Selective Gradient Diffusion for Localized Text-Guided Image Augmentation

1st G Sucharitha
dept of Computer Science and Engineering
Anurag University
Hyderabad, India
sucharithasu@gmail.com
0000-0001-7861-2451

2nd Harun Elkiran
Research Institute
Istanbul Medipol University
Istanbul 34810, Turkey
harun.elkiran@medipol.edu.tr
0000-0002-5834-6210

3rd Jawad Rasheed
dept of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul 34303, Turkey
jawad.rasheed@izu.edu.tr
0000-0003-3761-1641

***Abstract*—Diffusion models have demonstrated strong capabilities in text-guided image editing; however, most existing approaches update the global parameters of the generative network, leading to unintended modifications in non-target regions and reduced structural fidelity. This work introduces a Selective Gradient Diffusion (SGD) framework for text-guided image-to-image augmentation, designed to achieve localized modifications while preserving the integrity of the surrounding content. The proposed architecture leverages a latent diffusion backbone in which low-rank adapters (LoRA) are embedded within cross-attention layers of the U-Net to enable parameter-efficient fine-tuning. To further constrain edits, a region-weighted noise prediction loss emphasizes modifications within specified masks, and a gradient-masking strategy restricts weight updates to selected neurons. This combination ensures that edits driven by natural language prompts are confined to semantically relevant regions without disturbing unrelated pixels. Experiments conducted on interior design datasets demonstrate that the proposed method achieves higher edit precision, improved structural preservation, and reduced parameter overhead compared to state-of-the-art diffusion-based editing approaches. The results suggest that selective neuron updating in diffusion models offers an effective direction for controllable and efficient text-guided image augmentation. The supportive code is available at: https://sucharithasu.github.io/SGDWeb/,**

***Index Terms*—Selective gradient diffusion, LoRA, U-Net, Text-Guided, image augmentation..**

I. Introduction

In the image synthesis process, recent advancements in diffusion models have demonstrated their superiority over existing GAN-based approaches. Diffusion models belong to the family of generative models and are capable of producing high-quality, semantically coherent images conditioned on natural language prompts. The fundamental principle of diffusion involves gradually corrupting data with Gaussian noise during a forward diffusion process and then training a neural network to learn the reverse process, step by step, in order to reconstruct the original signal or generate novel images from pure noise [1]- [3]. Recent research has explored text-guided image manipulation and augmentation using diffusion models, demonstrating significant advances over traditional methods [4]- [6]. These diffusion models represent a state-of-the-art approach in generative artificial intelligence, particularly excelling in image generation and serving as key components in text-to-image generators and large language models [7]. These models operate by systematically adding noise to training data and then learning to reverse this process, enabling the generation of new synthetic outputs from random data [8]- [9]. The inversion process has recently drawn a considerable attention over GAN models [10]- [12], but not addressing the problems with text guided diffusion models. Gallon *et al.,* [13] provide a mathematically rigorous framework for denoising diffusion probabilistic models (DDPMs), covering training procedures, generation methods, and extensions including improved DDPMs, denoising diffusion implicit models, classifier-free diffusion guidance, and latent diffusion models. The mathematical foundations connect to partial differential equation diffusion models, making them relevant for courses in stochastic processes, inference, machine learning, and scientific computing [14]. Although DDPM has achieved high quality image generation without adversarial training, yet they require simulating a Markov chain for many steps to produce a sample, making computationally expensive. To address efficiency concerns, Denoising Diffusion Implicit Models (DDIMs) [15] construct non-Markovian diffusion processes that maintain the same training objective while enabling 10× to 50× faster sampling. It is a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs. Low-Rank Adaptation (LoRA) has emerged as a significant fine-tuning technique for diffusion models, allowing adaptation to specific domains, characters, styles, or concepts using limited context examples [16]- [17]. LoRA's importance lies in its ability to efficiently customize pre-trained diffusion models like Stable Diffusion for specialized tasks. Advanced guidance techniques like AutoLoRA further enhance LoRA-fine-tuned models by balancing domain consistency with sample diversity, improving both quality and variability in generated images. However, most existing approaches update global network parameters, often causing unintended alterations in non-target regions and loss of structural fidelity. Furthermore, current inpainting methods lack fine-grained control over where and how edits are applied, limiting their applicability in domains such as interior design or medical imaging where contextual preservation is critical. To overcome these issues, we introduce a Selective Gradient

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

Diffusion (SGD) approach that incorporates low-rank adapters (LoRA) within the cross-attention units of the U-Net so as to facilitate parameter-efficient fine-tuning without full retraining of the diffusion backbone. In our model, a region-weighted noise prediction loss is utilized to target updates within user-specified or automatically identified masks such that edits are confined to the desired region and minimize changes in background regions. To complement further selectivity, we propose a gradient-masking strategy that limits parameter updates to the most important neurons, avoiding irrelevant drift in global representations. Unlike traditional inpainting approaches that use pixel-level blending alone, our framework is fully conducted in the latent space, where semantic structures are maintained, and edits can be effectively controlled. This blend of LoRA-based adaptation, spatially sensitive loss, and gradient control at the neuron level enables SGD to attain text-guided, fine-grained image-to-image augmentation that is efficient in terms of computation, robust, and effective in interior design, where structural fidelity preservation is of the highest concern.

II. Related Work

Kim *et al.,* [18], introduced the DiffusionCLIP model, which performs text-driven image manipulation by combining diffusion models with CLIP loss. This approach achieves performance comparable to GAN-based methods while providing superior inversion capabilities and enabling manipulation across unseen domains. The approach successfully handles diverse real images from ImageNet and supports multi-attribute manipulation through novel noise combination techniques. For data augmentation applications, diffusion models have proven effective in generating diverse, contextually rich training data. Shin *et al.,* [19], demonstrate that text-to-image diffusion models using rich-text prompts, multi-object generation, and inpainting techniques significantly improve classification accuracy on Oxford-IIIT Pets and Caltech-101 datasets, with inpainting particularly excelling at handling class imbalances. Custom-Edit addresses precision issues in text-guided editing by customizing diffusion models with reference images, discovering that customizing only language-relevant parameters with augmented prompts maintains source similarity while improving reference similarity [20]. Dong *et al.,* [21] introduced two-stage text-drive image editing model. In the first stage, they represented the input image as a learnable conditional embedding by prompt tuning inversion, and in the second stage they used classifier-free guidance to sample the edited image. In this the conditional embedding is computed by linearly interpolating between the target embedding and the optimized embedding achieved in the first stage. In the extension of this research, Ruiz *et al.,* [22] proposed DreamBooth technique for text-to-image diffusion models. This technique fine-tunes pretrained models using just a few reference images to bind a unique identifier with a specific subject. The method employs an autogenous class-specific prior preservation loss to enable synthesis of subjects in diverse contexts while preserving key features. DreamBooth often overlooks learned concepts when integrating them into new prompts, attributed to incorrect learning of embedding alignment and it also suffering with overfitting. Brooks *et al.,* [23], introduced a novel approach called InstructPix2Pix for image editing from human instructions by combining GPT-3 and Stable Diffusion to generate training data, enabling quick edits without per-example fine-tuning. The model performs edits in a single forward pass without requiring per-example fine-tuning, enabling quick image editing in seconds. Several extensions have been developed to improve its capabilities like InstructPix2Pix [24]. Building on these models, Instruct-NeRF2NeRF extended instruction-based editing to 3D scenes by iteratively editing input images with InstructPix2Pix while optimizing the underlying NeRF representation in [25]. The recent research noted with numerous approaches to text-guided image editing using diffusion models, each addressing different aspects of control and precision. Zhang *et al.,* [26] introduced a new approach, Forgedit: a vision-language joint optimization framework that reconstructs images in 30 seconds while proposing a vector projection mechanism in text embedding space to separately control identity similarity and editing strength. The method discovers that U-Net encoders learn space and structure while decoders learn appearance and identity, leading to forgetting mechanisms that tackle overfitting issues, but lacks explicit spatial masking and region-level control, limiting precise localized modifications compared to newer selective diffusion approaches. Liu *et al.,* [27], proposed S²Edit, which focuses on personalized editing with precise semantic and spatial control by fine-tuning models to embed identity information into learnable text tokens while enforcing orthogonality constraints to disentangle identity from editable attributes. Hu *et al.,* [28] introduced DCEdit, it is a Precise Semantic Localization strategy using visual and textual self-attention to enhance cross-attention maps, coupled with a Dual-Level Control mechanism operating at both feature and latent levels. Li *et al.,* [29] proposed a mask matching module with fusion-diffusion technique that creates masks corresponding to textual descriptions to guide localized editing.

III. Proposed Framework

The over all flow of proposed selective gradient diffusion (SGD) for text guided image to image augmentation is shown in figure 1. This architecture builds upon a latent diffusion backbone, but introduces three novel components to achieve localized and identity-preserving edits. First, a region-weighted noise prediction loss emphasizes modifications within specified masks while suppressing changes outside the target area. In this approach, masks are generated automatically using rectangular regions whose size and position vary across samples. These masks serve only as guidance to localize the editable region during training. While rectangular masks are uneven, the SGD framework remains robust because the region-weighted loss and gradient-masking mechanism restrict updates strictly to the masked area, preventing unintended global changes. Second, low-rank adapters (LoRA) are embedded within the cross-attention layers of the U-

Net, enabling parameter-efficient fine-tuning conditioned on text prompts. Finally, a gradient masking strategy constrains weight updates to selected neurons, ensuring that edits remain spatially localized and preventing unintended global modifications. Together, these components form a lightweight yet effective pipeline for precise, text-driven image augmentation.

Let $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ denote the input image, and $E_\phi(\cdot)$ and $D_\phi(\cdot)$ represent the encoder and decoder of the VAE, respectively. The image is first projected into the latent space as:

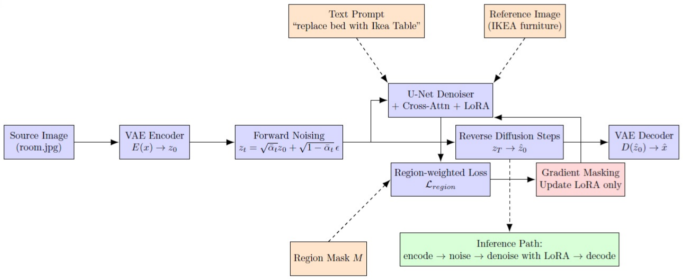

Fig. 1. Overview of the selective gradient diffusion. The three major innovation, specified mask design, LoRA embedded U-Net and weight updates for selective neurons.

$$\mathbf{z}_0 = E_\phi(\mathbf{x}), \quad \mathbf{z}_0 \in \mathbb{R}^{h\times w\times d}$$

where $h < H$, $w < W$, and d is the latent dimension.

In the forward diffusion stage, Gaussian noise is gradually added to the latent space $\mathbf{z}_0$ over T steps according to the following equation:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad t \in \{1, \ldots, T\}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative product of noise scheduling coefficients.

For each time step, the added noise to the image is shown in Figure 2, and the reconstruction errors in terms of MSE are reported in Table 1.

The Mean Squared Error (MSE) is computed as:

$$\text{MSE} = \frac{1}{N}\sum_i (z_{0,i} - \hat{z}_{0,i})^2$$

The total number of time-steps utilized to gradually add and eliminate noise during the forward and reverse procedures is indicated by the parameter T. To guarantee that the model learns to denoise over the whole range of noise levels, from mildly affected images to almost pure Gaussian noise, a large value of T=1000 is frequently used during training. However, it is computationally prohibitive to carry out the opposite procedure throughout all 1000 stages at inference. Rather, effective samplers like DDIM enable precise reconstruction with fewer steps, usually between 25 and 50, maintaining output quality while drastically cutting down on generation time.

During the reverse process in the second stage, a U-Net denoiser parameterized by θ predicts the added noise.

$$\hat{\epsilon}_\theta(z_t, t, c) \tag{1}$$

Conditioned on timestep t and the text embedding c obtained from a transformer based text encoder.

To achieve localized editing, we introduce a region-weighted noise prediction loss.

$$\begin{aligned} L_{\text{region}} = \sum_{i,j} \Big[& M_{(i,j)} \,\|\epsilon_{(i,j)} - \hat{\epsilon}_\theta(z_t, t, c)_{(i,j)}\|^2 \\ & + \lambda(1 - M_{(i,j)})\,\|\epsilon_{(i,j)} - \hat{\epsilon}_\theta(z_t, t, c)_{(i,j)}\|^2 \Big] \end{aligned} \tag{2}$$

where $M \in \{0,1\}^{h\times w}$ is the binary mask specifying the edit region, and $\lambda \ll 1$ controls the preservation of non-target regions.

For parameter-efficient fine-tuning, we embed low-rank adapters (LoRA) into the cross-attention layers of the U-Net. Given an attention projection weight $W \in \mathbb{R}^{d\times d}$, LoRA reparametrizes it as,

$$\begin{aligned} & W' = W + \Delta W, \qquad \Delta W = \frac{\alpha}{r} AB \\ & A \in \mathbb{R}^{d\times r}, \quad B \in \mathbb{R}^{r\times d}, \quad A, B \text{ are low-rank matrices} \\ & r \ll d, \quad r \text{ is rank and } \alpha \text{ is LoRA scaling factor} \end{aligned} \tag{3}$$

Such that only A and B are updated during training, while W remains unchanged.

In the third stage, a gradient masking strategy is applied to restrict weight updates to neurons associated with the masked region. Formally, after completing the gradient $\Delta_\theta L$, we apply,

$$\Delta_\theta^{\text{masked}} = G \odot \Delta_\theta L \tag{4}$$

$G \in \{0,1\}^\theta$ is a binary mask over parameters, and $\odot$ denotes element-wise multiplication. This ensures that updates are applied only to selected LoRA parameters, preventing unintended global modifications.

These three elements work together to create a lightweight yet powerful pipeline for accurate, text-driven image augmentation: region-weighted loss, LoRA-based parameter-efficient adaptation, and gradient masking. By combining these strategies, the model is able to produce localized edits that accurately represent the user's textual instructions while maintaining global image reliability. This differentiates this suggested approach apart from previous diffusion-based editing techniques that depend on global or unconstrained parameter updates. LoRA-based parameter-efficient adaptation, region-weighted loss, and gradient masking work together to produce a lean yet effective pipeline for precise, text-guided image augmentation. The suggested method differs from current diffusion-based editing techniques that engage in unconstrained or global parameter updates because of the complementarity of these methods, which allow the model to realize localized edits that faithfully capture the user's text inputs while preserving global image faithfulness.

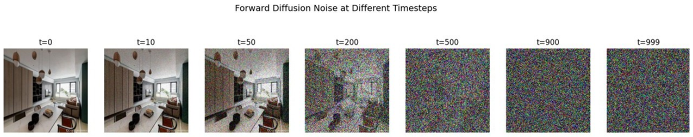

Fig. 2. Forward diffusion noise at different timesteps from 0 to 1000, at t=0 is original image from database and t=999 is the high noised image.

TABLE I
RECONSTRUCTION ERRORS AT EACH TIME-STEP

Index	t	$\bar{\alpha}$	MSE_recon
0	0	0.9999	1.97×10^{-17}
1	10	0.997806573	6.86×10^{-17}
2	50	0.969951494	2.58×10^{-16}
3	200	0.656347006	9.40×10^{-16}
4	500	0.077796658	8.26×10^{-15}
5	900	0.000270245	2.33×10^{-12}
6	999	0.0000404	1.53×10^{-11}

IV. EXPERIMENTAL ANALYSIS

The proposed approach selective gradient diffusion (SGD) framework on text-guided image-to-image augmentation tasks using the interior style [28] dataset. This dataset consists of 7233 interior design images with text description; sample images are shown in figure 3. Experiments are conducted with Stable Diffusion v1.5 as the backbone, and our framework is compared against representative diffusion-based editing methods, including Stable Diffusion Inpainting, InstructPix2Pix, and DiffusionCLIP [18]. Evaluation metrics include the Fréchet Inception Distance (FID) for image realism, Learned Perceptual Image Patch Similarity (LPIPS) for perceptual consistency, and CLIP-based similarity scores for text-image alignment. In addition, qualitative comparisons and a user study are conducted to assess the fidelity of localized edits. The Fréchet Inception Distance (FID) [30] evaluates the similarity between real and generated images by comparing the mean and covariance of their feature distributions extracted using the Inception-V3 model. Lower FID scores indicate that generated images are closer to real images and thus more realistic. LPIPS [31] is metric developed by NVIDIA to measure the perceptual distance between images using deep neural networks. The lower score indicates two images are similar. CLIP based similarity [32] calculates the similarity between CLIP embedding for an image to CLIP embedding to the Text prompt. It uses the cosine similarity, and the score ranges between 0 to 100, higher score indicates more similarity. The results demonstrate that the proposed SGD framework achieves precise object-level modifications (e.g., replace a bed with a sofa) while preserving the integrity of the surrounding content. Unlike existing methods, which frequently cause unintended changes in lighting, background textures, or other objects, our approach confines modifications to the masked region through region-weighted loss and gradient masking. Furthermore, LoRA-based parameter-efficient fine-tuning enables effective adaptation using significantly fewer trainable parameters compared to full-model fine-tuning strategies. The results are shown in figure 4.

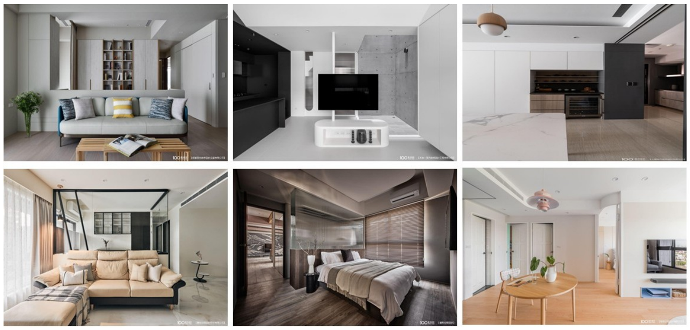

Fig. 3. Sample images from Interior-design dataset. It has around 7.3k images of interior designs.

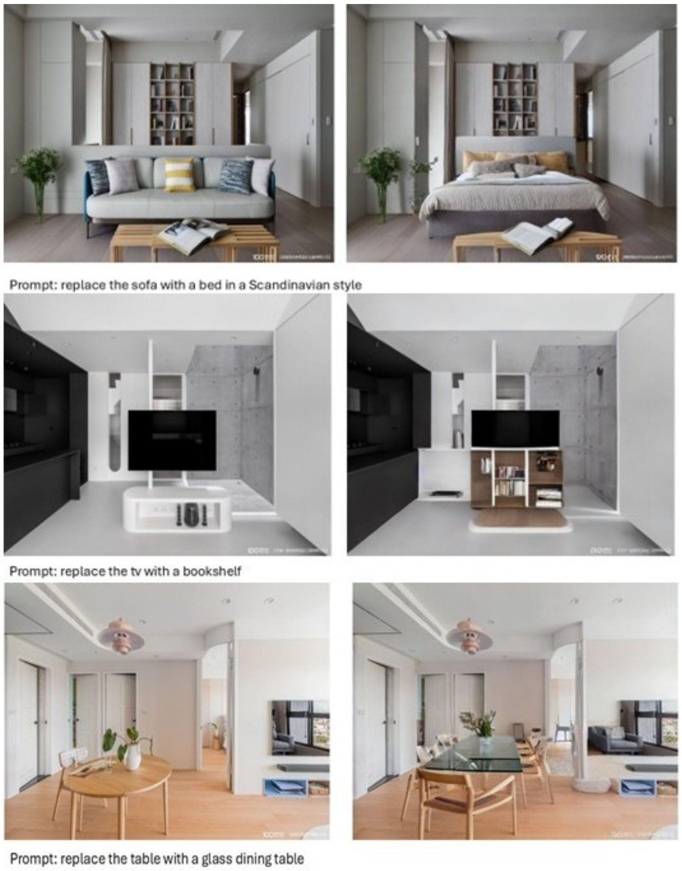
Prompt: replace the sofa with a bed in a Scandinavian style
Prompt: replace the tv with a bookshelf
Prompt: replace the table with a glass dining table

Fig. 4. Results for the SGD for a given prompt, the left side image is an original image and the right image is augmented image with respect to the prompt given.

The results showed in Table 2, on Interior dataset demonstrated that the proposed framework achieves superior perceptual quality and text alignment, outperforming existing baselines such as Stable Diffusion Inpainting, InstructPix2Pix, DiffusionCLIP, and DreamBooth. Quantitatively, SGD achieved an FID of 24.3, LPIPS of 0.165, and a CLIP score of 0.82, indicating significant improvements in both realism and semantic consistency. Subjective user studies further validated the perceptual advantages, with 72% user preference for the generated results. A detailed graphical representation for each metric is shown in figure 5.

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS

Method	FID ↓	LPIPS ↓	CLIP Score ↑	User Preference (%) ↑
SD Inpainting [2]	32.1	0.218	0.71	45
InstructPix2Pix [23]	28.7	0.224	0.75	52
DiffusionCLIP [24]	27.5	0.210	0.77	56
DreamBooth [22]	26.2	0.235	0.79	58
SGD	24.3	0.165	0.82	72

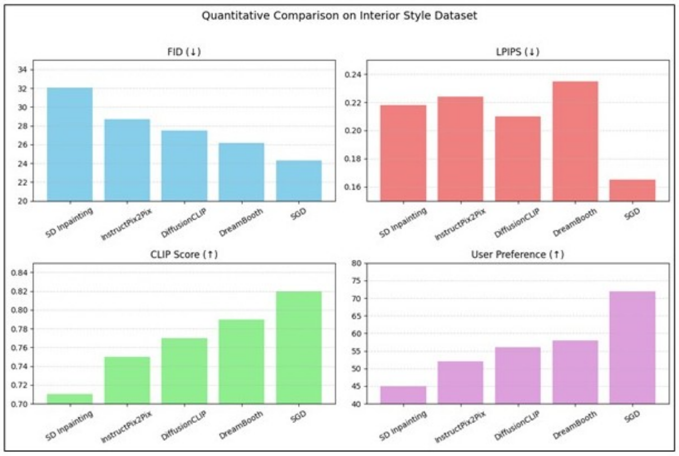

Fig. 5. Quantitative comparison with existing approaches to SGD.

TABLE III
COMPUTATIONAL EFFICIENCY COMPARISON

Method	**Params**	**Train**	**Infer**
	(M)	Time (hr)	Time (s)
SD Inpainting	860	6.8	1.92
DreamBooth	900+	7.2	1.97
InstructPix2Pix	860	–	2.01
SGD–LoRA	**8–12**	**2.1**	**1.21**

V. CONCLUSION

In this paper, a new Selective Gradient Diffusion (SGD) framework was introduced to obtain localized, text-guided image-to-image augmentation with minimal structural distortion. Unlike traditional diffusion-based editing techniques that update model parameters globally, the introduced method incorporates Low-Rank Adaptation (LoRA) modules into cross-attention layers of the diffusion U-Net to support parameter-efficient fine-tuning. A region-weighted noise prediction loss and gradient masking process were added to selectively limit updates to the target region so that edits are kept spatially localized while maintaining the integrity of the regions around it.

REFERENCES

[1] Dhariwal, Prafulla, and Alexander Nichol. ”Diffusion models beat gans on image synthesis.” Advances in neural information processing systems 34 (2021): 8780-8794.

[2] Rombach, Robin, et al. ”High-resolution image synthesis with latent diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[3] Ding, Zheng, et al. ”Patched denoising diffusion models for high-resolution image synthesis.” (2024).

[4] Chen, Haoxing, et al. ”Diffute: Universal text editing diffusion model.” Advances in Neural Information Processing Systems 36 (2023): 63062-63074.

[5] Avrahami, Omri, Dani Lischinski, and Ohad Fried. ”Blended diffusion for text-driven editing of natural images.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[6] Zhang, Zhixing, et al. ”Sine: Single image editing with text-to-image diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[7] Higham, Catherine, Desmond J. Higham, and Peter Grindrod. ”Diffusion models for generative artificial intelligence: An introduction for applied mathematicians.” Siam review 67.3 (2025): 607-623.

[8] Sucharitha, G., et al. ”Mutual contextual relation-guided dynamic graph networks for cross-modal image-text retrieval.” Scientific Reports 15.1 (2025): 1-14.

[9] Graikos, Alexandros, et al. ”Diffusion models as plug-and-play priors.” Advances in Neural Information Processing Systems 35 (2022): 14715-14728.

[10] Park, Taesung, et al. ”Semantic image synthesis with spatially-adaptive normalization.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[11] Patashnik, Or, et al. ”Styleclip: Text-driven manipulation of stylegan imagery.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[12] Vinker, Yael, et al. ”Image shape manipulation from a single augmented training sample.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[13] Gallon, Davide, Arnulf Jentzen, and Philippe von Wurstemberger. ”An overview of diffusion models for generative artificial intelligence.” arXiv preprint arXiv:2412.01371 (2024).

[14] Higham, Catherine, Desmond J. Higham, and Peter Grindrod. ”Diffusion models for generative artificial intelligence: An introduction for applied mathematicians.” Siam review 67.3 (2025): 607-623.

[15] Zeng, Yan, Masanori Suganuma, and Takayuki Okatani. ”Inverting the generation process of denoising diffusion implicit models: Empirical evaluation and a novel method.” 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

[16] Kasymov, Artur, et al. ”Autolora: Autoguidance meets low-rank adaptation for diffusion models.” arXiv preprint arXiv:2410.03941 (2024).

[17] Ćulafić, Igor, et al. ”Output manipulation via LoRA for generative AI.” 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE, 2024.

[18] Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. ”Diffusionclip: Text-guided diffusion models for robust image manipulation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[19] Shin, Jeongmin, and Hyeryung Jang. ”Data Augmentation Techniques Using Text-to-Image Diffusion Models for Enhanced Data Diversity.” 2024 15th International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2024.

[20] Choi, Jooyoung, et al. ”Custom-edit: Text-guided image editing with customized diffusion models.” arXiv preprint arXiv:2305.15779 (2023).

[21] Dong, Wenkai, et al. ”Prompt tuning inversion for text-driven image editing using diffusion models.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[22] Ruiz, Nataniel, et al. ”Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[23] Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. ”Instructpix2pix: Learning to follow image editing instructions.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[24] An, Zifeng, et al. ”Fine-Tuning InstructPix2Pix for Advanced Image Colorization.” arXiv preprint arXiv:2312.04780 (2023).

[25] Haque, Ayaan, et al. ”Instruct-nerf2nerf: Editing 3d scenes with instructions.” Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[26] Zhang, Shiwen, Shuai Xiao, and Weilin Huang. ”Forgedit: Text guided image editing via learning and forgetting.” arXiv preprint arXiv:2309.10556 (2023).

[27] Liu, Xudong, et al. ”S ^{2} Edit: Text-Guided Image Editing with Precise Semantic and Spatial Control.” arXiv preprint arXiv:2507.04584 (2025).

[28] Hu, Yihan, et al. ”DCEdit: Dual-Level Controlled Image Editing via Precisely Localized Semantics.” arXiv preprint arXiv:2503.16795 (2025).

[29] Li, Jungang, et al. ”Text-Guided Local Control for Regional Editing.” Proceedings of the 2024 2nd Asia Symposium on Image and Graphics. 2024.

[30] Yu, Yu, Weibin Zhang, and Yun Deng. ”Frechet inception distance (fid) for evaluating gans.” China University of Mining Technology Beijing Graduate School 3.11 (2021).

[31] Zhang, Richard, et al. ”The unreasonable effectiveness of deep features as a perceptual metric.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[32] Peng, Yuxin, and Chong-Wah Ngo. ”Clip-based similarity measure for query-dependent clip retrieval and video summarization.” IEEE Transactions on Circuits and Systems for Video Technology 16.5 (2006): 612-627.