Stable Diffusion Image Generation System Optimized with Variational Autoencoders and Low-Rank Adaptation

ChunLing Zhang
Weifang University of Science and Technology, School of Computer Science, Shouguang, 262700, China
E-mail: zhangcl0115@wfust.edu.cn

Keywords: image generation system, fréchet inception distance (fid), variational autoencoder (vae), low-rank adaptation (lora), stable diffusion (sd)

Recieved: July 23, 2025

Text-to-image generation has quickly evolved with diffusion-based generative models that combine semantic conditioning and latent-space denoising, allowing machines to generate high-quality visuals from natural language prompts. Despite these developments, existing diffusion systems still face challenges in prompt clarification accuracy, model adaptability, and computational efficacy, which limit their performance in real-time and resource-limited settings. The research aims to design and optimize an image generation framework based on Stable Diffusion (SD) that improves prompt processing, improves image quality, and enables lightweight fine-tuning. The system utilizes the LAION-Aesthetics v2 4.5 dataset, which contains high-quality text–image pairs suitable for visual generation tasks. Preprocessing involves text cleaning, tokenization, and semantic structuring, utilizing a transformer-based tokenizer to ensure accurate language-to-visual mapping. The architecture integrates Stable Diffusion, Variational Autoencoder (VAE) for latent-space decoding, and Low-Rank Adaptation (LoRA) for efficient fine-tuning with minimal computational cost. Results show that SD-VAE-LoRA achieved a PSNR of 33.7 dB, SSIM of 93 %, FID of 17.8, Inception Score of 36.02, and R-Precision of 90 %, superior to baseline SD and advanced diffusion models such as Latent Diffusion Method (LDM) [24], Menstrual Cycle-Inspired Latent Diffusion Method (MCI-LDM) [24], and Conditional Generative Adversarial Networks, Attention mechanisms, and Contrastive Learning (C-GAN+ATT+CL). The optimized system advances semantic alignment, decreases training time, and preserves image realism, confirming its strength for scalable, adaptive, and high-fidelity image generation applications.

Povzetek: Prispevek predstavlja optimiziran sistem za generiranje slik iz besedila, ki z nadgradnjo Stable Diffusion in uporabo VAE ter LoRA izboljša semantično ujemanje, kakovost slik in učinkovitost učenja ob nizkih računskih stroških.

1 Introduction

The rise of Artificial Intelligence (AI) tools has significantly transformed content generation, with text-to-image generation being a notable application. This system uses computation to generate images based on textual input, enabling machines to visualize human language and revolutionize the content generation landscape [1-2]. Text-to-image generation transforms text-based language prompts to high-quality images to bridge the gap between language and images and be used in design, art, education, and entertainment [3]. Deep Learning (DL) and vision-language understanding combine to create photorealistic and diverse images, a trend in fields like game development, animation, advertising, and personal media generation, enabling diverse styles [4-5]. Text-to-image generation is expected to expand into digital education, medical visualization, and e-commerce, with real-time art editing and AI-driven design assistance promising future domains [6]. Contemporary systems offer advantages due to automation of creative tasks, quick generation of visual imagery by audiences, and the ability of non-experts to create high-quality imagery [7]. It also fosters accessibility by providing visual content to non-image-makers [8]. Contemporary image generation systems use pre-trained methods to interpret text semantics, encode them, and decode them into visual artistic forms, but their development is influenced by various factors [9]. The systems of today rely on enormous images, leading to visual bias and expensive resources. The compromise between the complexity of a model and the user experience is a key to effective and efficient performance [10]. Diffusion models are effective in generating images based on text, but have difficulties in prompt accuracy, image quality, and resource efficiency, which restricts scalability [11]. The lightweight framework is needed to improve

alignment, fidelity, and computational properties [12]. The SD-VAE-LoRA model addresses these issues through a lightweight Stable Diffusion framework integrating SD for latent-space denoising, VAE for structured semantic decoding, and LoRA for efficient fine-tuning. Together, these modules enable resource-efficient, context-consistent image generation from diverse text prompts. The research aims to design a lightweight SD-VAE-LoRA model, which enhances image realism, semantic matching, and fine-tuning performance. It combines VAE as a structured representation and LoRA as an adaptation resource-efficiently. It is to generate high-quality images that are semantically consistent at a lower cost of computation.

1.1 Research question

1. Can Low-Rank Adaptation (LoRA) reduce training cost and resource consumption while maintaining or improving image generation quality?
2. Does integrating Variational Autoencoder (VAE) enhance latent-space decoding accuracy and semantic coherence in text-to-image synthesis?
3. How does the proposed SD-VAE-LoRA compare with existing methods (LDM, MCI-LDM, C-GAN+ATT+CL) in terms of realism, fidelity, and training efficiency?

2 Related work

The research [13] showed that SD methods, optimized for XL-XDXL architectures, can generate high-detail images with limited VRAM, improving quality by 35%, but lack real-time performance testing and comparisons with commercial alternatives. The investigation [14] improved segmentation performance and expert-rated image quality by using diffusion probabilistic techniques to produce realistic 3D Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images. While there was improvement in Dice scores with synthetic data (0.91 to 0.95), there has been no clinical validation or generalizability of results across conditions. The research [15] evaluated the educational value of text-to-image AI in visual art, analyzing over 72,000 prompts. It found transformative teaching possibilities and cost savings, but identified gaps in legal ownership and economic models for AI-generated artwork in education. SynDiff, a confrontational circulation method based on unpaired health image conversion [16], showed improved performance over GANs in MRI and CT tasks by estimating conditional mappings from source noise to target, but the research failed to address real-world validation and runtime viability for clinical settings. The evaluation synthesized anonymized chest X-rays using a privacy-preserving sampling approach and trained classifiers on synthetic data using a latent diffusion approach [17]. The model achieved only a 3.5% Area under the Curve (AUC) gap from benchmarks on real data, but further validation was needed for rarer conditions and for robustness from a privacy perspective. Related works are included in Table 1.

Table 1: Comparison of other related methods

Reference	Objective	Dataset	Method	Numerical Results	Limitations
Yang et al. [18]	To enhance prompt privacy and safety in diffusion-based text-to-image generation systems	Internal prompt dataset (NSFW – Not Safe for Work filtered text corpus)	Prompt obfuscation and text-filtering mechanism for diffusion models	Qualitative improvement in safe image generation; no FID (Fréchet Inception Distance)	Lacks ethical safeguard validation
Paananen et al. [19]	To evaluate creative engagement and user experience in prompt-based generative systems	User-generated prompts via Midjourney, Stable Diffusion, and DALL-E	Prompt-based creativity analysis through diffusion model outputs	Improved user creativity ratings;	Absence of long-term validation, no real-world application testing
Lian et al. [20]	To improve multilingual prompt comprehension and scene layout accuracy in diffusion models	Custom multilingual prompt dataset	Scene layout prediction using a multilingual-aware diffusion framework	Achieved ~2× improvement in scene layout accuracy vs baseline models	Lacks usability and real-world validation;
Chen et al. [21]	To achieve energy-efficient and high-	LAION-Aesthetics v1.5 dataset	Optimized low-energy diffusion model	FID ≈ 18.5, PSNR ≈ 33 dB, and ~90% reduction in	Limited to specific domains;

	quality text-to-image generation			CO_2 emissions vs Stable Diffusion v1.5	reduced robustness to noise
Guo et al. [22]	To ensure temporal consistency in text-to-image video and animation generation.	Community text-to-image animation datasets	Temporal diffusion model enforcing frame-to-frame coherence	FID ≈ 19.2; achieved smooth motion continuity	Restricted to short-duration animations; lacks extreme motion
Huang et al. [23]	To enhance artistic fidelity and stylistic coherence in image generation	Custom artistic image corpus	Artistic diffusion model with SSIM (Structural Similarity Index Measure)	SSIM ≈ 90%; improved user preference and stylistic realism	Lacks evaluation on real-time and large-scale art generation;

2.1 Research gap

Diffusion models are effective in generating images through text prompts, but their current systems have issues with prompt handling accuracy, image quality, and fine-tuning resources [11]. These issues negatively affect their scalability and flexibility in real-world operations, especially for high-quality, semantically correct images at fast rates. To improve alignment, operational efficiency, and fine-tuning, a lightweight framework is needed to optimize visual fidelity and assembly, ensuring better alignment and reduced computational expenditure [12]. The present research overcomes these limitations by presenting a lightweight Stable Diffusion-based image generation model (SD-VAE-LoRA) that can enhance semantic correspondence, operating performance, and fine-tuning capabilities. The model combines three complementary features, which are; Stable Diffusion (SD) denoising in latent space and high-level image generation, Variational Autoencoder (VAE) structured latent representation and correct semantic decoding, and Low-Rank Adaptation (LoRA) efficient fine-tuning with a low computational cost. Collectively, these modules form a flexible resource efficient pipeline that can generate visual compatible and context consistent images using varied text prompts. The research aims to design a lightweight SD-VAE-LoRA model, which enhances image realism, semantic matching, and fine-tuning performance. It combines VAE as a structured representation and LoRA as adaptation resource-efficiently. It is to generate high-quality images that are semantically consistent at a lower cost of computation.

3 Methodology

This research aims to create and optimize an image generation system based on the SD method that increases prompt handling, as well as generation performance, and facilitates lightweight fine-tuning. The text data is cleaned and tokenized through a transformer-based tokenizer for semantic clarity. The SD method starts image generation by injecting random noise into a refined latent space of lower dimensions. The refined latent representation is decoded into an image that can be seen with a Variational Autoencoder (VAE). Low-Rank Adaptation (LoRA) is used to improve versatility and make method training less computationally intensive, allowing one to fine-tune a model without having to retrain the entire method again. Figure 1 represents the entire process of the research in a pictorial representation.

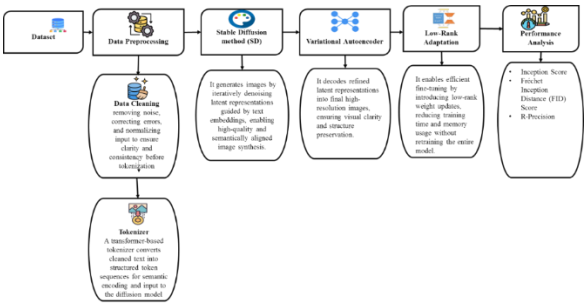

Figure 1: Process flow of the proposed methodology

3.1 Dataset

The system architecture commences with the acceptance of a user input request, where the users are required to input a text description, upon which they retrieve their corresponding images via the website's interface https://huggingface.co/datasets/laion/aesthetics_v2_4.5 . The dataset specifications and implementation of LoRA in the SD-VAE-LoRA framework are displayed in Table 2.

Table 2: Dataset and LoRA implementation details

Aspect	Description
Dataset Name	LAION-Aesthetics v2 4.5 (via Hugging Face)
Dataset Size	Approximately 12 million image–text pairs
Data Split	80% training, 10% validation, 10% testing

Data Filtering Criteria	Images filtered based on aesthetic score ≥ 4.5, NSFW content removed, and low-resolution (<512×512) images excluded
Data Cleaning	Duplicate captions and corrupted files were removed; normalized text encoding and resized all images to 512×512 pixels.
License and Usage	Dataset licensed under Creative Commons (CC-BY 4.0); used strictly for non-commercial, academic research

3.2 Pre-processing

Data preprocessing is the first stage of the system in which raw data is cleaned, formatted, and structured to help improve model performance. Data preprocessing involves cleaning, tokenizing, and transforming text prompts input from the user to make sure that they are semantically clear and compatible with the generation method.

Data cleaning is the method of discovering and repairing flaws or irregularities in data to ensure its quality and reliability. The data quality phase collects the raw data and performs context-aware analysis to measure the raw data across four dimensions of accuracy, completeness, consistency, and timeliness. In the second phase, the measurement phase, a specific set of metrics is applied to measure the quality of the data in the first phase. Next, the improvement phase allows practitioners to filter and sample data to remove a portion of the dataset using only quality, highly relevant data, and a reduced amount of data while maintaining the necessary quantity of data. Data cleansing organizes data effectively, allowing it to be used with accuracy and reliability by ridding the data of noise and errors while also improving system performance and the accuracy of generation.

Following the data cleaning, data tokenization is the next significant step for growth. Data tokenization takes the cleaned text and divides it into minor pieces, called tokens. This allows for the data to be managed much faster and in a way that it is understood semantically by the technique. Tokenization entails taking cleaned text and breaking it into evocative mechanisms for organization. Data tokenization assemblies cleaned text into specific tokens, allowing for accurate and effective interpretation and arranging to align correctly with the image generation process.

3.3 Image generation system based on stable diffusion method - variational autoencoder - low-rank adaptation (SD - VAE - LoRA)

This research aims to advance and optimize an image-generating classification based on the SD technique that advances quick handling, generation presentation, and allows for lightweight fine-tuning. SD plays a crucial part in changing text to images through improved denoising of the latent space. The VAE is a system that takes cleaned-up latent training and decodes it into visible images while ensuring high-quality reconstruction and semantically accurate image generation. LoRA proficiently updates parameters for fine-tuning, conserving training cost while still allowing the method to adapt to other tasks.

3.3.1 Stable diffusion (SD) method

SD allows for high-value, photorealistic images to be generated from text prompts effectively using latent-space denoising. SD is a DL method that generates realistic images from text prompts by denoising latent representations through iterative refinement. SD, a diffusion-based generative framework, produces superior images by step-by-step eliminating noise. The essential principles are as follows:

SD uses the Diffusion Method, which is qualified by mimicking an advancing procedure that gradually improves sound from a clear image until totally arbitrary noise is formed, known as the Forward Diffusion Process (FDP). SD acts in latent space, setting it apart from typical diffusion models. SD employs text-to-image generation to build plug-ins based on the provided text explanation. In particular, SD breaks down the model's weight matrix $X0 \in Qc \times l$ into two SD conditions: $\in Qq \times l \; and \; A \in Qc \times q$. The SD dimension parameter q denotes the rank number following dimensionality decrease. The form of this breakdown is as follows in equation (1).

$$X \approx B.A \tag{1}$$

Where X remains the resulting matrix, and $B.A$ signifies the matrix multiplication of B and A. Simultaneously, this strategy minimizes storage requirements, improving method finetuning efficiency. SD excels at producing superior images with effective modification, deprived of compromising method generation performance. SD allows for effective and high-quality image generation from text with the flexibility, scalability, and lower computational cost to be effective in creative use cases.

3.3.2 Variational autoencoder (VAE)

VAE serves as a fundamental part of the Stable Diffusion architecture in order to encode images into a lower-dimensional latent space and decode them into high-quality images. This implementation contrasts with another type of mathematical graph-based VAE, where latent representation is designed with a focus on efficient representation and reconstruction of visual features, which is inspired by the CompVis Stable Diffusion VAE design. The potential of the VAE is that it learns compressed latent representations that meaningfully encode the input, and

can generate images efficiently as well as in a high-quality and controlled manner, equation (2).

$$g_j^{(k+1)} = e_t^{(k)}\left(g_j^{(k)}, wj\right) + \sum_{i=1}^{M}\sum_{q=1}^{Q}\frac{B_{jiq}}{|T_j|}e_q^{(k)}(g_i^{(k)}, wj)$$

$$g_j^{(k+1)} = tang(g_j^{(k+1)} \tag{2}$$

Here, the feature vector of node j is obtained by combining self-information and aggregated data from neighboring nodes. Here, $g_i^{(k)}$ node j at layer k, wj is the input feature of node j from matrix W, $e_t^{(k)}$ is the self-connection function merging $g_i^{(k)}$ and wj, and $e_q^{(k)}$ is the edge-type–specific transformation aggregating information from neighbors. The adjacency tensor B_{jiq}represents the strength of the edge between nodes i and j under relation type q, normalized by T_j, the number of one-hop neighbors of node j. The indices $i = 1 \ldots M$ and $q = 1 \ldots Q$ denote the total number of nodes (M) and relation types (Q), while Tj is the neighbor set of node j. $g_j$applies the $tanh$ activation to introduce non-linearity and stabilize feature propagation. After multiple levels of propagation, aggregate nodes into a graph-level representation vector equation (3).

$$g'_H = \sum_{V\in U}\sigma\left(mm\left(g_v^{(K)}, w_v\right)\right) \odot tang\left(mm\left(g_V^{(K)}, w_V\right)\right)$$

$$g_H = tang\ g'_H \tag{3}$$

Where g'_H is the final graph-level embedding vector capturing the global context. The summation $\sum_{V\in U}$ aggregates feature from all nodes V within the set U, representing the collection of all nodes in the graph. σ (·) is the sigmoid function, mm is a linear neural network, and $\odot$ signifies element-wise multiplication. To approximate a normal distribution, it applies a linear neural network to g_H, yielding a mean vector μ_y and a diagonal covariance matrix Σ_y as in equation (4).

$$\mu_y = mm(g_H);\ \Sigma_y = mm(g_H) \tag{4}$$

Here, mm is a learnable matrix multiplication layer, where μ_y defines the distribution's center and Σ_y represents its variance, ensuring stable and continuous sampling for image reconstruction. Next, using the reparameterization approach, equation (5) yields the latent vector Y.

$$Y = \mu_Y + \in \odot \Sigma_Y \tag{5}$$

Where $\in$ is a Gaussian random variable. The VAE in the proposed SD-VAE-LoRA implementation facilitates effective latent compression, semantic consistency, and visual realism by ensuring generated images are structurally accurate and fine-grained in detail and that they are computationally simpler to operate through diffusion and decoding.

3.3.3 Low-Rank Adaptation (LoRA)

LoRA is used to facilitate efficient model fine-tuning with less computational cost. LoRA is used to fine-tune pre-trained models with efficiency through low-rank trainable matrices, minimizing memory and computational cost. To alter the adaptation process, the LoRA introduces a small square matrix $Q \in \mathbb{Q}^{q\times q}$ between frozen LoRA matrices that are set of the pre-trained weight matrix $X \in \mathbb{Q}^{n\times m}$.

One way to express the conventional LoRA forward path with an input $w \in \mathbb{Q}^m$is as equation (6).

$$g = wX + w\,\Delta X = wX + w\,BA \tag{6}$$

where $w \in \mathbb{Q}^m$is the input vector, X $\in \mathbb{Q}^{nxm}$ is the frozen pre-trained heaviness matrix, and $\Delta X = BA$ is the low-quality update that includes $B \in \mathbb{Q}^{nxr}$ and $A \in \mathbb{Q}^{rxm}$. This formulation efficiently adapts model parameters by updating only the low-rank matrices A and B, reducing memory and computation while preserving performance. While keeping matrices B and A frozen, and changing the forward path to equation (7).

$$g = wX + w\,\Delta X = wX + wBQA \tag{7}$$

Where B and A are set using the shortened SVD of the original weight matrix X. The SVD of X is represented by equation (8).

$$X = V\Sigma U^S \tag{8}$$

Where $V\,\Sigma\,\mathbb{Q}^{n\times n}, \Sigma \in \mathbb{Q}^{m\times m}$, and $U \in \mathbb{Q}^{m\times q}$ define (frozen) matrices B and A as in equation (9).

$$B = V_q\Sigma_q \text{ and } A = U_q^S \tag{9}$$

Here, U_q and V_q are the left and right singular vector matrices of the original weight matrix, Σ_q contains the top q singular values, and S denotes the selected subset for dimensional reduction. This decomposition initializes LoRA's trainable matrices A and B, preserving key information while minimizing parameter size.

The Q matrix is initialized with a Gaussian distribution $M(0,\sigma^2)$, where σ is a tiny but non-zero value. This ensures that it starts fine-tuning with a model that is nearly

identical to the pre-trained model. Fine-tuning involves freezing matrices $B$and A and updating only Q, which reduces the number of trainable parameters

This flexibility is especially useful for larger models, when standard approaches are limited by hidden dimensions. LoRA, like its predecessors, does not add computational burden or latency during inference. This module can be incorporated into the original matrix after training. The importance of LoRA was its ability to adapt large models in an efficient way using minimal parameters to preserve performance and adapt the model using very limited resource requirements. Algorithm 1 and figure 2 shows the process of SD - VAE - LoRA

Algorithm 1: Stable Diffusion method - Variational Autoencoder - Low-Rank Adaptation (SD - VAE - LoRA)

Step 1: Initialize Stable Diffusion (SD)
$SD\_model = initialize\_SD(parameters)$
Print ("Stable Diffusion initialized")
Step 2: Initialize Variational Autoencoder (VAE)
$VAE\_model = initialize\_VAE(latent\_dim, encoder\_params, decoder\_params)$
Print ("VAE initialized")
Step 3: Initialize Low-Rank Adaptation (LoRA)
$LoRA\_matrices = initialize\_LoRA(SD\_model.weights, rank\_q)$
Print ("LoRA initialized")
Step 4: Process SD-VAE-LoRA
$Input\_text$ = "Enter text prompt here"
Step 4a: Generate latent representation from SD
$latent\_X0 = SD\_model.encode\_text(Input\_text)$
Step 4b: Apply VAE encoding and decoding
$if\ latent\_X0\ is\ not\ None$:
$latent\_Y = VAE\_model.encode(latent\_X0)$
$reconstructed\_image = VAE\_model.decode(latent\_Y)$
Print ("Image reconstructed via VAE")
else:
print ("Error: SD latent representation not generated")
Step 4c: Apply LoRA fine-tuning
$if\ LoRA\_matrices\ is\ not\ None$:
Forward pass with LoRA adaptation
$adapted\_output = latent\_Y * LoRA\_matrices.B * LoRA\_matrices.Q * LoRA\_matrices.A$
$final\_image = decode(adapted\_output)$
Print ("LoRA fine-tuning applied")
else:
print ("Error: LoRA matrices not initialized")
Step 5: Output final image
$display(final\_image)$

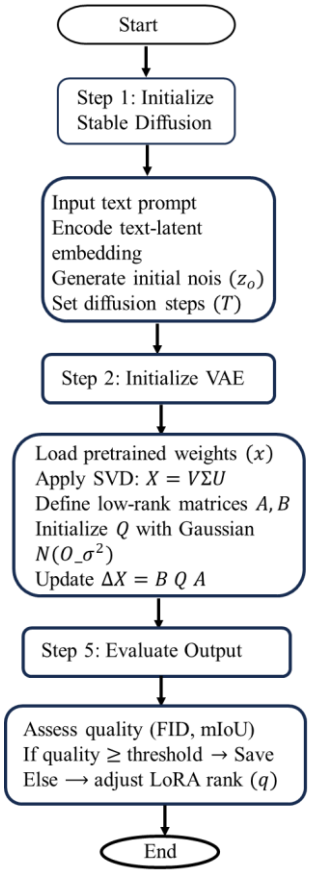

Figure 2: Flowchart for suggested SD-VAE-LoRA image generation evaluation system

4 Result and discussion

This research generates an optimized SD system to boost quick processing, image generation quality, and efficient fine-tuning. This research requires a pre-trained SD model, a GPU, a multi-core CPU, 16 GB RAM, Python 3.8+ compatibility, PyTorch, and LoRA tools. The presentation of the system is estimated using **user satisfaction evaluation** of its content and training time (in hours) to assess how it works with different image generation conditions.

4.1 Experimental setup

Table 3 indicates the system setup and testing environment to implement and test the suggested SD-VAE-LoRA-based image generation model.

Table 3: Experimental Setup for SD-VAE-LoRA Framework.

Component	Specification
Operating System	Ubuntu 22.04 LTS
CPU	Intel Xeon Gold 6258R, 24-core @ 2.7 GHz
RAM	64 GB DDR4
GPU	NVIDIA RTX A6000 (48 GB VRAM)
Programming Language	Python 3.10

Frameworks	PyTorch, Hugging Face Diffusers, Transformers, Accelerate
Dataset	LAION-Aesthetics v2 4.5 (Text–Image Pairs)
Model Components	Stable Diffusion (SD), Variational Autoencoder (VAE), Low-Rank Adaptation (LoRA)
Fine-Tuning Steps	20,000 iterations / 15 epochs
Training Time per Epoch	≈ 1.2 hours
Total Training Time	≈ 18 hours
Batch Size	16
Learning Rate	2×10^{-4}
Optimizer	AdamW
Evaluation Metrics	PSNR, SSIM, FID, Inception Score, R-Precision
Visualization Tools	Matplotlib, Seaborn, TensorBoard
Performance Monitors	NVIDIA NVML, PyTorch Profiler

4.2 Hyper parameter for the proposed model

Table 4 summarizes the hyperparameters of the suggested SD-VAE-LoRA image generation architecture.

Table 4: Parameter Setup for SD-VAE-LoRA Framework

Hyperparameter	Typical Values / Settings
Latent Dimension Size	512
Number of Epochs	15 (up to 20 for convergence)
Batch Size	16
Learning Rate	2×10^{-4}
Optimizer	AdamW
Dropout Rate	0.2
LoRA Rank (r)	8
LoRA Scaling Factor (α)	16
VAE Latent Channels	4
Diffusion Steps	1,000
Noise Scheduler	Cosine
Activation Function	GELU
Fine-Tuning Iterations	20,000
Evaluation Frequency	Every 1,000 steps
LoRA Adaptation Layers	Applied to cross-attention, U-Net transformer, and text encoder layers of Stable Diffusion
LoRA Rank (r)	8 (balancing efficiency and representational power)
Number of LoRA Modules	12 total modules integrated within the U-Net and text encoder blocks
Regularization Method	L2 weight decay and dropout (p = 0.2) to prevent overfitting
Optimization Objective	Minimized reconstruction loss and alignment error between text and image embeddings

- **User satisfaction evaluation:** It determines the similarity between the images generated and the expectations of the users in one aspect: realism, creativity, and immediate alignment; a high score signifies a higher perceived image quality and acceptance by the user, whereas a low score depicts a lack of satisfaction or visual quality.

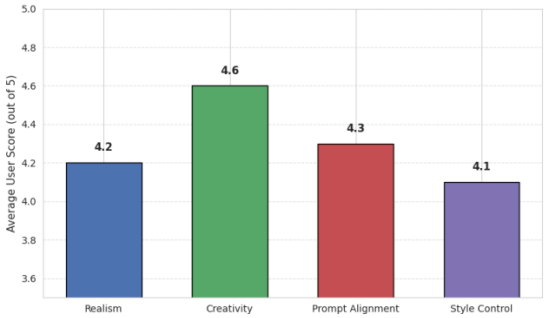

Figure 3: User satisfaction evaluation across image quality dimensions

Figure 3 illustrates user satisfaction with four criteria, such as realism, creativity, prompt alignment, and style control, with the highest score of creativity (~4.6), then prompt alignment (~4.3), and realism (~4.2), which is a good overall performance. Style control was rated slightly lower (~4.1), which implies that there is a little area to work on with regard to aesthetic consistency.

The proposed SD-VAE-LoRA and SD-VAE-baseline SD were evaluated in a double-blind context with 30 participants (15 experienced in digital design/AI art, 15 general users) through a structured user evaluation and a random selection of 50 image-prompt pairs used (25 pairs of image-prompts of each shape, 5 shape options, 25 neutral stimuli). The rating of images was done on a 5-point Likert scale on Realism, Creativity, Prompt Alignment, and Style Control. Mean +- standard deviation scores were obtained according to each criterion, and the overall score was obtained as a weighted mean (Realism 0.3, Prompt Alignment 0.3, Creativity 0.2, Style Control 0.2) to indicate visual fidelity and semantic accuracy.

- **FID Vs. Training time (in hours)**

FID is a performance metric that measures similarities between generated images and real images by comparing the distributions of their respective features. Training time is the number of days that it takes to train or even fine-tune a model up to the optimal performance and desired quality

of output on a given task. Figure 4 displays the FID scores over time of baseline SD vs. optimized SD-VAE-LoRA.

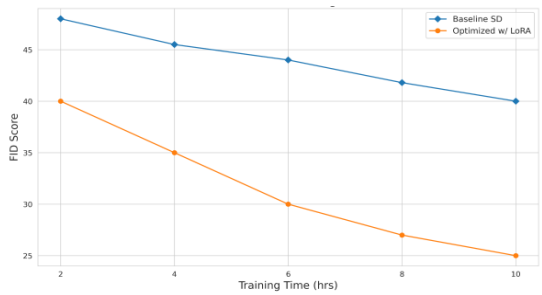

Figure 4: FID score comparison between baseline and optimized models

Figure 4 compares the FID scores with training time in Baseline SD and Optimized SD-VAE-LoRA models. The optimized model exhibits a more pronounced reduction in FID of 40 to 25 during 10 hours, which is faster converging and more faithful to the images than the baseline SD, which declined only by 48 to 40.

4.3 Comparison phase

Research compares the proposed SD-VAE-LoRA framework with existing, diffusion-based image generation methods, specifically Latent Diffusion Method (LDM) [24], Menstrual Cycle-Inspired Latent Diffusion Method (MCI-LDM) [24], and Conditional Generative Adversarial Networks, Attention mechanisms, and Contrastive Learning (C-GAN+ATT+CL) [25]. Table 5 illustrates the efficacy of the proposed SD-VAE-LoRA method, allowing it to generate images of higher quality.

Table 5: Quantitative comparison of image generation quality using PSNR and SSIM metrics

Methods	PSNR	SSIM (%)
LDM [24]	28.5dB	78
MCI-LDM [24]	32.7dB	92
SD-VAE-LoRA [Proposed]	**33.7dB**	**93**

Table 5 compares image quality measures between LDM [24], MCI-LDM [24], and the suggested SD-VAE-LoRA model. The offered method provides the best PSNR of 33.7 dB and SSIM of 93 percent, which is better than MCI-LDM (32.7 dB, 92 percent) and LDM (28.5 dB, 78 percent). These findings indicate that SD-VAE-LoRA has a higher reconstruction performance and structural fidelity.

- **Peak Signal-to-Noise Ratio (PSNR)** and **Structural Similarity Index (SSIM):** PSNR compares the quality of reconstruction between generated and reference images; the higher the PSNR, the higher the image fidelity and the lower the distortion, whereas the lower the PSNR, the higher the noise and the lower the reconstruction accuracy. SSIM compares the perceptual similarity of the generated image and the original image; a value of SSIM means the image has more structural and visual fidelity, whereas a lower value of SSIM means that the generated image loses some details or distorts the generated output.

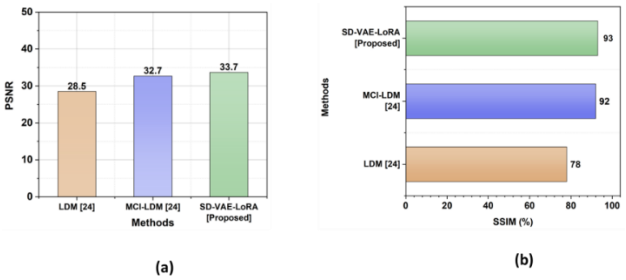

Figure 5: (a) PSNR Comparison of Existing Methods and Proposed Model, (b) SSIM performance across different diffusion models.

Figure 5 (a) shows the PSNR values of three methods, LDM [24], MCI-LDM [24], and the proposed SD-VAE-LoRA framework. The proposed model has the best PSNR of 33.7 dB, which is better than that of MCI-LDM (32.7 dB) and LDM (28.5 dB), and implies high accuracy in reconstruction and quality of the image. Figure 5 (b) shows the performance of SSIM in LDM [24], MCI-LDM [24], and the suggested SD-VAE-LoRA model. The SD-VAE-LoRA records the best SSIM value of 93% compared to MCI-LDM (92%) and LDM (78%), and this demonstrates that it has a higher level of structural consistency and visual fidelity when it comes to generated images.

Table 6: Methodological Comparisons Based on Inception Score, FID Score, and R-Precision

Methods	Inception Score	FID Score	R-Precision
C-GAN+ATT+CL [25]	35.23	18.2	89.14
SD-VAE-LoRA [Proposed]	**36.02**	**17.8**	**90**

Table 6 compares the Inception Score, FID Score, and R-Precision of C-GAN+ATT+CL [25] with that of the suggested SD-VAE-LoRA framework. The proposed model had a higher Inception Score (36.02), R-Precision (90), and a lower FID Score (17.8), which meant that it was more realistic, diverse, and accurately expressed the semantics. The reported findings support the enhanced perceptual quality and text-image correspondence.

- **Inception Score and Fréchet Inception Distance (FID) Score**: Inception Score evaluates the quality and diversity of generated images; a higher IS corresponds to more lifelike and varied image generation, and a lower IS represents a lack of diversity or less persuasive images. FID score compares the similarity between generated pictures and actual ones; a lower FID is used to mean more realistic and higher feature matching, whereas a higher FID is used to mean higher visual or semantic error.

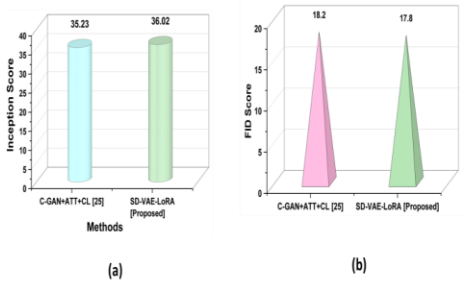

Figure 6: Comparative Analysis of (a) Inception Scores (b) FID Scores for Different Models

According to Figure 6 (a), SD-VAE-LoRA had a higher Inception Score (36.02) than C-GAN+ATT+CL (35.23), which implies that it has better image diversity and visual quality. This illustrates that the model is more generative and realistic. Figure 6 (b) demonstrates that SD-VAE-LoRA (17.8) had a lower FID score than C-GAN+ATT+CL (18.2), which represents greater image fidelity and semantic compatibility. Its better generative accuracy and efficiency is confirmed by the outcome.

- **R-Precision**: It quantifies the similarity between the most privileged generated images and their corresponding text prompts. High R-Precision indicates better text-image alignment, and a low value implies poor semantic correspondence.

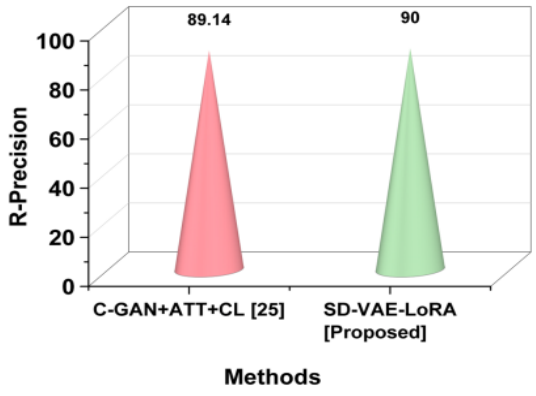

Figure 7: Assessment of R-precision across models

Figure 7 indicates that SD-VAE-LoRA achieved a higher R-Precision (90) compared to C-GAN+ATT+CL (89.14), which means that SD-VAE-LoRA has superior text-image semantic alignment. This enhancement shows greater consistency of generated visuals and input prompts.

- **Mean Intersection over Union (mIoU):** It calculates the overlap between generated and ground-truth regions, which translates into semantic correspondence and picture correctness. An increase in higher mIoU is an indication of a better prompt-to-image consistency delivered by the SD-VAE-LoRA setup.

- **Kernel Inception Distance (KID):** It gives the distribution of generated and real images' similarity using poly-kernel functions. The reduced KID is an indicator of better image realism and adherence to real-world visual semantics of the SD-VAE-LoRA model.

- **Learned Perceptual Image Patch Similarity (LPIPS):** It compares the feature of similarity in perception between generated and reference images on a deep feature representation. A reduced LPIPS signifies that the SD-VAE-LoRA model yields images with a higher visual fidelity and reduced perceptual variance with real images.

Table 7: Quantitative Comparison of Baseline and Proposed Model Performance

Method	mIoU [%	FID	KID	$LPIPS_{VGG}$
Stable Diffusion model [26]	67.89	54.57	0473 ± .0011	6281
SD-VAE-LoRA [Proposed]	72.54	42.16	0.0312 ± 0.0008	0.5124

Table 7 contrasts the results of the default Stable Diffusion model [26] and the suggested SD-VAE-LoRA framework. The proposed model obtains a larger mIoU of 72.54, lower FID (42.16), KID (0.0312 +- 0.0008), and LPIPSVGG (0.5124) with a better image realism, perceptual quality, as well as semantic alignment. These findings corroborate the fact that SD-VAE-LoRA has a higher visual fidelity and prompt consistency than the baseline models.

4.4 Dataset comparison

The optimized Stable Diffusion Image Generation System was trained and verified on the LAION-Aesthetics v2 4.5 dataset, and the generalization performance was tested

against the Text/Image GenAI dataset [27]. This comparative evaluation was aimed at evaluating the flexibility of the Proposed SD-VAE-LoRA architecture on large-scale aesthetic and small curated text-to-image datasets. The assessment involved standard metrics of performance, PSNR, SSIM, FID, Inception Score, and R-Precision, which are used to evaluate reconstruction fidelity, perceptual quality, and semantic consistency.

Table 8: Comparative performance metrics of Proposed datasets.

Metrics	LAION-Aesthetics v2 4.5 (Proposed)	Text/Image GenAI [27]
PSNR (dB)	33.7	32.8
SSIM (%)	93	91
FID	17.8	19.5
Inception Score	36.02	34.67
R-Precision (%)	90	88

Table 8 indicates that LAION-Aesthetics v2 4.5 (Proposed) data have better PSNR (33.7 dB) and SSIM (93%) than Text/Image GenAI [27]. It also reported a decreased FID (17.8) and increased Inception Score (36.02), which demonstrates a better image realism and variety. The R-Precision (90) also attests to the improved text-image alignment.

4.5 Statistical analysis

The research used Pearson correlation analyses to analyse the relationship between PSNR and SSIM, and FID and mIoU, and showed strong positive correlation rates between PSNR and SSIM and mIoU, but a negative correlation between PSNR and FID, indicating high image fidelity and semantic consistency.

Paired sample t-test: It compares the statistical means between two associated groups to ascertain whether the difference between the two groups is statistically significant. It ascertains that the SD-VAE-LoRA model has a larger enhancement in terms of image realism and semantic matching than the baseline model. It is represented as equation (10)

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{10}$$

Where $\dot{d}$is the mean of paired differences, $s_d$is the standard deviation of differences, and $n$is the number of paired observations. Table 9 shows the t-Test-Based Assessment of Model Performance

Table 9: Paired Sample t-Test comparison of model performance metrics

Metric	Baseline Mean	Proposed Mean	Mean Difference	t-value	p-value (Sig.)
mIoU (%)	67.89	72.54	4.65	5.14	0.0003
FID	54.57	42.16	-12.41	-4.88	0.0005
KID	0.0473	0.0312	-0.0161	-3.92	0.0012
LPIPS	0.6281	0.5124	-0.1157	-4.26	0.0008

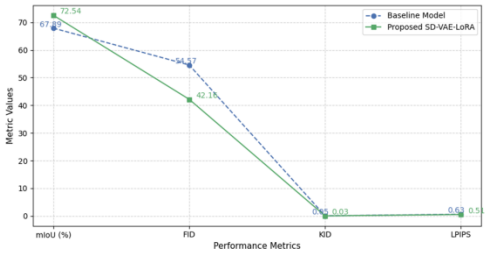

Figure 8: Evaluating Model Metrics Using Paired Sample t-Test

Figure 8 provides the results of a paired sample t-test that SD-VAE-LoRA performs significantly better than the baseline on all metrics, with a greater mIoU (72.54%), and a lower FID (42.16), KID (0.0312), and LPIPS (0.5124). The findings validate statistically significant increases ($p < 0.01$) in image quality and semantic performance.

5 Discussion

The research aims to design and optimize a Stable Diffusion-VAE-LoRA (SD-VAE-LoRA) model to increase prompts and understanding, as well as improve the quality of image generation and allow it to be fine-tuned efficiently and with lightweight. Previous researches found several limitations, such as the lack of usability and practical validation in multilingual diffusion models [20], the low adaptability of domains and weak resistance to noisy prompts in energy-efficient generation systems [21], poor management of extreme motion and short-duration focus in animation-based diffusion models [22], and the inability to test stylistic diffusion models in real-time and on a large scale [23]. The LDM [24] is afflicted with the multi-phase encoding-decoding and slower convergence. It has difficulties in preserving fine image details and semantic correspondence in large-scale generation tasks. MCI-LDM [24] is not very scalable and cannot be

generalized to various areas. It involves a lot of calculation and does not work systematically on abstract or dynamic prompts. C-GAN [25] has the problem of training instability and mode collapse during high-resolution image synthesis. It also requires efficient computational capabilities and hyperparameter sensitivity. ATT [25] models are memory-consuming and prone to overfitting. They occasionally deviate from concentration, decreasing text and image consistency. The CL [25] algorithms have issues with poor cross-modal alignments and sensitivity to noisy prompts. They cannot be used to ensure semantic consistency among various or ambiguous data. Error analysis reveals that prompt-to-image semantics mismatch occasionally, and finer details degrade slightly, which indicates places to consider how to optimize model robustness and visual accuracy. The optimized SD-VAE-LoRA framework overcomes these issues by integrating efficient latent-space denoising, semantic decoding through the VAE, and adaptive fine-tuning with LoRA. This combination reduces overfitting, enhances stability, and ensures faster convergence with improved visual fidelity and prompt accuracy. The system has allowed it to produce images in real-time, with high fidelity, creative, and educational uses, with minimal hardware and computational requirements.

6 Conclusion

Despite these progresses, existing diffusion systems still face challenges in prompt clarification accuracy, model flexibility, and computational efficacy, which limit their performance in real-time and resource-limited settings. The research aims to design and optimize an image generation basis based on Stable Diffusion (SD) that improves prompt processing, improves image quality, and enables lightweight fine-tuning. The optimized SD-VAE-LoRA framework significantly improved semantic alignment, image fidelity, and computational efficiency compared to baseline SD, SD-VAE, and other diffusion systems. On the basis of the LAION-Aesthetics v2 4.5 data, preprocessing based on transformer-based tokenization guaranteed accurate text-image association. VAE, combined with LoRA, allowed the efficient latent-space decoding and adaptive fine-tuning. The system attained PSNR = 33.7 dB, SSIM = 93, FID = 17.8, Inception Score = 36.02, and R-Precision = 90, verification of excellent performance. The constraints are that it is based on scaled curated datasets, limited generalization to abstract or low-context prompts, and relies on computational dependence during large-scale inference. Further research can be done to add multimodal input conditioning, adaptive prompt understanding, domain-specific training of specialized imagery, and real-time deployment in resource-constrained environments to further improve the scalability and generalization across domains.

References

[1] Vartiainen H, & Tedre M (2024). How text-to-image generative AI is transforming mediated action. IEEE Computer Graphics and Applications, 44(2), 12–22. https://doi.org/10.1109/MCG.2024.3355808

[2] Po R, Yifan W, Golyanik V, Aberman K, Barron JT, Bermano A, Chan E, Dekel T, Holynski A, Kanazawa A, & Liu CK (2024). State of the art on diffusion models for visual computing. Computer Graphics Forum, 43(2), e15063. https://doi.org/10.1111/cgf.15063

[3] Yoon J, Yu S, Patil V, Yao H, & Bansal M (2024). Safree: Training-free and adaptive guard for safe text-to-image and video generation. arXiv. https://doi.org/10.48550/arXiv.2410.12761

[4] Alhabeeb SK, & Al-Shargabi AA (2024). Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. IEEE Access, 12, 24412–24427. https://doi.org/10.1109/ACCESS.2024.3365043

[5] Habib MA, Wadud MAH, Patwary MFK, Rahman MM, Mridha MF, Okuyama Y, & Shin J (2024). Exploring progress in text-to-image synthesis: An in-depth survey on the evolution of generative adversarial networks. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3435541

[6] Jin H (2025). Cross-Modal Attention GAN for Text-to-Artistic Image Generation. Informatica, 49(20). https://doi.org/10.31449/inf.v49i20.8303

[7] Tang A, Wei L, Ni Z, & Huang Q (2025). Multi-Modal Modified U-Net for Text-Image Restoration: A Diffusion-Based Multimodal Information Fusion Approach. Informatica, 49(2). https://doi.org/10.31449/inf.v49i2.8245

[8] Arnaubec A, Ferrera M, Escartín J, Matabos M, Gracias N, & Opderbecke J (2023). Underwater 3D reconstruction from video or still imagery: Matisse and 3Dmetrics processing and exploitation software. Journal of Marine Science and Engineering, 11(5), 985. https://doi.org/10.3390/jmse11050985

[9] Tiribelli S, Pansoni S, Frontoni E, & Giovanola B (2024). Ethics of artificial intelligence for cultural heritage: Opportunities and challenges. IEEE Transactions on Technology and Society. https://doi.org/10.1109/TTS.2024.3432407

[10] Mittal S, Wittman RL, Gibson J, Huffman J, & Miller H (2023). Providing a user extensible service-enabled multi-fidelity hybrid cloud-deployable SoS Test and Evaluation (T&E) infrastructure: Application of modeling and simulation (M&S) as a service (MSaaS). Information, 14(10), 528. https://doi.org/10.3390/info14100528

[11] Yuan M, Chen J, Hu Y, Feng S, Xie M, Mohammadi G, Xing Z, & Quigley A (2024). Towards human-AI

synergy in UI design: Enhancing multi-agent-based UI generation with intent clarification and alignment. arXiv. https://doi.org/10.48550/arXiv.2412.20071

[12] Hariri W (2023). Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. arXiv. https://doi.org/10.48550/arXiv.2304.02017

[13] Kabir AI, Mahomud L, Al Fahad A, & Ahmed R (2024). Empowering local image generation: Harnessing stable diffusion for machine learning and AI. Informatica Economica, 28(1), 25–38. https://doi.org/10.24818/issn14531305/28.1.2024.03

[14] Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarburger C, Schulze-Hagen M, Schad P, Engelhardt S, Baeßler B, Foersch S, & Stegmaier J (2023). Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports, 13(1), 7303. https://doi.org/10.1038/s41598-023-34341-2

[15] Dehouche N, & Dehouche K (2023). What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon, 9(6). https://doi.org/10.1016/j.heliyon.2023.e16757

[16] Özbey M, Dalmaz O, Dar SU, Bedel HA, Özturk Ş, Güngör A, & Cukur T (2023). Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging, 42(12), 3524–3539. https://doi.org/10.1109/TMI.2023.3290149

[17] Packhäuser K, Folle L, Thamm F, & Maier A (2023). Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. IEEE International Symposium on Biomedical Imaging, 1–5. https://doi.org/10.1109/ISBI53787.2023.10230346

[18] Paananen V, Oppenlaender J, & Visuri A (2024). Using text-to-image generation for architectural design ideation. International Journal of Architectural Computing, 22(3), 458–474. https://doi.org/10.1177/14780771231222783

[19] Lian L, Li B, Yala A, & Darrell T (2023). LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv. https://doi.org/10.48550/arXiv.2305.13655

[20] Chen J, Yu J, Ge C, Yao L, Xie E, Wu Y, Wang Z, Kwok J, Luo P, Lu H, & Li Z (2023). Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv. https://doi.org/10.48550/arXiv.2310.00426

[21] Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, Agrawala M, Lin D, & Dai B (2023). AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv. https://doi.org/10.48550/arXiv.2307.04725

[22] Huang N, Zhang Y, Tang F, Ma C, Huang H, Dong W, & Xu C (2024). Diffstyler: Controllable dual diffusion for text-driven image stylization. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3342645

[23] Luccioni AS, Akiki C, Mitchell M, & Jernite Y (2023). Stable bias: Analyzing societal representations in diffusion models. arXiv. https://doi.org/10.48550/arXiv.2303.11408

[24] Mahmoud GM, Elbaz M, Said W, & Elsonbaty AA (2025). Menstrual cycle inspired latent diffusion model for image augmentation in energy production. Scientific Reports, 15(1), 16749. https://doi.org/10.1038/s41598-025-99088-4

[25] Habib MA, Wadud MAH, Pinky LY, Talukder MH, Rahman MM, Mridha MF, Okuyama Y, & Shin J (2023). GACnet: Text-to-image synthesis with generative models using attention mechanisms with contrastive learning. IEEE Access, 12, 9572–9585. https://doi.org/10.1109/ACCESS.2023.3342866

[26] Kaleta J, Dall'Alba D, Płotka S, & Korzeniowski P (2024). Minimal data requirement for realistic endoscopic image generation with stable diffusion. International Journal of Computer Assisted Radiology and Surgery, 19(3), 531–539. https://doi.org/10.1007/s11548-023-03030-w