captured with a Phase One XF IQ4 camera, 200 Mega Pixels, it features insane detailing and depth of field. The textures give a feeling of depth and richness, enhancing the overall beauty of the composition. The editorial photography and photoshoot elements are evident in the detailed and professional capture.
SC-general, Row 1, Col 3	White and grayish Persian cat with fluffy fur, vibrant green eyes, not a flat nose, has a distinct stop, looking directly into the camera, soft dramatic lighting, cinematic style, slightly backlit.
SC-general, Row 1, Col 4	an alien cyborg with eyes and oozing in the woods, in the style of rendered in cinema4d, undefined anatomy, tangled nests, dark white and crimson, eerily realistic, soft sculptures, made of mist.
SC-general, Row 1, Col 5	A luminous figure draped in glowing robes holds a radiant orb of light with plants and leaves on their shoulders, resembling the Keeper of the Light, Dota 2, in an enchanting, mystical forest ambiance.
SC-artext, Row 1, Col 1	The image presents a simple yet striking visual. Dominating the frame is the word “Average”, spelled out in capital letters. Each letter is identical in size and color, creating a sense of uniformity and balance. The letters are not solid but rather composed of small bumps, giving them a textured appearance that stands out against the stark white background. The word “Average” is centrally positioned, drawing the viewer’s attention immediately to it. Despite the simplicity of the elements involved, the image conveys a clear message: the word “Average”. The absence of any other elements or distractions underscores this message, making it the sole focus of the viewer’s attention.
SC-artext, Row 1, Col 2	The image presents a 3D rendering of the word “Average”. The word is written in a cursive font and is colored in a vibrant shade of blue. It’s slightly tilted to the right, adding a dynamic touch to the overall composition. Each letter is slightly larger than the last, creating a cascading effect that leads the viewer’s eye down the word. The background is a stark white, which contrasts sharply with the blue of the word, making it stand out prominently. The image does not contain any other objects or text, and the focus is solely on the word ”Average”. The simplicity of the image allows the viewer to clearly see and understand the meaning of the word.
SC-artext, Row 1, Col 3	The image presents a 3D rendering of the word “Average”. The word is written in a bold, sans-serif font and is colored in a vibrant shade of red. The letters are slightly tilted to the right, adding a dynamic touch to the overall composition. Each letter is enveloped in a ring of fire, with the letters “A”, “V”, and “R” being particularly noticeable due to their larger size. The background is a stark white, which contrasts sharply with the fiery red of the word, making it stand out prominently. The image does not contain any other discernible objects or text. The focus is solely on the word “Average” and its fiery presentation.
SC-artext, Row 1, Col 4	The image presents a 3D rendering of the word “Average” in a vibrant shade of green. The letters are intricately crafted from grass, giving them a natural and organic feel. Each letter is adorned with small white flowers, adding a touch of whimsy to the overall design. The letters are arranged in a staggered formation, creating a sense of depth and dimension. The word “Average” stands out prominently against the stark white background, making it the focal point of the image. The image does not contain any discernible text apart from the word “Average”.
SC-artext, Row 1, Col 5	The image presents a vibrant display of the word “Average” in a cursive font. The letters are filled with splashes of paint in a rainbow of colors, transitioning from red to orange, then to yellow, green, blue, and finally to purple. Each letter is slightly tilted, adding a dynamic feel to the overall composition. The background is a stark white, which contrasts with the colorful text and allows it to stand out prominently. The word “Average” is the only text present in the image. The relative positions of the letters suggest they are stacked on top of each other, further enhancing the visual impact of the image.

Table 8. Examples of texts in SC-general and SC-artext. Textual description of the first row in Figure 19.