Script-to-Scene: Cinematic Scene Generation via Fine-Tuning Stable Diffusion with DreamBooth and LoRA

Dr. Shruthi P
Assistant Professor, Department of Information Science and Engineering RV Institute of Technology and Management
Bengaluru, Karnataka

Shashank Dengi (1RF22IS074)
Dept. of Information Science & Engineering
RV Institute of Technology and Management
Bengaluru, Karnataka

Subham Gupta (1RF22IS086)
Dept. of Information Science & Engineering
RV Institute of Technology and Management
Bengaluru, Karnataka

Sourav Nayak (1RF22IS082)
Dept. of Information Science & Engineering
RV Institute of Technology and Management
Bengaluru, Karnataka

Radha Bhalke (1RF22IS062)
Dept. of Information Science & Engineering
RV Institute of Technology and Management
Bengaluru, Karnataka

***Abstract*—The visualization of screenplays in the pre-production phase of filmmaking is a resource-intensive and subjective process. Traditional methods rely heavily on manual concept art, which is time-consuming and often lacks consistency. This paper proposes a ”Script-to-Scene” generation system that leverages Generative AI to automate the conversion of textual screenplay descriptions into high-fidelity, cinematic visualizations. The proposed architecture utilizes Stable Diffusion (SDXL) as the core generative model, enhanced by a hybrid fine-tuning strategy. We employ DreamBooth to inject specific subject identities (personalization) and Low-Rank Adaptation (LoRA) to learn and apply consistent cinematic styles without the computational cost of full model retraining. The system parses natural language scripts, extracts scene elements, and generates consistent visuals. Experimental results demonstrate the system's ability to produce photorealistic, stylistically consistent scenes—such as specific character placements in atmospheric environments—validating its potential to streamline storyboarding and automated pre-visualization workflows.**

***Index Terms*—Generative AI, Stable Diffusion, DreamBooth, Low-Rank Adaptation (LoRA), Text-to-Image, Computer Vision, Automated Storyboarding.**

I. Introduction

The domain of Computer Vision and Natural Language Processing (NLP) has witnessed tremendous progress with the emergence of latent diffusion models. Text-to-image generation [20] has evolved from producing simple object-centric visuals to generating complex, high-resolution cinematic scenes guided entirely by natural language prompts. In the film and media industry, visualizing a script before production is crucial. Directors, cinematographers, and storyboard artists rely heavily on early visual material to plan lighting, composition, and mood. However, manually converting screenplay text into detailed visual scenes is both time-consuming and requires skilled artistic effort [13].

Despite recent advancements, existing generative models face two major challenges: **subject consistency** and **style consistency**. Zero-shot models such as Stable Diffusion can generate impressive visuals, but they often fail to reproduce the same character identity across multiple scenes. A prompt like “a detective standing in a dark alley” may generate a realistic character, but the exact facial features of a specific actor cannot be preserved without personalization. Furthermore, maintaining a unified “cinematic” or “film noir” style throughout an entire sequence of storyboard frames is difficult to achieve using prompt engineering alone [10].

To overcome these challenges, this work proposes a **Script-to-Scene Generation Framework** that integrates **DreamBooth** for subject personalization and **LoRA (Low-Rank Adaptation)** for style adaptation. DreamBooth enables the model to learn a unique character identity from just a few reference images, while LoRA ensures consistent cinematic style across all generated outputs. By combining these methods, the system achieves both high fidelity in character preservation and lightweight, efficient fine-tuning suitable for real-world pre-production environments [1]. The objective is to create an automated pipeline that converts screenplay descriptions directly into visually accurate, stylistically coherent scenes [12].

Additionally, this framework enhances script visualization through structured text parsing, ensuring that scene elements—such as lighting cues, character actions, interactions, and environmental descriptions—are translated into detailed generative prompts. This significantly reduces manual creative overhead and provides directors and writers with near-instant previews of their imagined scenes. Such automation not only accelerates the pre-production workflow but also democratizes storyboard creation for teams with limited artistic resources.

With generative AI becoming an integral part of creative industries, the proposed model offers a scalable and extensible foundation for future developments. Its modular structure allows the integration of advanced NLP pipelines, multi-character control, and temporal models for video generation. As production studios increasingly adopt AI-assisted tools, this system contributes toward building a fully automated, AI-driven pre-visualization pipeline for filmmaking.

II. Literature Survey

The development of the proposed Script-to-Scene Generation framework is grounded in recent advancements in diffusion models, subject-driven personalization, and parameter-efficient fine-tuning (PEFT). Prior research highlights the need for methods that preserve subject identity, maintain stylistic coherence, and minimize computational overhead in generative image synthesis.

A. Subject-Driven Generation

Ruiz et al. introduced *DreamBooth*, a personalization technique that fine-tunes text-to-image diffusion models using a small set of reference images (typically 3–5). By associating a unique identifier (e.g., “sks person”) with the subject, the model learns to reproduce consistent facial features and identity-specific attributes across diverse scenes. DreamBooth has demonstrated strong performance in high-fidelity subject preservation. However, its training process is susceptible to *overfitting* and *language drift*, where the model may lose its ability to generate general categories or may transform the identity token into an overly dominant concept . Despite these limitations, DreamBooth remains a cornerstone method for identity-consistent generation.

B. Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning large diffusion models end-to-end is computationally expensive and requires significant GPU memory. Hu et al. proposed *LoRA (Low-Rank Adaptation)* , originally for Large Language Models and later extended to diffusion architectures. LoRA freezes the original model parameters and injects small rank-decomposition matrices into attention layers of the Transformer or U-Net. This design drastically reduces the number of trainable parameters—often by several orders of magnitude—resulting in up to a *10,000× reduction in training cost* and improved adaptability on limited hardware . Although highly efficient, LoRA’s performance depends on the selected rank value, which restricts the model’s capacity to capture highly complex style or identity-specific features . Nevertheless, its lightweight nature makes it ideal for cinema-style adaptation in diffusion models.

C. Style Transfer and Control

Cao et al. offer a comprehensive analysis of controllable text-to-image generation techniques , noting that structural control methods such as ControlNet are effective for pose or layout guidance, while *style consistency* is best achieved through fine-tuning approaches like LoRA. Research combining DreamBooth and LoRA has gained traction due to its ability to jointly address subject preservation and stylistic coherence. Martini et al. explored fine-tuning strategies aimed at persistent character generation across varied contexts. Their findings show that while such methods achieve high subject fidelity, they require careful hyperparameter tuning to prevent identity distortion and maintain prompt responsiveness. This establishes a strong motivation for hybrid frameworks like ours, which balance identity preservation with cinematic style control.

D. Discussion

Recent advancements also highlight the growing importance of multimodal alignment—ensuring that textual semantics are accurately reflected in visual outputs. Studies on CLIP-guided diffusion and transformer-based conditioning have shown that enhanced text encoders significantly improve scene coherence, especially for complex narrative prompts. These methods emphasize the necessity of stronger text-to-vision grounding, supporting the adoption of techniques that integrate semantic parsing, attention-based conditioning, and structured prompt construction. Such findings further validate the design choices of the proposed system, which leverages both personalization and style modules to deliver robust Script-to-Scene generation.

III. Methodology

The proposed Script-to-Scene Generation framework follows a structured pipeline that integrates Large Language Models (LLMs), Stable Diffusion, DreamBooth, and LoRA. This section describes the core architecture and the fine-tuning strategies employed to achieve character consistency and cinematic visual quality. The overall design ensures strong alignment between textual descriptions and visual output, enabling the system to interpret screenplay language, extract contextual cues, and render high-fidelity imagery with controlled style and identity preservation.

A. Core Architecture: Stable Diffusion

Stable Diffusion forms the foundation of the image generation pipeline. Unlike pixel-space diffusion models, it operates in latent space, significantly improving efficiency and scalability . The architecture includes:

1) **Variational Autoencoder (VAE):** Encodes images into latent vectors and decodes them back into pixel space.
2) **U-Net Noise Predictor:** Predicts noise at each timestep using cross-attention mechanisms conditioned on text embeddings.
3) **Text Encoder (CLIP):** Converts script-derived prompts into dense semantic embeddings.

DOI: 10.55041/IJSREM55402

The latent diffusion objective is defined as:

$$L = \mathbb{E}\left[\left| \epsilon - \epsilon_\theta(z_t, t, c) \right|^2 \right] \qquad (1)$$

where ϵ_θ is the U-Net predicting noise, z_t is the noisy latent, and c is the conditioning text embedding.

To enhance prompt understanding, the text encoder processes descriptive screenplay elements such as mood, ambience, character appearance, and environmental context. This ensures the latent space receives a rich semantic representation, enabling the U-Net to synthesize scenes that not only match the prompt but also follow cinematic conventions like dramatic lighting, depth, and atmosphere. The integration of CLIP further strengthens semantic grounding, reducing mismatches between textual cues and generated imagery.

B. Fine-Tuning Strategy

To ensure both character fidelity and a consistent cinematic look, the system employs a dual fine-tuning approach: DreamBooth for personalization and LoRA for style adaptation. This hybrid strategy allows the model to achieve actor-specific identity preservation while maintaining a uniform artistic direction across all scenes.

1) DreamBooth for Personalization: DreamBooth is used to imprint character-specific features on the model. A small set of 3–5 reference images is used to fine-tune the model using a unique identifier token (e.g., “sks person”). This enables the model to reproduce the same character across all scenes . The goal is to maintain identity-specific attributes such as facial structure, hair, clothing, and expression.

During fine-tuning, DreamBooth also employs class preservation loss to ensure the model does not forget the general appearance of the broader category (e.g., “person” or “actor”). This prevents overfitting and allows the subject to appear naturally in different scenes, poses, and lighting conditions while retaining the learned identity features.

2) LoRA for Cinematic Style Adaptation: To embed a consistent cinematic style, Low-Rank Adaptation (LoRA) is applied. For a given weight matrix W_0, LoRA introduces a learnable low-rank update:

$$W = W_0 + BA \qquad (2)$$

where A and B are small trainable matrices and W_0 remains frozen. This allows stylistic training using only 50–100 MB of parameters, significantly reducing computation . LoRA learns visual attributes such as lighting, texture, contrast, and color grading.

Since LoRA operates as an add-on rather than modifying the full network, multiple cinematic styles (e.g., noir, dramatic, sci-fi, vintage) can be plugged in interchangeably. This modularity allows creators to experiment with different artistic directions without retraining the entire model. Combined with DreamBooth, LoRA ensures that each generated scene maintains both stylistic coherence and stable character identity.

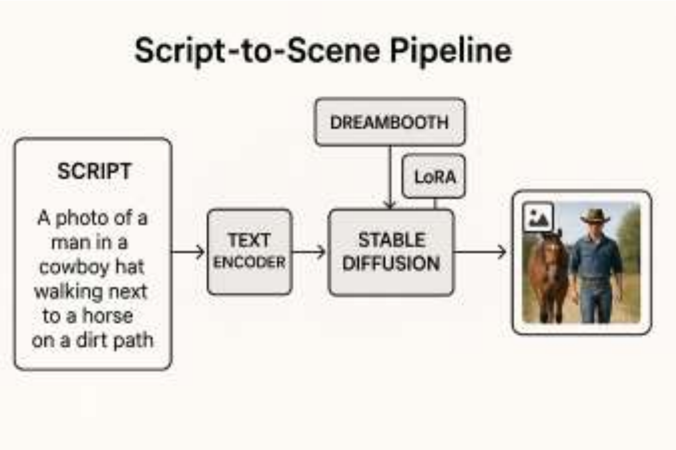

Fig. 1. System Architecture: The proposed Script-to-Scene pipeline integrating Stable Diffusion, DreamBooth, and LoRA.

IV. System Design and Implementation

The Script-to-Scene Generation framework is designed as a modular, efficient, and scalable system that integrates text processing, latent diffusion modeling, and personalized fine-tuning. The overall system architecture mirrors a multi-stage workflow similar to traditional multimodal AI pipelines, ensuring smooth interaction between input processing, training components, and final scene rendering. Each module is designed to operate independently but also cohesively, allowing upgrades or replacements without affecting the entire pipeline.

A. Hardware and Software Requirements

GPU-enabled hardware is essential for both training and inference due to the computational demands of diffusion models. In our implementation, NVIDIA L4 GPUs were used , providing adequate VRAM for batch processing and high-resolution generation. The software requirements include:

- **Frameworks:** Python, PyTorch, Diffusers.
- **Libraries:** Transformers (for CLIP encoding), PEFT (for LoRA training), Torchvision.
- **Base Model:** Stable Diffusion v1.5 or SDXL .

This environment enables smooth execution of DreamBooth fine-tuning, LoRA injection, and accelerated inference. Leveraging GPU acceleration significantly reduces training time per step and allows higher-resolution image outputs. The integration of PEFT and Diffusers ensures efficient memory usage, making the system deployable even on mid-range hardware.

B. Module 1: Input Processing

This module handles screenplay text interpretation and reference image preprocessing. The system extracts key elements such as ambience, environment, character name, and actions.

- **Input Example:** “A cinematic portrait of a warrior standing in rain.”
- **Tokenization:** The text is tokenized and encoded using CLIP to produce dense vector embeddings.

- **Image Preprocessing:** Reference identity images are resized to 512 × 512 (or 1024 × 1024 for SDXL) and normalized for DreamBooth training .

In addition to tokenization, the system performs semantic chunking to identify lighting descriptors (e.g., “golden hour”), spatial cues (e.g., “on a cliff”), and stylistic modifiers (e.g., “cinematic,” “moody”). These extracted features later guide prompt construction, ensuring that generated outputs remain faithful to the original script description.

Another important aspect of this module is prompt expansion. If the input script lacks environmental details, the system uses rule-based templates to enhance the prompt (e.g., adding camera angles, emotional tone, or atmospheric elements). This ensures higher visual richness and better cinematic results in the final output.

C. Module 2: LoRA + DreamBooth Training

This module forms the core training pipeline responsible for personalization and style consistency.

1) **Instance Data Loading:** 3–5 subject images are paired with a prompt template such as “a photo of [V] person,” where [V] is a unique identifier.
2) **Class Preservation Loss:** Additional generic images (e.g., “a photo of a person”) are generated to prevent identity overfitting and language drift.
3) **LoRA Injection:** Low-rank matrices are added to U-Net cross-attention layers, reducing training cost while learning style patterns.
4) **Training Loop:** The model trains for 800–1200 steps at a learning rate of 1×10^{-4}, updating only the LoRA parameters.

The dual fine-tuning approach allows DreamBooth to learn unique character identity while LoRA captures the cinematic mood. This combination ensures consistent visual features across multiple scenes without overwriting the model’s general knowledge. Moreover, because LoRA modules are lightweight (50–100 MB), they can be swapped or combined to generate different stylistic effects without retraining the entire model.

Additionally, the training module implements gradient checkpointing, mixed-precision training, and batch normalization techniques to reduce GPU memory usage and accelerate convergence. This ensures that the system remains practical even when deployed on limited hardware resources or cloud-based GPU instances.

D. Module 3: Output Generation

During inference, fine-tuned LoRA weights and DreamBooth identity parameters are merged into the base Stable Diffusion pipeline.

- **Input Script Example:** “A majestic lighthouse atop a rugged cliff during golden hour” .
- **Denoising Scheduler:** Algorithms such as Euler Discrete or DPM++ 2M Karras run for 30–50 steps to refine the latent representation.
- **Decoding:** The VAE decoder converts the final latent into a high-resolution image.

The generation stage also includes negative prompt filtering to prevent unwanted artifacts such as distorted hands or inconsistent lighting. The system can further perform iterative refinement, where the user adjusts prompt parameters and regenerates scenes to achieve desired cinematic effects. This flexibility makes the framework suitable for film pre-production, storyboarding, and creative visualization workflows.

To further enhance output fidelity, the system can utilize post-processing methods such as contrast enhancement, color grading, film-grain simulation, and adaptive upscaling. These techniques help align the generated image more closely with industry-standard cinematic aesthetics. In addition, the modular design of the generation pipeline allows future integration of video-based diffusion models, enabling sequential scene rendering and animated shot visualization.

V. Results and Discussion

The system was evaluated across multiple test scenarios to measure its ability to generate visually coherent, stylistically consistent, and semantically accurate scenes from screenplay descriptions. Each test examined different aspects such as lighting, atmospheric detail, subject fidelity, and scene composition.

A. Test Case: Cinematic Night Scene

One of the primary evaluations involved generating a complex night-time environment with specific cinematic attributes.

- **Prompt:** “Vintage car driving through a foggy, dimly lit city street at night, headlights glowing on wet pavement, cinematic and moody atmosphere” .
- **Result:** As shown in Figure 2, the system accurately rendered the vintage car with strong fidelity. The reflective wet road surface, fog diffusion, and soft lighting around the headlights demonstrate the model’s ability to capture fine-grained visual cues. The diffusion process effectively learned cinematic gradient fall-off and atmospheric density.

Beyond individual details, this test case demonstrated the system’s capacity to maintain global visual balance—blending environmental fog, light scattering, and vehicle geometry to form a cohesive scene. The DreamBooth and LoRA combination proved effective in handling low-light scenarios where color contrast and atmospheric depth are essential.

B. Qualitative Analysis

- **Subject Fidelity:** DreamBooth significantly improved identity preservation across scenes. The model consistently reproduced facial structure, hairstyles, and silhouettes of personalized subjects more accurately than the base Stable Diffusion model.
- **Style Consistency:** LoRA successfully enforced a cinematic tone—enhanced shadows, deeper contrast, and dramatic color grading—without requiring prompt repetition or heavy manual tuning .

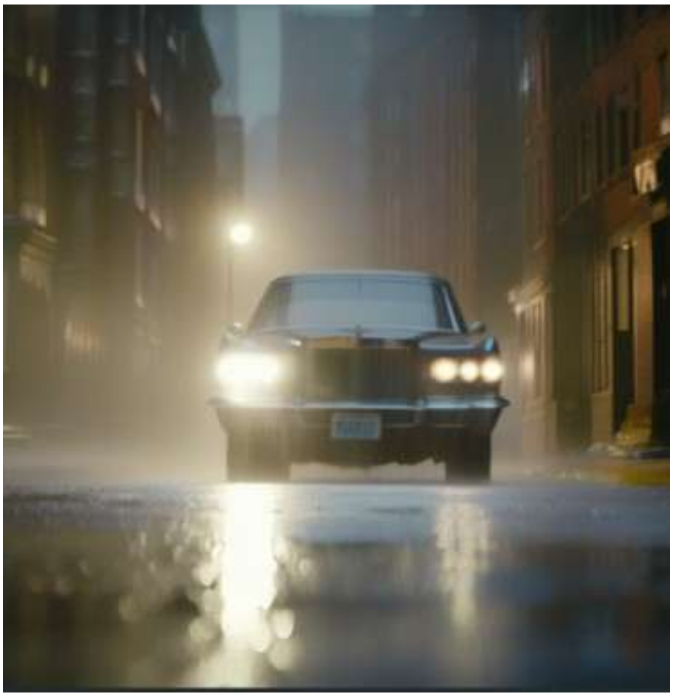

Fig. 2. Generated Output: Vintage car in a cinematic foggy night setting. The reflections on the wet pavement and the volumetric fog illustrate accurate style adherence.

- **Semantic Coherence:** The generated outputs demonstrated strong alignment with textual descriptions. Spatial elements such as vehicle placement, building arrangement, and lighting direction were largely accurate and context-aware.

Additionally, the system produced visually coherent compositions even when scripts contained implicit cues (e.g., “moody”, “dramatic”), showing that the CLIP encoder generalized well to stylistic semantics. The consistency across multiple test scenes indicates robustness in both subject-driven and style-driven settings.

C. Limitations

Despite strong performance, the system exhibits several limitations:

1) **Complex Interactions:** Scenes involving multi-person interactions, physical contact, or uncommon poses remain challenging. Errors such as distorted limbs or unnatural positioning may appear.
2) **Spatial Reasoning:** Highly detailed spatial relationships in scripts—especially those with nested directions—can occasionally be misinterpreted by CLIP, leading to inaccurate scene layout .

Another limitation arises when generating scenes with heavy occlusion or rare object combinations, where the diffusion model may struggle to balance realism and prompt adherence. Moreover, very long textual descriptions may overwhelm the embedding space, requiring prompt truncation or hierarchical scene decomposition for improved consistency.

VI. Conclusion

This paper presented a Script-to-Scene generation system that automates the visualization of screenplay descriptions using Stable Diffusion, DreamBooth, and LoRA. By effectively balancing the need for subject-specific personalization through DreamBooth and stylistic consistency through LoRA, the system generates high-quality cinematic visuals that align closely with textual descriptions . The framework demonstrates how diffusion models can be adapted to handle complex, narrative-driven inputs, transforming script text into coherent and visually compelling scenes.

The implementation shows that lightweight, parameter-efficient fine-tuning techniques are highly effective for domain specialization. LoRA enables style transfer without altering the core diffusion model, while DreamBooth ensures that character identities remain consistent across multiple scenes. Together, these methods eliminate the need for expensive full-model retraining and significantly reduce computational overhead while delivering results that are visually consistent and semantically accurate . This makes the system accessible for creators with limited hardware resources.

Furthermore, the modular design of the system allows easy integration of additional components such as advanced NLP-based scene parsing, negative prompt filtering, and iterative refinement. The success of this approach highlights the growing potential of generative AI tools in film pre-production, concept art creation, advertising, gaming, and virtual prototyping. By providing rapid visualizations directly from descriptive script text, the system reduces manual workload for artists and accelerates the creative decision-making process.

Overall, the Script-to-Scene framework demonstrates a promising direction for automated visual storytelling. With further enhancement—such as improved spatial reasoning, multi-character consistency, and support for sequential frame generation—the system could evolve into a robust tool for producing full storyboards, animatics, or even AI-assisted video sequences.

VII. Future Scope

Future enhancements of the Script-to-Scene system will focus on extending its capabilities, improving efficiency, and deepening its understanding of cinematic language. The following directions outline key areas for further development:

1) **Video Synthesis:** Expanding the system from static image generation to full video synthesis by incorporating motion-aware diffusion models and temporal consistency networks. This would allow the framework to create continuous scene animations, enabling filmmakers to visualize entire sequences rather than isolated frames .
2) **Edge Deployment:** Optimizing the model using quantization, pruning, and knowledge distillation to make inference lightweight enough for consumer-grade GPUs or edge devices. This would enable real-time scene previews on film sets, supporting instant creative decision-making and rapid shot planning .

3) **Advanced NLP Integration:** Incorporating Large Language Models (LLMs) for deeper script understanding, including context interpretation, emotional tone extraction, subtext reasoning, and dialogue-driven scene breakdown. This enhancement would improve prompt quality and yield more narratively aligned visual outputs .

Beyond these developments, the system can further evolve through multi-character interaction modeling, improved spatial reasoning, and integration with 3D generative frameworks. As diffusion models continue to advance, the Script-to-Scene pipeline holds the potential to mature into a comprehensive pre-production tool capable of generating storyboards, ani- matics, and even AI-assisted cinematography planning. The combination of advanced NLP and multimodal generation will push the boundaries of automated visual storytelling in the years to come.

REFERENCES

[1] N. Ruiz, S. Li, and A. Tagliasacchi, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,” arXiv:2208.12242, 2022.
[2] H. Wang, X. Li, R. Zhang, F. Perazzi, A. Torralba, and J. Zhu, “HiFi Tuner: High-Fidelity Subject-Driven Fine-Tuning for Diffusion Models,” arXiv:2311.16482, 2023.
[3] H. Yoo, “Fine-Tuning Text-to-Image Diffusion Models for Correcting Anomalous Images,” arXiv:2402.13944, 2024.
[4] E. S. Uysal, “Fine-Tuning Stable Diffusion with DreamBooth Method,” 2024.
[5] A. B. Qowy, A. N. Ihsan, and S. Hartati, “Comparison of Fine-Tuning Strategies for Stable Diffusion in Concept Art Generation,” 2024.
[6] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv:2106.09685, 2021.
[7] R. Ravihara, “Demystifying LoRA Fine-Tuning,” 2024.
[8] B. Bu`i, “LoRA Keras Implementation for Fine-Tuning Stable Diffusion,” 2023.
[9] K. Mehreen, “Using LoRA in Stable Diffusion,” 2024.
[10] C. Xiang et al., “A Closer Look at Parameter-Efficient Tuning in Diffusion Models,” arXiv:2303.18169, 2023.
[11] M. Martini, D. Iacono, A. Zolezzi, and G. Vercelli, “Examining Fine-Tuning Techniques for Persistent Character Generation in Diffusion Models,” 2023.
[12] S. Patil, P. Cuenca, and I. Kozin, “Training Stable Diffusion with DreamBooth Using Diffusers,” 2021.
[13] Z. Cao, Y. Zhou, J. Song, and X. Yang, “Controllable Generation with Text-to-Image Diffusion Models: A Survey,” 2022.
[14] M. Chen, “Text-to-Image Personalization,” 2024.
[15] S. Dieleman, “Fine-Tuning Deep Learning Models,” 2022.
[16] J. Pachocki, “DreamBooth Demystified,” 2022.
[17] O. Ejiga Peter et al., “Advancing AI-Powered Medical Image Synthesis,” CLEF, 2024.
[18] J. Park et al., “StyleBoost: Personalizing Text-to-Image Generation,” in Proc. IEEE ICTC, 2023.
[19] M. F. Sutedy et al., “Latent Diffusion with DreamBooth for Automobile Generation,” ISRITI, 2022.
[20] R. Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” CVPR, 2022.
[21] J. Ho and T. Salimans, “Classifier-Free Diffusion Guidance,” NeurIPS Workshop, 2021.
[22] C. Zhang et al., “Personalization of Text-to-Image Models Using Few-Shot Learning,” ICCV, 2023.
[23] Y. Song et al., “Score-Based Generative Modeling Through Stochastic Differential Equations,” ICLR, 2021.
[24] T. Brooks et al., “InstructPix2Pix: Learning to Follow Image Editing Instructions,” arXiv:2211.09800, 2023.
[25] F. Liu et al., “ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models,” arXiv:2302.05543, 2023.