Text Image Generation for Low-Resource Languages with Dual Translation Learning

Supplementary Material

Diffusion Steps (for Training)	1000
Diffusion Steps (for EDB Guidance)	100
Diffusion Steps (for FE Guidance)	200
Noies Schedule	cosine
Channels	192
Channel Multiplier	1,1,2,2,4,4
Number of Heads	3
Number of ResBlocks	3
Batch Size	128
Learning Rate	1e-4
Dropout	0.1
Iterations	1.5M
Embedding Dimension	768
Attention Resolution	32,16,8

Table 6. Hyperparameters for the diffusion model in our proposed framework.

A. Dataset Details

We selected English for the source language due to the widespread availability of datasets. As described in the main text, we incorporated 12 publicly available real datasets, culminating in a total of 2.56M text images aling with their corresponding text labels.

Regarding the target languages, we selected five languages: Arabic, Bengali, Chinese, Japanese, and Korean. To produce synthetic and plain text images for these five languages, as well as English, we utilized SynthTIGER [69], producing 2M images per language. For producing these synthetic text images, we utilized SynthTIGER's default settings, including background image color, texture, text layouts, text styles, midground text, geometric transformation, postprocessing.

The word set for each language was derived from the EasyOCR repository [1] and Wikipedia pages using an API. These word sets consisted of 36 Arabic, 74 Bengali, 6,614 Chinese, 2,100 Japanese, and 1,471 Korean unique characters, respectively. We randomly choose one word from the word set to create the corresponding synthetic and plain text images. We sourced free font files primarily from Google Fonts. However, due to a scarcity of Chinese and Japanese font files in Google Fonts, additional collections were made from various online sources. The final count of font files obtained for each language was 64 for Arabic, 20 for Bengali, 24 for Korean, 33 for Chinese, and 98 for Japanese.

B. Implementation Details

The implementation of the diffusion model in our proposed framework was based on the code released by the authors in [13, 43]. For the training process, we employed the AdamW [29] optimizer with a learning rate 1×10^{-4}. Our diffusion model underwent training for 1.5M iterations using a batch size of 128. During training, the diffusion time step T was set to 1000. At inference, T was set to 100 for the FDB Guidance and 200 for the FE Guidance. The hyperparameters of the model architecture are presented in Tab. 6.

We selected $w_{min} = 1$, $w_{max} = 6$ as the hyperparameters for FDB Guidance across all languages. For the hyperparameters of FE Guidance, we selected $n = 120$. In addition, for k_1 and k_2, we set $(k_1, k_2) = (2, 1)$ for Arabic, $(3, 1)$ for Bengali, $(1, 6)$ for Chinese, $(1, 6)$ for Japanese, and $(1, 1)$ for Korean. During the inference phase with the FE Guidance, we also utilized the FDB Guidance to schedule the guidance scales. All models utilized the same hyperparameters, with the exception of those related to the two guidance techniques. All images, both for training and inference, were resized to a resolution of 32×128.

As explained in the main text, the diffusion model in our framework is conditioned on two types of inputs. The first is a plain image y, and the second is a binary variable c with two states: *synth* and *real*. To condition the diffusion model on a plain image y, the tth input image x_t is replaced by $[x_t, y]$, where $[\cdot, \cdot]$ stands for the operation of concatenation in the channel dimension. For the binary variable, we first represented it as a one-hot vector. This vector is then embedded to match the dimension of the time-embedding vectors. Subsequently, the diffusion model received this embedded one-hot vector in two ways: by adding it to the time-embedding vector, and through the use of cross-attention modules.

The diffusion model in our proposed framework adopts classifier-free guidance [19] to generate text images that more accurately reflect the textual content of the input plain text images. Utilizing classifier-free guidance necessitates the use of an unconditional denoising model $\epsilon_\theta(x_t, t)$ during the inference phase. Following prior studies, this unconditional model is acquired through the joint training of both unconditional model $\epsilon_\theta(x_t, t)$ and conditional model $\epsilon_\theta(x_t, t, c, y)$. Specifically, the conditions c and y were probabilistically omitted at a consistent rate during the training. In our framework, these conditions were omitted with a 10% probability. For omitting condition c, we used

all-zero vectors, and for omitting condition y, we used completely white images.

C. Qualitative Evaluation for the Effectiveness of FE Guidance

In Figs. 7-11, we showcase examples to qualitatively evaluate the efficacy of FE Guidance. These figures display samples of Arabic, Bengali, Chinese, Japanese, and Korean text images, in that order. The first row depicts plain text images, the second row presents examples prior to the application of FE Guidance, and the third row illustrates examples following the application of FE Guidance.

The efficacy of the FE Guidance in correcting the textual content of generated images improves as the value of k_1 increases and the value of k_2 decreases. As described in the main text, we set higher k_1 values and smaller k_2 values for Arabic and Bengali languages. The Arabic and Bengali text images shown in Figs. 7 and 8 reveal that, even through the textual contents of the text images before applying FE Guidance significantly deviate from the intended output, the FE Guidance is capable of successfully rectifying it. Nonetheless, as explained in the main text, this effectiveness is achieved at the expense of the style of real text images. Indeed, after applying the FE Guidance, the styles of these text images tend to resemble those of synthetic text images more closely. Conversely, for Chinese and Japanese, where we used lower k_1 and higher k_2 values, the rectification capability is not as pronounced as for Arabic and Bengali. Our experiments indicate that our framework, without FE Guidance, generates more accurate text images for Chinese and Japanese than for Arabic and Bengali. Thus, in many instances, the lower k_1 and higher k_2 settings are sufficient. The benefit of using lower k_1 and higher k_2 values are that our framework can create text images while maintaining the style of real text images.

Figure 7. Arabic text images before and after applying FE Guidance. The top row shows plain text images, the middle row displays examples before FE Guidance is applied, and the bottom row demonstrates examples after the application of FE Guidance.

Figure 8. Bengali text images before and after applying FE Guidance. The top row shows plain text images, the middle row displays examples before FE Guidance is applied, and the bottom row demonstrates examples after the application of FE Guidance.

Figure 9. Chinese text images before and after applying FE Guidance. The top row shows plain text images, the middle row displays examples before FE Guidance is applied, and the bottom row demonstrates examples after the application of FE Guidance.

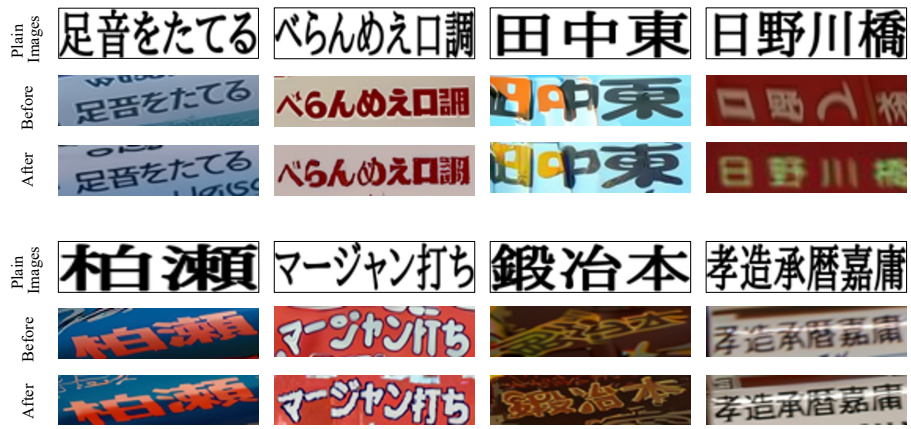

Figure 10. Japanese text images before and after applying FE Guidance. The top row shows plain text images, the middle row displays examples before FE Guidance is applied, and the bottom row demonstrates examples after the application of FE Guidance.

Figure 11. Korean text images before and after applying FE Guidance. The top row shows plain text images, the middle row displays examples before FE Guidance is applied, and the bottom row demonstrates examples after the application of FE Guidance.