TIP-Sen: Text-to-Image Generation Pipeline for Korean Seniors with Senior-specific Adaptation and Prompt Assistant

Wooseok Song, Ah Reum Oh, Joongyong Choi, Hyunjoo Kim

Electronics and Telecommunications Research Institute (ETRI)
Email: {wooseok.song, aro1116, choijy725, hjookim}@etri.re.kr

***Abstract*—Text-to-image models have demonstrated strong capabilities in generating high-quality 2D images from text prompts. However, existing models often struggle to accurately generate images based on senior keywords, such as traditional foods or architectures from past decades. These limitations stem from a lack of domain-specific data and the difficulty seniors face in crafting effective prompts. We introduce TIP-Sen, an image generation pipeline tailored for Korean seniors. TIP-Sen addresses these challenges by first extending a Korean senior-focused text-image dataset. It then incorporates senior-related knowledge into a diffusion model through parameter-efficient fine-tuning. Finally, our Large Language Model (LLM) based prompting assistant refines simple senior input prompts into more detailed prompts. Our experiments demonstrate that TIP-Sen effectively generates high-fidelity images that faithfully reflect senior-specific keywords, highlighting its potential not only to improve inclusiveness in generative models but also to empower seniors to participate in creative content authoring through text-based image generation actively.**

***Index Terms*—Generative model, Korean senior dataset, Personalized model, Diffusion model, Large Language Model, Image generation.**

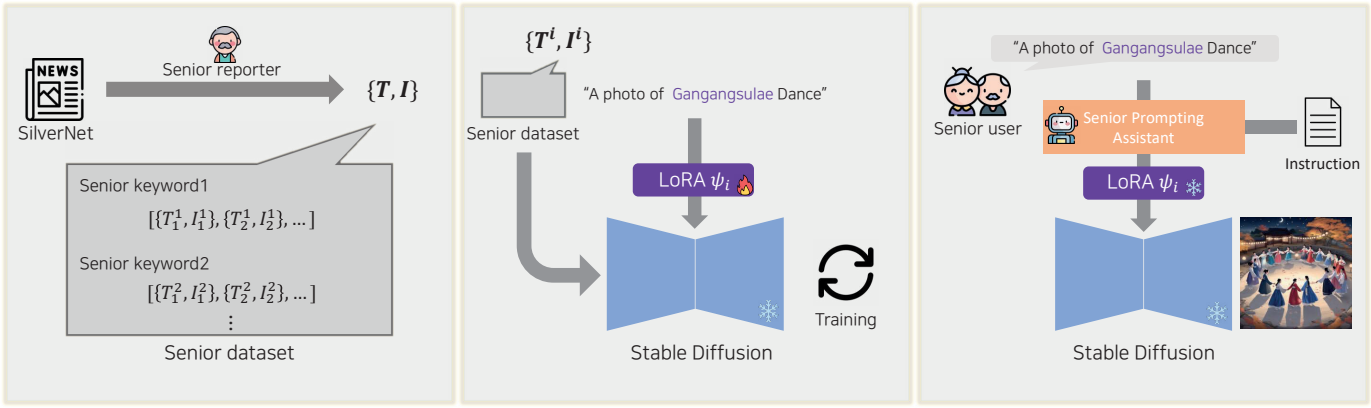
Recent advancements in text-to-image generation methods [1]–[6] have significantly enhanced the ability to create high-quality 2D images from a single text prompt. These advancements have enabled not only specialists but also common users to visualize their imaginations without requiring specialized knowledge. However, current text-to-image generation models [1]–[6] struggle to accurately generate images based on prompts related to Korean senior-specific keywords, such as traditional food or unique architecture from past decades. Seniors often wish to recreate images that reflect their memories using text-based prompts, yet existing models frequently fail to generate accurate depictions, as illustrated in Figure 1. This limitation arises from two key challenges: 1) the lack of senior-specific knowledge in current text-to-image models, stemming from the scarcity of senior-specific data, and 2) the difficulty common seniors face in crafting effective prompts.

To address this issue, we propose TIP-Sen, an image generation pipeline for Korean seniors. TIP-Sen enables high-quality image generation that accurately reflects Korean senior-related keywords. Our method effectively addresses two key challenges with three main stages: 1) Korean Senior Dataset Collection, 2) Korean Senior Keyword Diffusion Finetuning, and 3) Senior Prompting Assistant. First, we extend a text-image pair dataset focused on Korean senior-related content. Second, we use this dataset to finetune a text-to-image diffusion model with parametric-efficient fine-tuning methods to effectively integrate senior-related knowledge. Finally, we introduce a Large Language Model (LLM) [7]–[9]-based senior prompting assistant to help seniors craft more effective prompts, thereby improving the fidelity of generated images. Through extensive experiments, TIP-Sen demonstrates its ability to generate high-fidelity images that faithfully reflect senior-specific keywords. Moreover, our TIP-Sen can empower seniors to actively engage in creative content authoring through our senior-friendly image generation pipeline.

Fig. 1. **Failure cases in generating Korean senior keywords.** (a) Example of Korean senior keywords and corresponding images: "Yakgwa" refers to traditional Korean food, and "Cheomseongdae" is an ancient Korean astronomical observatory. (b) Current text-to-image diffusion model (e.g. Stable Diffusion v1.4, v2.1, and XL) fails to generate images that accurately reflect these senior-specific keywords.

Our contributions are as follows:

- We introduce TIP-Sen, the first image generation pipeline tailored for Korean seniors, addressing the underrepresentation of senior-related content in existing text-to-image models.
- We extend a Korean Senior Dataset and introduce Ko-



(a) Korean Senior Dataset Collection (b) Korean Senior Diffusion Finetuning (c) Senior Prompting Assistant

Fig. 2. **The overall pipeline of TIP-Sen.** (a) A Korean senior dataset is collected from SilverNet news by senior reporters. (b) The diffusion model is finetuned with LoRA layers to incorporate senior-specific knowledge. (c) The Senior prompting assistant takes simple input text from seniors and refines it into detailed prompts, which are then used by the diffusion model to generate high-quality images.

rean Senior Keyword Diffusion Finetuning, a parameter-efficient approach to effectively inject senior-related knowledge into a diffusion-based generative model.

- We propose a Senior Prompting Assistant based on LLM, which helps seniors craft effective prompts and improves the relevance and fidelity of the generated images.

I. Related Work

A. Text-to-Image Diffusion Model

Text-to-image diffusion models aim to generate images from a single text prompt through a denoising process. Diffusion models [2], [3] introduce a forward and reverse noise process to model data distributions for image generation. Classifier guidance [1] enables conditional generation by incorporating an external classifier to control the sampling process. In contrast, classifier-free guidance [10] allows conditional generation without additional classifiers, by interpolating conditional and unconditional score predictions. Latent Diffusion Model (LDM) [4] improves efficiency and scalability by operating in a compressed latent space using a VAE [11] encoder and an unet-based diffusion model. LDM also incorporates cross-attention layers to support various conditioning modalities, such as text prompts, segmentation maps, or other structured inputs. Based on LDM, many open-sourced diffusion models (e.g. Stable Diffusion) have been developed and widely adopted for high-resolution text-to-image generation tasks. However, existing text-to-image diffusion models [4], [5] still struggle to generate images that accurately reflect Korean senior-related content and concepts. In this work, we adopt SDXL [5] as the baseline model, as it produces higher-resolution and higher-quality images compared to SD v1.4 and SD v2.1.

B. Diffusion Personalization

Diffusion personalization involves capturing unique, user-defined keywords or concepts, enabling text-to-image diffusion models to generate personalized images that reflect those specific inputs. Recent research has focused on two primary approaches: finetuning diffusion model [12]–[14] and optimizing text embedding [15]. Finetuning-based methods [12]–[14] finetunes diffusion model to reconstruct an image of the unique keyword. DreamBooth [12] selects null text token and finetunes diffusion model to reconstruct user-defined images using diffusion loss. CustomDiffusion [13] introduces an efficient finetuning strategy by solely updating cross-attention layers in the diffusion model. On the other hand, optimization-based methods [15] optimize text embedding to reconstruct images of unique keywords. Textual Inversion [15] introduces text embedding optimization to represent the target concept using diffusion loss. Recent approaches leverage Low-Rank Adaptation (LoRA) [16], a matrix decomposition method for parametric efficient finetuning. These LoRA layers allow the diffusion model to adapt to new concepts by fine-tuning only the additional LoRA parameters.

II. Method

The overall pipeline of our method is illustrated in Figure 2. Our approach consists of three main stages: 1) Korean Senior Dataset Collection, 2) Korean Senior Keyword Diffusion Finetuning, and 3) Senior Prompting Assistant. First, we collect and extend text-image datasets specifically tailored for Korean seniors. Second, we finetune the text-to-image diffusion model using efficient parametric training techniques for each senior keyword, ensuring better integration of senior-related knowledge into the text-to-image diffusion model. Finally, we leverage our Korean Senior Prompting Assistant (SPA) to help

seniors craft more effective prompts, improving the overall fidelity of generated images.

A. Korean Senior Dataset Collection

Current text-to-image diffusion models [4], [5], [10] often fail to generate images that accurately reflect Korean senior-specific keywords. This limitation stems from the lack of senior-specific knowledge within these models, primarily due to the scarcity of dedicated datasets for senior-related content. To address this issue, we searched for datasets related to the interests of Korean seniors and identified the Korean Senior Knowledge, Experience, Interest, and Time-period Background Dataset Version 1.0 (KS-KEIT-V1) [17], which was collected from a Korean senior news platform called SilverNet. The KS-KEIT-V1 is an initial version of a Korean senior dataset that includes 30 keywords related to Korean senior interests, such as traditional culture, food, and historical figures, across multiple modalities including text, images, videos, and images. However, the dataset contains a limited number of image samples per keyword, which is insufficient for robust fine-tuning of text-to-image diffusion models. To overcome this limitation, we introduce KS-KEIT-V1.2, an upgraded version of KS-KEIT-V1 enriched with additional text-image pairs tailored for senior-specific text-to-image model training. Furthermore, since KS-KEIT-V1 was collected from wild environments and contains images of varying quality and resolution, we carefully curated 100 text-image pairs for each of the 30 keywords to ensure consistency and semantic relevance. In Figure 3, we present the structure of the dataset along with an example: "Cheomseongdae," a historic astronomical observatory.

Fig. 3. **The structure and examples from the Korean Senior Dataset**. (a) The dataset consists of 30 predefined cultural keywords, each representing distinct Korean traditions. (b) Each keyword is associated with multiple image-text pairs, where the text provides a detailed description of the corresponding image.

B. Korean Senior Keyword Diffusion Finetuning

Once the Korean senior dataset is collected, we finetune the text-to-image diffusion model. Specifically, we finetune a diffusion model using DreamBooth [12] with LoRA [16] layers to adapt each individual keyword. For the i-th senior-specific keyword, we train the corresponding LoRA layers ψ_i by minimizing the following objective:

$$\mathbb{E}_{t,\epsilon}\left[\left\|w(t)\left(\epsilon_{\theta+\psi_i}(x_i, y_i, t) - \epsilon_k\right)\right\|_2^2\right] \tag{1}$$

Here, $w(t)$ denotes a timestep-dependent weighting function, and $\epsilon_{\theta+\psi_i}(\cdot)$ represents the predicted noise by the diffusion model with the adapted parameters. x_i and y_i denote the image and text prompt corresponding to the i-th keyword, respectively. For y_i, we can either use the original prompt from the Korean senior dataset or apply a formatted template. In our case, we use the formatted template prompt: ``A photo of [*senior_keyword*] [*cls_token*].'', where [*cls_token*] is a class prompt that describes [*senior_keyword*]. After training each senior keyword-specific LoRA, the diffusion model can faithfully generate images that reflect senior-specific keywords without additional finetuning. During inference, such images can be generated by applying the corresponding LoRA layers to the diffusion model and using prompts that include both [*senior_keyword*] and [*cls_token*].

C. Senior Prompting Assistant

To generate high-quality images using a text-to-image diffusion model, input prompts need to be detailed and descriptive. However, many seniors struggle to create such prompts due to their unfamiliarity with the process. To address this, we propose a LLM [8] based Senior Prompting Assistant (SPA), which refines simple input prompts from seniors into more detailed and descriptive prompts. Specifically, our assistant takes a simple input text prompt and outputs refined prompts followed by well-crafted instructions. The instruction consists of two main components: 1) task description, and 2) in-context examples, as illustrated in Table I. The task description provides the language model with explicit guidance on the format of the expected output. In particular, the task description includes detailed elements such as people, clothing, actions, environments, lighting, and atmosphere to enrich the generation. Then in-context examples serve to refine the input prompts into more structured and effective queries. These examples are constructed using simple formatted prompts (``A photo of [*senior_keyword*] [*cls_token*]'') and corresponding text descriptions extracted from our Korean senior dataset. The final refined prompt is then passed to the text-to-image diffusion model to generate high-quality, senior-specific images.

TABLE I
THE INSTRUCTIONS AND EXAMPLES USED BY THE SENIOR PROMPTING ASSISTANT.

Component	Prompting Guidelines and Examples
LLM Instruction	**[Task description]** You are a prompt engineer that creates high-quality prompts for image generation. Please write a detailed prompt for an image generation model that describes this scene in a realistic photographic style. - Include "A realistic photo of" in front of the prompt. - Include exact input description in the prompt. - Include details about the people, clothing, actions, environment, lighting, and atmosphere. - Use natural language that models can easily understand. - Make sure prompt should not exceed 30 words. **[In-context examples]** **User Prompt:** "A photo of a Sebae Tradition." **Refined Prompt:** "This image depicts a Korean traditional bowing ceremony, Sebae Tradition, with individuals in Hanbok showing respect in a temple setting." **User Prompt:** "A photo of a kimbap Food" **Refined Prompt:** "This image depicts kimbap Food, a popular Korean dish, consisting of cooked rice and various other ingredients rolled in seaweed and sliced into bite-sized pieces." ⋮
LLM Input	**User Prompt:** "A photo of Gimjang Tradition."
LLM Output	**Refined Prompt:** "A realistic photo of a group of Korean women gathered around a large, earthenware jar, preparing kimchi for Gimjang Tradition, surrounded by autumn foliage and traditional Korean houses."

III. EXPERIMENTS

A. Implementation Details

The experiments are conducted based on Stable Diffusion XL (SDXL) [5] and run on a single NVIDIA A100 GPU. For senior-specific keyword learning, we adopt DreamBooth-LoRA [12] with a rank of 20. The LoRA [16] parameters are optimized using the AdamW [18] optimizer with a learning rate of 2e-4, a batch size of 2, and 500 training steps. Each senior keyword LoRA takes approximately 20 $\sim$ 30 minutes to train. For the senior prompting assistant, we utilize LLaMA 3-Instruct [8] as the LLM, with a temperature setting of 0.8. During image generation, we generate a single image using 28 sampling steps, with a guidance scale of 5 and a LoRA scale of 0.7.

B. Evaluation Metrics

We evaluate our method using the CLIP [19] image-image alignment score to assess how faithfully the generated images reflect senior-specific keyword images. The alignment score is computed as the cosine similarity between CLIP image embeddings of the generated and a real image representing the target senior keyword. The alignment score follows:

$$\text{Image-align} = \frac{\langle \mathbf{e_{gen}}, \mathbf{e_{real}} \rangle}{|\mathbf{e_{gen}}| \cdot |\mathbf{e_{real}}|} \tag{2}$$

Here, $\mathbf{e_{gen}}$ and $\mathbf{e_{real}}$ denote the CLIP image embedding vectors of the generated and real images, respectively. The operator $\langle \cdot, \cdot \rangle$ denotes the dot product, and $|\cdot|$ represents the L2 norm.

TABLE II
QUANTITATIVE RESULTS.

Method	Image-align	
	Mean ↑	Std↓
SD v1.4	0.6406	0.1152
SD v2.1	0.6508	0.1044
SDXL	0.6347	0.1026
SDXL + LoRA (Ours)	**0.7358**	0.0783
SDXL + LoRA + SPA (Ours)	0.7138	**0.0762**

For each senior keyword, we generate 100 images and compute the alignment scores by comparing them with 100 real images. The final score is obtained by averaging the alignment scores across all image pairs.

C. Quantitative Results

In Table II, we report the CLIP [19] image alignment scores of the generated outputs. Each value represents the mean and standard deviation of alignment scores across 30 senior keywords. Our SDXL [5] with LoRA [16] achieves the highest alignment score among all baselines, indicating our method can faithfully generate images reflecting senior-specific keywords. When incorporating our Senior Prompting Assistant (SPA), the mean alignment score slightly decreases compared to SDXL [5] with LoRA [16], however, the standard deviation is the lowest among all methods. This indicates SPA contributes to more stable and consistent image generation while maintaining a comparable level of alignment with senior-related content.

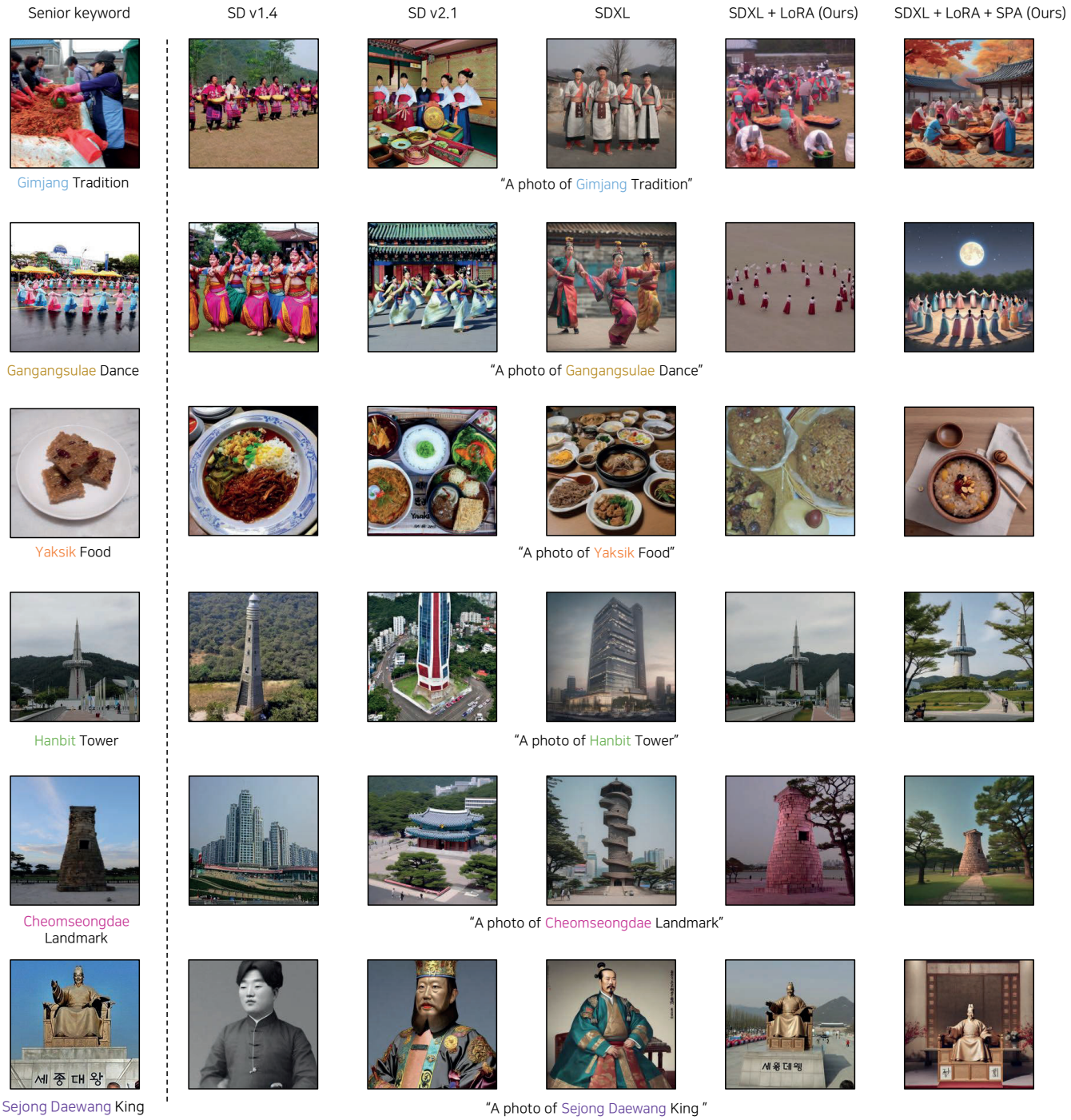


Fig. 4. **Qualitative results.** Row 1: "Gimjang Tradition" refers to the traditional Korean Gimchi-making process. Row 2: "Ganggangsullae Dance" refers to a traditional Korean group circle dance. Row 3: "Yaksik Food" is a traditional Korean sweet rice dish made with glutinous rice, nuts, and jujubes. Row 4: "Hanbit Tower" denotes a traditional architectural structure built for the Daejeon Expo. Row 5: "Cheomseongdae Landmark" refers to an ancient astronomical observatory. Row 6: "Sejong Daewang King" represents the historical figure who created the Korean script, Hangeul.

D. Qualitative Results

In Figure 4, we compare our method with several baselines across various senior-specific keywords. We evaluate the outputs of SDXL [5] with LoRA [20] and SDXL with LoRA combined with the Senior Prompting Assistant (SPA), and compare them against SD v1.4 [4], SD v2.1 [4], and SDXL [5]. Our method demonstrates a superior ability to faithfully reflect senior-specific concepts compared to the baselines. Notably, while SDXL with LoRA alone tends to overfit to the training images of senior keywords, the inclusion of SPA leads to more

enriched and diverse generations, showing better generalization and semantic alignment.

IV. Conclusion

In this work, we proposed TIP-Sen, an image generation pipeline tailored for Korean seniors, addressing two key challenges, the lack of domain-specific data and knowledge, and the difficulty seniors face in crafting effective prompts. We extended a text-image pair dataset focused on senior-related content and curated 30 representative senior-specific keywords. Using this dataset, we finetuned a text-to-image diffusion model with a parameter-efficient approach to incorporate senior-related knowledge. In addition, we introduced a Senior Prompting Assistant that refines simple input prompts into more detailed and descriptive ones using a large language model guided by well-crafted instructions. Experimental results show that TIP-Sen can faithfully generate high-quality images that reflect senior-specific concepts.

V. Future Work

Future work includes the development of a diffusion-based generative model capable of producing Korean content that reflects distinctive traditional features of Korean culture from simple prompts. The proposed pipeline can serve as a foundation for a content authoring tool, enabling seniors to easily generate items aligned with their personal interests and cultural backgrounds. This, in turn, may provide retired elderly individuals with more opportunities to engage with society and share their experiences across generations. Ultimately, such interactions are expected to promote social cohesion by fostering mutual understanding and bridging the generational gap.

Acknowledgment

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of generative AI-based content creation platform technology to improve content creation accessibility for senior, Project Number: RS-2024-00340342, Contribution Rate : 100)

References

[1] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[7] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[8] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[9] W. Feng, W. Zhu, T.-j. Fu, V. Jampani, A. Akula, X. He, S. Basu, X. E. Wang, and W. Y. Wang, "Layoutgpt: Compositional visual planning and generation with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[11] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational bayes," 2013.

[12] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[13] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1931–1941.

[14] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu *et al.*, "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[16] S. Ryu, "Low-rank adaptation for fast text-to-image diffusion fine-tuning. 2022," *URL https://github. com/cloneofsimo/lora*.

[17] A. R. Oh, J. Choi, and H. J. Kim, "Dataset structure design of korean senior data for adjusting generative ai model," in *Proceedings of the 1st International Conference on Artificial Intelligence Computing and Systems*, 2024, pp. 113–116.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[19] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.