Journal of Computer Games and Contents Vol. 38, No. 4, Aug, 2025, pp.1-1
ISSN 3091-7409(Print) ISSN 2288-5528(Online) Copyright ⓒ 2016 KSCG

Style–Preserving Prompt Optimization for Game Item Images via Image–Based Prompt Extraction and Genetic Algorithms

이미지 기반 프롬프트 추출과 유전 알고리즘을 이용한 게임 아이템 이미지 스타일 유지형 프롬프트 최적화

ABSTRACT

This paper proposes a prompt–optimization framework for generating style–consistent game images using Stable Diffusion XL. Given a reference game item image, the system first extracts an initial prompt using a vision–language captioner and a domain–specific prompt bank. The extracted prompt is converted into a list of noun-like elements, and a genetic algorithm searches for compact combinations of these elements under dual gating based on SSIM and CLIP scores. The best combinations are treated as a “style template” that can reproduce the reference icon with high structural and semantic similarity. We then investigate whether this template can be reused when the main object is changed while preserving the original visual style. Experiments on fantasy-style item images show that the framework reconstructs reference images using only 8–9 automatically discovered prompt elements, and that changing the main object token together with associated detail elements yields image sets that share a consistent visual style. In contrast, naïvely replacing only the main object token often produces visually ambiguous or stylistically inconsistent images. These results demonstrate that combining automatic prompt extraction from images with evolutionary optimization provides a concrete example of style–preserving prompt design for game item image generation.

Key words: Game Item Image, Stable Diffusion XL, Automatic Prompt Extraction, Prompt Optimization, Genetic Algorithm, CLIP, SSIM, Style Consistency

1. Introduction

Recent text-to-image diffusion models make it increasingly feasible to automatically produce game graphics such as item images. Even though these images are small and compositionally simple, they still need to share a consistent palette, shading, outlines, and background treatment to be perceived as one visual set. Designing prompts that both reproduce a given style and allow the depicted object to change is therefore an important practical problem.

Two questions arise in this context. First, how should prompts for text-to-image models be organized? The output depends sensitively on the wording, and both style and shape can fluctuate greatly with small changes. Second, how can we maintain stylistic consistency when depicting different objects, for example when expanding from a single reference image to a family of items?

Given a single reference game item image, this paper automatically extracts prompt candidates from the image, optimizes them to obtain a compact "style template," and then examines whether this template can be reused while changing the main object. We first collect textual descriptions using a vision–language captioner and a PromptBank, and convert them into a list of noun-like prompt elements. A genetic algorithm with SSIM- and CLIP-based dual gating then searches for combinations that faithfully reproduce the reference image, and the resulting combination is treated as the style template. Finally, we keep this template fixed while replacing only the main-object-related elements with those of other items, and evaluate the visual consistency of the resulting set of item images.

The contributions of this paper are threefold. (1) We present a framework that combines image-based automatic prompt extraction with a genetic algorithm to learn a compressed prompt structure that reproduces the style of a reference item image. (2) By separating style-template elements from object-specific elements, we show that it is possible to generate images that remain stylistically consistent even when the depicted object changes. (3) Using fantasy-style game item images as a case study, we provide a concrete example of style-preserving prompt design for game graphics.

2. Related Works

2.1 Game Images and Stylistic Consistency

In game interfaces, item images are used in various contexts such as inventories, skill trees, and shop UIs. Within a small canvas, these images form a visual style through a combination of color, shading, outlines, and background treatment. For multiple images to be perceived as a single set within the same project, these elements need to be kept consistent. This study addresses the problem of image style consistency from the perspective of text-to-image prompt design.

2.2 Text-to-Image Generation and Image-Based Prompt Extraction

Stable Diffusion-family models can generate images in diverse styles from text prompts and are increasingly considered for game illustrations and 2D item images [1][2]. In parallel, BLIP-family captioners [5] and CLIP-based vision-language models [15] can describe a given image in natural language and compute image-text similarity. In this work, we use such models to obtain initial prompt candidates from a reference item image, which then serve as the starting point for our optimization pipeline.

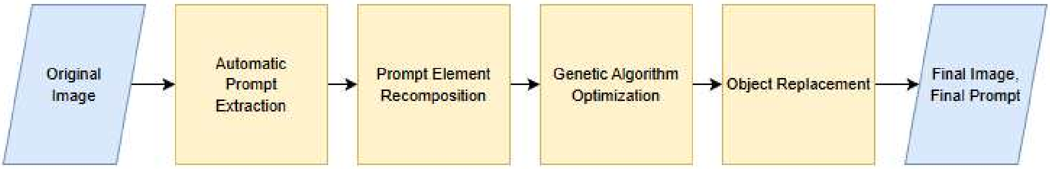
2.3 Automatic Prompt Optimization and Evolution-Based Search

Recent work in natural language processing has systematized strategies for writing and refining prompts [3]. Methods such as Automatic Prompt Optimization (APO), EvoPrompt, and Promptbreeder automatically edit and evolve prompts for language-model tasks [4][6][7], typically using whole sentences as the search unit and targeting accuracy or text quality. Evolutionary operators have also been applied to prompt search [8]. Our approach is related but differs in that (1) prompts are automatically extracted from an image, (2) they are represented as a list of noun-based elements, (3) visual similarity is directly evaluated via SSIM and CLIP, and (4) the target domain is game item images.

3. Proposed System

3.1 Overview

The overall pipeline of this study is shown in [Fig. 1].


Original Image
Automatic Prompt Extraction
Prompt Element Recomposition
Genetic Algorithm Optimization
Object Replacement
Final Image, Final Prompt

[Fig. 1] Overall pipeline

First, a reference game item image is provided as the input (original image), and a textual description of this image is automatically obtained using a vision-language model and PromptBank (automatic prompt extraction). In the next step, noun-based elements among the extracted expressions are organized into a list and restructured as prompt elements to be used for subsequent search.

Based on this element list, we apply a genetic algorithm to find a prompt combination that faithfully reproduces the reference image; this corresponds to the GA optimization stage. Finally, while preserving the resulting style template, we perform object (prompt) replacement by substituting only the main-object-related elements with those of another item, and evaluate the stylistic consistency of the final generated image and the final prompt.

3.2 Image-Based Automatic Prompt Extraction

First, we input the reference item image into a BLIP-family captioner and a

CLIP-based vision-language model to obtain multiple textual descriptions [5][15]. These include both style expressions such as “fantasy style, dark background, glowing, digital illustration” and object nouns such as “book, helmet, potion.” In parallel, we query PromptBank, which is built from public prompt galleries such as DiffusionDB [13], to retrieve n-grams frequently co-occurring with similar images. After merging these candidates, we apply part-of-speech tagging and keep only nouns and noun phrases, yielding an element set such as “book, cover, gold rim, glowing rim, dark background, soft light, fantasy style” [9].

3.3 Prompt Element List and Genetic Algorithm

We assign indices to the automatically extracted element set to construct a prompt element list:

$$E = \{e_0, e_1, \ldots, e_{N-1}\} \tag{1}$$

$$x = (x_0, x_1, \ldots, x_{N1}),\ x_i \in \{0, 1\} \tag{2}$$

$$P(x) = concat\ \{e_i \mid x_i = 1\} \tag{3}$$

E : a set of prompt elements obtained through automatic extraction and filtering

x : a binary mask (chromosome) indicating whether each element is selected

$P(x)$: final prompt formed by concatenating only the elements with xi = 1 in order

We set e_0 as the main object (e.g., book, helmet, potion) and use the remaining elements as style/detail elements that describe the background, material, color tone, lighting, and ornamentation, while enforcing that e_0 is always included.

For each individual in every generation, we generate an image using an SDXL pipeline [10] and compute SSIM and CLIP scores with respect to the reference image [14][15].

For the reference image I_{ref} and the image I_x generated from the prompt $P(x)$:

$$s(x) = SSIM(I_{ref}, I_x), \qquad c(x) = CLIP(I_{ref}, I_x) \tag{4}$$

Let the SSIM and CLIP thresholds be τs and τc respectively. The dual-gating function $g(x)$ is defined as:

$$g(x) = \begin{cases} 1, s(x) \geq \tau s \text{ and } c(x) \geq \tau c \\ 0, otherwise \end{cases} \tag{5}$$

Finally, the fitness function $f(x)$ maximized by the GA is:

$$f(x) = g(x)(\lambda s(x) + (1 - \lambda)c(x)) \tag{6}$$

$\lambda \in [0, 1]$: a hyperparameter that controls the weighting between SSIM and CLIP

The dual gating $g(x)$ treats a candidate as valid only when both metrics exceed their thresholds, filtering out combinations that are only structurally or only semantically similar. We iterate selection, crossover, and mutation over multiple generations and take high-scoring combinations in the final generation as style template candidates [8].

3.4 Style Template and Object Replacement

In the object replacement stage, we fix the style-template elements and change only the object-specific slot.

- Strategy A : replaces only the 0th element (the main object) while keeping all other elements unchanged.
- Strategy B replaces the 0th element together with related detail elements, while preserving the style-template elements.

We compare the images generated by each strategy using CLIP and qualitative evaluation.

4. Experiments and Results

4.1 Experimental Setup

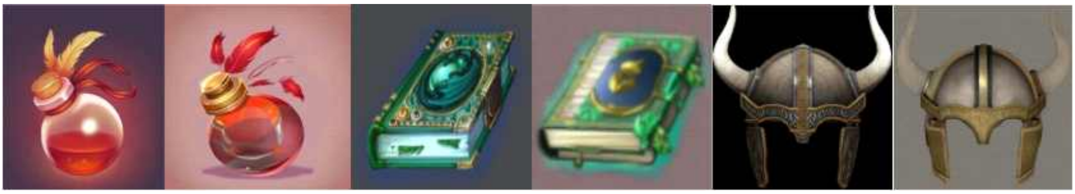
We used three reference images (a spellbook, a helmet, and a potion bottle). For each reference image, we ran the image-based prompt extraction procedure to obtain 12–15 initial elements, which were then directly used as the search space of the genetic algorithm without manual tuning. The GA population size was 15 and the maximum number of generations was 6.

Final images were generated at a resolution of 1024×1024. ControlNet and IP-Adapter were employed to partially preserve the layout and color tone of the reference item image [11][12]. We evaluated similarity to the reference image using SSIM and CLIP text–image similarity [14][15], and additionally conducted a qualitative evaluation of whether the generated images were perceived as belonging to the same visual set.

4.2 Reference Image Reproduction Results

As shown in [Fig. 2], for all three reference images, the genetic algorithm found prompt combinations that reproduce the reference image in a similar manner using only 8–9 elements in the final stage. As generations progressed, the mean SSIM and CLIP values and the pass rate increased or remained above a certain level, while cases of structural collapse or semantic mismatch observed in early generations decreased.

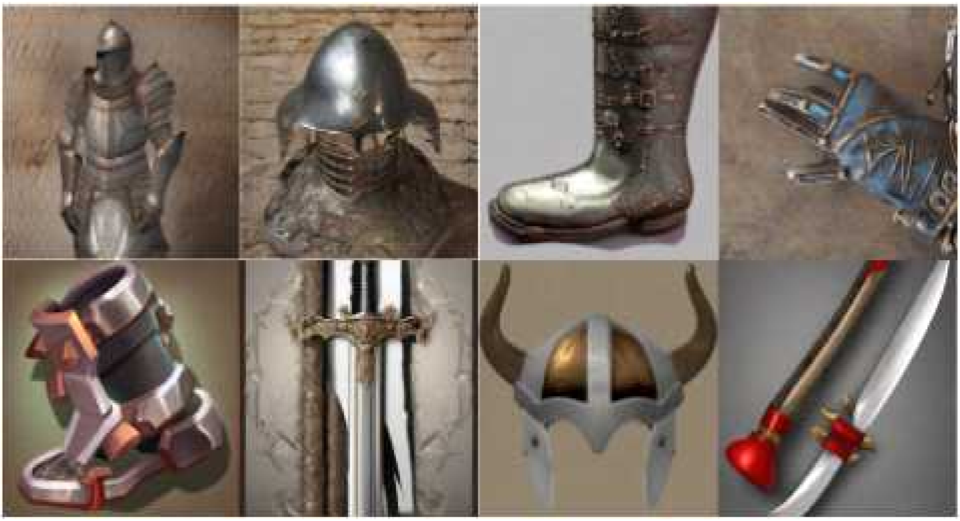
This suggests that combining image-based automatic extraction with evolutionary selection yields a compressed set of core elements that represent the style. In qualitative evaluation, final-generation images tended to share background tone, lighting, and coloring style with their respective reference images.



[Fig. 2] Image Reproduction Results

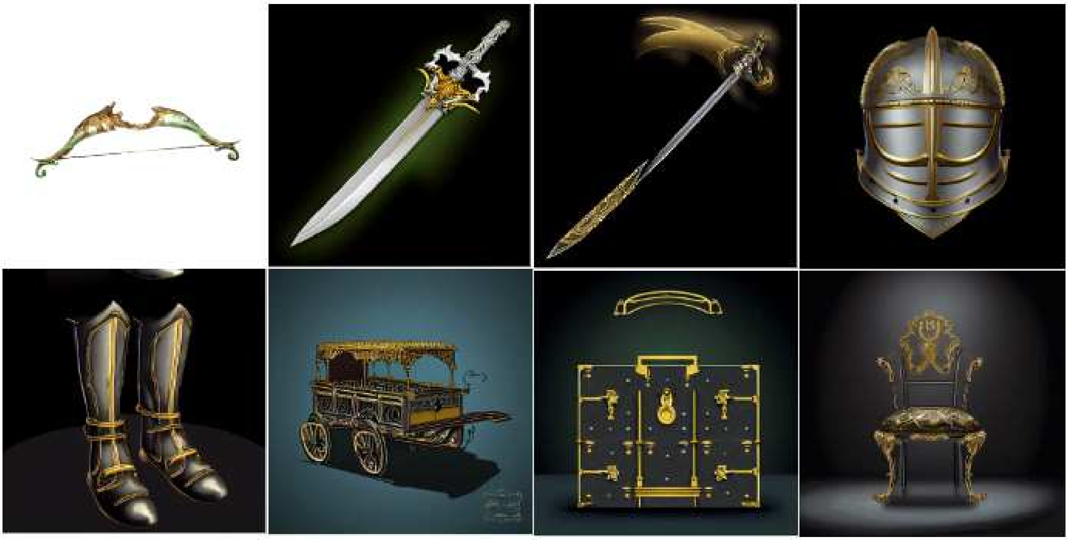
4.3 Object Replacement Results

In the object replacement experiments, Strategy A maintained stylistic consistency when changing to a similar category. However, when changing to a completely different category, mixed images were frequently generated: characteristics of the book and the sword were blended or unnecessary details remained, and CLIP similarity was low in such cases ([Fig. 3]).



[Fig. 3] Results When Only the Object Is Replaced

In contrast, Strategy B fixed the style-template elements while replacing the main object together with its related details. This preserved background, lighting, and coloring style similar to the reference item image, while clearly expressing the shape and function of each object ([Fig. 4]). Qualitative evaluation indicated that images generated with Strategy B were more likely to be perceived as a coherent set of item images despite depicting different objects.



[Fig. 4] Results When Both the Object and Elements Are Replaced

5. Conclusion and Future Work

This paper proposed a prompt optimization framework that combines image-based automatic prompt extraction with a genetic algorithm and SSIM/CLIP dual gating to address stylistic consistency in game item image generation. Given a single reference image, we automatically extracted initial prompt candidates using vision-language models and PromptBank, organized them into a noun-based element list, and obtained a style template that reproduces the reference image through a genetic algorithm. We then showed that, by replacing the main object and its related detail elements while preserving the style template, it is possible to generate visually consistent sets of item images for different objects.

Future work includes extending this framework to other game-graphics domains and implementing it as an interactive tool that allows users to edit the element list and style template. In addition, PromptBank can be expanded into genre- and style-specific datasets, and an ontology-based part structure may provide more fine-grained element recommendations for categories such as weapons, armor, and consumables.

References

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 6840 - 6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685, 2022.

[3] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” ACM Comput. Surveys, vol. 55, no. 9, Art. 195, pp. 1–35, 2023.

[4] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng, “Automatic prompt optimization with ‘gradient descent’ and beam search,” in Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 7957–7968, 2023.

[5] J. Li, Y. Li, T. Xiong, and S. C. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in Proc. 39th Int. Conf. on Machine Learning (ICML), Proc. Mach. Learn. Res., vol. 162, pp. 12888–12900, 2022.

[6] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang, “EvoPrompt: Connecting LLMs with Evolutionary Algorithms Yields Powerful Prompt Optimizers” arXiv preprint, arXiv:2309.08532, 2023.

[7] C. Fernando, D. S. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel, “Promptbreeder: Self-referential self-improvement via prompt evolution,” in Proc. 41st Int. Conf. on Machine Learning (ICML), Proc. Mach. Learn. Res., vol. 235, pp. 13481–13544, 2024.

[8] T. Alam, M. S. Siddiqui, and S. M. Sait, “Genetic algorithm: Reviews, implementations, and applications,” Int. J. Eng. Pedagogy (iJEP), vol. 10, no. 6, pp. 57–77, 2020.

[9] J. Oppenlaender, “A taxonomy of prompt modifiers for text-to-image generation,” Behav. Inf. Technol., vol. 43, no. 15, pp. 3763–3776, 2024.

[10] 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 Proc. Int. Conf. on Learning Representations (ICLR 2024), 2024.

[11] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 3813–3824, 2023.

[12] H. Ye, H. Chen, W. Zhang, Y. Zhang, and J. Xu, “IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models,” arXiv preprint, arXiv:2308.06721, 2023.

[13] Z. J. Wang, A. Narayan, A. Pu, N. E. Roth, M. AlKhamissi, A. Ghandeharioun, and A. Liu, “DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models,” in Proc. 61st Annu. Meet. of the Assoc. for Comput. Linguistics (ACL), pp. 893–911, 2023.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.

[15] A. Radford, J. W. Kim, C. Hallacy et al., “Learning transferable visual models from natural language supervision,” in Proc. 38th Int. Conf. on Machine Learning (ICML), Proc. Mach. Learn. Res., vol. 139, pp. 8748–8763, 2021.