K-StyleLoRA: Information-Guided Image Generation via Selective Feature Learning

Soyoung Lee[1*] Hyoungseo Cho[2*] Myungjoo Kang[2,3] YoungJoon Yoo[1,†]
[1]Department of Artificial Intelligence, Chung-Ang University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]Department of Mathematical Sciences and Research Institute of Mathmatics, Seoul National University
soyounglee@cau.ac.kr {hscho96, mkang}@snu.ac.kr yjyoo3312@cau.ac.kr

Abstract

*Despite remarkable advances in image generation, existing diffusion models struggle to capture diverse cultural aesthetics. While Low-Rank Adaptation (LoRA) enables efficient fine-tuning, conventional approaches lack semantic awareness. They apply uniform adaptations across all features, leading to suboptimal cultural representation. To address these limitations, we introduce **K-StyleLoRA**, a novel framework that leverages CLIP's cross-modal capabilities for cultural image generation. Our approach introduces two key contributions. First, CLIP-Guided Information Gating dynamically modulates LoRA adaptations based on cultural relevance scores, selectively enhancing relevant features while suppressing irrelevant ones. Second, Cultural Semantic Loss provides additional semantic guidance by optimizing CLIP-based similarity to cultural concepts. Extensive experiments on Korean traditional culture show superior cultural fidelity while maintaining generation quality and diversity. In particular, K-StyleLoRA demonstrates exceptional cultural transfer capability on generic prompts that require implicit cultural understanding. Our method achieves a Cultural Similarity Score of 0.274, representing a 9.6% improvement over the vanilla SDXL baseline (0.250). Our framework establishes semantic-aware adaptation as a powerful paradigm for cultural representation. This scalable approach can be extended to diverse cultural contexts and generation tasks beyond Korean aesthetics.*

1. Introduction

Image generation has undergone remarkable advancement with the emergence of large-scale diffusion models [1–5]. These models demonstrate exceptional capability in producing high-quality, diverse images from various input modalities. However, these models exhibit pronounced cultural bias, reflecting the dominant cultural contexts in their training data while struggling to authentically represent diverse cultural traditions [6–9]. This bias stems from training data imbalances and the difficulty of accurately representing diverse cultural nuances.

This limitation becomes particularly evident when examining culturally-specific concepts. Consider the prompt *"traditional clothes"* across different cultural contexts. Each culture has different interpretations, such as hanbok in Korea, lehenga in India, kimono in Japan, or ball gowns in Europe. Despite these diversity, current models often default to the most represented cultural interpretation in their training data. This reveals a fundamental gap in cultural understanding. These models require explicit cultural cues in prompts to generate culturally-appropriate content.

To address these cultural biases, various approaches have been explored. Recent attempts have focused on data augmentation and prompt engineering techniques [10–13], but these approaches face significant limitations. They either require extensive data collection or depend on carefully crafted prompts. This can limit their scalability and practical applicability. These challenges highlight the need for more fundamental approaches that can enable models to understand cultural semantics without relying on extensive external resources.

In parallel, efficient fine-tuning approaches have emerged as a promising strategy to adapt pre-trained models to specific cultural domains. Specifically, Low-Rank Adaptation (LoRA) [14–17] has shown outstanding performance, enabling customization of large diffusion models with minimal computational overhead. However, conventional LoRA approaches face two critical limitations when applied to cultural adaptation. First, they apply uniform adaptations across all features without considering semantic relevance. This becomes particularly problematic for cultural adaptation, where traditional cultural data is in-

[0]*Equal contribution †Corresponding author

herently scarce. To achieve effective learning with limited datasets, semantic-aware adaptation that selectively focuses on culturally-relevant features is crucial. Second, they lack mechanisms to understand cultural context and meaning, resulting in insufficient semantic guidance to ensure cultural authenticity. These limitations motivate the need for a more sophisticated approach to cultural adaptation.

To address these challenges, we introduce **K-StyleLoRA**, a novel framework that addresses cultural bias through semantic guidance. Our approach leverages the rich cross-modal knowledge embedded in pre-trained vision-language models like CLIP [18] for cultural understanding. Specifically, we show that selective feature enhancement based on cultural relevance achieves superior authenticity compared to uniform adaptation methods. Importantly, this selective strategy proves particularly effective in culture-specific domains where training data is inherently limited.

Our method integrates two complementary mechanisms into the LoRA adaptation process. *CLIP-Guided Information Gating* leverages CLIP's cross-modal understanding to compute cultural relevance scores for input features. This mechanism dynamically modulates LoRA adaptations to enhance culturally-relevant aspects while suppressing irrelevant ones. The selective approach enables more efficient adaptation and proves particularly effective for cultural domains where large-scale training data may be limited. Complementing this, our *Cultural Semantic Loss* provides explicit training guidance by encouraging generated images to align with cultural concepts through CLIP-based similarity optimization. This ensures the model captures meaningful cultural representations rather than superficial visual patterns.

We validate our approach through comprehensive experiments on Korean traditional image generation. This domain requires detailed understanding of aesthetic principles, color palettes, and compositional elements. Our evaluation covers five distinct categories: architecture, fashion, art, portrait, and landscape. This comprehensive assessment allows us to evaluate the model's cultural understanding across diverse visual domains.

The main contributions of this work are summarized as follows:

- We propose K-StyleLoRA, a novel framework that addresses cultural bias in diffusion models through selective feature learning guided by cultural relevance.
- We introduce CLIP-Guided Information Gating, a mechanism that dynamically modulates LoRA adaptations based on cultural relevance scores. This enables targeted parameter updates while preserving general knowledge capabilities.
- We design a Cultural Semantic Loss that leverages CLIP-based similarity optimization to ensure generated images align with cultural concepts. This approach proves particularly effective when working with limited cultural training data.

Our framework addresses an important limitation in current generative AI systems by providing an effective approach to mitigate cultural bias. We demonstrate this approach using Korean traditional culture as our evaluation domain. The methodology, however, is designed for broader applicability across diverse cultural contexts. This work contributes to the development of more inclusive generative models that better represent cultural diversity.

2. Related Work

2.1. Text-to-Image Generation and Cultural Bias

Recent years have witnessed remarkable progress in text-to-image generation, with the emergence of diffusion models [19–23]. Latent Diffusion Models (LDMs) [1] significantly improved computational efficiency by operating in the latent space of pre-trained autoencoders. Building on this work, Stable Diffusion further advanced this field with impressive text-to-image generation capabilities [24–26]. More recently, Stable Diffusion XL (SDXL) [27] introduced significant architectural improvements including a larger UNet and refined conditioning mechanisms. This achieved state-of-the-art performance in high-resolution image generation. Other notable models like DALL-E 2 [3] and Imagen [2] have also achieved high-quality image generation.

Despite these advances, recent studies have revealed significant cultural biases in AI models [6, 7, 28, 29]. Moreover, limited work has explicitly addressed cultural representation in generative models. Early approaches focused on data augmentation strategies [30, 31] or bias mitigation through careful training data curation. Other methods employed prompt engineering techniques [10] or style transfer approaches [32] to improve cultural representation. However, these methods often require extensive manual annotations and lack sufficient cross-modal understanding for effective cultural adaptation. Our work addresses this limitation by developing selective adaptation techniques that leverage CLIP's capability for cultural adaptation.

2.2. Parameter-Efficient Fine-tuning (PEFT)

Fine-tuning large-scale generative models requires substantial computational resources, motivating research into parameter-efficient adaptation methods [33]. Among these approaches, Low-Rank Adaptation (LoRA) [14] emerged as a particularly effective method that decomposes weight updates into low-rank matrices. These decomposed updates can then be efficiently merged with pre-trained parameters. Building on this foundation, AdaLoRA [34] extends this approach by adaptively allocating ranks across different lay-

ers based on importance scores. Recent work has further explored block-wise adaptation strategies within the LoRA framework. B-LoRA [35, 36] introduces selective application of LoRA to specific transformer blocks, while Frenkel et al. [37] employ B-LoRA for style-content separation in diffusion models.

Several works have applied LoRA for domain adaptation of diffusion models [38, 39]. However, these approaches focus on general artistic styles rather than culture-specific adaptations. Additionally, they treat all features uniformly without considering semantic relevance to the target domain. Recognizing these issues, recent work has shown that selective adaptation approaches are particularly effective for few-shot learning scenarios with limited training data. Our K-StyleLoRA addresses these limitations through CLIP-guided selective feature learning specifically designed for cultural adaptation. This approach demonstrates superior performance even with limited cultural training data.

2.3. Vision-Language Models for Guidance

Vision-language models, particularly CLIP [18], have demonstrated strong capabilities in understanding semantic relationships across modalities. Early works leveraged CLIP for guiding image generation and editing during inference. For example, CLIP-Guided Diffusion [40] uses CLIP to guide the generation process toward desired text prompts. Similarly, CLIPStyler [41] employs CLIP for artistic style transfer by optimizing images to match style descriptions.

Recent approaches have explored CLIP for guidance in various contexts. Notably, InstructPix2Pix [42] uses CLIP features to guide instruction-based image editing, while other works explore CLIP-based loss functions for improved text-image alignment during training. However, these approaches do not specifically address cultural adaptation or employ CLIP for selective feature learning in parameter-efficient fine-tuning.

3. Method

We present K-StyleLoRA, a novel framework that integrates CLIP's cross-modal understanding with Low-Rank Adaptation for cultural image generation. As illustrated in Figure 1, our approach addresses the limitation of conventional LoRA methods that apply uniform adaptations without considering semantic relevance. K-StyleLoRA achieves this through two main components: CLIP-guided information gating for selective feature modulation and cultural semantic loss for semantic consistency.

3.1. CLIP-Guided Information Gating

We employ a shared CLIP ViT-B/32 model to maintain consistent cultural understanding. Target cultural concepts are pre-encoded and normalized:

$$T_{\text{cultural}} = \{\mathcal{F}_{\text{norm}}(\text{CLIP}_{\text{text}}(c_k))\}_{k=1}^{K} \in \mathbb{R}^{K\times 512}. \quad (1)$$

For input features $h \in \mathbb{R}^{B\times S\times D}$ from UNet attention layers, we compute spatial attention weights and apply multi-scale feature aggregation:

$$w_{\text{spatial}} = \sigma(\text{MLP}_{\text{attn}}(h)) \in \mathbb{R}^{B\times S\times 1}, \quad (2)$$

$$h_{\text{att}} = \frac{\sum_{s=1}^{S} w_s \odot h_s}{\sum_{s=1}^{S} w_s + \epsilon}, \quad (3)$$

$$h_{\text{combined}} = \lambda h_{\text{att}} + (1-\lambda)\max_s h_s, \quad (4)$$

where w_s, h_s denote the s-th spatial element, and $\lambda = 0.7$.

Cultural relevance scores are computed by:

$$v = \mathcal{F}_{\text{norm}}(\phi_{\text{proj}}(h_{\text{combined}})) \in \mathbb{R}^{B\times 512}, \quad (5)$$

$$s = vT_{\text{cultural}}^{T} \in \mathbb{R}^{B\times K}, \quad (6)$$

where ϕ_{proj} is a projection layer.

3.2. Hierarchical Cultural Modulation

The hierarchical modulation system generates three types of weights based on cultural relevance:

$$w_{\text{down}} = \sigma(\phi_{\text{down}}(s)) \in \mathbb{R}^{B\times r}, \quad (7)$$

$$w_{\text{up}} = \sigma(\phi_{\text{up}}(s)) \in \mathbb{R}^{B\times r}, \quad (8)$$

$$\alpha_{\text{intensity}} = \sigma(\phi_{\text{gate}}(s)) \in \mathbb{R}^{B\times 1}, \quad (9)$$

where ϕ_{down}, ϕ_{up}, and ϕ_{gate} are lightweight MLPs.

The culturally-guided LoRA adaptation is applied as:

$$\Delta h = \alpha_{\text{intensity}} \cdot \lambda_{\text{scale}} \cdot \mathbf{W}_{\text{up}} \left((\mathbf{W}_{\text{down}} h) \odot w_{\text{down}} \odot w_{\text{up}} \right), \quad (10)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D\times r}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r\times D}$ are standard LoRA matrices, and λ_{scale} combines network scaling and guidance strength.

3.3. Cultural Semantic Loss

To maintain cultural consistency during training, we introduce a Cultural Semantic Loss. Given predicted latents $z_{\text{pred}} \in \mathbb{R}^{B\times C\times H\times W}$ from the diffusion process, we first downsample and apply a learnable projection:

$$z_{\text{down}} = \mathcal{D}(z_{\text{pred}}) \in \mathbb{R}^{B\times C\times H'\times W'}, \quad (11)$$

$$z_{\text{flat}} = \text{flatten}(z_{\text{down}}) \in \mathbb{R}^{B\times D_{\text{latent}}}, \quad (12)$$

$$f_{\text{proj}} = \mathcal{F}_{\text{norm}}(\phi_{\text{cultural}}(z_{\text{flat}})) \in \mathbb{R}^{B\times 512}. \quad (13)$$

The projected features are compared against pre-encoded cultural concept embeddings:

$$s_{\text{cultural}} = f_{\text{proj}} T_{\text{cultural}}^{T} \in \mathbb{R}^{B\times K}, \quad (14)$$

$$w_{\text{concepts}} = \text{softmax}(s_{\text{cultural}}) \in \mathbb{R}^{B\times K}, \quad (15)$$

$$\mathcal{L}_{\text{cultural}} = -\mathbb{E}_i \left[\sum_{k=1}^{K} s_k^{(i)} w_k^{(i)} \right]. \quad (16)$$

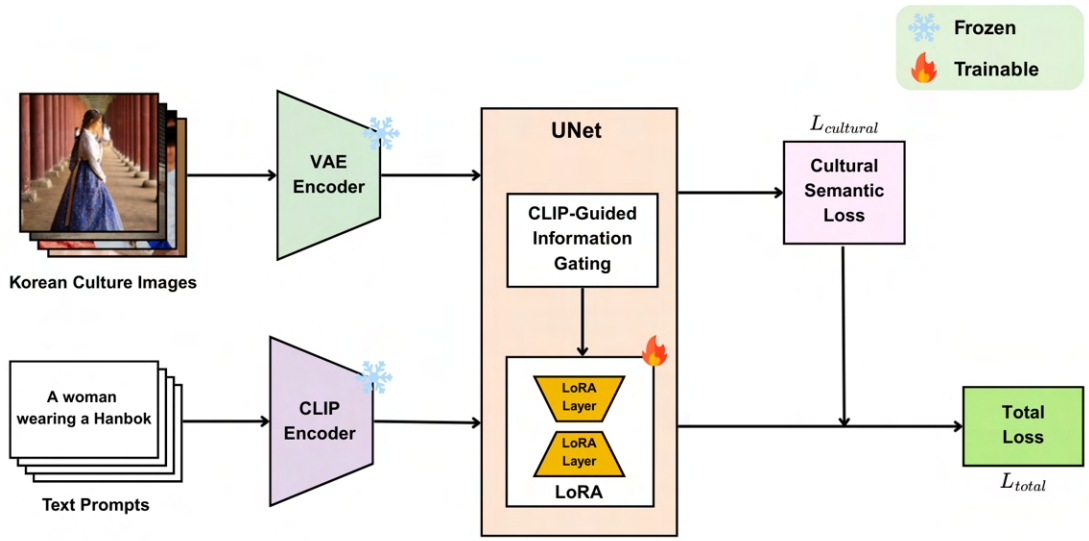

Figure 1. Overview of K-StyleLoRA framework. Our method integrates CLIP-guided information gating with cultural semantic loss for cultural adaptation. (a) CLIP-Guided Information Gating dynamically modulates LoRA adaptations based on cultural relevance scores. (b) Cultural Semantic Loss provides semantic guidance by optimizing CLIP-based similarity between generated images and targeted cultural concepts. (c) The combined approach enables cultural adaptation for both explicit cultural keywords and standard text prompts without cultural markers.

This approach enables efficient cultural semantic guidance while reducing memory consumption compared to pixel-space alternatives.

3.4. Training Objective

Our complete training objective combines standard diffusion loss with cultural semantic guidance:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{cultural}}, \tag{17}$$

where $\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,\epsilon}\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2$.

This approach enables K-StyleLoRA to achieve parameter-efficient cultural adaptation while maintaining high generation quality. By combining selective feature modulation with semantic guidance, our method provides an effective solution for culturally-aware image generation.

4. Experiments

We evaluate K-StyleLoRA through comprehensive experiments on culturally-aware image generation. Specifically, we focus on Korean traditional image generation to demonstrate effective cultural adaptation.

4.1. Experimental Setup

4.1.1. Dataset

We collect a dataset of Korean traditional cultural images consisting of 128 high-quality images from copyright-free online sources. The training data consists of Korean cultural images from publicly available repositories and cultural archives.

4.1.2. Implementation Details

We implement K-StyleLoRA on top of Stable Diffusion XL using PyTorch and the Diffusers library. Training is performed on a single GPU with a batch size of 1 and gradient accumulation of 4 steps. The image resolution is set to 1024×1024 pixels to match SDXL's native input resolution.

We use the ViT-B/32 architecture as our shared CLIP model. The LoRA rank is set to $r = 4$ across all experiments, while the cultural loss weight λ is set to 0.1. Training is performed for 10 epochs using the AdamW optimizer with a learning rate of 1×10^{-4}.

4.1.3. Baseline Methods

We compare K-StyleLoRA against several strong baseline methods:

Vanilla SDXL: The original Stable Diffusion XL model without any cultural adaptation, serving as the baseline for

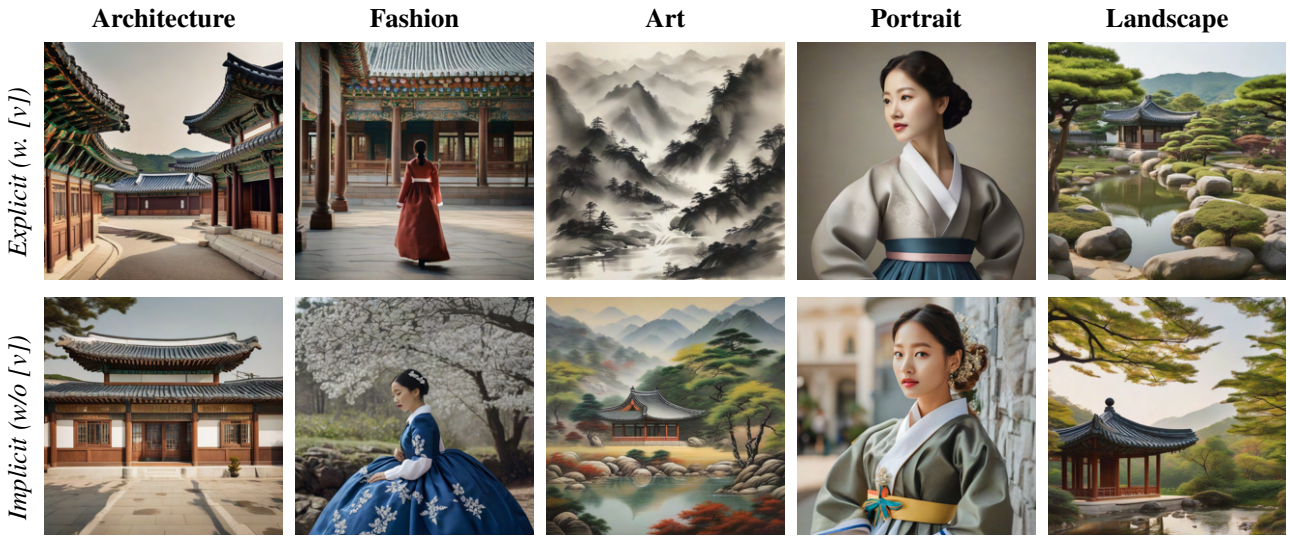

Figure 2. K-StyleLoRA qualitative results organized by Korean cultural categories. Each column represents a different cultural domain. Top row shows explicit cultural activation using [v] token representing "Korean traditional" (e.g., "a woman wearing [v] dress"), while bottom row demonstrates implicit cultural transfer for equivalent prompts without tokens (e.g., "a woman wearing a traditional dress"). Our method successfully applies Korean aesthetics in both scenarios while maintaining high visual quality and prompt adherence.

cultural representation capability.

Standard LoRA: Conventional LoRA fine-tuning without CLIP guidance or cultural semantic loss.

LoRA + Cultural Loss: Standard LoRA with cultural semantic loss ($\lambda = 0.1$) but without CLIP-guided information gating.

All baseline methods are trained on the same dataset and identical experimental settings to ensure fair comparison.

4.2. Quantitative Results

4.2.1. Evaluation Metrics

We use quantitative metrics to objectively evaluate generation quality and cultural authenticity:

Cultural Similarity Score (CSS): For each generated image, we compute the average CLIP similarity between generated images and Korean cultural concept embeddings.

$$\text{CSS} = \frac{1}{N}\sum_{i=1}^{N}\max_{k=1}^{K}\text{sim}(\text{CLIP}(I_i), T_k) \tag{18}$$

where I_i are generated images, T_k are Korean cultural concept embeddings, and N is the number of generated images.

CLIP Score: We measure CLIP similarity between generated images and prompts to evaluate text-image alignment.

LPIPS: we calculate average LPIPS distance between image pairs to measure generation diversity.

4.2.2. Main Results

Table 1 presents the quantitative comparison of K-StyleLoRA against baseline methods. We evaluate cultural alignment using our Cultural Similarity Score (CSS) across different prompt categories: all prompts (CSS-All), explicit cultural prompts (CSS-Exp), and generic prompts (CSS-Gen). K-StyleLoRA demonstrates consistent improvements across all cultural similarity metrics. Notably, our method achieves a 9.6% improvement over vanilla SDXL on generic prompts, showing effective implicit cultural transfer.

Method	CSS(All) ↑	CSS(Exp) ↑	CSS(Gen) ↑	CLIP ↑
Vanilla SDXL	0.291±0.04	0.299±0.02	0.250±0.04	0.324±0.02
Standard LoRA	0.290±0.04	0.295±0.03	0.255±0.05	0.323±0.02
LoRA + CSL	0.293±0.04	0.302±0.03	0.261±0.04	0.327±0.02
K-StyleLoRA	**0.298±0.04**	**0.309±0.03**	**0.274±0.03**	**0.327±0.02**

Table 1. Quantitative evaluation results on Korean traditional image generation. CSS denotes Cultural Similarity Score and CSL denotes Cultural Semantic Loss.

4.3. Ablation Studies

Method	CSS(All) ↑	CSS(Gen) ↑	CLIP Score↑	LPIPS ↑
Standard LoRA	0.290±0.04	0.255±0.05	0.323±0.02	0.725±0.06
+ CLIP Loss Only	0.293±0.04	0.261±0.04	0.327±0.02	0.755±0.07
+ CLIP Gating Only	0.293±0.04	0.262±0.04	0.325±0.02	0.724±0.06
K-StyleLoRA ($\sigma = 0.5$)	0.289±0.04	0.250±0.03	0.324±0.02	0.761±0.08
K-StyleLoRA ($\sigma = 1.0$)	0.292±0.04	0.257±0.04	**0.330±0.02**	0.741±0.06
K-StyleLoRA ($\sigma = 1.5$)	**0.298±0.04**	**0.274±0.03**	0.327±0.02	0.749±0.06

Table 2. Ablation Study on K-StyleLoRA Components and Guidance Strength.

We conduct ablation studies examining our two key con-

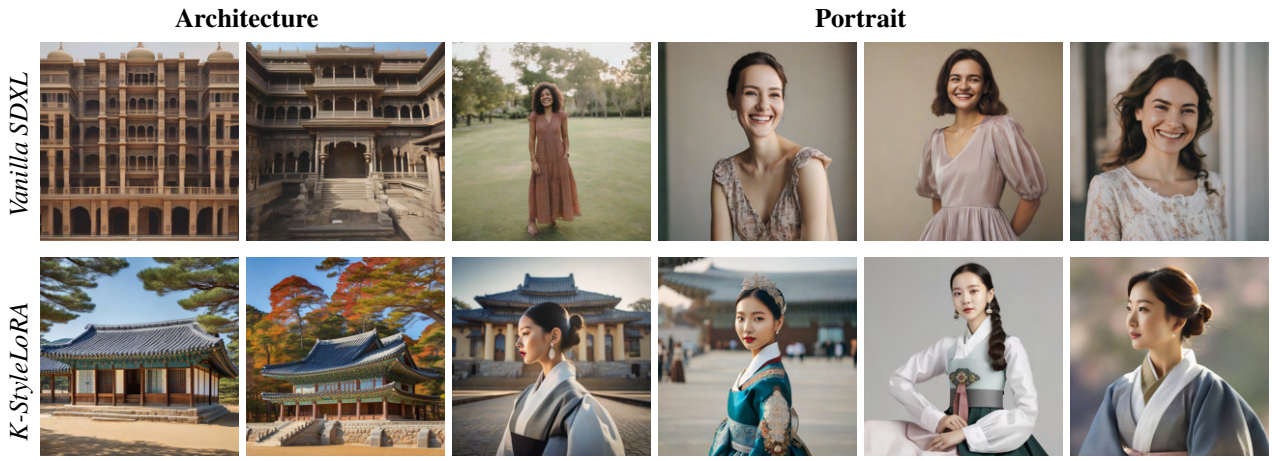

Figure 3. Qualitative comparison between Vanilla SDXL (top) and K-StyleLoRA (bottom). This comparison demonstrates K-StyleLoRA's implicit cultural adaptation capabilities. Architecture examples use prompts like "traditional building" and "traditional architecture." Portrait examples use prompts such as "a woman wearing a dress" and "a woman in silk dress." K-StyleLoRA naturally applies Korean cultural elements without explicit cultural keywords.

tributions, CLIP-guided information gating and cultural semantic loss. We also analyze the effect of different guidance strength parameters (σ). The ablation study reveals that both components contribute effectively to cultural representation. The guidance strength analysis shows that stronger guidance achieved stronger cultural similarity scores.

4.4. Qualitative Results

Figure 2 presents a comprehensive showcase of K-StyleLoRA's generation capabilities across different prompt categories. K-StyleLoRA maintains cultural authenticity while producing high-quality outputs. Figure 3 compares vanilla SDXL and K-StyleLoRA to show how our method applies Korean cultural aesthetics to various prompt types.

Cultural Authenticity: K-StyleLoRA consistently generates culturally authentic Korean images across all prompt categories. These images accurately reflect traditional Korean aesthetics.

Implicit Cultural Transfer: The method demonstrates strong capabilities even when cultural elements are not explicitly mentioned in the prompt. K-StyleLoRA incorporates Korean aesthetic elements while preserving strong text-image alignment.

4.5. Limitations

While K-StyleLoRA significantly improves cultural representation, several limitations remain. First, our method is inherently dependent on CLIP's performance and cultural understanding capabilities. If CLIP has biases or limitations in recognizing certain cultural elements, these may affect our generation system's performance. Future work could explore developing culture-aware vision-language models to address CLIP's cultural limitations. Additionally, developing more comprehensive cultural evaluation metrics to assess cross-cultural generation quality. Second, the method requires careful curation of cultural concept embeddings, which may not capture all nuances of a cultural domain. Very abstract or conceptual cultural elements may still be challenging to represent accurately. Adaptive cultural embedding techniques could address this limitation. Third, our experiments are conducted on a limited dataset and specific experimental settings. More comprehensive evaluation across diverse cultural contexts and experimental configurations could better validate the generalizability and robustness of our approach.

5. Conclusion

In this paper, we presented K-StyleLoRA, a novel framework for culturally-aware image generation that addresses cultural bias in existing diffusion models. Our approach introduces two key innovations for cultural adaptation. CLIP-Guided Information Gating enables selective feature learning based on cultural relevance, while Cultural Semantic Loss provides global semantic guidance through CLIP-based similarity optimization. Extensive experiments on Korean cultural image generation show that K-StyleLoRA significantly outperforms vanilla SDXL in cultural similarity while maintaining text-image alignment and generation diversity. Our framework demonstrates effective semantic-aware adaptation for cultural representation. This approach can be extended to other cultural domains and generation tasks. Overall, this work contributes to developing more culturally-aware and parameter-efficient generative models.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)]. This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] This work is supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2024-00421203, No. RS-2024-00406127, RS-2021-NR059802)

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 1, 2

[2] 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. 36479–36494, 2022. 2

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022. 2

[4] P. Pernias, D. Rampas, M. L. Richter, C. J. Pal, and M. Aubreville, "Würstchen: An efficient architecture for large-scale text-to-image diffusion models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[5] H. Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng, and L. Zhang, "Comprehensive exploration of diffusion models in image generation: a survey," *Artificial Intelligence Review*, vol. 58, no. 99, 2025. 1

[6] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023. 1, 2

[7] J. Cho, A. Zala, and M. Bansal, "Dall-e 2 is seeing double: Flaws in word-to-image generation models," *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1–14, 2023. 2

[8] Y. Wan, A. Subramonian, A. Ovalle, Z. Lin, A. Suvarna, C. Chance, H. Bansal, R. Pattichis, and K.-W. Chang, "Survey of bias in text-to-image generation: Definition, evaluation, and mitigation," *arXiv preprint arXiv:2404.01030*, 2024.

[9] Z. Bayramli, A. Suleymanzade, N. M. An, H. Ahmad, E. Kim, J. Park, J. Thorne, and A. Oh, "Diffusion models through a global lens: Are they culturally inclusive?," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. 1

[10] J. Liu, C. Wang, S. Joty, C. Xiong, and R. Socher, "Cultural prompting: Activating cultural knowledge in language models for bias reduction," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 4556–4571, 2023. 1, 2

[11] R. Skilton and A. Cardinal, "Inclusive prompt engineering: A methodology for hacking biased ai image generation," in *The 42nd ACM International Conference on Design of Communication*, 2024.

[12] P. W. Shin, J. J. Ahn, W. Yin, J. Sampson, and V. Narayanan, "Can prompt modifiers control bias? a comparative analysis of text-to-image generative models," *arXiv preprint arXiv:2406.05602*, 2024.

[13] S. Jeong, I. Choi, Y. Yun, and J. Kim, "Culture-trip: Culturally-aware text-to-image generation with iterative prompt refinement," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. 1

[14] 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," *International Conference on Learning Representations*, 2022. 1, 2

[15] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models," 2023.

[16] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *International Conference on Learning Representations (ICLR)*, 2023.

[17] S. Wang, L. Jiang, L. Yang, F. Fang, S. Li, B. Qin, R. Zhang, and Z. Lin, "Lora-ga: Low-rank adaptation with gradient approximation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*, pp. 8748–8763, 2021. 2, 3

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. 2

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *International Conference on Learning Representations*, 2021.

[21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 16784–16804, 2022.

[22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in*

Neural Information Processing Systems, vol. 34, pp. 9694–9705, 2021.

[23] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021. 2

[24] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. 2

[25] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, "Controlnet++: Improving conditional controls with efficient training," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[26] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 4296–4304, 2024. 2

[27] 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," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. 2

[28] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in nlp," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020. 2

[29] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021. 2

[30] J. An *et al.*, "Adaptive augmentation for effectively mitigating dataset bias," in *Asian Conference on Computer Vision (ACCV)*, 2022. 2

[31] S. Sharma *et al.*, "Data augmentation for discrimination prevention and bias disambiguation," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020. 2

[32] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. 2

[33] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, 2023. 2

[34] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," *International Conference on Learning Representations*, 2023. 2

[35] Y. Wang, Z. Chen, Q. Liu, M. Zhang, and J. Li, "B-lora: Block-wise low-rank adaptation for efficient fine-tuning of large language models," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5847–5859, 2023. 3

[36] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Block-wise lora: Switchable and efficient fine-tuning for large foundation models," *International Conference on Machine Learning*, pp. 21540–21558, 2023. 3

[37] Y. Frenkel, Y. Vinker, A. Shamir, and D. Cohen-Or, "Implicit style-content separation using b-lora," in *ECCV*, pp. 181–198, Springer, 2024. 3

[38] 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," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. 3

[39] 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," *European Conference on Computer Vision*, pp. 208–224, 2022. 3

[40] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021. 3

[41] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17292–17301, 2022. 3

[42] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023. 3