

Application of human-machine collaborative creative generation process in industrial design

Hansheng Yao¹, Wei Yu^{1}*

¹School of Art, Design and Media, East China University of Science and Technology, Shanghai, China

*Corresponding Author. Email: weiyu@ecust.edu.cn

Abstract. This paper proposes a closed-loop human-machine co-creation process suitable for the early stages of industrial design. By integrating the Stable Diffusion model with the Low-Rank Adaptation (LoRA) fine-tuning strategy, and constructing an image quality evaluation mechanism based on the dual metrics of Contrastive Language-Image Pretraining (CLIP) and CLIP Maximum Mean Discrepancy (CMMD), the system guides designers in filtering and providing feedback on generated outputs to iteratively optimize prompts. The system integrates automatic scoring, manual filtering, and keyword clustering recommendation to form a collaborative closed loop of “generation-selection-optimization.” In a desk lamp design task, experiments demonstrate that this process significantly enhances the consistency of image styles and the quality of creative expression. The study verifies the feasibility of the human-machine collaboration mechanism in complex design tasks and offers a new paradigm for the application of generative AI in industrial product design.

Keywords: human-machine collaborative design, generative artificial intelligence, prompt optimization, product appearance design

1. Introduction

Contemporary industrial design is undergoing a transformative shift with the deep integration of generative Artificial Intelligence (AI). Generative technologies such as diffusion models can rapidly produce large volumes of high-quality design images, offering designers a rich pool of creative resources [1]. However, selecting design options from this bulk output that align with intended design objectives and stylistic requirements remains a challenge: automated algorithms struggle to fully grasp designers’ aesthetic preferences and creative intentions [2], while manual selection from a vast array of options can lead to cognitive overload and decision fatigue [3].

To address these issues, human-AI co-creation has emerged in recent years as a prominent research focus in academia. Song et al. proposed a unified and comprehensive classification scheme for AI roles and developed an AI design framework [4]. This framework details AI’s expected capabilities (such as analysis and synthesis), interaction attributes (such as real-time feedback), and trust-enhancing factors (such as explainability) across different collaborative scenarios. Their study emphasizes that designing AI systems to enhance human team collaboration is crucial—especially in the context of engineering design and innovation. Wang et al. pointed out that traditional Human-AI Interaction (HAI) differs fundamentally from Human-AI Collaboration (HAIC) [5]. True collaboration requires shared understanding of goals, co-management of tasks, and synchronized progress tracking, rather than AI merely serving as a tool. They advocate incorporating the perspective of Computer-Supported Cooperative Work (CSCW) into the design of AI algorithms, with the aim of building trustworthy collaboration models between humans and AI in future applications. Jiang et al. conducted a systematic review of the Human-AI Interaction and Integration (HAI) field, identifying collaboration as one of the core research themes [6]. They recommend that future research expand to include diverse user groups, AI roles, and tasks, while integrating interdisciplinary theories—from communication studies, psychology, and sociology—to support sustained development in the field. Puranam regarded collaborative decision-making between humans and AI as a matter of organizational design and proposed a typology of human-AI task division and learning configurations [7].

Although these studies have laid a theoretical foundation for human-machine collaboration, there remains a lack of actionable, iterative mechanisms aimed at optimizing design quality in the practical implementation of generative design. This is particularly evident during the early stage of appearance ideation, where a gap often exists between the image quality, style coherence, and design logic of AI-generated outputs, limiting their applicability in real-world industrial design workflows.

In response, this study proposes a closed-loop human-machine collaborative creative generation process. Based on the Stable Diffusion model, the system employs LoRA fine-tuning to enhance style controllability and constructs an automatic scoring mechanism that integrates CLIP and CMMD metrics, enabling dual-dimensional evaluation of semantic alignment and stylistic coherence in the generated images. Building on this foundation, the system introduces a designer-in-the-loop filtering and feedback mechanism, enabling efficient iteration and targeted evolution of creative outputs through multi-round prompt optimization and semantic clustering recommendations. Using a household desk lamp appearance design task as a case study, this paper demonstrates the practical effectiveness of the proposed workflow and explores its potential in advancing human-AI co-creation mechanisms and promoting intelligent transformation in industrial design.

2. Automated scoring system based on CLIP and CMMD

Batch image generation using diffusion models forms the foundation for obtaining diverse design concepts. However, a subsequent challenge arises: how to objectively evaluate the quality of the generated outputs. To address this, we propose a dual-metric automated scoring system that integrates CLIP and CMMD, enabling semantic and stylistic evaluation of large-scale generated images. The two scores are normalized and weighted to produce a composite score that serves as the basis for subsequent filtering.

CLIP Score (Semantic Similarity): We adopt the Contrastive Language-Image Pretraining (CLIP) model developed by OpenAI to compute the semantic similarity between each generated image and its corresponding text prompt [8]. Specifically, the image and the prompt are each projected into a shared embedding space, and their cosine similarity is calculated as the score:

$$S_{CLIP} = \cos(\text{Image_Embedding}, \text{Text_Embedding}) \quad (1)$$

A higher score indicates that the image more faithfully reflects the prompt in terms of material, form, and style. Prior research suggests that CLIP scores are highly correlated with human judgments and can effectively assess text-image alignment.

CMMD Score (Distribution Consistency / Style Credibility): To assess whether a generated image's style aligns with that of real-world product distributions, we incorporate the recently proposed CLIP-based Maximum Mean Discrepancy (CLIP-MMD) metric [9], hereafter referred to as CMMD. This metric first extracts feature embeddings of both generated and real product images using CLIP, then calculates the Maximum Mean Discrepancy (MMD) between the two distributions in the embedding space:

$$MMD^2(P, Q) = \mathbb{E}_{x, x' \sim P} [k(x, x')] + \mathbb{E}_{y, y' \sim Q} [k(y, y')] - 2\mathbb{E}_{x \sim P, y \sim Q} [k(x, y)] \quad (2)$$

Smaller differences indicate that the overall style of the generated images is closer to that of real-world products, and therefore more credible. For intuitive interpretation, the MMD distance is inversely normalized into a CMMD score ranging from 0 to 1, where a higher score suggests a more natural and authentic design style. Compared to traditional evaluation metrics (e.g., Fréchet Inception Distance, FID), CMMD leverages richer semantic information from CLIP and avoids the biases introduced by Gaussian distribution assumptions, offering more reliable assessment of generated image quality.

Composite Score: Since design tasks often require balancing semantic fidelity and stylistic realism, we introduce a weighted composite scoring function that fuses the CLIP and CMMD scores with a tunable parameter α .

$$S_{total} = \alpha \cdot S'_{CLIP} + (1 - \alpha) \cdot S'_{CMMD} \quad (3)$$

By default, $\alpha=0.5$ (equal weighting), which suits most scenarios. If semantic alignment with the prompt is prioritized, α can be increased; if stylistic realism is emphasized, it can be decreased. The composite score enables preliminary sorting and selection of generated images, with higher-scoring images prioritized for designer review.

In practice, we collected 60 SKUs of desk lamps with strong recent sales from the official websites of major domestic lighting brands (Philips, Oppl, Yeelight, Panasonic), as well as third-party shopping platforms. After removing duplicates, blurry images, and distorted angles, and cleaning complex backgrounds using Photoshop, we curated an initial dataset of 40 samples, as shown in Figure 1.

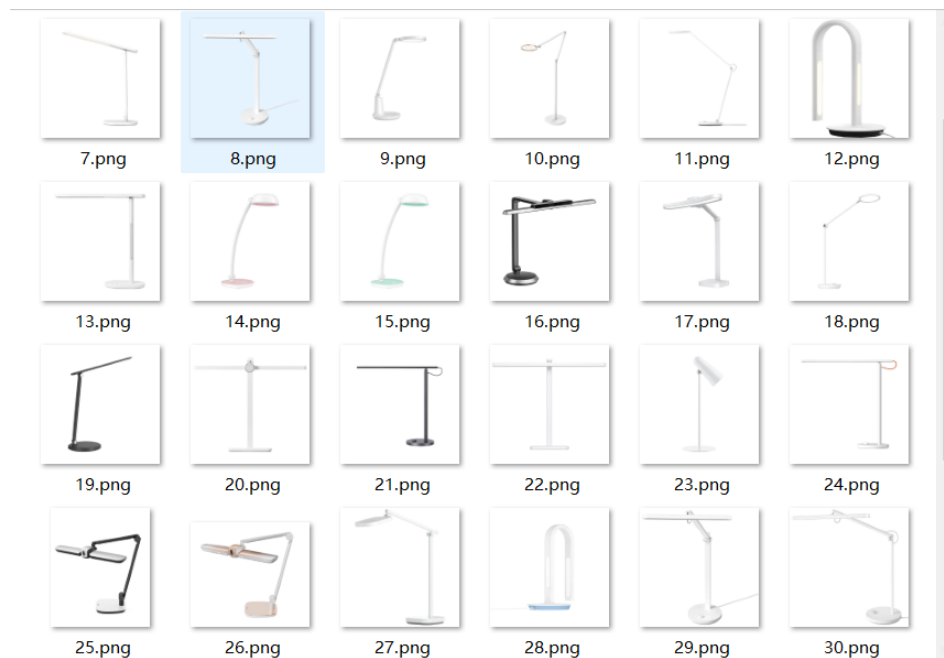


Figure 1. Selected experimental samples (image source: author-illustrated)

Using the fine-tuned diffusion model, we generated 128 conceptual images of desk lamps in a single batch and calculated the CLIP and CMMD scores for each. The resulting composite scores and evaluation labels are presented in Table 1. The automated scoring results provided a streamlined and organized candidate set for manual selection, allowing designers to focus their efforts on high-potential concepts.

Table 1. Composite scoring results of selected generated images (table source: author-illustrated)






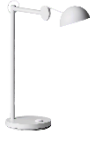


Image ID	Image	Original CLIP Score	Normalized CLIP Score	CMMD Distance	Normalized CMMD Score	Composite Score	System Evaluation Notes
006		0.92	0.92	0.18	0.82	0.87	High text-image alignment; realistic appearance with consistent style
017		0.91	0.91	0.41	0.59	0.75	Semantically precise, but proportions are slightly distorted
037		0.85	0.85	0.24	0.76	0.81	Reasonable form; details meet design requirements
048		0.58	0.58	0.16	0.84	0.71	Style is realistic, but lamp arm representation is insufficient
060		0.39	0.39	0.34	0.66	0.53	Vague expression; lacks realism in form

Table 1. Continued

069		0.78	0.78	0.55	0.45	0.62	Moderate alignment; stylistic inconsistency
086		0.67	0.67	0.19	0.81	0.74	Overall reasonable; some details require adjustment
104		0.52	0.52	0.22	0.78	0.65	Stylistically coherent; some elements not clearly expressed

3. Manual screening and multi-round optimization feedback mechanism

Although automated scoring efficiently filters out inferior options on a rational level, aesthetic judgments in creative design still require human oversight. Many key design attributes—such as creativity, novelty, aesthetic value, and brand alignment—are difficult for current algorithms to quantify or accurately capture. Therefore, after the automated scoring stage, this study introduces a manual screening phase in which professional designers perform a secondary evaluation of high-scoring candidate images to ensure the final proposals meet both functional logic and aesthetic standards. This human-AI collaborative mechanism iteratively refines the design outcomes through multiple rounds of feedback, balancing AI’s efficiency with human creative judgment.

3.1. Process and role of manual screening

In each iteration, the system first ranks the generated images based on their composite scores and selects a batch of high-alignment candidate proposals. Designers then evaluate each candidate individually, judging from the perspectives of functional feasibility, structural soundness, stylistic coherence, and creative uniqueness. For proposals that meet the requirements, designers mark them as “Confirmed for Adoption”; those with potential but in need of improvement are labeled “Conditionally Adopted” with remarks; unsatisfactory ones are eliminated. After each screening round, the system records designers’ selection preferences and comments. These data are then used to analyze design inclinations and guide the next generation cycle, as shown in Table 2. It is worth noting that manual screening is not merely a revalidation of the automated scores—it also identifies creative elements missed by the algorithm and corrects potentially overestimated suboptimal results. By integrating human and machine evaluations, the screening outcomes more reliably reflect the design objectives.

Table 2. Designers’ decision records for selected generated images (table source: author-illustrated)





Rank	Image ID	Image	Composite Score	Designer’s Decision	Designer’s Remark
1	006		0.87	Confirmed for Adoption	High alignment and strong feasibility for implementation
2	037		0.81	Conditionally Adopted	Proportions between lamp arm and pivot require further refinement

Table 2. Continued

3	017		0.75	Conditionally Adopted	Distortion in proportion details is obvious; style may be retained
4	086		0.74	Rejected, Archived	Elements meet design requirements but contain structural errors

3.2. Multi-round iteration and feedback loop

This section establishes a cyclical process of “generation-selection-optimization,” enabling progressive enhancement through human-AI collaboration. Specifically, after each round of manual screening, the system analyzes the high-value images selected by designers and summarizes their common features and semantic preferences. On the one hand, these shared attributes are used to adjust the parameters or prompts for the next round of image generation, guiding the AI to explore in directions aligned with designer preferences. On the other hand, designers can also refer to the experience from the previous round to adjust their design strategies. For example, if the previous set of generated designs lacked a certain stylistic element, the designer may include relevant descriptions in the new prompt. Through this bidirectional feedback, designers and the AI model adapt to each other in every iteration: the AI gradually becomes attuned to the designer’s preferences, while the designer can more efficiently expand the boundaries of creativity with AI support. After multiple rounds of iteration, the system ultimately converges on a set of outstanding design solutions that are both creative and feasible, completing the design process from divergence to convergence.

It is worth emphasizing that this multi-round optimization loop is not limited to desk lamp design tasks. In complex product design, the designer’s aesthetic judgment always plays a crucial role. Automated scoring provides objective benchmarks, while manual selection ensures the final decisions align with human aesthetic standards. The two complement each other, making the human-machine collaboration process both efficient and reliable.

4. Scoring clustering and prompt optimization mechanism

To further leverage the value of scoring data, this study introduces semantic clustering analysis to extract prompt optimization suggestions from designer preferences. This mechanism can recommend keywords for the next round of generation based on the commonalities of high-scoring designs, thus forming an intelligent prompting function that facilitates human-AI collaboration.

4.1. Clustering analysis of high-scoring samples

After each generation and screening round, this study clusters the set of high-scoring images either selected or retained by the designer, based on their scoring vectors and semantic labels. The “scoring vector” here refers to the combination of each image’s CLIP score and CMMD score, which can be further expanded to include other semantic features of the image. The goal of clustering is to identify underlying categories in designer preferences: for instance, one cluster may represent a preference for “structurally innovative and functionally sound” designs, while another reflects “stylistically unique yet brand-consistent” designs.

Taking the minimalist desk lamp design experiment as an example, the study uses the GPT-4o model to label each image with 20 English keywords capturing semantic information, covering dimensions such as color, material, finish, and geometry. The labels are manually reviewed and corrected by a design focus group to serve as structured semantic representations of the images.

Based on the semantic tags, a term-frequency matrix is built using the TF-IDF vectorization method. Principal Component Analysis (PCA) is then used to reduce the vector dimensions to three for visual analysis and clustering. Finally, K-Means clustering is applied to the samples, with the number of clusters $k \in [2, 5]$. The silhouette score is used to evaluate clustering performance, and $k=5$ is selected as the optimal partition. Each cluster corresponds to a group of semantically similar design schemes. The central features of these clusters can be described using a set of keywords, as shown in Table 3. This clustering analysis structures dispersed design preferences and lays the foundation for automated prompt generation.

Table 3. Keyword clustering results (table source: author-illustrated)

Cluster	Keywords	Frequency	Number of Images
Cluster 1	<i>plastic material, white color, modern style, neutral tone, matte finish</i>	18, 15, 12, 11, 9	25
Cluster 2	<i>functional design, adjustable arm, minimalist aesthetic, simple geometry, ergonomic design</i>	21, 16, 15, 15, 14	30
Cluster 3	<i>task lighting, pivot joint, modern style, soft edges, sleek body</i>	19, 17, 13, 11, 9	29
Cluster 4	<i>LED light source, neutral tone, lightweight, pivot joint, clean silhouette</i>	19, 10, 8, 8, 7	20
Cluster 5	<i>minimalist aesthetic, simple geometry, retro elements, plastic material, modular parts</i>	15, 14, 11, 9, 8	19

4.2. Prompt recommendation and generation

Based on the clustering results, a method is designed to generate prompt optimization suggestions. First, the focus is placed on the image cluster(s) with high designer acceptance, and high-frequency keywords are extracted from their semantic labels. In this implementation, each image is pre-labeled with a set of keywords generated by the GPT model, describing its features across dimensions such as material, structure, and style. For images in the selected clusters, this study calculates the TF-IDF weight of each keyword and selects the top five as the core semantics representing the cluster. Next, methods such as variance thresholding are applied to filter out keywords that contribute little or have low distinctiveness within the cluster, ensuring that the retained keywords effectively highlight the characteristics of the cluster. These remaining keywords are then categorized into three groups according to design semantics: material, structure, and stylistic descriptors. Using the example keywords above, the filtering results are shown in Table 4.

Table 4. Keyword filtering results (table source: author-illustrated)

Material	Structure	Stylistic Descriptors
<i>None</i>	<i>adjustable arm</i>	<i>functional design, minimalist aesthetic, simple geometry, ergonomic design, LED light source</i>

4.3. Collaborative recommendation and human decision-making

The system presents the generated prompt words in the form of a recommendation list to the designer. In the interactive interface, each recommended word is labeled with its source (e.g., “High-frequency keyword from high-rated images”) and its corresponding preference cluster category (e.g., “Structural Preference”). A thumbnail of a related high-rated image is also displayed as reference. Designers are free to select which recommendation words to adopt: they can include them in a new prompt with a single click, or they may edit the words by adding, removing, or modifying them. It is important to note that the suggested prompts are auxiliary rather than mandatory—they serve to inspire and provide direction, but the ultimate creative control remains with the designer. Through this interpretable and controllable recommendation approach, designers can experiment with different prompt adjustments at lower cost, thereby improving the efficiency of the generation-iteration process.

4.4. Dynamic optimization and co-evolution

The prompt recommendation mechanism also includes a feedback-driven adaptive module. As multiple iterations progress, the system monitors the effectiveness of the recommended words: if, in a certain round, images generated with a specific recommended word frequently receive high scores and are selected by the designer, the system will increase the weight of that word in subsequent iterations or continue to recommend related vocabulary. Conversely, if a recommended word repeatedly leads to rejected outputs, the system will reduce its weight or remove it from the candidate list. Designers’ evaluations of the generated results are also used to refine the recommendation algorithm through regression optimization, gradually aligning the system’s suggestions with the

designer's long-term preferences. This “human choice → system adaptation” mechanism ensures that the recommended prompts co-evolve with the designer's taste, allowing the AI assistant to increasingly “understand” the user while keeping creative leadership in the designer's hands.

In summary, the scoring-based clustering and prompt optimization module establishes an intelligent semantic bridge within the human-AI collaboration system: it mines design preferences from data and feeds them back into prompt generation, thus enabling a transformation from passive output to active assistance in the creative process.

5. Closed-loop human-AI co-creation mechanism and experimental validation

By integrating the modules described above, this study constructs a complete closed-loop design workflow for human-AI co-creation. Starting from the designer's initial prompt and the model-generated outputs, the process goes through automatic scoring and filtering, manual refinement and feedback, prompt optimization, and then enters the next round of creation—repeating iteratively until a satisfactory design is achieved. This workflow tightly couples the speed of AI with human intelligence, significantly improving both the efficiency and quality of creative generation. To validate the effectiveness of this closed-loop mechanism, a comparative experiment was conducted, evaluating design outcomes with and without the involvement of recommended prompts.

Using a desk lamp appearance design task as the test scenario, the same base model and parameters were applied (a LoRA-fine-tuned Stable Diffusion model with identical random seeds, etc.). Control Group A used only the designer's original prompt to generate images; Experimental Group B appended system-recommended keywords to the end of the prompt in each round, as shown in Table 5. Each group generated 50 images, which were evaluated across multiple dimensions. Evaluation metrics included: (1) text-image consistency (CLIP score); (2) structural and stylistic consistency (CMMD score); and (3) subjective design quality scores (blind ratings by five designers, scoring 1-5 based on visual quality, stylistic fit, and design value). To ensure reliability, independent samples t-tests were conducted to assess the significance of differences.

Table 5. Recommended prompts for both groups (table source: author-illustrated)

Group	Prompt
Group A (No Recommendation)	<i>A modern minimalist desk lamp with white matte plastic finish, circular base and lampshade, adjustable arm.</i>
Group B (With Recommendation)	<i>A modern minimalist desk lamp with white matte plastic finish, circular base and lampshade, adjustable arm, functional design, simple geometry, ergonomic structure, LED light source, minimalist aesthetic.</i>

Results and Analysis: As shown in Table 6, the prompts incorporating system-recommended keywords (Experimental Group B) significantly outperformed those of Control Group A across all evaluation dimensions. Specifically, the average CLIP score of Group B improved by approximately 14.6%, the average CMMD score increased by around 22.4%, and the average subjective rating from designers rose by about 19.6%. All differences passed significance testing ($p < 0.05$). These findings demonstrate that the scoring-driven prompt optimization mechanism effectively enhances semantic alignment and stylistic naturalness in generated images, providing more valuable references for designers. Several designers involved in the blind review commented: “Prompts with recommended keywords produced outputs with greater visual integrity and stylistic coherence; the system's suggested keywords clarified the design direction and reduced trial-and-error.” Such feedback underscores the positive role of human-AI collaboration in minimizing repeated creative missteps—AI offers rational suggestions based on data, while designers adjust their thinking accordingly. Their synergy ensures that each iteration evolves toward better outcomes.

Table 6. Experimental comparison between two groups (table source: author-illustrated)

Evaluation Metric	Group A (No Recommendation)	Group B (With Recommendation)	Improvement	Significance (p-value)
Average CLIP Score	0.534	0.612	14.6%	0.021
Average CMMD Score	0.478	0.585	22.4%	0.008
Average Human Score	3.36	4.02	19.6%	0.014

Through the above experiment, this study verifies the performance gains of the proposed closed-loop workflow in real design tasks. Designers, aided by multi-dimensional evaluations and intelligent prompts from the system, can explore creative spaces more efficiently; meanwhile, AI, guided by human feedback, produces outputs that better align with design intentions. In essence, this human-AI co-creation mechanism achieves simultaneous improvement in design efficiency and quality, marking a shift in generative design workflows from “one-time output” to “continuous optimization.” This outcome introduces a new paradigm for

integrating AI into industrial design practice—where AI is no longer merely a passive tool, but gradually becomes a “creative assistant” that understands design.

6. Discussion on generalizability and scalability

This study uses a desk lamp as an example to illustrate a closed-loop human-AI collaborative creative generation process, yet the underlying design philosophy is broadly generalizable. First, the mechanism of scoring, filtering, and feedback is not limited to a specific product category. By adjusting evaluation metrics and reference data according to the task, the process can be transferred to other fields of industrial design. For instance, in furniture design, reference images of various furniture styles can be collected and used with CLIP+CMMD to evaluate generated chair or table designs. In automotive styling, similar semantic scoring and manual filtering can be applied to guide AI in progressively aligning with designers' concepts. The core lies in establishing a domain-specific generative design loop system through a customized evaluation framework (e.g., using CLIP to assess semantics, CMMD for style, or integrating other domain-specific metrics) and human-AI collaborative decision-making.

Secondly, the process is highly scalable. In its current implementation, the study focuses primarily on image-level generation and feedback, but it can be expanded to incorporate more modalities and stages. For example, a user interaction module could be added to involve end-users in evaluating conceptual designs, thereby integrating market preferences into the iteration process. CAD modeling and rapid prototyping phases could also be introduced, transforming 2D image concepts into 3D models for engineer review, thus forming a design-engineering integrated collaboration loop. Furthermore, the scoring mechanism could be extended to cover additional dimensions, such as automated evaluation of novelty, aesthetics, or green design indicators, to meet the unique needs of various projects. Since the entire system is modular—comprising clearly interfaced modules for generation, scoring, filtering, and recommendation—replacing or adding modules is relatively straightforward. This openness makes systematic implementation feasible: design teams can tailor and assemble the workflow according to their specific needs, progressively building an intelligent in-house design platform.

7. Conclusion and outlook

Focusing on the task of desk lamp appearance design, this paper proposes a closed-loop human-AI collaborative generation process, incorporating modules such as automated semantic scoring, manual filtering and feedback, preference clustering analysis, and prompt optimization suggestions. The process effectively integrates AI's objective evaluation with designers' subjective creativity, forming a cycle from image generation to evaluation and filtering, and finally to prompt optimization—thereby enhancing the efficiency and quality of generative design. Experimental results validate the significant improvement brought by the system's recommendation mechanism, further demonstrating the feasibility and value of the human-AI co-creation paradigm in design practice.

Future research and application efforts may be strengthened in the following aspects: (1) Data-wise, the process can be applied to a broader range of product categories and larger sample sizes across other industrial design domains to validate its generalizability. Accumulating more designer interaction data can also support the optimization of recommendation algorithms. (2) Experiment-wise, larger-scale user studies involving designers from diverse backgrounds can be conducted to assess the system's impact on the creative process—such as whether it reduces design time or enhances novelty. Feedback collected will help improve the human-AI interaction interface and collaboration strategy. (3) Technology-wise, integrating new generative models and evaluation metrics could be explored—for example, adopting large-scale multimodal models to enhance text comprehension and generation quality, or using human visual perception models to assess image aesthetics—in order to build a more comprehensive evaluation and feedback system.

In summary, the human-AI collaborative generation process represents a significant direction for the integration of industrial design and artificial intelligence. It enables AI to evolve from a design tool into a creative partner, simultaneously freeing productivity and sparking greater innovation. As technology and practice continue to advance, this collaborative paradigm is poised to flourish across broader design domains, driving the future of design processes toward intelligence, efficiency, and human-centeredness.

References

- [1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper
- [2] Lawson, B. (2010). *How designers think: The design process demystified* (Reprint). Elsevier Architectural Press.
- [3] Cross, N. (2001). Designerly Ways of Knowing: Design Discipline Versus Design Science. *Design Issues*, 17(3), 49–55. <https://doi.org/10.1162/074793601750357196>
- [4] Song, B., Zhu, Q., & Luo, J. (2024). Human-AI collaboration by design. *Proceedings of the Design Society*, 4, 2247–2256. <https://doi.org/10.1017/pds.2024.227>

- [5] Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–6. <https://doi.org/10.1145/3334480.3381069>
- [6] Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human-AI interaction research agenda: A user-centered perspective. *Data and Information Management*, 8(4), 100078. <https://doi.org/10.1016/j.dim.2024.100078>
- [7] Puranam, P. (2021). Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design*, 10(2), 75–80. <https://doi.org/10.1007/s41469-021-00095-2>
- [8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [9] Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2023). Supplementary Material for Rethinking FID: Towards a Better Evaluation Metric for Image Generation. arXiv. <https://doi.org/10.48550/arXiv.2401.09603>