

Research on Building K-Contents Specialized Text2Image Pipeline

Seungeun Kang*, Dahyun Lee*, Sangbum Han*, Seokjae Lim**, Namgi Kim*

*Division of AI Computer Science and Engineering, Kyonggi University, 154-42 Suwon, Republic of Korea

** AI Tech Team, LOTTE INNOVATE, 179 Seoul, Republic of Korea

haun620@kyonggi.ac.kr, idahyun22@kyonggi.ac.kr, hsb422@kyonggi.ac.kr

seokjae.lim@lotte.net, ngkim@kyonggi.ac.kr

Abstract—Recently, there has been a lot of research on the Text2Image generative model as social and technological interest in generative models has increased. In addition, the interest in Korean contents (K-Contents) has increased. In this paper, we propose “K-Contents Specialized Text2Image Pipeline”. “K-Contents Specialized Text2Image Pipeline” is specialized for understanding Korean prompts and aims to generate images specialized for various K-Contents. To build the pipeline, we collect and preprocess text and image data. Then, we train the Korean Text Encoder on the text data and train three specialized K-Contents generation pipelines on the image data. We also propose to use the SDXL model’s Img2Img Refiner with the SD model for time-efficient image generation. After that, we use the learned pipelines to generate real images and analyze their performance. Finally, we discuss the utilization of the above pipelines, limitations, and future research.

Keywords—Image Generation, Stable Diffusion, K-Contents, Korean text encoder

I. INTRODUCTION

With the success of Chat GPT, global interest in generative AI has increased, and interest in the Text2Image generative model has increased. In particular, the Text2Image generative model is utilized in various fields and is used as an innovative tool in fields that require creativity, such as design development, advertising, storyboarding, and virtual reality content creation. In addition, the recent “K-Contents Boom” has increased interest in Korean culture and created a global demand for traditional Korean art, hanbok, Korean animation, etc. We were interested in building a pipeline that understands the Korean language and generates Korean-specific images in this technological and cultural context. Therefore, we propose a Text2Image pipeline based on the Stable Diffusion pipeline that can generate various K-Contents images using Korean text prompts. In this paper, we collect text and image data to build a K-Contents Specialized Text2Image Pipeline, and preprocess them according to each data type. Then, a Korean Text Encoder that understands Korean prompts for image generation is trained, and the pipeline is trained to generate three types of K-Contents well. During training, we added an img-to-img process using the SDXL [1] Refiner model to improve the quality of the generated images in a time-efficient manner to solve the problem that the existing pipeline generates strange faces, arms, etc. In addition, we conduct a comparison experiment with the existing Korean Text Encoder

by generating images through Korean prompts and experiment with image generation for K-Contents.

II. RELATED WORK

A. Sentence Embedding Models with Contrastive Learning

The Text2Image Pipeline requires a text encoder to convert natural language sentences into vectors. Recently, sentence embeddings have been the focus of much research, particularly methods utilizing Contrastive Learning (CL). CL learns embeddings by positioning similar sentences closer together and dissimilar sentences farther apart in the embedding space. [2] This approach helps to better capture the meaning of sentences and has been shown to perform well on a variety of NLP tasks. In this regard, Contrastive Language-Image Pre-Training (CLIP) [2] is a prime example of contrastive learning using text-image pairs. However, CLIP, which is pre-trained in a specific language, cannot be used in other languages. Therefore, it is necessary to tune the text encoder to use it in other languages. Multilingual using Knowledge Distillation (N. Reimers et al) [3] has been proposed as a method for this.

B. Diffusion Based Text2Image Pipelines

Diffusion-based models are an emerging approach in Text2Image generation that focuses on generating high-resolution images through progressive denoising during the image generation process. Diffusion models are effective at learning the distribution of data to generate new samples, and are particularly good at achieving both image quality and diversity. The Stable Diffusion model [4] is an example of this evolution of diffusion models. Stable Diffusion 2.0 (SD 2.0) [5] is a further refinement of this performance, providing more sophisticated text-to-image mapping and enabling custom image generation. Stable diffusion XL (SDXL) [1] has nearly tripled the number of parameters from 890 million to 2.6 billion, so the detail is much better than in SD 1.5, and the latent vectors that have been unset can be passed back through the Refiner Unet to produce higher quality images.

III. DATA PROCESSING

A. Collecting Dataset

In this research, we used two types of data. First, we used Korean-English paired corpus data (Corpus Dataset) for training the text encoder model. The Corpus Dataset was

created by merging three datasets from AI hub: “Korean-English Translation (Parallel) Corpus” [6], “Korean-English Translation Corpus (Technical Science)” [7], and “Korean-English Translation Corpus (Social Science)” [8]. The Corpus Dataset has a total of 4,308,894 Korean-English translation corpus pairs. We divided the Corpus Dataset into 3,878,005 training data and 430,889 validation data to train the text encoder model. Second, we used various image data for Stable Diffusion Fintuning. We prepared a total of three types of image data to build the K-Contents Text2Image Pipeline. For “ink painting” fintuning, we used AI hub’s “Korean traditional ink painting production data by style”(Ink Painting Dataset) [9], for “Hanbok” fintuning, we used web crawling data, and for “Korean animation” we used AI hub’s “Character Face Landmark Data in Animation”(Character Dataset) [10]. As shown in Table 1, we extracted data from Ink Painting Dataset [9] for 2,574 objects, people, and landscapes, and used the “Hanbok” dataset, which contains 120 web-crawled realistic portraits. In addition, the “Korean animation” dataset extracts 1,470 photos of specific characters from Character Dataset [10].

TABLE 1. Description of Used Dataset

Data Type	Dataset Name	Number of Data	Description
Text Data	Korean-English Corpus (Parallel)	4,308,894	includes "Korean - English Translation Corpus" data
	Korean-English Corpus (Technical Science)		
	Korean-English Corpus (Social Science)		
Image Data	Korean Traditional Ink Painting	2,574	includes "object", "people", "landscape" data
	Hanbok	120	includes "web-crawled photo" data
	Korean Animation	1,470	includes "korean animation characters" data

B. Preprocessing Dataset

We preprocessed the data from the datasets introduced above for use in model training. We applied data-specific preprocessing methods to different types of data: text data and image data. Corpus Dataset was in json format. We merged three json dataset files (Parallel, Tech Science, Social Science) and converted them into a csv file by setting one Korean-English pair as one row. However, for deep learning model training, text data should be represented as vectors rather than natural language. So, as shown in Figure 1, we implemented tokenize to convert natural language data into vector data. In the Korean-English pair data, we used the tokenizer of huggingface’s “Bingsu/clip-vit-large-patch14-en” [11] for Korean, and the tokenizer of huggingface’s “openai/clip-vit-base-patch32” [12] for English. For Low-

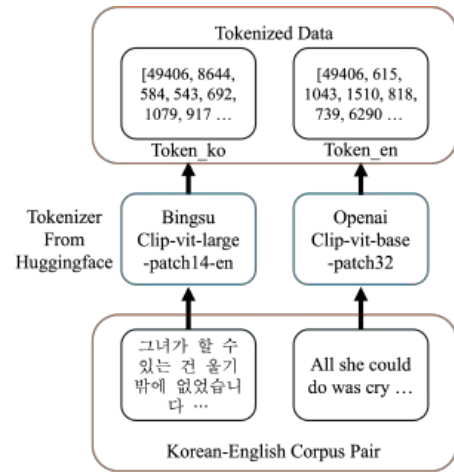


Figure 1. Tokenizing Process

Rank Adaptation(Lora) [13] finetuning of the Stable diffusion model, a square (512*512) image is required. However, the image dataset has a wide range of image aspect ratios, and some images are “extremely vertical”. As shown in Figure 2, we used resizing and random cropping to solve these problems. We resize the shorter part of the image’s width and height to 512px in length. Then, we randomly cropped the long part at a random location to extract a total of two images from one image. After preprocessing the images, we used the Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) [14] model to generate descriptions of the images for image capturing. The “English” descriptions generated by the BLIP model were translated into Korean using the English-Korean translator API. Then, we checked the translated sentences for errors and saved them as image and caption pairs.

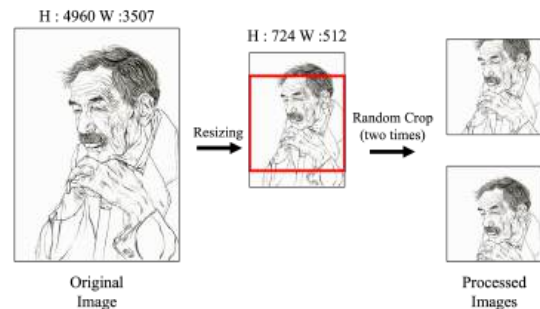


Figure 2. Image Processing

IV. METHODOLOGY

A. Korean Text Encoder

Research on pre-learning a text encoder with a stable diffusion model in Korean has been ongoing, especially in the development online community, and for example, Koclip in 2021 [15] and bingsu/koclip in 2022 [11] have been distributed. However, these models lacked understanding of

“Korean sentences” due to problems such as learning time and data volume. Therefore, we retrained a text encoder that “understands Korean better” using the bingsu/koclip model as a baseline. We used the N. Reimers et al (2020) [3] method, a learning method in bingsu/koclip, to learn text encoders. This learning method uses a “well trained” English text encoder as a teacher model and a “to-be-learned” Korean text encoder as a student model. Given a preprocessed English token and a Korean token, find the embedding values for the English token in the teacher model and the embedding values for the English token and the Korean token in the student model to get the respective MSE (Mean Squared Error). Add the two MSE values to get the training loss. Shown in Table 2, We used batch size 1200, lr 5e-4, epoch 50, MSE loss, cosine scheduler as hyperparameters for training, AdamW [16] as optimizer for updating weights, and 1e-4 as weight decay for optimizer.

TABLE 2. Used Hyperparameter in Text Encoder Training

hyperparameter	value
loss function	MSE
optimizer	AdamW
batch size	1200
learning rate	0.0005
weight decay	0.0001
epochs	50
scheduler	cosine scheduler with warmup

As shown in Figure 3, the Training Loss continues to converge during training, but overfitting after about 10 epochs. However, we saved the text encoder model from the epoch with the smallest validation loss to create a “general” text encoder. After training the text encoder, we build a “Stable Diffusion Pipeline that understands Korean” using the text encoder and use it to train Lora on K-Contents data.

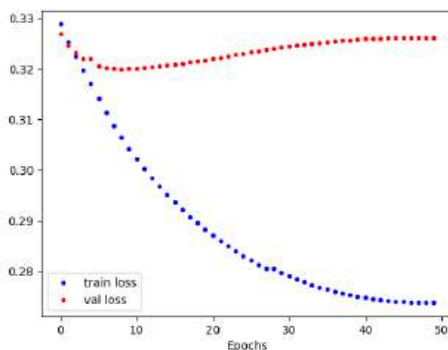


Figure 3. Training and Validation Loss Graph

B. Image Generation Pipeline

In this research, we use the huggingface “runwayml/stable-diffusion-v1-5” [17] pipeline (SD 1.5), a model of the Stable Diffusion method, to build the “K-Contents Specialized

Text2Image Pipeline”. We replaced the text encoder in SD 1.5 with the trained Korean Text Encoder and the tokenizer with the trained Korean tokenizer. Then, we train Lora on the unet model of SD 1.5 with the preprocessed image-caption pairs. Lora training was performed with lr 1e-4, batch size 1. The Ink Painting Dataset was trained with 24,000 steps, the Hanbok Dataset with 3,000 steps, and the Korean Animation Dataset with 12,000 steps. After training Lora weights, we

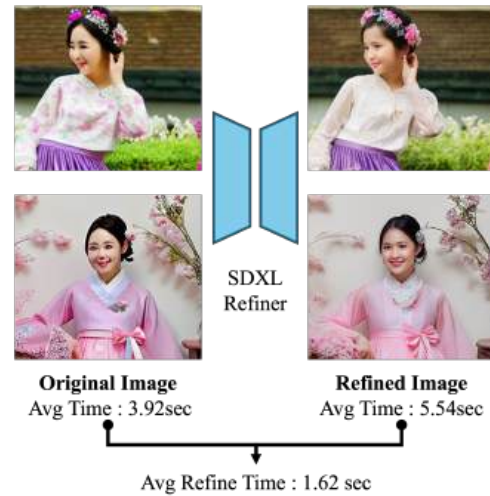


Figure 4. Use SDXL Refiner to improve output Image Quality

loaded them into the SD 1.5 pipeline and generated a few images, we realized a big problem with SD 1.5: it generates strange hands, face shapes, and postures. In particular, strange face generation is a difficult problem to solve with prompting methods that utilize negative prompts. Therefore, in this study, we use the Image2Image Refiner of the SDXL [1] model to improve SD 1.5 images. Shown in Figure 4 this method is time-efficient and effective in improving image quality. The overall pipeline organized by synthesizing the above methods is shown in Figure 5. Enter the Korean prompt to utilize the embedding value of the Korean Text Encoder as the condition of the SD 1.5 Latent Unet, select the Lora Weight and merge it with the weight of the SD 1.5 Latent Unet, and generate the original image. Afterward, SDXL’s Image2Image Refiner generates a high-quality refined image by entering the pre-written Base Prompt. In the next chapter, we will analyze the image generation capabilities of the pipeline built above and compare it to the text encoder.

V. EXPERIMENTS

A. Experiment on Korean Text Encoder

We experimented with image generation using the Korean Text Encoder and the baseline bingsu/koclip. We generated images for 50 full text Korean prompts generated by OpenAI GPT 4.0 from various keywords in various fields and judged the image quality. As shown in Figure 6, the experimental results show that our Korean Text Encoder reflects the sentence data that the existing text encoders could not reflect.

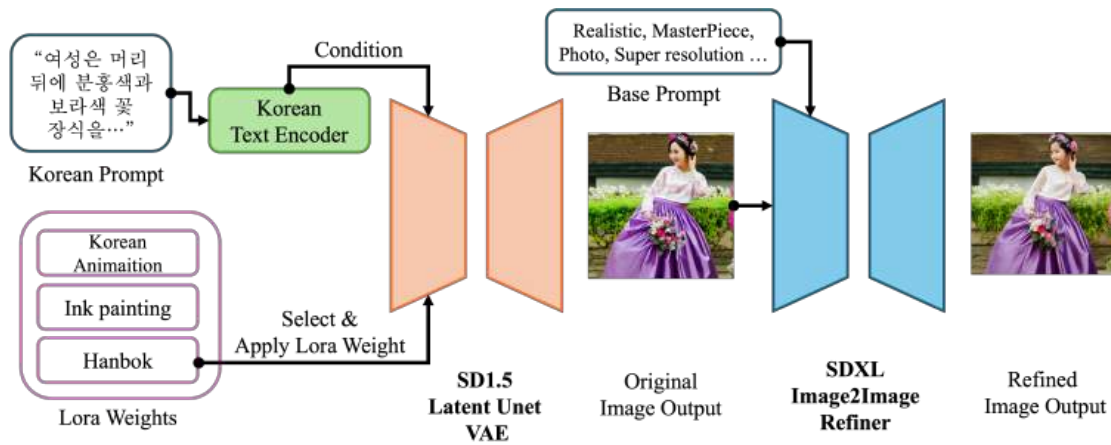


Figure 5. Overall Structure of Text2Image Pipeline

Also, when we translated Korean prompts into English and compared the images generated for the existing SD 1.5 English text encoder, we found that “people,” “seafood,” and “fish” were well represented in the English encoder and ours, which were not reflected in the baseline.

B. Experiment on K-Contents Image Generation

As shown in the bottom of Figure 6, we also conducted image generation experiments on three datasets (Ink Painting, Hanbok, and Korean Animation). First, for Ink Painting, we generated images that reflected the trigger words such as paper type, painting style, and picture type (portrait, object, landscape). Also, the overall art style was specialized in ink painting and black and white painting, so we were able to create various types of ink painting images. Secondly, for Hanbok, image generation was done by reflecting trigger words such as “hanbok” and “woman” for short Korean prompts. It also reflected the color changes of hanbok such as “purple” and “red” well. However, when we entered a very long prompt, we found that some parts of the prompt were not reflected. Third, in the case of Korean Animation, the image was generated by reflecting the trigger words such as “cartoon,” “character,” and “bug” well, and the shape of the character was slightly different compared to the training data, which can be seen to meet our intention of “generating K-Contents images”.

VI. LIMITATIONS

But, this research has some limitations. First, the Korean text encoder we trained does not understand sentences as well as the English text encoder. Although the loss value kept decreasing during training, the validation loss did not go down to less than 0.32. This indicates that there is a “semantic gap” between the vectors encoded by the English text encoder and the Korean text encoder. Second, the number of Lora weights is insufficient. We have trained three Lora weights, but we expect that more Lora weights are needed to implement the “K-Content image generation pipeline”.

VII. CONCLUSIONS AND FUTURE WORK

In this study, we used Korean-English corpus dataset and K-Contents image dataset for preprocessing, trained the Korean Text Encoder with the corpus dataset, and learned using Lora method to build a Korean-specific image generation pipeline with the K-Contents image dataset. In addition, we added SDXL’s Refiner to solve the problem of SD 1.5, which is that “face”, “hand”, etc. are generated abnormally. As an experiment, we compared the images generated using the existing Korean encoder as a baseline and the Korean encoder we trained, and analyzed the images generated by the Lora-trained pipeline for three datasets (Ink Painting, Hanbok, and Korean Animation). As a follow-up research, we would like to build a complete and well-made “K-Content Image Generation Pipeline” by learning more Lora weights to compensate for the limitations.

ACKNOWLEDGMENT

This paper was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number 2020R1A6A1A03040583) and the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-00436954) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

We sincerely thank LOTTE INNOVATE for their generous support and equipment provision, which greatly contributed to this research. Thank you.



Experiment on K-Contents Specialized Text2Image Generation



Figure 6. 1. Image generation experiment results for Text Encoder with fixed seed. This is an image generation comparison experiment for Baseline (bingsu/koclip), Ours (Korean Text Encoder), and English text encoder in combination with Korean prompt. 2. Images generated by Lora training on three datasets with Korean prompts as input. In the prompt to the right of the image, the red text is what is not reflected in the image, and the bold text is what is reflected well in the image.

REFERENCES

- [1] PODELL, Dustin, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint, 2023, arXiv:2307.01952.
- [2] RADFORD, Alec, et al. "Learning transferable visual models from natural language supervision. In: International conference on machine learning." PMLR, 2021. p. 8748-8763.
- [3] N. Reimers, I. Gurevych "Making monolingual sentence embeddings multilingual using knowledge distillation." arXiv preprint, 2020, arXiv:2004.09813.
- [4] ROMBACH, Robin, et al. "High-resolution image synthesis with latent diffusion models." in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 10684-10695.
- [5] (2022) Huggingface Model, "stabilityai/stable-diffusion-2" [Online]. Available: <https://huggingface.co/stabilityai/stable-diffusion-2>
- [6] (2019) AI hub Dataset, "Korean-English Translation (Parallel) Corpus" [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=126>
- [7] (2020) AI hub Dataset, "Korean-English Translation (Technical Science) Corpus" [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=124>
- [8] (2020) AI hub Dataset, "Korean-English Translation (Social Science) Corpus" [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=125>
- [9] (2022) AI hub Dataset, "Korean traditional ink painting production data by style" [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71380>
- [10] (2021) AI hub Dataset, "Character Face Landmark Data in Animation" [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=570>
- [11] (2022) Huggingface Model, "Bingsu/clip-vit-large-patch14-ko" [Online]. Available: <https://huggingface.co/Bingsu/clip-vit-large-patch14-ko>
- [12] (2021) Huggingface Model, "openai/clip-vit-base-patch32" [Online]. Available: <https://huggingface.co/openai/clip-vit-base-patch32/>
- [13] HU, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint, 2021, arXiv:2106.09685.
- [14] J. Li, D. Li, C. Xion and S. Hoi "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." in International conference on machine learning. PMLR, 2022, p. 12888-12900.
- [15] (2021) Github Repository, "jaketae/koclip" [Online]. Available: <https://github.com/jaketae/koclip>
- [16] LOSHCHILOV, Ilya HUTTER, Frank. "Decoupled weight decay regularization." arXiv preprint, 2017, arXiv:1711.05101.
- [17] (2022) Huggingface Model, "runwayml/stable-diffusion-v1-5" [Online]. Available: <https://huggingface.co/runwayml/stable-diffusion-v1-5>
- [18] TURIAN, Joseph; RATINOV, Lev; BENGIO, Yoshua. "Word representations: a simple and general method for semi-supervised learning." in Proceedings of the 48th annual meeting of the association for computational linguistics. 2010. p. 384-394.



Dahyun Lee is expected to receive a bachelor's degree in Computer Engineering from Kyonggi University, in 2026. She is Undergraduate student in the Kyonggi University, Suwon, Korea. Her research interests include Computer Vision.



Sangbum Han is expected to receive a bachelor's degree in Computer Engineering from Kyonggi University, in 2024. He is Undergraduate student in the Kyonggi University, Suwon, Korea. His interests include ML Engineering, Optimization



Seokjae Lee received the B.S. degree in the Department of Electronics Engineering and the M.S. degree in the Department of Electronic, Information, and Communication Engineering, Konkuk University, Seoul, South Korea in 2019 and 2021, respectively. His research interests include image enhancement, computer vision, deep learning, and biometrics



Namgi Kim received the B.S. degree in Computer Engineering from Sogang University, Seoul, Korea, in 1997 and received the M.S. and Ph.D. degrees in Computer Engineering from Korea Advanced Institute of Science and Technology, Daejeon Metropolitan City, in 2005. Until 2007, he was a Senior researcher at Samsung Electronics Telecommunication Research Institute. Since 2009, he has been a Professor with the Computer Engineering Department, Kyonggi University, Suwon. His research interests include computer network, anomaly detection, optimization of recommendation system and multicast routing tree.



Seungeun Kang is expected to receive a bachelor's degree in Computer Engineering from Kyonggi University, in 2026. He is Undergraduate student in the Kyonggi University, Suwon, Korea. His research interests include Edge Intelligence, AI with Internet of Things (AIoT).