Multimedia Tools and Applications
https://doi.org/10.1007/s11042-025-20873-4

A novel approach with vision-language models for custom e-commerce product listings

Y Huynh Ngoc Nhu[1,2] · Quoc-Dung Nguyen[3] · Cherdsak Kingkan[2]

Received: 11 November 2024 / Revised: 6 April 2025 / Accepted: 16 April 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract
This study introduces an innovative approach to enhancing e-commerce product listings through subject-driven text-to-image generation, leveraging advanced AI technologies. Focused on transforming consumer first impressions, it blends personalized visual styles with online retail needs, striking a balance between standardization and customization. The research develops a unique method for image synthesis, improving upon existing AI models such as DreamBooth and Textual Inversion. This work not only equips online sellers with dynamic visual tools but also significantly enriches AI applications in e-commerce, offering both practical and academic contributions to the field. Our proposed model is evaluated based on various numerical and human-based evaluation metrics. The experimental results show that our model achieves a significant performance compared to other baseline models. Our model is further analyzed and discussed under correlation analysis, visual quality assessment, and ablation study to ensure its practical applicability and user satisfaction.

Keywords E-commerce · Visual styles · Vision-language models · Generative models · Human-computer interaction · Personalization in retail · Subject-driven text-to-image synthesis

1 Introduction

In the fast-evolving landscape of e-commerce, the role of Artificial Intelligence (AI), particularly Generative AI, has become increasingly pivotal. This study delves into the intersection of e-commerce and AI, with a focus on Generative AI's application in subject-driven text-to-

✉ Quoc-Dung Nguyen
dung.nguyen@vlu.edu.vn

Y Huynh Ngoc Nhu
y.huynh@deakin.edu.au

Cherdsak Kingkan
cherdsak.kingkan@gmail.com

[1] Applied Artificial Intelligence Institute (A2I2), Deakin University, Burwood, Melbourne, Australia

[2] School of Engineering and Technology, Asian Institute of Technology, Khlong Nueng, Thailand

[3] Faculty of Mechanical - Electrical and Computer Engineering, School of Technology, Van Lang University, Ho Chi Minh City, Vietnam

Published online: 30 April 2025

image generation. While existing methodologies like Textual Inversion [1] and DreamBooth [2] represent significant advancements in this field, they are not without limitations. Textual Inversion, for instance, demands extensive fine-tuning time and often yields images with poor fidelity. Similarly, DreamBooth, despite its potential, is constrained by high fine-tuning time requirements.

AI's integration into e-commerce has predominantly revolved around modifying specific areas of an image [3], such as the background, without facing significant challenges in subject fidelity. However, this approach is limited in scope and fails to address the comprehensive needs of image generation in e-commerce settings.

As far as we know, no studies using AI technologies, especially using recent generative models, have been proposed for creating realistic and diverse synthetic images for displaying products on e-commerce platforms.

This study introduces a novel approach to subject-driven text-to-image generation, extending beyond mere modifications to encompass the creation of entire images. Our methodology integrates the vision-language model BLIP-2 [4] with Stable Diffusion [5], aiming to enhance both fine-tuning time and fidelity metrics. This approach not only improves subject fidelity but also facilitates the generation of complete images, marking a significant advancement in the application of Generative AI in e-commerce.

Furthermore, we introduce a set of internal metrics, more extensive than those used in previous works, to aid in the selection and upscaling of images. This is crucial to meet the evolving requirements of e-commerce platforms, which increasingly demand high-quality, high-resolution product visuals.

To comprehensively evaluate our model, we adopt a dual assessment approach, combining these new internal metrics with empirical Human-Computer Interaction (HCI) experimentation. This blend of quantitative and qualitative evaluations allows us to thoroughly assess the model's effectiveness, both in terms of technical performance and user-centric experience, paving the way for more advanced and practical applications in e-commerce image generation.

In summary, this study makes the following main contributions:

- We introduce a pioneer work in subject-driven text-to-image generation for e-commerce product listings.
- We propose an efficient model combining two state-of-the-art vision-language models BLIP-2 and Stable Diffusion that improves fine-tuning time and image fidelity. We also employ two text prompt templates for mitigating overfitting and optimizing fine-tuning time of the proposed model.
- We provide a strategic approach and guidance for diverse and representative dataset selection and preparation.
- We suggest varied key performance indicators for evaluating our proposed model from both quantitative and qualitative perspectives as well as comparing it with other baseline models.
- Under efficient settings of the model parameters, our proposed model outperforms the baseline models on the various performance metrics.

The rest of the paper is structured as follows. Section 2 provides an overview of the recent AI technologies applied to e-commerce and the generative models. In Section 3, our proposed subject-driven text-to-image generation model and its architecture are presented. Section 3 also describes the dataset, the selection criteria, and the data preprocessing steps in detail. The various evaluation metrics with respect to both computational assessments and human judgments are provided in Section 4. In this section, the experimental settings are shown and the experimental results of the proposed model are demonstrated and compared to the other baseline models. Section 5 discusses the impact of input image diversity and the image selection criteria based on the Image-Text Matching score to further analyze subject fidelity, prompt fidelity, and visual integrity of the generated images. Finally, Section 6 gives conclusions and future work.

2 Related works

2.1 AI in e-commerce

E-commerce, defined as the buying and selling of products and services via the Internet [6], has become increasingly significant for businesses. This is largely driven by growing customer demand for online services and the competitive advantages it offers [7]. The integration of rapidly evolving information technology, which is both easily adaptable and cost-effective, has led to a competitive e-commerce landscape. Businesses are compelled to continuously adapt their business models to meet evolving customer needs [7, 8].

Artificial Intelligence (AI) represents the latest technological advancement impacting e-commerce. It is revolutionizing the industry with its ability to accurately interpret external data, learn from it, and use the acquired knowledge to achieve specific goals and tasks through adaptive flexibility [9]. AI in e-commerce can be defined as the application of AI techniques, systems, tools, or algorithms to support the activities related to buying and selling products or services online [10]. It presents opportunities for businesses to gain competitive advantages by leveraging big data to uniquely meet customer needs through personalized services [11].

Current AI applications [12] in e-commerce primarily focus on recommendation systems [13–15]. Key research themes include sentiment analysis [16, 17], optimization related to the accuracy of recommendation systems and predictions [18, 19], trust and personalization [20, 21], and the technologies and tools of AI [22, 23].

Our review indicates a lack of research utilizing AI, particularly recent generative models, in creating realistic and diverse synthetic images for e-commerce product displays. This area remains unexplored, signifying a potential opportunity for innovative applications of AI in enhancing the e-commerce experience.

2.2 Generative models

In the realm of artificial intelligence and machine learning, generative models stand as a pivotal innovation, particularly in their application to e-commerce, where the visual presentation of product listings plays a crucial role in influencing consumer behavior. This literature review explores various approaches, ranging from alternative segmentation methods beyond Mask R-CNN to the advancements in generative models, emphasizing their role in text-to-image synthesis to enhance e-commerce product listings.

The Mask R-CNN architecture [24] has been a cornerstone in the field of instance segmentation, providing robust performance in delineating object boundaries in images. In the realm of multimodal learning, the authors of [25] introduce a multimodal model capable of understanding diverse commerce topics through image-text pairs. The ability to generalize across various tasks, including image-to-product retrieval, is critical for e-commerce platforms aiming to provide personalized recommendations. While Mask R-CNN remains a powerful tool for segmentation tasks that could enhance e-commerce product listings, the integration of multimodal learning frameworks further enriches this landscape, enabling a more nuanced understanding and representation of e-commerce products. As the field continues to evolve, it is essential to explore these approaches to optimize the effectiveness of vision-language models in e-commerce.

Proposed by Goodfellow et al. in 2014, Generative Adversarial Networks (GANs) have revolutionized generative modeling [26]. The GAN framework comprises two neural networks: a generator that synthesizes data and a discriminator that differentiates between real and generated data. These networks engage in an adversarial process, where the generator aims to produce data indistinguishable from real samples. By 2019, advancements in GANs enabled the generation of high-resolution, detailed images, as demonstrated in "Large Scale GAN training for High fidelity natural image synthesis" [27]. Despite their remarkable success, GANs face notable challenges. Google has identified issues like vanishing gradients, where an overly strong discriminator hinders generator improvement, as well as persistent problems such as mode collapse and convergence difficulties.

Recent developments in neural module networks (NMNs) have demonstrated the potential for explicit multi-hop reasoning through handcrafted neural modules, which can be combined to achieve complex visual reasoning tasks [28, 29]. This modular approach not only enhances interpretability but also allows for the efficient processing of visual data. However, most NMN networks suffer from two major drawbacks, scalability and generalizability, due to task-specific training. To avoid these drawbacks, visual programming, particularly in the context of compositional visual reasoning, has gained traction as a method that allows the interpretation and generation of visual data [30]. The concept of compositionality in visual reasoning posits that complex visual tasks can be decomposed into simpler, interpretable components. By leveraging these compositional visual reasoning techniques, e-commerce platforms can dynamically generate product images and descriptions that are tailored to specific consumer preferences, thereby improving user engagement and conversion rates.

As GANs gained prominence, another approach, known as Denoising Diffusion Probabilistic Models, emerged in 2020. This method reinterprets data generation as a gradual denoising process, where an original data sample is progressively corrupted with noise and then reconstructed [31]. The 2021 paper "Improved Denoising Probabilistic Models" further refined this technique, enhancing performance metrics and stability [32]. By mid-2021, certain benchmarks indicated that diffusion models outperformed GANs in specific image synthesis tasks [33].

Recent advancements have shifted focus to text-to-image synthesis, transforming textual descriptions into corresponding images. Models like DALL-E2 [34], which integrates CLIP latent for improved text-to-image conversion, and Stable Diffusion [5], which synthesizes high-resolution images, have gained prominence. Imagen [35], combining deep language understanding with diffusion models for photorealistic images, is another notable contribution. However, these models often lack precise control over the generated images, particularly in maintaining consistency in subject identity across different images.

In addition to these advancements, the exploration of semantic control in image synthesis by Liu et al. [36] further illustrates the potential for fine-grained control in text-to-image generation. Their unified framework allows for diverse applications, including text-guided and image-guided synthesis, which require specific visual styles or attributes.

Significant progress has been made in subject-driven text-to-image generation with methodologies like Textual Inversion [1] and DreamBooth [2]. Textual Inversion, utilizing Stable Diffusion, dynamically adjusts token embeddings for contextual image synthesis, but struggles with maintaining subject fidelity and requires extensive fine-tuning. DreamBooth, on the other hand, personalizes diffusion models for specific image generation needs and shows better performance in subject fidelity, though it also demands considerable fine-tuning time.

In the broader context of AI in marketing, our study stands out by focusing on complete image generation, diverging from methods that primarily enhance specific image areas. We aim for dynamic image synthesis, generating subjects with various movements and interactions, thus offering a flexible and innovative solution for e-commerce visual presentations. This literature review thus sets the groundwork for our exploration into custom visual styles in e-commerce, leveraging the latest advancements in AI and generative models.

3 Proposed model and dataset

3.1 Proposed model

In this study, we introduce BLIP2Booth, a novel vision-language model that builds upon the foundation of DreamBooth and integrates the BLIP-2 model. Our approach is designed to address the challenges of high fine-tuning time and limited generalization in existing models, while enhancing image fidelity for e-commerce applications. In our proposed model, we implement a three-phase approach to enhance the visual representation of e-commerce products:

- Prompt Generation by BLIP-2: The first phase utilizes the BLIP-2 model for generating detailed text prompts based on input images. This process involves creating descriptive, context-aware captions that accurately reflect the content and setting of each product image. These generated prompts are crucial as they are paired with the corresponding input images for the next phase.
- Enhanced DreamBooth with Stable Diffusion Integration: In this phase, the prompts generated by BLIP-2, coupled with the input images, undergo fine-tuning in the DreamBooth framework integrated with Stable Diffusion version 2.1. This combination addresses the challenges of lengthy fine-tuning times and limited generalization capabilities seen in traditional models. By leveraging the strengths of both DreamBooth and Stable Diffusion, we achieve high-quality image generation that is particularly tailored for e-commerce applications. The output from this phase is a set of refined images, each with a resolution of 768 x 768 pixels, ready for the final enhancement step.
- Image Upscaling with ESRGAN [37]: The concluding phase involves upscaling the generated 768 x 768 pixel images to a higher resolution of 2048 x 2048 pixels using Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN). This upscaling is essential for meeting the demanding visual standards of e-commerce platforms, ensuring that the product images are not only of high fidelity but also of a resolution that enhances their visual appeal and effectiveness in an online retail environment.

3.2 Model architecture

In alignment with our proposed model structure, the BLIP2Booth architecture (illustrated in Fig. 1) initiates with the BLIP-2 model for prompt generation. This initial phase involves BLIP-2 creating detailed, context-aware captions from images using a custom text prompt template. For example, given an image of a sneaker, the prompt might be structured as, "Question: Based on the image, describe the context starting with ’a sneaker’?", BLIP-2 responds with a descriptive caption like "a sneaker sitting on the top of a rock". This caption

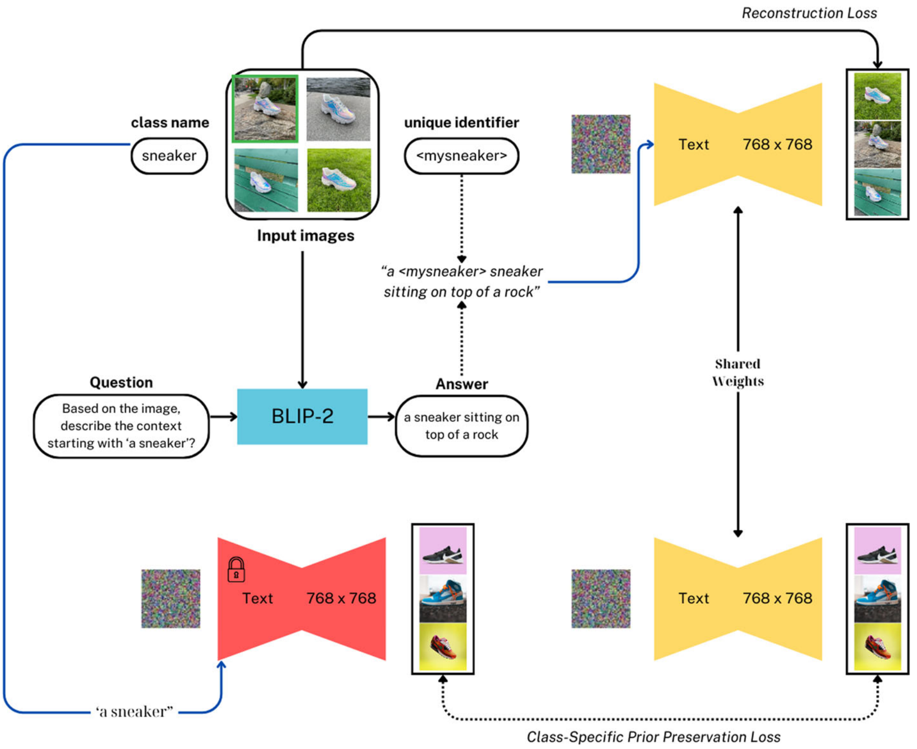

Reconstruction Loss
class name
sneaker
Input images
unique identifier
<mysneaker>
"a <mysneaker> sneaker sitting on top of a rock"
Text
768 x 768
Shared Weights
Question
Based on the image, describe the context starting with 'a sneaker'?
BLIP-2
Answer
a sneaker sitting on top of a rock
Text
768 x 768
Text
768 x 768
'a sneaker"
Class-Specific Prior Preservation Loss

Fig. 1 Fine-tuning of the BLIP2Booth (our proposed method)

is then enriched with a unique identifier "<mysneaker>", forming a comprehensive prompt "a sneaker <mysneaker> sitting on the top of a rock" for the next phase.

In the subsequent phase, these refined prompts, in conjunction with their associated images, are fed into an enhanced DreamBooth framework integrated with Stable Diffusion version 2.1 for image synthesis. The integration of DreamBooth and Stable Diffusion, along with the incorporation of class-specific preservation loss, significantly augments the output diversity and mitigates language drift.

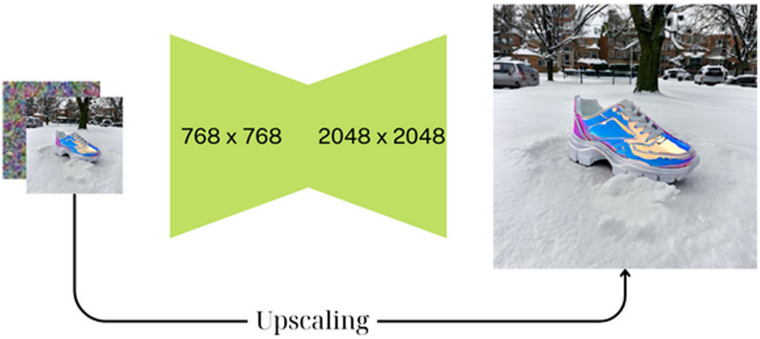


Fig. 2 Super-Resolution Components: Image Upscaling with ESRGAN

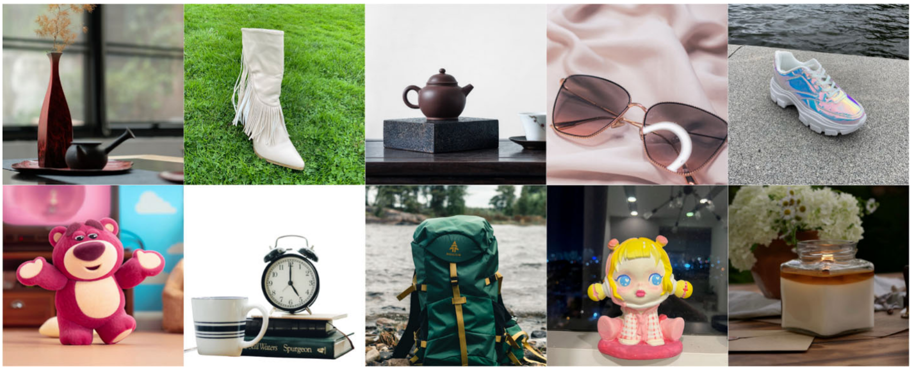


Fig. 3 Dataset. Example images for each subject in our dataset

The final phase of our model involves the crucial step of image upscaling (Fig. 2). Post the generation of images sized at 768 x 768 pixels, we employ the ESRGAN technique to upscale these images to a higher resolution of 2048 x 2048 pixels. This upscaling process utilizes pairs of low and high-resolution images from our dataset for fine-tuning, ensuring that the final product images not only align with but also surpass the resolution standards of e-commerce platforms, enhancing their utility and appeal in online retail environments.

3.3 Dataset

The dataset serves as the foundation of this study, crucial for model training and fine-tuning. We employed a strategic approach to assemble a dataset that showcases diversity and representativeness across various product categories. This selection was guided by the need to reflect the multifaceted nature of consumer goods, including aspects such as form and texture, to rigorously test the model's adaptability and performance.

3.3.1 Data selection

Our dataset encompasses ten distinct products (Fig. 3), categorized as follows:

- Fashion: This category includes a variety of items like sneakers, boots, sunglasses, and backpacks.
- Toys and Hobby: It consists of a teddy bear (specifically Lotso), a doll toy, and a candle.
- Furniture and Home Decor: This group features a clock, vase, and teapot.

The selection process was influenced by consumer trends reported in the Statista's 2022 report (Fig. 4), which analyzed e-commerce spending patterns in Vietnam. As Vietnamese researchers, we focused on products that reflect the purchasing behavior of local online consumers. However, the principles of our model design remain applicable beyond a single market.

In addition to ensuring consumer relevance, our dataset was designed to evaluate key challenges in subject-driven generation models:

- Complex objects like dolls and teddy bears were included to assess the model's ability to preserve subject identity, particularly for objects with human-like facial features - a known challenge in generative models.

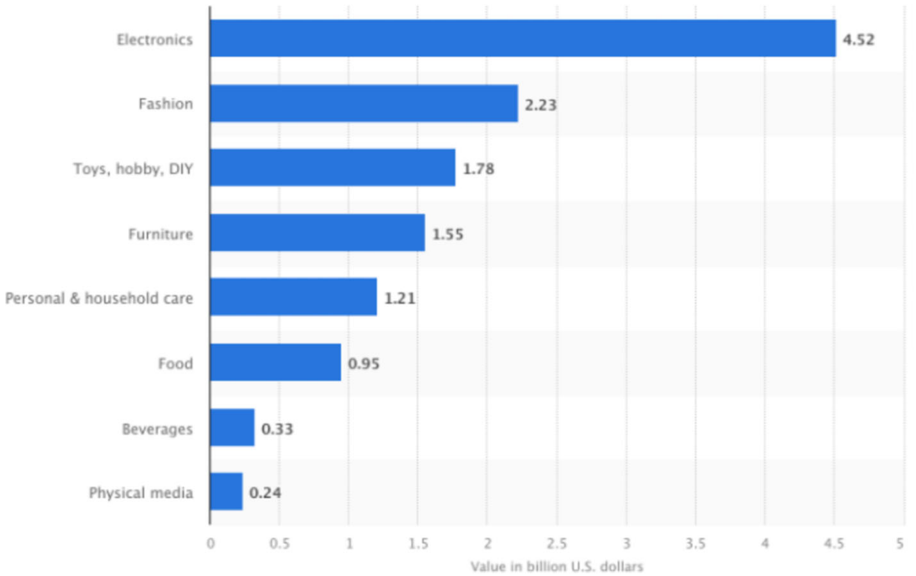


Fig. 4 E-commerce expenditure on consumer goods among internet users in Vietnam in 2022, by category [38]

- A diverse range of shapes and textures was incorporated to examine how well the model generalizes across different product types.
- Some items were sourced from DreamBooth's dataset (e.g., sneakers, vases, and teapots) to facilitate direct comparisons with prior work.

To maintain authenticity and variability, the images for the teddy bear, doll toy, candle, clock, and sunglasses were either captured by the author or sourced from https://unsplash.com. The remaining items - sunglasses, sneakers, boots, vases, and teapots - are from the dataset used in DreamBooth (https://github.com/google/dreambooth), offering a baseline for model comparisons.

For each product, 5-7 images were collected, capturing various angles and lighting conditions. This approach ensures a detailed representation of each product's attributes, which is crucial for effective model development and validation.

While this study focuses on e-commerce in Vietnam, future research will expand dataset selection to incorporate globally representative product categories, ensuring broader applicability of our approach.

3.3.2 Data preprocessing

Preprocessing is a critical step in preparing the dataset for effective model training and high-quality image generation. The following preprocessing steps were implemented:

- Image Cropping We cropped all input images to a standard size of 768 x 768 pixels. This uniformity in image size centers the focus on the product and creates a consistent input format for the model.
- Class Name Specification Each product was assigned a specific class name, such as 'sunglasses', 'clock', 'vase', 'teapot', 'sneaker', 'boot', 'backpack', 'teddy bear', 'doll', and 'candle'. These names categorize the images and facilitate model training.

- Unique Identifier Assignment For personalized model tuning, unique identifiers were assigned to each product class. Identifiers like <mysunglasses>, <myclock>, etc., are crucial for subject-driven text-to-image generation. They assist in the precise fine-tuning of the model for individual subjects in the dataset.

4 Evaluation metrics and experimental results

4.1 Evaluation metrics

In evaluating the performance of our generative models, we emphasize the importance of fidelity in both the subjects of the images and their alignment with the given prompts. Our evaluation strategy revolves around a set of meticulously chosen metrics, each offering unique insights into the models' capabilities.

4.1.1 Internal metrics

Cosine Similarity

Cosine similarity is a widely used metric for evaluating image feature similarity, particularly in subject-driven text-to-image generation models. Previous works, including Textual Inversion [1] and DreamBooth [2], have leveraged cosine similarity to measure the alignment between generated images and their original subjects, as well as prompt fidelity validation. These studies applied cosine similarity to various aspects of evaluation, including both subject fidelity and text-image alignment.

As a fundamental metric, we utilize cosine similarity to gauge the likeness between high-dimensional vectors, particularly those representing features extracted from images. This metric calculates the cosine of the angle between two vectors $\mathbf{A}$ and $\mathbf{B}$ is defined as:

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

where $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (or norms) of the vectors $\mathbf{A}$ and $\mathbf{B}$, respectively. Thus, we employ cosine similarity to evaluate two key metrics - DINO and CLIP-I. The cosine similarity metric effectively quantifies the degree of alignment or correlation between two vectors, with values ranging from 0 to 1, where a higher value indicates a greater degree of alignment.

Subject Fidelity

This evaluates how well the model preserves the core characteristics of the subject in the generated images. We use the following metrics:

- DINO [39]: This metric calculates the average pairwise cosine similarity between the Vit-S/16 DINO embeddings of the generated and original images. It is particularly effective in identifying unique features of the subjects.
- CLIP-I: It measures the average pairwise cosine similarity between CLIP [40] embeddings of the generated and original images, denoted as CLIP-I, providing insights into the visual integrity of the subject in the generated images.

Prompt Fidelity

This assesses the alignment of generated images with the provided text prompts. The metrics used are:

- CLIP-T: This represents the average cosine similarity between the embeddings of the image and the associated prompt as interpreted by CLIP [40], denoted as CLIP-T.
- Image-Text Matching (ITM) Score [41]: Derived from the BLIP model, ITM is a binary classification metric evaluating the direct matching between an image and its associated text. A high ITM score indicates a precise and contextually relevant connection between the image and text.
- Image-Text Contrastive (ITC) Score [41]: Also from the BLIP model, ITC measures the contrastive alignment between image-text pairs. It quantifies the model's ability to differentiate relevant pairs from irrelevant ones. Higher ITC scores signify a stronger model discriminative capability, reflecting effective correlation of images with accurately matched texts. Both ITM and ITC metrics range from 0 to 1.

Limitations of Cosine Similarity

While cosine similarity is widely used, it has known limitations in high-dimensional spaces, including potential biases and inconsistencies when comparing complex image embeddings. To mitigate these challenges, we incorporate Human-Computer Interaction (HCI) evaluations to ensure that our similarity metrics align with human perception.

4.1.2 Human-computer interaction (HCI) evaluation

The HCI evaluation focuses on comparing our proposed BLIP2Booth (BB) model with the established DreamBooth (DB) model. This study aims to assess the models' performance in terms of subject and prompt fidelity based on human evaluations and to explore the relationship between these evaluations and internal model metrics.

The HCI evaluation is designed as follows:

The study involved a diverse demographic group aged 20-50, encompassing all genders and varied levels of expertise with e-commerce platforms. Procedure: Participants evaluated images generated by both BB and DB models. The evaluation was divided into two tasks:

- Task 1 - Subject Fidelity: Participants rated images on a Likert scale (1-5) based on how well they preserved the subject's identity and characteristics.
- Task 2 - Prompt Fidelity: Participants ranked images on their alignment with the given text prompts.

The ratings from both tasks were normalized to a range of (0,1), denoted as `task1_normalized` and `task2_normalized`, to facilitate a consistent and comparative analysis.

A mixed-design approach was employed, featuring:

- Within-subject variables: Model and Prompt.
- Between-subject variables: Product category.
- Order Effect: The sequence of conditions was randomized to prevent bias.

Participants reviewed 8 images per product, covering 4 prompts (2 general, 2 specific) for both models. It took approximately 4 minutes per participant, including 60 seconds for instruction review. Each image was evaluated within a 10-second window.

4.1.3 Ablation studies

Our research includes ablation studies aimed at assessing the impact of specific model components and configurations on performance. This analysis focuses on optimizing the image generation process and involves two key variables:

Training Steps for Text Encoder

We investigate the effect of the number of training steps on the text encoder's output quality. The study compares the standard DreamBooth approach of numerous training steps against our proposed reduction to 200 steps. This evaluation will help determine if fewer steps can still achieve results comparable to or better than the traditional method.

Prompt Specificity

The role of prompt specificity is critically examined to understand its influence on the model's performance (details in Appendix A). We analyze how general versus specific prompts affect the scores in subject and prompt fidelity. This will clarify the impact of prompt granularity on the model's capacity to generate images that accurately represent both the subject and the context of the prompt.

These variables are tested against a set of internal metrics, including CLIP-I, DINO, CLIP-T, ITM, and IMC, to measure their effects on image generation quality. The outcomes of these ablation studies are integral to refining our model's performance and enhancing its overall efficiency.

4.2 Experimental settings

To ensure reproducibility and assess the feasibility of our proposed approach, we outline the computational resources and training protocols used in our study.

Training Platform and Hardware

Our experiments were conducted on Google Colab, utilizing an NVIDIA A100 GPU. This choice was driven by the high computational demands of BLIP-2 and Stable Diffusion models, particularly for subject-driven fine-tuning tasks. The training time for our model remains within practical constraints, making it feasible for real-world deployment.

BLIP-2 Implementation

The pre-trained BLIP-2 captioning model serves as a pivotal component in our pipeline, generating descriptive text prompts that are later fed into the Stable Diffusion model. This step requires significant computational power due to the need for efficient vision-language embedding generation.

Stable Diffusion Training Protocol

The fine-tuning of Stable Diffusion was conducted with approximately 1500 UNet [42] training steps, using a learning rate of 2e-6. To mitigate overfitting, particularly for simpler subjects, we adjusted the number of training steps accordingly. The text encoder was fine-tuned separately for 200 steps at a learning rate of 1e-6.

Inference

Table 1 Comparison of Training Efficiency Between DreamBooth and BLIP2Booth

Model	Training Steps	Learning Rate	Training Time (A100 GPU)	Dataset Size
DreamBooth	~1000 iterations	5×10^{-6}	~5 minutes	3-5 images
BLIP2Booth (Ours)	~1500 UNet steps, 200 Text Encoder steps	2×10^{-6} (UNet), 1×10^{-6} (Text Encoder)	**Less than 5 minutes** (adaptive tuning)	5-7 images

These bold entries indicate the best-performing model

During inference, we employed *DPMSolverMultistep* [43] with Stable Diffusion, setting a default step count of 50 for all products, as per DreamBooth's recommendations. To optimize computational time for less complex subjects, we reduced the step count to around 20, thus enhancing efficiency without compromising image quality.

***Computational Efficiency and Comparision with DreamBooth* [2]**
A key advantage of our BLIP2Booth model over DreamBooth is its optimized training efficiency while maintaining high subject fidelity (Table 1):

4.3 Experimental results

4.3.1 Internal metric results

The internal metrics results reveal that BB demonstrates superior subject fidelity. This is evidenced by its highest mean scores in both DINO and CLIP-I metrics, coupled with the lowest variability (Table 2), suggesting its superior capability in maintaining subject integrity and the stability of this particular generative model. DB follows closely, indicating competitive performance. Textual Inversion (TI), while lagging slightly, shows potential in certain aspects of image generation. In prompt fidelity, the metrics show a closer performance among the models, with minor variations in CLIP-T and ITC scores (Table 3).

In our analysis, we employed bar charts (Fig. 5), box plots (Fig. 6), and histograms (Fig. 7) to compare the performance of BB, DB, and TI across various fidelity metrics. The bar charts revealed that BB excels in subject fidelity (CLIP-I and DINO), while DB slightly edges out in prompt fidelity (ITC), with TI generally trailing in most metrics. Box plots indicated a consistent range of high scores for BB in subject fidelity, showcasing less

Table 2 Subject fidelity scores across models

Model	Metric	Mean	Std Deviation	Min	Max
BB	DINO	**0.778**	0.065	0.541	0.973
DB	DINO	0.710	0.074	0.471	0.973
TI	DINO	0.663	0.098	0.369	0.921
BB	CLIP-I	**0.883**	0.051	0.622	0.978
DB	CLIP-I	0.860	0.066	0.591	0.979
TI	CLIP-I	0.836	0.113	0.381	0.973

These bold entries indicate the best-performing model

Table 3 Prompt fidelity scores across models

Model	Metric	Mean	Std Deviation	Min	Max
BB	CLIP-T	0.242	0.028	0.158	0.331
DB	CLIP-T	0.239	0.031	0.143	0.328
TI	CLIP-T	0.238	0.033	0.159	0.332
BB	ITM	0.495	0.498	0.000	1.000
DB	ITM	0.494	0.500	0.000	1.000
TI	ITM	0.497	0.499	0.000	0.999
BB	ITC	0.415	0.046	0.232	0.518
DB	ITC	0.424	0.044	0.250	0.529
TI	ITC	0.385	0.064	0.197	0.545

variability compared to TI. However, a broader spread in prompt fidelity scores across all models suggested influence by factors of greater variability. Histograms further highlighted this trend, with BB showing a skewed distribution towards higher scores in subject fidelity metrics, whereas the prompt fidelity metrics displayed more variation and overlap among the models. These findings suggest BB's robust performance in subject fidelity but also point towards potential areas for improvement in prompt alignment for all models.

4.3.2 HCI results

Subject and Prompt Fidelity

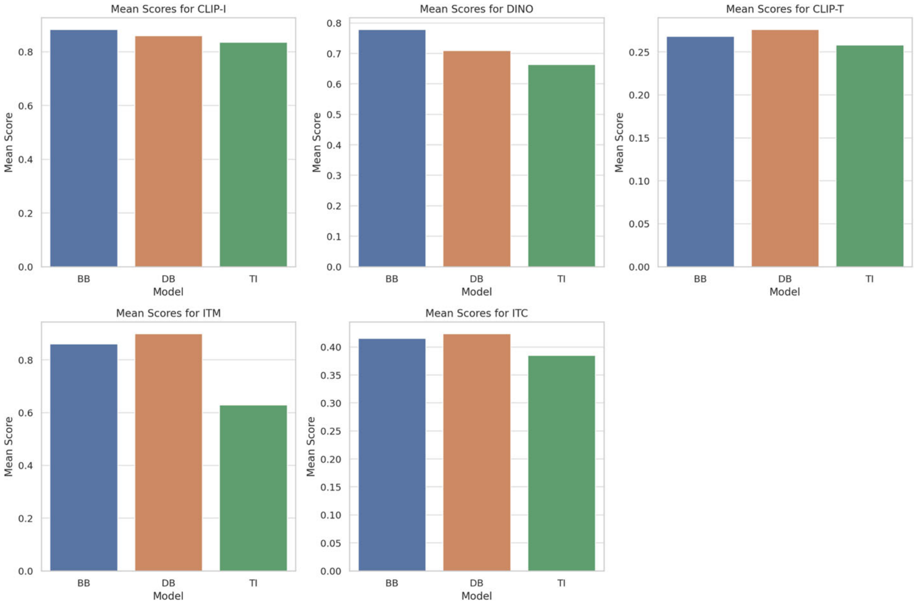


Fig. 5 Mean Scores for all scores by Model

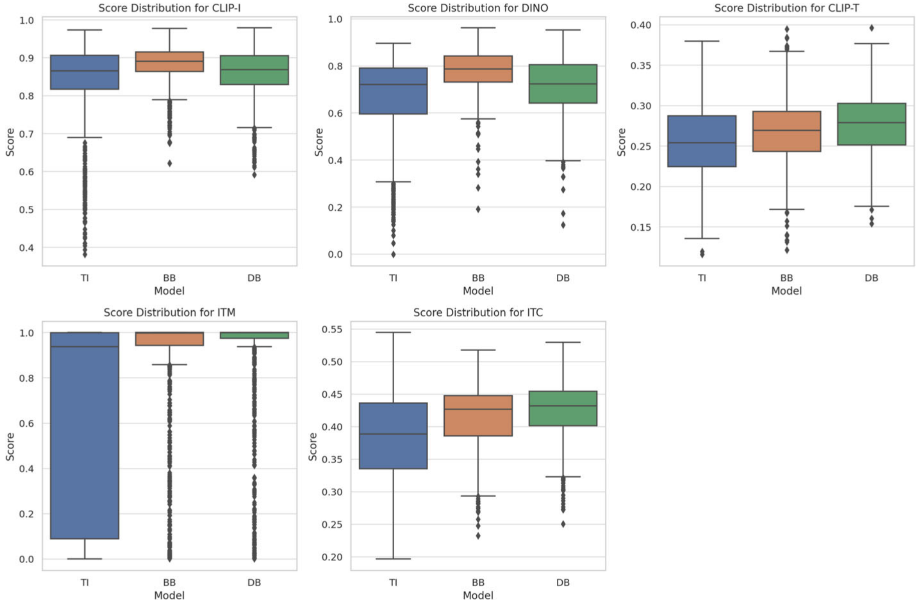


Fig. 6 Score Variability for all scores Across Models

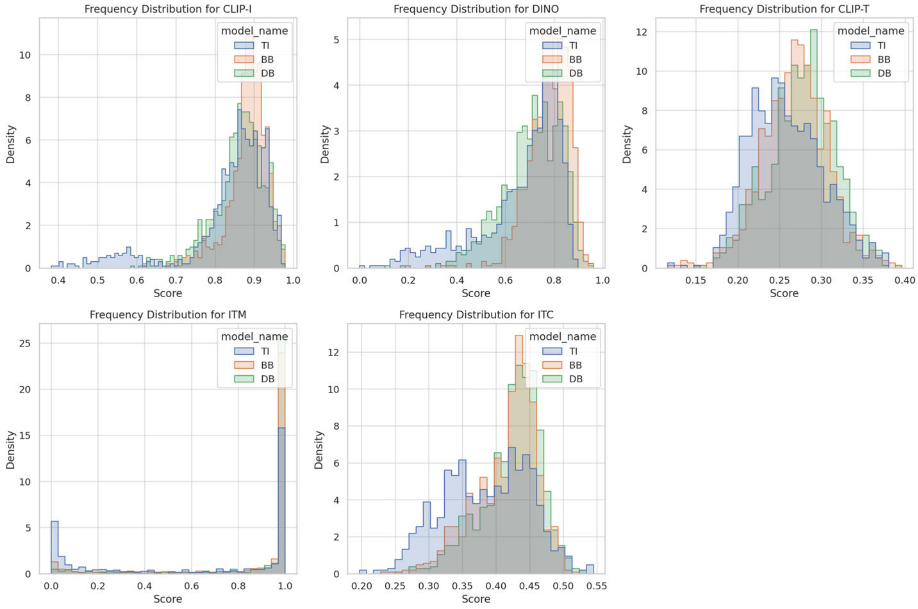


Fig. 7 Frequency Distribution for 5 Scores Across Models

Table 4 HCI Average Normalized Fidelity Scores Across Models

Model	subject fidelity	prompt fidelity
BB	**0.7603**	**0.8351**
DB	0.6286	0.8262

These bold entries indicate the best-performing model

Table 4 shows that BB received higher average scores in subject fidelity compared to DB, indicating a preference for BB in preserving subject characteristics. Both models showed similar performance in prompt fidelity, with BB having a slight advantage.

Correlation Analysis

Our analysis aims to uncover the relationship between human judgments and the metrics generated by our model, including CLIP-I, DINO, CLIP-T, ITM, and ITC. This understanding is crucial to evaluate how well the computational assessments align with human perceptions and preferences. The method employed in our analysis was the use of Spearman correlation coefficients, as depicted in Fig. 8, which enabled us to evaluate monotonic relationships between the computational assessments generated by our model and human evaluations.

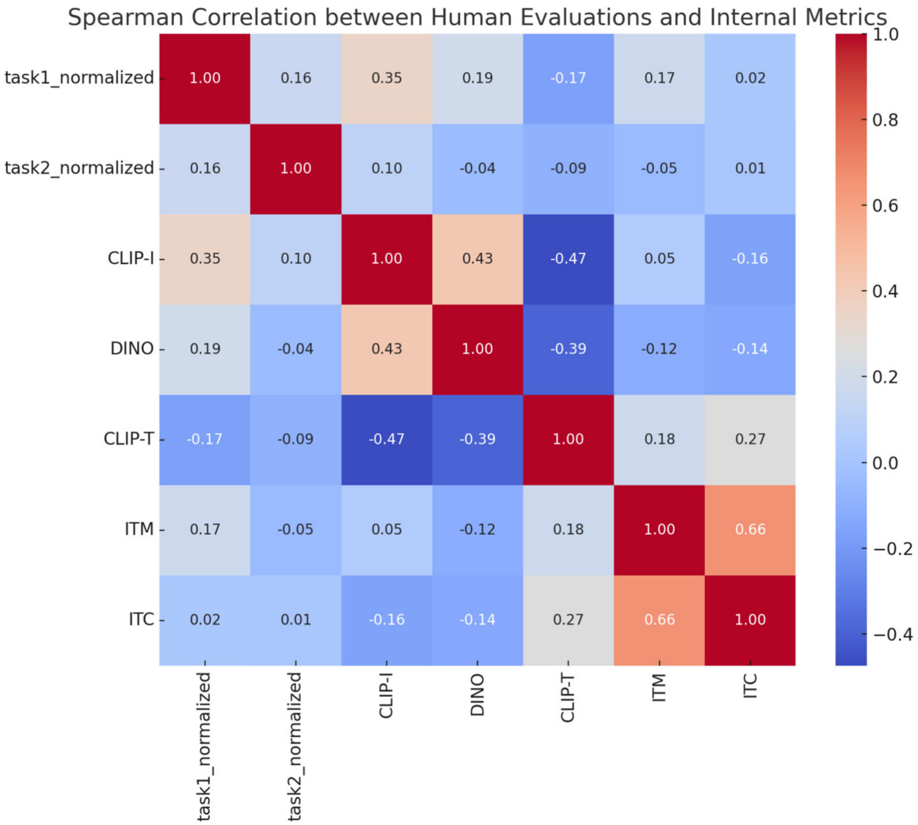


Fig. 8 Spearman Correlation between Human Evaluation and Internal Metrics

Our observations from this analysis highlighted several key relationships:

- Subject Fidelity: We observed a moderate positive correlation with CLIP-I (0.35). However, the correlations with DINO and ITM were weaker (0.19 and 0.17 respectively), and there was a slight negative correlation with CLIP-T (-0.17).
- Prompt Fidelity: In this category, correlations were generally weaker. Notably, there was a slight positive correlation with CLIP-I (0.1).
- Relationship Among Internal Metrics: A moderate positive correlation was observed between CLIP-I and DINO (0.43). Additionally, a strong correlation was noted between ITM and ITC (0.66).

The positive correlation observed can be attributed to the fact that both CLIP-I and DINO metrics are employed to evaluate subject fidelity, while ITM and ITC metrics focus on assessing prompt fidelity. This alignment in their respective purposes contributes to the observed correlation.

Figure 9 is the scatter plot visualizing the relationship between `task1_normalized` and CLIP-I (the p-value is approximately 4.67×10^{-7}). This plot illustrates the moderate positive correlation between the human evaluation of subject fidelity and the internal CLIP-I metric, as indicated by the upward trend in the scatter of points.

The implications of these findings are significant. The HCI study underscored that while internal metrics are instrumental in gauging model performance, they do not entirely correspond with human evaluations. This disparity underscores the necessity of integrating human perspectives into the assessment of generative models, ensuring their relevance and applicability in practical scenarios. The insights gleaned from this correlation analysis are crucial for understanding the model's performance from a human-centric viewpoint and will be invaluable in guiding future enhancements and research directions.

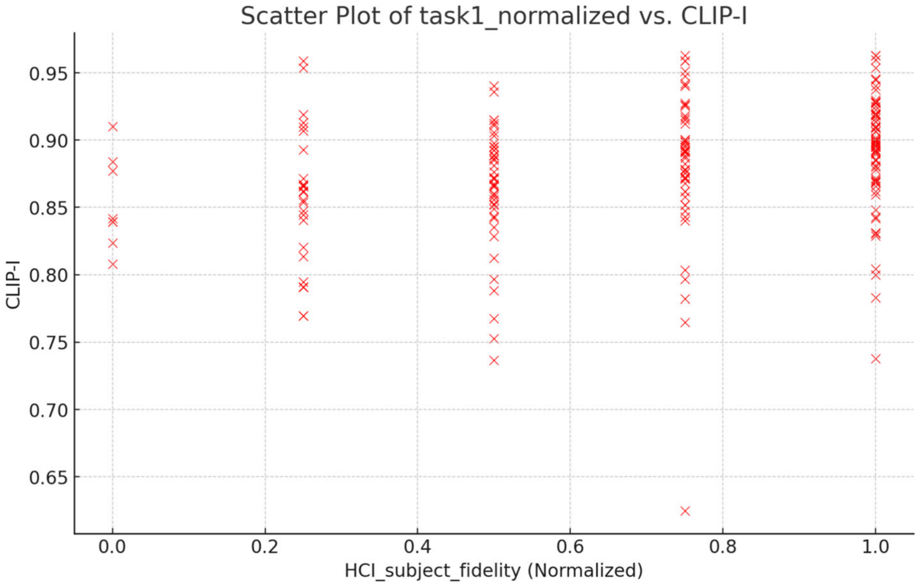


Fig. 9 Scatter Plot of Human Evaluation (Subject Fidelity - Task 1) vs. CLIP-I

4.3.3 Qualitative results of generated images

In the domain of generative models, the balance between quantitative evaluation through numerical metrics and the perceptual assessment of image quality is crucial. To harmonize these elements, we selected a range of generated images exceeding an ITM score of 0.9, as per our established selection criteria in Section 5.3. These images are showcased in a series of tables, providing a qualitative insight into the capabilities of the three models we evaluated. This approach allows us to assess the models' performance not only through data but also through visual analysis.

- From DreamBooth's Dataset: Table 5 features high-fidelity images from the DreamBooth dataset. These images, selected for their exceptional alignment with text prompts (ITM scores), are further analyzed using DINO and CLIP-I scores for subject fidelity.
- From Our Dataset: Table 6 showcases exemplary images from our own dataset, illustrating the models' ability to adapt to new, unseen inputs while maintaining high subject fidelity. This is particularly notable in complex subjects, like dolls with human-like features, where the challenge lies in accurately rendering these subjects with visual authenticity.
- Unconditional Upscaling: Table 7 presents a comparative view of original images sized at 768x768 pixels and their enhanced versions upscaled to 2048x2048 pixels through ESR-GAN. This process highlights the capability the model not just to maintain the integrity and details of the original images but also to significantly enhance their resolution. The upscaled images demonstrate a clear improvement in visual fidelity, crucial for meeting the rigorous high-resolution standards of e-commerce platforms. This aspect of our study underscores the models' utility in real-world e-commerce scenarios, where image quality and resolution play a pivotal role in customer engagement and product presentation.

Through these sections, we bridge quantitative evaluations with the aesthetic appreciation of the models' outputs. This comprehensive approach offers insights into the visual and quantitative capabilities of the generative models, enhancing the understanding of their potential applications and performance.

4.3.4 Ablation study results

Our ablation studies focuses on two key variables: the number of training steps for the text encoder and the specificity of prompts. The study aims to optimize image generation and assess the impact of these variables on model performance.

Training Steps for Text Encoder

In exploring the optimal training protocol for our text encoder, we initially reduced the traditional 350 training steps used in DB to 200 within our BB model, attaining comparable subject and prompt fidelity. Subsequent increments to 350 steps highlighted a pivotal trade-off: an increase in subject fidelity was coupled with a decrease in prompt fidelity, as showed in Fig. 10. This phenomenon indicates an overemphasis on subject features, potentially overshadowing the prompt's context. We propose a calibrated approach to training, starting at 100 steps and incrementally increasing by 50 steps, fine-tuning the balance between subject representation and contextual adherence with each evaluation.

Prompt Specificity

The influence of prompt specificity on model performance was also scrutinized. Our findings, illustrated in Fig. 11, reveal that BB and DB models favor general prompts for

Table 5 Selected Examples of Generated Images from Dataset of DreamBooth by Model

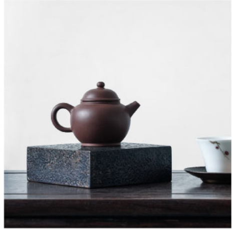
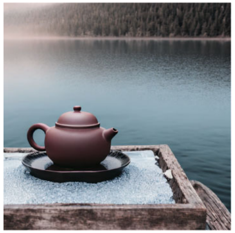
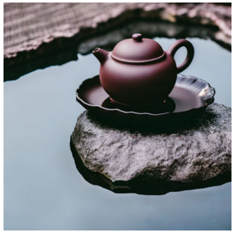
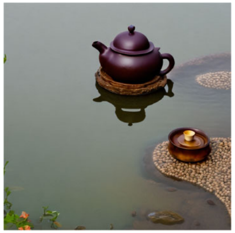
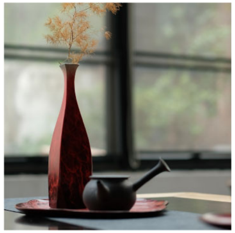
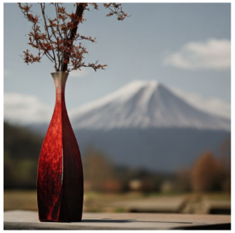
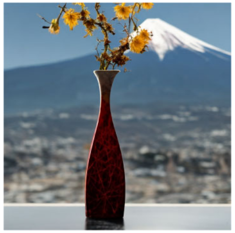
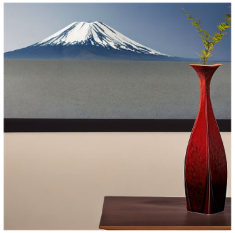
Product	BLIP2Booth	DreamBooth	Textual Inversion
Teapot	Prompt: a teapot peacefully drifting on the tranquil surface of a crystal-clear lake		
			
	DINO: 0.677 - CLIP-I: 0.897	DINO: 0.713 - CLIP-I: 0.954	DINO: 0.547 - CLIP-I: 0.859
Vase	Prompt: a vase with a Fuji mountain in the background		
			
	DINO: 0.711 - CLIP-I: 0.882	DINO: 0.575 - CLIP-I: 0.839	DINO: 0.527 - CLIP-I: 0.808
Fancy boot	Prompt: a fancy boot stands elegantly amidst the historic charm of a cobblestone street		

Table 5 continued

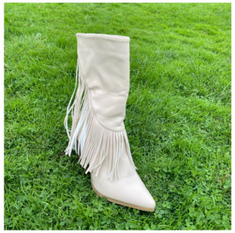
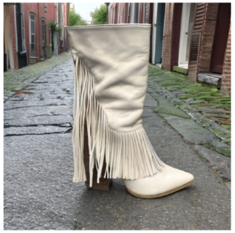
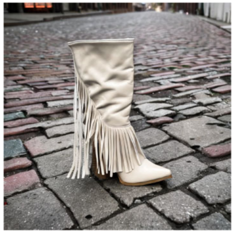
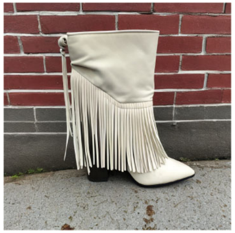
Product	BLIP2Booth	DreamBooth	Textual Inversion
	 DINO: 0.852 - CLIP-I: 0.929	 DINO: 0.795 - CLIP-I: 0.843	 DINO: 0.656 - CLIP-I: 0.765

Table 6 Selected Examples of Generated Images from our Dataset by Model

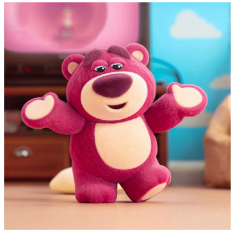
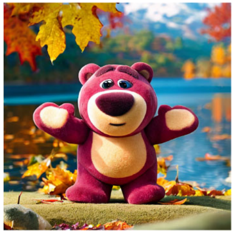
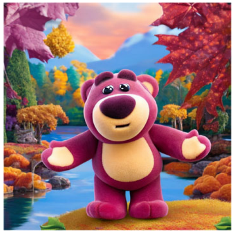
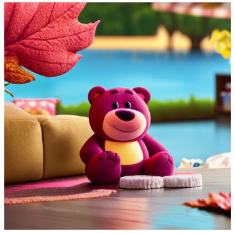
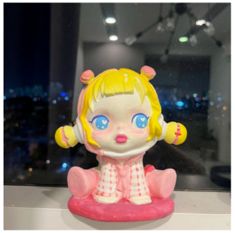
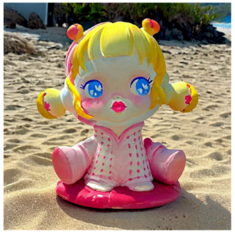
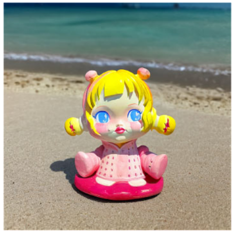
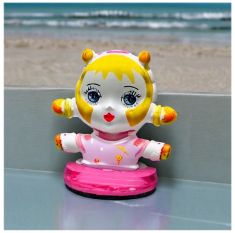
Product	BLIP2Booth	DreamBooth	Textual Inversion
Teddy bear	Prompt: a teddy bear with a tree and autumn leaves in the background		
			
	DINO: 0.779 - CLIP-I: 0.897	DINO: 0.765 - CLIP-I: 0.873	DINO: 0.757 - CLIP-I: 0.842
Doll	Prompt: a doll on the beach		
			
	DINO: 0.752 - CLIP-I: 0.808	DINO: 0.790 - CLIP-I: 0.758	DINO: 0.720 - CLIP-I: 0.746
Clock	Prompt: a clock in the snow		

Table 6 continued

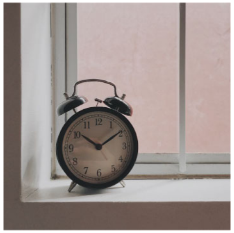
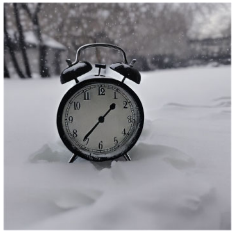
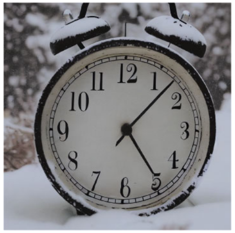
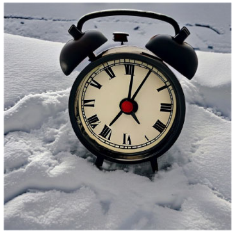
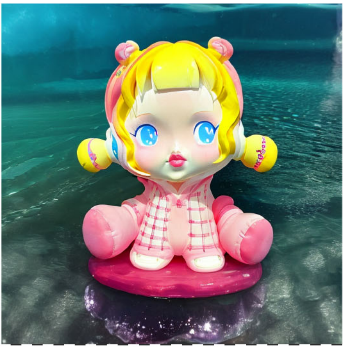
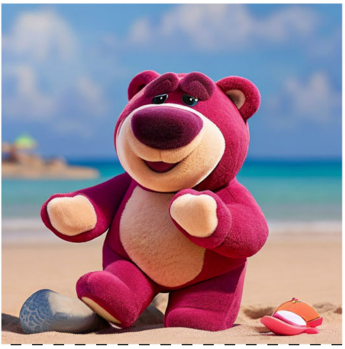
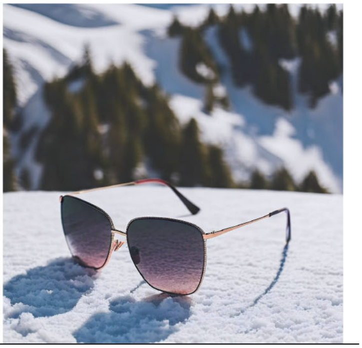
Product	BLIP2Booth	DreamBooth	Textual Inversion
			
	DINO: 0.785 - CLIP-I: 0.869	DINO: 0.87 - CLIP-I: 0.925	DINO: 0.709 - CLIP-I: 0.918

Table 7 Comparison of Image Resolution Enhancement from 768x768 to 2048x2048

subject fidelity, likely due to the broader interpretative scope they offer. However, specific prompts significantly bolster prompt fidelity across all models, enhancing the alignment with detailed textual descriptions. These insights underscore the critical role of prompt design in model output quality, dictating whether the emphasis falls on subject likeness or adherence to textual context.

These findings are instrumental in guiding future training and development strategies for generative models, highlighting the importance of balancing training parameters and prompt design for optimal output quality.

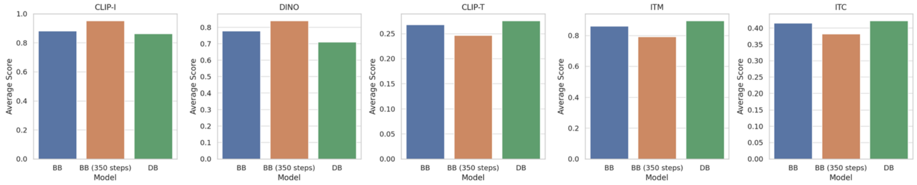


Fig. 10 Comparison of DreamBooth, BLIP2Booth (200 steps) and BLIP2Booth (350 steps) with five fidelity scores

5 Discussions

5.1 Limitations of cosine similarity

While cosine similarity (DINO, CLIP-I, CLIP-T, ITM, ITC) provides a computational measure of alignment, it has known limitations when applied to high-dimensional representations:

- Curse of dimensionality: As the number of dimensions increases, cosine similarity may become less meaningful due to reduced angular separation between vectors.
- Lack of perceptual awareness: Cosine similarity does not always align with human perception, as small angular differences might not correspond to meaningful visual differences.
- Feature sparsity and embedding drift: Two images with distinct structural details may still produce similar embeddings, leading to misleading similarity scores.

To address these issues, we integrate HCI-based human evaluations and correlation analysis between numerical scores and subjective ratings. This ensures that the generated images not only meet computational alignment requirements but also align with real-world human perception.

For future work, we propose exploring alternative similarity measures such as:

- Learned Perceptual Image Patch Similarity (LPIPS), which aligns better with human perception.
- Structural Similarity Index Measure (SSIM) for assessing pixel-level structural consistency.

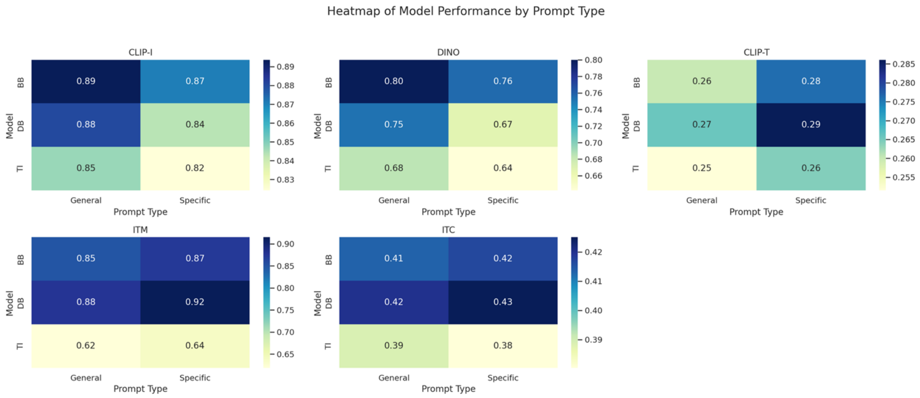


Fig. 11 Heatmap of Model Performance by Prompt Specificity

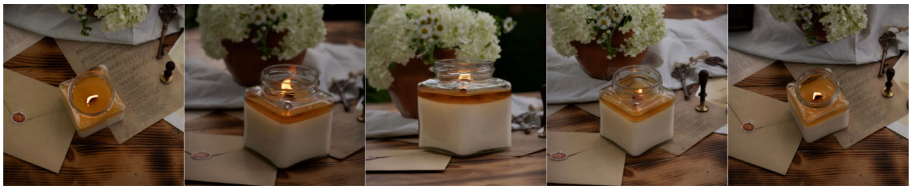


Fig. 12 Five input images of subject ’candle’

- Contrastive learning-based embeddings (e.g., CLIP ITM/ITC scores) to enhance perceptual relevance in image similarity evaluation.

By incorporating both computational and perceptual analysis, our methodology ensures a more holistic assessment of similarity, reinforcing the reliability of our findings beyond numerical similarity scores.

5.2 Impact of input image diversity on model performance and overfitting

In the realm of generative AI models, the selection of input images plays a pivotal role in determining the model’s ability to generalize and produce diverse outputs. This section delves into the significance of background diversity in input images and its correlation with model performance and susceptibility to overfitting. We utilize the example of the ’candle’ from our dataset to elucidate this phenomenon.

Importance of Background Diversity in Input Images

Our study underscores the necessity of incorporating a wide range of backgrounds in the training set. Diverse backgrounds enable the model to semantically understand and adapt to varying contexts, thereby enhancing its generative capabilities. This diversity is not merely a visual aspect but a fundamental factor that influences the model’s ability to interpret and respond to different prompts effectively.

Analysis and Quantitative Evidence

In our dataset, the ’candle’ images had uniform backgrounds (Fig. 12), leading to a notable form of overfitting. When generating images with the prompt ’a <mycandle> candle with a tree and autumn leaves in the background’ (Fig. 13), the model produced outputs that

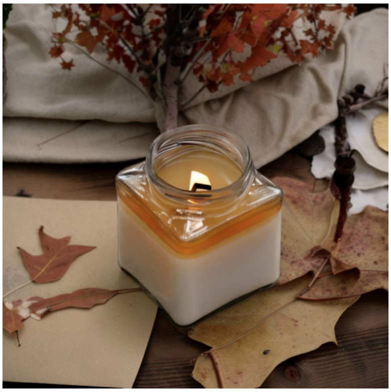


Fig. 13 Generated Candle Image - Overfitting Example

closely resembled the input set, despite achieving high scores in key metrics like CLIP-T (0.29901513), ITM (0.990453124), and ITC (0.4279243648), all of which exceed the average scores across all products. This pattern indicates the model's struggle to innovate beyond the learned background patterns, emphasizing the need for diverse training data to avoid overfitting and enhance output variability.

Broader Implications and Recommendations

This phenomenon is a microcosm of a broader trend in generative AI, where the diversity of training data profoundly impacts model performance. A varied set of input images, particularly in terms of backgrounds, is crucial for the model's ability to produce contextually varied outputs. For future model developments and training methodologies, a strategic approach to input image selection, prioritizing diversity, is essential for both academic research and practical applications where adaptability and innovation in image generation are key.

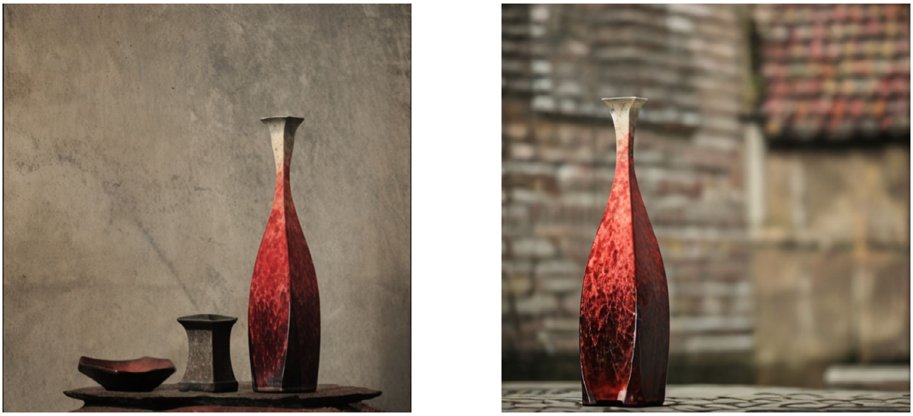
5.3 Image selection and qualitative analysis based on ITM scores

In our research, the ITM scores are crucial for selecting images for deeper analysis within HCI evaluations and qualitative assessments. Understanding the methodology and reasoning behind this selection is essential to guarantee that the images used in our evaluations are both relevant and of high quality.

Image Selection Based on ITM Scores

In our analysis, ITM scores predominantly clustered at the extremes, as shown in Fig. 7. The bimodal distribution, with scores predominantly near 0 or 1, indicated that generated images either closely aligned with the text prompts or significantly diverged. Consequently, we established an ITM threshold greater than 0.9 for image selection. Figure 14 also demonstrates the impact of ITM scores on visual alignment with prompts. It contrasts an image with a low ITM score, where the subject's integration into the context is weak, against an image with a high ITM score that accurately reflects the prompt.

Link to HCI and Qualitative Results



(a) ITM = 0,046 (b) ITM = 0,997

Fig. 14 Comparison of Generated Images with Low and High ITM Scores

In our study, images with ITM scores over 0.9 were essential for HCI evaluations and qualitative analysis. This high threshold ensured that the images used in our user satisfaction studies and task-based fidelity assessments were of the highest quality and alignment with text prompts. These same images, presented as Curated Image Examples in the subsequent section, not only illustrate the model's performance but also contextualize its capabilities within our study. This approach was key in fostering focused discussions on the model's effectiveness, providing a balanced view of its output quality that is crucial for both academic research and practical HCI applications.

6 Conclusions

In this research, we hypothesized that integrating a vision-language model like BLIP-2 with the stable diffusion techniques of DreamBooth could enhance subject-driven text-to-image generation, particularly in e-commerce applications. Our major findings confirmed this hypothesis, as the developed model, BLIP2Booth, demonstrated improved subject fidelity and prompt alignment compared to existing methods. The addition of image upscaling with ESRGAN further ensured that the images met the high-resolution standards of online marketplaces.

The relevance and added value of our work lie in its ability to generate high-fidelity images that accurately reflect both the subject's characteristics and the nuances of textual prompts. This is particularly significant for e-commerce, where visual representation is crucial. The novel internal metrics, including ITM and ITC scores, and the empirical HCI experimentation provide a comprehensive evaluation framework, enhancing our understanding of the correlation between computational assessments and human perception.

One of the key contributions of this study is addressing the computational efficiency of fine-tuning subject-driven models. By leveraging BLIP-2 for textual prompt enhancement and optimizing UNet and text encoder fine-tuning, our approach significantly reduces training overhead compared to traditional methods like DreamBooth. Our model maintains training efficiency while achieving high subject fidelity with adaptive tuning, highlighting the feasibility of our approach in real-world applications and ensuring accessibility to researchers and businesses with limited computational resources.

However, the study has limitations, notably in text retention within images. While numerical text in products like clocks was well-preserved, textual elements on other products posed challenges. Additionally, cosine similarity, while widely used for evaluating both subject fidelity and prompt fidelity, has known limitations in high-dimensional spaces. Cosine similarity scores do not always align perfectly with human perception, necessitating HCI-based evaluations to validate computational assessments. Our findings indicate that while cosine similarity remains a useful metric, incorporating learned perceptual similarity measures (e.g., LPIPS, SSIM, CLIP ITM/ITC scores) could enhance evaluation robustness in future research.

Moreover, the research primarily focused on product categories relevant to the Vietnamese e-commerce market, which may limit the generalizability of the findings. While the dataset was carefully curated to include diverse subject types ranging from structured objects (e.g., sneakers, boots, sunglasses) to complex textures (e.g., teddy bears, dolls), further studies could incorporate a broader range of global e-commerce products to improve the model's applicability across different cultural contexts.

Future research should focus on improving text preservation in generated images, particularly for small or embedded text elements, which remain a challenge in subject-driven text-to-image models. Expanding the dataset to include a more diverse range of global e-commerce products would enhance the model's generalizability and applicability across different cultural contexts. Additionally, optimizing computational efficiency by incorporating techniques such as model quantization, distillation, and low-rank adaptation could make the approach more accessible for real-world use. Beyond e-commerce, exploring applications in digital marketing, virtual reality, and interactive AI-driven design tools could significantly broaden the impact of this research. Finally, refining evaluation methods by incorporating alternative similarity metrics, such as perceptual similarity measures (LPIPS, SSIM), would improve the alignment between computational assessments and human perception, leading to more reliable image generation models.

Appendix A Prompt List for Model Evaluation

This appendix provides a comprehensive list of prompts used to evaluate the generative models in our study. These prompts are designed to assess the models' ability to generate images that are both accurate to the subject and contextually relevant.

The first ten prompts are deliberately broad and open-ended, allowing the models greater flexibility in interpreting the subject matter without strict contextual constraints. This approach is designed to assess the models' ability to preserve subject fidelity when given minimal specific guidance.

The next ten prompts are markedly more specific, incorporating detailed descriptors that establish a precise context for the model to follow. These prompts are designed to test the models' ability to maintain prompt fidelity, ensuring a closer alignment between the generated images and the detailed nuances of the text.

To demonstrate the distinction between general and specific prompts, consider pairs of prompts that are thematically linked yet differ in their level of descriptiveness and detail:

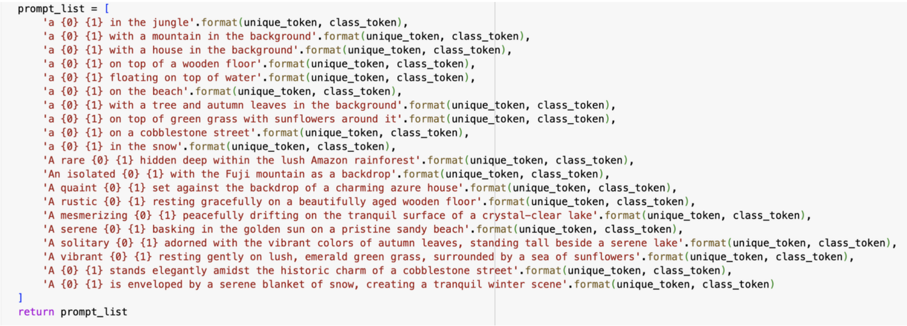
- General Prompt (Prompt 6): "with a tree and autumn leaves in the background"
- Specific Prompt (Prompt 16): "adorned with the vibrant colors of autumn leaves, standing tall beside a serene lake"

and

- General Prompt (Prompt 8): "on a cobblestone street"
- Specific Prompt (Prompt 18): "stands elegantly amidst the historic charm of a cobblestone street"

The prompt numbers (0-9 for general and 10-19 for specific) reflect the degree of specificity, with 0 being the most general and 19 the most specific. Figure 15 visually illustrates these prompts, showcasing the range from general to specific:

This systematic methodology enables a nuanced assessment of the models' performance across varying prompt types, offering a comprehensive insight into their strengths and areas for improvement in both general and highly detailed image generation tasks.



```
prompt_list = [
    'a {0} {1} in the jungle'.format(unique_token, class_token),
    'a {0} {1} with a mountain in the background'.format(unique_token, class_token),
    'a {0} {1} with a house in the background'.format(unique_token, class_token),
    'a {0} {1} on top of a wooden floor'.format(unique_token, class_token),
    'a {0} {1} floating on top of water'.format(unique_token, class_token),
    'a {0} {1} on the beach'.format(unique_token, class_token),
    'a {0} {1} with a tree and autumn leaves in the background'.format(unique_token, class_token),
    'a {0} {1} on top of green grass with sunflowers around it'.format(unique_token, class_token),
    'a {0} {1} on a cobblestone street'.format(unique_token, class_token),
    'a {0} {1} in the snow'.format(unique_token, class_token),
    'A rare {0} {1} hidden deep within the lush Amazon rainforest'.format(unique_token, class_token),
    'An isolated {0} {1} with the Fuji mountain as a backdrop'.format(unique_token, class_token),
    'A quaint {0} {1} set against the backdrop of a charming azure house'.format(unique_token, class_token),
    'A rustic {0} {1} resting gracefully on a beautifully aged wooden floor'.format(unique_token, class_token),
    'A mesmerizing {0} {1} peacefully drifting on the tranquil surface of a crystal-clear lake'.format(unique_token, class_token),
    'A serene {0} {1} basking in the golden sun on a pristine sandy beach'.format(unique_token, class_token),
    'A solitary {0} {1} adorned with the vibrant colors of autumn leaves, standing tall beside a serene lake'.format(unique_token, class_token),
    'A vibrant {0} {1} resting gently on lush, emerald green grass, surrounded by a sea of sunflowers'.format(unique_token, class_token),
    'A {0} {1} stands elegantly amidst the historic charm of a cobblestone street'.format(unique_token, class_token),
    'A {0} {1} is enveloped by a serene blanket of snow, creating a tranquil winter scene'.format(unique_token, class_token)
]
return prompt_list
```

Fig. 15 Prompts. Evaluation prompts

Funding This research is funded by Van Lang University, Vietnam under grant number VLU-RP-202307.

Data Availability The authors declare that data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of Interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical Approval This study does not violate and does not involve moral and ethical statement.

Consent for Publication The authors were aware of the publication of the paper and agreed to its publication.

References

1. Gal R, et al (2022) An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv:2208.01618
2. Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (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, pp 22500–22510
3. Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R, Van Gool L (2022) Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11461–11471
4. Li J, Li D, Savarese S, Hoi S (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597
5. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
6. Holsapple CW, Singh M (2000) Electronic commerce: From a definitional taxonomy toward a knowledge-management view. J Organization Comput Electron Commerce 10(3):149–170
7. Gielens K, Steenkamp J-BEM (2019) Branding in the era of digital (dis)intermediation. Int J Res Marketing 36(3):367–384
8. Klaus T, Changchit C (2019) Toward an understanding of consumer attitudes on online review usage. J Comput Inf Syst 59(3):277–286
9. Kaplan A, Haenlein M (2019) Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. Business Horizon 62(1):15–25
10. Bawack RE, Wamba SF, Carillo K (2021) A framework for understanding artificial intelligence research: insights from practice. J Enterprise Inf Manage 34(2):645–678

11. Deng S, Tan CW, Wang W, Pan Y (2019) Smart generation system of personalized advertising copy and its application to advertising practice and research. J Advert 48(4):356–365
12. Bawack RE, Wamba SF, Carillo KDA et al (2022) Artificial intelligence in e-commerce: a bibliometric study and literature review. Electron Markets 32:297–338
13. Lee D, Hosanagar K (2021) How do product attributes and reviews moderate the impact of recommender systems through purchase stages? Manage Sci 67(1):524–546
14. Stöckli DR, Khobzi H () Recommendation systems and convergence of online reviews: The type of product network matters! Decision Support Syst 142:113475
15. Jannach D, Manzoor A, Cai W, Chen L (2021) A survey on conversational recommender systems. ACM Comput Surv (CSUR) 54(5):1–36
16. Al-Natour S, Turetken O (2020) A comparative assessment of sentiment analysis and star ratings for consumer reviews. Int J Inf Manage 54:102132
17. Da'u A, Salim N, Rabiu I, Osman A (2020) Recommendation system exploiting aspect-based opinion mining with deep learning method. Inf Sci 512:1279–1292
18. Xia H, Wei X, An W, Zhang ZJ, Sun Z (2021) Design of electronic-commerce recommendation systems based on outlier mining. Electron Market 31(2)
19. Praet S, Martens D (2020) Efficient parcel delivery by predicting customers' locations. Decision Sci 51(5):1202–1231
20. Barzegar Nozari R, Koohi H (2020) A novel group recommender system based on members' influence and leader impact. Knowl-Based Syst 205:106296
21. Dong M, Zeng X, Koehl L, Zhang J (2020) An interactive knowledge-based recommender system for fashion product design in the big data environment. Inf Sci 540:469–488
22. Gupta S, Kant V (2020) Credibility score based multi-criteria recommender system. Knowledge-Based Syst 196:105756
23. Pourgholamali F, Kahani M, Bagheri E (2020) A neural graph embedding approach for selecting review sentences. Electron Commer Res Appl 40:100917
24. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: 2017 IEEE international conference on computer vision (ICCV), pp 2980–2988
25. Yu L, Chen J, Sinha A, Wang M, Chen Y, Berg TL, Zhang N (2022) Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pp 4433–4442
26. Goodfellow I, et al (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27
27. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. CoRR arXiv:1809.11096
28. Hudson DA, Manning CD (2018) Compositional attention networks for machine reasoning. In: International conference on learning representations
29. Cadene R, Ben-younes H, Cord M, Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1989–1998
30. Chen W, Gan Z, Li L, Cheng Y, Wang W, Liu J (2021) Meta module network for compositional visual reasoning. In: 2021 IEEE winter conference on applications of computer vision (WACV), pp 655–664
31. Choi J, Kim S, Jeong Y, Gwon Y, Yoon S (2021) ILVR: conditioning method for denoising diffusion probabilistic models. CoRR arXiv:2108.02938
32. Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: International conference on machine learning, pp 8162–8171. PMLR
33. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst 34:8780–8794
34. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125
35. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inf Process Syst 35:36479–36494
36. Liu X, Park DH, Azadi S, Zhang G, Chopikyan A, Hu Y, Shi H, Rohrbach A, Darrell T (2023) More control for free! image synthesis with semantic diffusion guidance. In: 2023 IEEE/CVF winter conference on applications of computer vision (WACV), pp 289–299
37. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C (2018) Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the european conference on computer vision (ECCV) workshops, pp 0–0
38. Statista (2022) Vietnam: E-commerce Spend by Category. https://www.statista.com/statistics/1112147/vietnam-e-commerce-spend-by-category. Accessed: 11 November 2023

39. Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660
40. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR pp 8748–8763.
41. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning, PMLR pp 12888–12900.
42. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Med Image Comput Comput-Assist Intervent - MICCAI 2015:234–241
43. Lu C, Zhou Y, Bao F, Chen J, Li C, Zhu J (2022) Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv:2211.01095

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.