iCONTRA: Toward Thematic Collection Design Via Interactive Concept Transfer

Dinh-Khoi Vo*
vdkhoi20@clc.fitus.edu.vn
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Duy-Nam Ly*
ldnam@selab.hcmus.edu.vn
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Khanh-Duy Le
lkduy@selab.hcmus.edu.vn
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Tam V. Nguyen
tamnguyen@udayton.edu
Department of Computer Science
University of Dayton
Ohio, United States

Minh-Triet Tran
tmtriet@fit.hcmus.edu.vn
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Trung-Nghia Le†
ltnghia@fit.hcmus.edu.vn
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

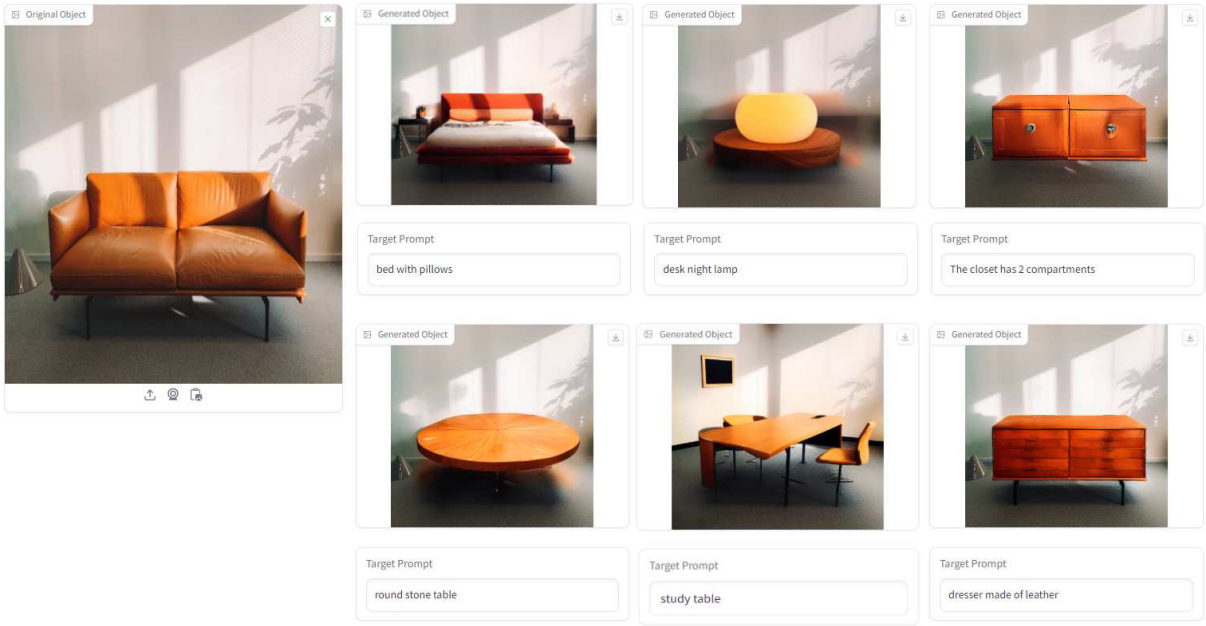

Figure 1: Given a single object image (i.e., left image), when users input brief prompts, our proposed iCONTRA system unleashes the power of generative AI, recommending objects that have the same theme as the input object (i.e., right images). The system also preserves the original structure and scene layout with unwavering precision, which is helpful for designers when visualizing new objects in the same environment, such as background and light.

ABSTRACT

Creating thematic collections in industries demands innovative designs and cohesive concepts. Designers may face challenges in maintaining thematic consistency when drawing inspiration from existing objects, landscapes, or artifacts. While AI-powered graphic design tools offer help, they often fail to generate cohesive sets

*Equal contribution
†Corresponding author. Email address: ltnghia@fit.hcmus.edu.vn

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650788

based on specific thematic concepts. In response, we introduce iCONTRA, an interactive CONcept TRAnsfer system. With a user-friendly interface, iCONTRA enables both experienced designers and novices to effortlessly explore creative design concepts and efficiently generate thematic collections. We also propose a zero-shot image editing algorithm, eliminating the need for fine-tuning models, which gradually integrates information from initial objects, ensuring consistency in the generation process without influencing the background. A pilot study suggests iCONTRA's potential to reduce designers' efforts. Experimental results demonstrate its effectiveness in producing consistent and high-quality object concept transfers. iCONTRA stands as a promising tool for innovation and creative exploration in thematic collection design. The source code will be available at: https://github.com/vdkhoi20/iCONTRA.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Interactive systems and tools**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Thematic collection design, Zero-shot image editing, Diffusion model

ACM Reference Format:
Dinh-Khoi Vo, Duy-Nam Ly, Khanh-Duy Le, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2024. iCONTRA: Toward Thematic Collection Design Via Interactive Concept Transfer. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3650788

1 INTRODUCTION

Crafting objects with a unified concept or theme is a critical aspect of various industries, heavily dependent on the creativity of designers for innovative designs, particularly to meet the demand for cohesive concepts in areas such as fashion and interior decor. Designers often face challenges where they must draw inspiration from existing objects, landscapes, or artifacts to achieve thematic cohesion in their design collections. For example, a designer may aspire to fashion a clothing line influenced by the intricate patterns found in a vintage piece of furniture or the harmonious color palette of a serene natural landscape. In such cases, the ability to seamlessly transfer these inspirations into cohesive and stylish fashion designs becomes crucial. This intricate process underscores the designer's proficiency in conceptualization and innovation.

More specific, when crafting thematic objects, designers can explore different websites, such as Pinterest, Instagram, using keyword descriptions to find inspiration and reference for creating a collection with a unified design theme. For instance, given a sofa placing in a glass window showroom, designers need to create other furniture, including bed, table, wardrobe that share the same design concept as the sofa, as illustrated in Fig. 1. However, this process can be time-consuming and challenging, often resulting in difficulty finding suitable objects while potentially losing the feeling of layout and environment of the original image.

AI-driven graphic design tools redefine the creative landscape by leveraging AI to streamline and enhance the design process [10]. These tools not only expedite tasks but also serve as innovative aids, saving time and providing inspiration. With capabilities ranging from swift task execution to offering creative suggestions, AI becomes a valuable partner for designers, unlocking unlimited possibilities and pushing creative boundaries beyond traditional approaches. While existing text-driven image manipulation methods excel in tasks like translation [5, 12, 19] and style transfer [6, 11], challenges arise due to the lack of specific object shape targets, making the process time-consuming and requiring fine-tuning efforts. On the other hand, AI-powered commercial softwares such as Adobe Firefly [1] , Midjourney [2] , DALL·E 2 [3] , DALL·E 3 [4] , Stable Diffusion [5] are not freely available for everyone and may be difficult to use for people without AI knowledge. Moreover, these tools cannot directly support designing thematic collections. These tools tend to produce entirely different objects, resulting in a loss of the original environmental background and a deviation from the desired thematic continuity.

To overcome these challenges, we introduce a novel interactive CONcept TRAnsfer (iCONTRA) system, designed to assist both experienced designers and individuals without prior design skills. iCONTRA aims to provide a user-friendly platform for creative exploration and expression in the design process of thematic collection. Our system can address limitations of existing systems in creating cohesive sets of objects. The user interface provides a seamless experience, allowing users to upload the original image and provide textual descriptions in each generation cell for the desired object (See Fig. 2). With six cells available, users can generate multiple objects iteratively. After generating initial objects, users can further refine them using convenient import and edit prompt features. They can also observe the generated object alongside the edited prompt, facilitating the achievement of desired designs. This iterative and user-friendly approach enhances the overall design exploration process, providing creative freedom and ease of use while reducing the effort required to find similar patterns, thus enhancing the overall user experience.

Our proposed iCONTRA system is built on a cutting-edge generative model [17], indicating its capacity to generate diverse and high-quality images. This model is tailored to achieve consistent and intricate non-rigid object generation while preserving overall textures and identity from the original object. When replacing an object in an image by another one, the central challenge is preserving the original information of the real object, in which previous methods encountered difficulties [2–4, 15]. To offer more robust solution, we develop a novel zero-shot image editing model allowing replacing objects in an image without the need for fine-tuning or specific training. Our algorithm exploits information from the original image, preventing abrupt changes during the generation process. Additionally, we automatically mask the foreground object from the background, minimizing the impact of background adjustment. This enables querying correlated local structures and textures

[1] https://www.adobe.com/products/firefly.html
[2] https://www.midjourney.com/
[3] https://openai.com/dall-e-2
[4] https://openai.com/dall-e-3
[5] https://platform.stability.ai/sandbox/text-to-image

from the original object, ensuring consistency in generating new objects.

We conducted a pilot study to elicit preliminary insights and feedback on the current state of our system. Through the study sessions, we garnered positive feedback regarding the efficacy of iCONTRA in assisting users to create related objects with the same concept. Additionally, participants highlighted numerous valuable possibilities for the incoming improvement version. The source code will be available at: https://github.com/vdkhoi20/iCONTRA.

Our contributions can be summarized as follows:

- We propose interactive CONcept TRAnsfer (iCONTRA), an intuitive application designed to effortlessly generate a sequence of cohesive objects with shared conceptual bases, offering a versatile and efficient tool for creative design exploration.
- We develop a novel zero-shot image editing algorithm that eliminates the need for training, allowing for the consistent and intricate generation of non-rigid object images. Our approach can preserve the object's characteristics and texture seamlessly without influencing the background.
- Insights from a pilot study with participants proficient in design suggest that our system reduces designers' effort in generating desired objects. Experimental results demonstrate the effectiveness of iCONTRA in transferring consistent and high-quality object concepts, showcasing its potential for innovative design applications.

2 RELATED WORK

2.1 Text-to-Image

The realm of text-to-image generation witnessed significant progress due to diffusion-based techniques, such as GLIDE [14], DALL·E 2 [16], and Imagen [20]. These models employed text embeddings from large language models, exhibiting the ability to generate diverse, high-quality images that aligned with intricate textual prompts. GLIDE and DALL·E 2 are conditioned on CLIP textual embeddings while DALL·E 2 generates image embeddings from the input text CLIP embedding, followed by image generation through another diffusion model. To handle high-resolution image generation, both GLIDE and Imagen generated low-resolution text-conditioned images using cascaded diffusion models. Rombach et al. [18] proposed conducting conditional text-to-image diffusion in a reduced-dimensional latent space for efficient training and sampling. Building upon LDM, Stable Diffusion [17] introduced a substantial text-to-image model, trained on a vast dataset, and made available for open research.

Recent computer graphic tools utilizing advanced generative models for design generation and editing have gained significant attention due to their impressive features and user-friendly interfaces. Midjourney represents an advanced AI image generator that opened up artistic possibilities. By accepting text prompts and utilizing a Discord bot, it facilitates the creation of detailed and high-quality graphics suitable for both personal and professional projects. DALL·E 3 , developed by OpenAI, stands out as a sophisticated image generation model. This tool assists users in creating a diverse range of visuals, from realistic images to stylized illustrations, translating textual descriptions into captivating artwork. Jasper Art [6] , a creative tool that turned ideas into visual representations, leverages cutting-edge AI technology to produce unique and one-of-a-kind art based on written input. Adobe Firefly is at the forefront of generative AI tools, aiming to revolutionize how creators, designers, and artists engaged with digital content creation. This innovative platform seamlessly transforms textual descriptions into vibrant images, turned sketches into fully realized pictures, and interpreted 3D models into stunning visuals. While existing tools generally performs well in typical cases, in this particular task, users often need to provide detailed descriptions or require visual guidance to generate the desired object, a scenario where these tools commonly face challenges.

2.2 Text-based Image Editing

While DiffusionCLIP [8] introduced step-by-step diffusion inversion for text-guided image editing, relying on diffusion model refinement, Prompt-to-Prompt [4] achieved comprehensive text-guided image editing without diffusion model refinement. It incorporated both global and local editing without predefined masks, primarily focusing on generated image editing due to the unreliability of step-by-step inversion for real images, especially with larger classifier-free guidance scales. Some methods utilized cross-attention or spatial features for editing but often preserved the original layout and struggled with non-rigid transformations. Null-text inversion (NTI) [13] proposed optimal image-specific null-text embeddings for accurate reconstruction, combined with PTP techniques for real image editing. Imagic [7], a related work, facilitated various non-rigid image editing by altering prompts directly, demanding meticulous optimization of textual embeddings and model fine-tuning, making it less user-friendly for ordinary users.

In contrast to Masactrl [1], a tuning-free method achieved complex non-rigid and consistent text-guided image editing by replacing self-attention with mutual self-attention, allowing it to query correlated local structures and textures from a source noise for consistency. Building upon the Masactrl approach, our innovative method incorporates a FiOII attention mechanism, enabling consistent and detailed non-rigid image synthesis and editing. Unlike Masactrl, our approach effectively addresses the issue of information loss caused by DIM inversion in real images. Moreover, it can alter a wide range of object attributes, such as pose, shape, and color, by simply modifying the text prompt. Remarkably, these modifications occur without any changes to the model configuration or system architecture, eliminating the need for fine-tuning.

3 PROPOSED SYSTEM

3.1 User Interface

In the development of the iCONTRA prototype, we explore several designs, including a conversation chat design like Midjourney or BingAI. Nevertheless, this design is ill-suited for our target users, who seek to create numerous designs from an original image. When entering various prompts to generate multiple designs, users may face the challenge that the back-end responses from the aforementioned tools no longer rely on information from the image users initially input but rather on the context of recent prompts. For

[6] https://www.jasper.ai/art

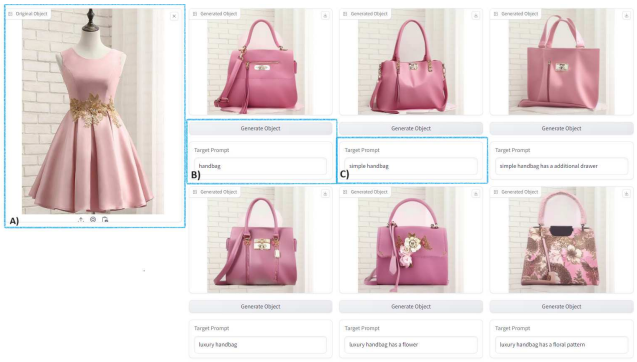

Figure 2: iCONTRA interface. a) Upload the original object image, b) Generate the Object via prompt, c) Modify the prompt until obtain the desired object.

instance, an individual employs BingAI to locate a handbag that complements her dress. She uploads the original photo of the dress and inputs numerous prompts to generate multiple designs for consideration. However, after a certain number of prompts, BingAI's responses may stray from the original image information she initially uploaded and instead be influenced by recent prompts. Consequently, she needs to repeatedly re-upload the original image to complete her task.

Therefore, we propose a web-based user interface which offers a seamless experience, ensuring simplicity and efficiency. Users can effortlessly navigate the platform by uploading the original object image on the left-hand side and providing a textual description in each cell on the right-hand side for the desired object (refer to Fig. 2). The interface comprises six cells on the right-hand side, allowing users to generate multiple objects iteratively. The generated objects are guaranteed to be based on the original image and information from the prompts that the user has entered. This leads to users avoiding the necessity of repeatedly re-uploading the original image multiple times, unlike the process with other tools implementing a conversational chat interface. After users generate their initial objects, they can choose to retain and refine them further using the convenient import and edit prompt feature. Additionally, they can observe the generated object alongside the edited prompt to achieve their desired object more effectively. For instance, if a user generates a bag and wishes to add patterns or modify its size, they can seamlessly continue the design process by importing the initial bag and refining it with additional prompts. The intuitive design and straightforward process make it accessible for both experienced designers and those with a limited design background, enhancing the overall user experience. This simplicity reduces the effort needed to find similar patterns on the internet or through other tools.

To expedite the development of a web interface for the initial iCONTRA prototype, we employed Gradio as the foundational technology to construct an interactive interface for users. This interface operates under the logic of the backend, powered by our proposed zero-shot image editing algorithm.

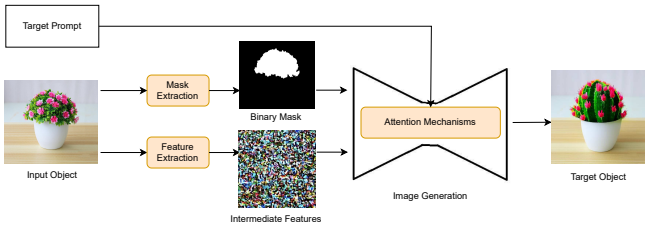

Figure 3: The proposed zero-shot image editing pipeline consists of two phases: data extraction and image synthesis.

3.2 Zero-Shot Image Editing Algorithm

The primary objective is to create a target object that conforms to the edited text prompt, maintaining the object contents from while keeping the background environment intact, all without the need for training models. Our proposed zero-shot image editing aims to accomplish spatially edited image synthesis using an input object image, a target prompt, and an automatically derived object mask, as depicted in Fig. 3. We build our algorithm on the Stable Diffusion [17] model. The input image in this task often contains only one object or a salient object, so we decide to obtain the object mask using the Rembg library in Python or by allowing the user to type additional prompts corresponding to the desired object and then automatically obtaining the object mask using a LangSam model [9]. Similar to Stable Diffusion [17], we also generate an attention mask based on the target prompt, namely target mask. In the generation phase, the algorithm controls the fade-in of original object's content by gradually interpolating intermediate features guided by the object mask and the target mask via attention mechanisms. As the result, we can edit the object without effecting to background.

We present qualitative images to showcase the performance of iCONTRA. As depicted in Fig. 4, iCONTRA demonstrates its ability to generate multiple objects with the same concept while preserving the background environments effectively. The results underscore the system's capacity to produce diverse and coherent designs, indicating its potential for creative applications in design spaces. iCONTRA can produce different shapes, colors, etc., consistent with the original object. As seen in the first row with the lamp, it seamlessly integrates the lampshade into the background, ensuring consistent, non-rigid generation across various categories and shapes.

3.3 Implementation

We apply our proposed method to the cutting-edge text-to-image Stable Diffusion model, utilizing publicly available version 1.5 checkpoints. Initially, we transform the object image into its base noise map using the deterministic inversion technique of DDIM with null-text guided because problem of image reconstruction problem [13]. During the sampling process, we employ DDIM sampling with 50 denoising iterations, and the classifier-free guidance is set at 7.5 and control with new mechanism fade-in of original object's content by gradually interpolating intermediate features guided by masks.

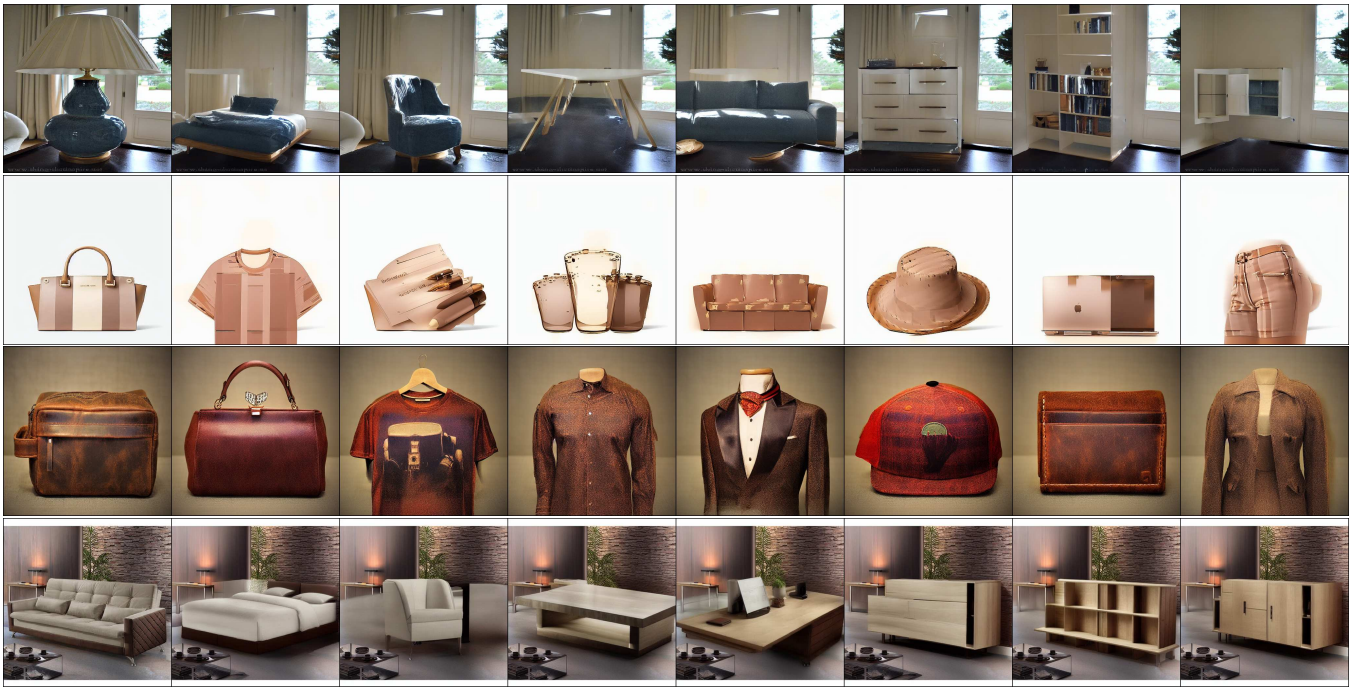

Figure 4: The first column represents the original objects (lamp, bag, wallet and sofa). The series of objects generated share the same concept, starting from the second column.

4 PILOT STUDY

We conducted a preliminary user study to explore how our system might assist users to create families of related objects, especially in comparisons with existing AI-based generation tools. This should provide us early insights on the merits and drawbacks of the system and how we can improve it in the next steps.

4.1 Participants

We invited 4 students (2 males, 2 females, average age: 19) from our local university's design club to participate in our study. They are all fresh designers and regularly join in the design activities organized by their club. Three of them are knowledgeable and have used AI tools several times, while the remaining man frequently uses the Midjourney tool to get ideas for his work.

4.2 Baseline condition and Measurement

During our research, we identified BingAI, Midjourney, and Photoshop as relevant tools with functionalities that assist users in generating images based on prompts, which are related to our work. Among these, BingAI stands out as a widely used and free tool, well-known to our participants. Notably, BingAI enables users to upload photos for queries. Consequently, we select BingAI as the baseline for comparison with our system. To assess the effectiveness of iCONTRA in aiding users to create a new object related to an existing one, we recorded the number of prompts users entered until achieving the desired result under both conditions.

4.3 Tasks

Before starting the pilot study, we group-interviewed four participants to discuss the typical tasks they are required to undertake. P1 highlighted that in the realm of car racing, it is customary to witness athletes embracing a consistent thematic approach. Additionally, he noted that individuals often acquire accessories that harmonize with the color of their cars. Meanwhile, P2 mentioned that for girls, encountering the challenge of finding a handbag that complements their dress is a common issue. To meet the number of tasks we need, we proposed exploring two topics: jewelry and interior. The consensus among everyone was in favor of these topics. Subsequently, we have designed five tasks as outlined below:

- Task 1: Generate a new ring which is related to the given watch picture.
- Task 2: Generate a new motorbike which is related to the given car picture.
- Task 3: Generate a new helmet which is related to the given motorbike picture.
- Task 4: Generate a new bag which is related to the given dress picture.
- Task 5: Generate a new bed which is related to the given sofa.

4.4 Apparatus and Procedure

Our pilot study took place in our lab, the participants performed the tasks we gave using the laptop we provided under our observation. To minize the learning effect in our study, we required two users to try our system first, while the remaining two started with

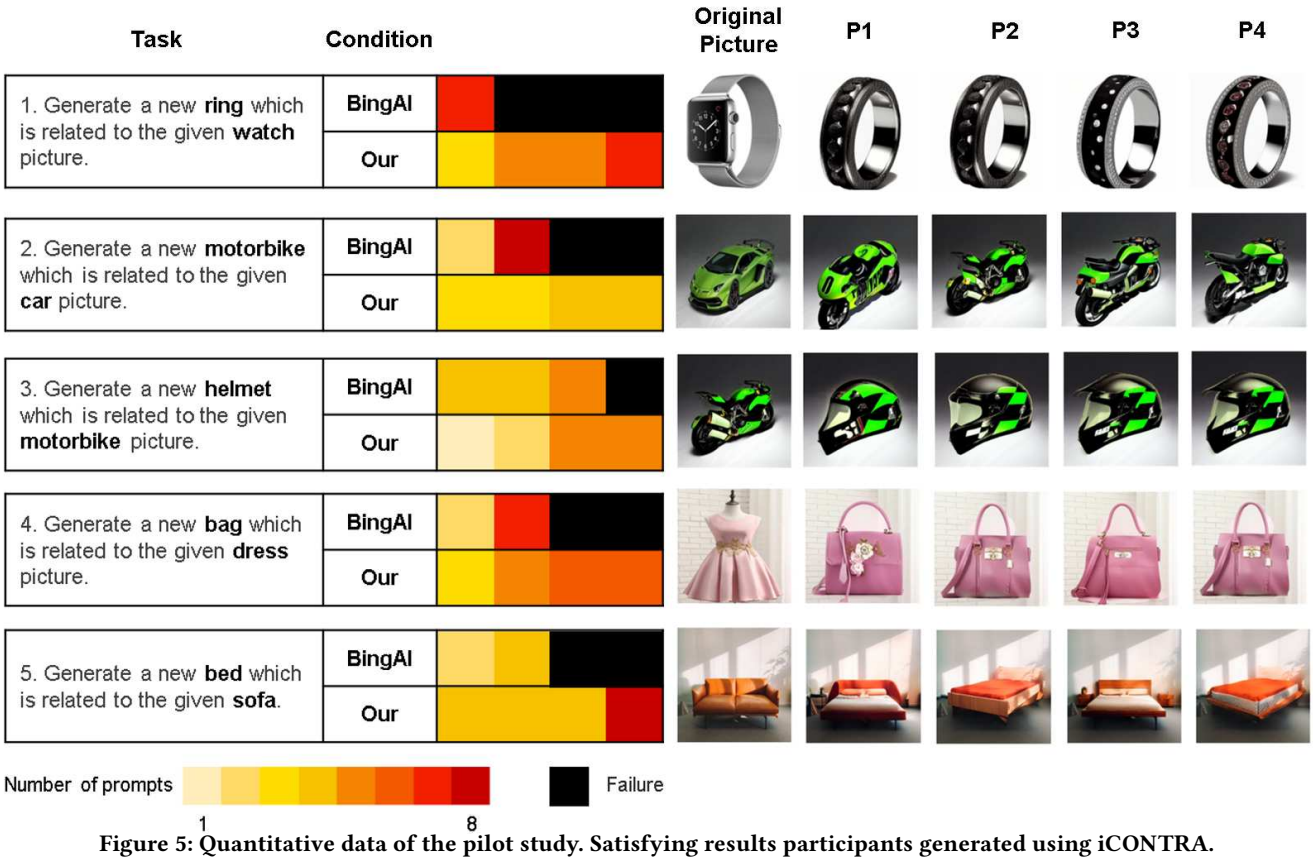

Figure 5: Quantitative data of the pilot study. Satisfying results participants generated using iCONTRA.

BingAI for their study sessions. At the end of each session, we also gathered feedback and suggestions for the future version of iCONTRA. The overall time for these study sessions was around 40 minutes (including post-interview). The whole pilot study was video-recorded for our data analysis.

4.5 Quantitative result and Qualitative feedback

The table above shows that it is difficult for participants to find designs that meet their expectations using BingAI. To explain this, users clearly outlined it in the post-interview. Participants stated that the results response from BingAI are mostly unrealistic. To elicit desired outcomes, participants often needed to provide detailed descriptions of various attributes such as colour and material for the input image. Notably, only in three cases did users achieve the desired results after making two queries with BingAI. Based on our observations, when users repeatedly input prompts and received results that deviate from the original image, they often resorted to re-uploading the initial image. Alternatively, if the expected results were not achieved after a considerable number of prompt entries (at least 8 prompts), users stopped and expressed that they failed in their attempts.

Meanwhile, all users verbally reported a satisfaction level exceeding 7 out of 10 with the final output image generated using iCONTRA. Through the utilization of iCONTRA, users found themselves requiring less effort to contemplate and describe extensive information about the original object, in contrast to their experience with the baseline tool. However, there were still many cases where users needed to enter prompts at least five times, surpassing our initial expectations. Reviewing the recorded video we found that there were many prompts for participants who wanted to change the perspective or transform a specific part of the object in the resulting image. That led users to spend more prompts on iCONTRA to complete the tasks. We discussed with participants how to tackle these problems and get more insights for the next improvements of iCONTRA. For the need to modify a specific area in the picture, participants expressed a preference for the ability to select the region for alteration before entering the prompt. They also agreed that it would be great if they could convert the output image into an editable format file in Adobe Illustrator or Photoshop.

4.6 Limitations and Failure Cases

Many prompts are required by participants who want to change the perspective or transform a specific part of the object in the resulting image. This leads users to spend more prompts on iCONTRA to complete the tasks. Although we edit the image guided by masks, we still fail when editing the specific region of the object, requiring more complex details and unable to generate a different pose, view of the generated object or retain the original object. Additionally, users provide extremely detailed descriptions of the object, but the model does not completely understand them (as illustrated in Fig. 6).

A) Clutch small handbag,chain strap,yellow chain.

B) Motorcycle racing helmet, front face shield, front view perspective,white helmet.

C) … front view…

D) Same watch,color yellow

Figure 6: Failure cases which users face. A) The user desires a handbag with a yellow chain, but the model inadvertently changes the entire handbag to yellow. B) The model currently struggles to fully understand the prompt, particularly when it includes a description of a white color helmet at the end of the input text. C) The model is unable to generate different of pose or view of object. D) The model is unable to retain the original object.

Our method inherits most of the limitations of the pretrained Stable Diffusion in generating desired images and suffers from the following main aspects. We are unable to edit the specific region of objects, even small or non-salient ones. This problem arises because the model is trained on datasets like image-captioning, where the text prompt only attends to the salient object and generates only one object corresponding to the text prompt. Another issue arises from our heavy dependence on the image layout generated from the given prompt P_t. If the Stable Diffusion model struggles to produce the desired layout or shape, our method encounters difficulties. Additionally, the dataset on which the model is trained lacks details of description, so the model is unable to understand prompt completely.

5 CONCLUSION AND FUTURE WORK

We discussed with participants how to tackle these problems and get more insights for the next improvements of iCONTRA. For the need to modify a specific area in the picture, participants expressed a preference for the ability to select the region for alteration before entering the prompt. They also agreed that it would be great if they could convert the output image into an editable format file in Adobe Illustrator or Photoshop.

We introduced iCONTRA, a system enabling users to upload an input object image and generate related objects based on that input. Insights and feedback from the pilot study suggest that iCONTRA has partially demonstrated its capability to assist users in minimizing cognitive effort and formulating prompts for generating desired objects in the mentioned task.

In our forthcoming endeavors, our focus is on enhancing the visual guidance offered to users during image manipulation. This enhancement aims to provide users with greater control and personalization options, allowing them to select specific regions they want to modify or change the perspective they want to see. By enabling users to exert more influence over the image generation and editing process, we anticipate that the resulting outputs will be better aligned with their creative vision. To change the pose, view, edit specific regions and understand completely prompt, we must fine-tune the model on a new dataset that describes more complex details of objects. Ultimately, these improvements will contribute to a more tailored and satisfying experience for users.

ACKNOWLEDGMENTS

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant Number 102.05-2023.31.

Additionally, Dr. Tam V. Nguyen is supported by National Science Foundation (NSF) under Grant Number 2025234.

REFERENCES

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22560–22570.

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

[3] Adham Elarabawy, Harish Kamath, and Samuel Denton. 2022. Direct inversion: Optimization-free text-driven real image editing with diffusion models. *arXiv preprint arXiv:2211.07825* (2022).

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626* (2022).

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189.

[6] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics* 26, 11 (2019), 3365–3385.

[7] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[8] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2426–2435.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[10] Minh-Hien Le, Chi-Bien Chu, Khanh-Duy Le, Tam V Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2023. VIDES: Virtual Interior Design via Natural Language and Visual Guidance. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 689–694.

[11] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. 2017. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036* (2017).

[12] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems* 30 (2017).

[13] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[15] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arXiv 2022. *arXiv preprint arXiv:2204.06125* (2022).

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[19] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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 (CVPR)*. 22500–22510.

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.