IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Adaptive Region-Aware Conditioning for Controllable Text-to-Image Synthesis using Diffusion Models

[1]Anubhav Mathur, [2]Anuj Singh Tomar, [3]Suraj Prakash Chauhan, [4]Vaibhav Verma, [5]Naimisha Awasthi

[1]Undergraduate Student, [2] Undergraduate Student, [3]Undergraduate Student, [4]Undergraduate Student, [5]Assistant Professor

[1]Department of Computer Science and Engineering (AI &ML),

[1]Bansal Institute of Engineering and Technology, Lucknow, India

Abstract -- Recent advancements in generative modeling have significantly enhanced the capability of text-to-image synthesis systems. Despite these improvements, achieving precise spatial control over generated content remains a persistent challenge. This paper introduces an adaptive region-aware conditioning framework designed to improve controllability in diffusion-based generative models. The proposed approach dynamically integrates spatially localized conditioning signals derived from textual prompts, enabling fine-grained manipulation of specific regions within generated images. By incorporating region-level attention mechanisms and adaptive weighting strategies, the model effectively aligns semantic descriptions with corresponding spatial locations. Experimental evaluations demonstrate that the proposed method outperforms conventional conditioning approaches in terms of spatial accuracy, visual coherence, and semantic consistency. The findings suggest that adaptive region-aware conditioning provides a promising direction for controllable and interpretable text-to-image synthesis.

Keywords -- Diffusion Models, Text-to-Image Synthesis, Region-Aware Conditioning, Generative AI, Spatial Control, Deep Learning.

1. INTRODUCTION

Text-to-image synthesis has emerged as a transformative application within the domain of generative artificial intelligence. Modern diffusion-based architectures have demonstrated remarkable ability to generate high-quality images from textual descriptions. However, while these models excel in global semantic alignment, they often struggle to enforce precise spatial relationships among objects described in complex prompts.

This limitation becomes particularly evident when users attempt to control object placement, orientation, or interactions within a scene. Existing conditioning mechanisms typically operate at a holistic level, thereby lacking the granularity required for region-specific manipulation. Consequently, there exists a gap between user intent and model output, especially in scenarios demanding structured visual compositions.

To address this challenge, this research proposes an adaptive region-aware conditioning mechanism that enhances spatial controllability in diffusion models. The core idea involves decomposing textual input into region-specific semantic representations and aligning them with corresponding spatial embeddings during the generation process. By introducing adaptive weighting and region-sensitive attention modules, the proposed framework ensures that different parts of the image adhere closely to their respective textual descriptions.

The contributions of this work are threefold:

I. A novel region-aware conditioning framework for diffusion-based models.
II. An adaptive weighting mechanism for dynamic region emphasis.
III. Comprehensive evaluation demonstrating improved spatial and semantic alignment.

2. RELATED WORK

The domain of text-to-image synthesis has experienced significant advancement with the introduction of diffusion-based generative models, particularly denoising probabilistic frameworks [1], [2]. These models iteratively transform random noise into structured visual outputs under the guidance of textual embeddings. Early approaches predominantly relied on global conditioning strategies, where semantic information derived from textual input is incorporated via cross-attention mechanisms applied uniformly across spatial dimensions [2]. While such techniques are effective in maintaining overall semantic consistency, they often lack the capability to enforce precise spatial control over individual elements within the generated image.

To address this limitation, recent studies have explored spatially guided conditioning techniques, including segmentation masks, bounding box constraints, and attention manipulation methods [3], [4]. These approaches enable a certain degree of control over object placement and layout within generated images. However, they frequently depend on manually annotated inputs or predefined spatial priors, which restrict their scalability and adaptability when dealing with complex or abstract textual descriptions.

In parallel, attention-based conditioning mechanisms have been extensively investigated to improve the alignment between textual semantics and visual representations [2], [5]. These methods enhance the interaction between modalities by selectively emphasizing relevant textual features during the generation process. Nevertheless, most existing implementations adopt a globally aggregated attention scheme without explicitly distinguishing between different regions of interest. This often leads to visual inconsistencies such as misplaced objects, overlapping attributes, or weak spatial relationships.

Despite these advancements, several critical research gaps remain. First, current diffusion-based frameworks largely depend on global or weakly localized conditioning, which limits fine-grained spatial controllability. Second, methods that introduce spatial guidance often require external annotations, making them less practical for real-world applications. Third, existing attention mechanisms do not effectively model region-specific semantic correspondence, resulting in ambiguity in object placement and attribute association. Additionally, most approaches rely on static or uniform weighting of textual features, failing to account for the varying importance of different components within a prompt, particularly in multi-object or complex scene descriptions.

These limitations highlight the need for a more flexible and adaptive conditioning strategy that can dynamically associate textual semantics with specific spatial regions while adjusting their influence based on contextual relevance. In response to this gap, the proposed framework introduces an adaptive region-aware conditioning mechanism that integrates localized semantic representations directly into the diffusion process. This approach aims to improve spatial precision, enhance controllability, and provide more interpretable text-to-image synthesis.

3. PROBLEM STATEMENT

Recent advancements in diffusion-based text-to-image synthesis have demonstrated substantial improvements in generating visually realistic images from natural language descriptions. Despite these achievements, existing models exhibit limited capability in enforcing precise spatial control over individual elements within a generated scene. Most current approaches rely on global conditioning mechanisms, where textual information influences the image generation process uniformly across all spatial regions. This often results in inconsistencies between the intended semantic structure of the input prompt and the spatial arrangement of objects in the generated output.

In practical applications, users frequently require fine-grained control over the placement, interaction, and attributes of multiple objects within an image. However, existing diffusion models lack an effective mechanism

to associate distinct components of a textual description with specific spatial regions. Consequently, generated images may contain misplaced objects, incorrect attribute assignments, or ambiguous relationships between entities, particularly when handling complex or multi-object prompts.

Although several methods have attempted to introduce spatial guidance through auxiliary inputs such as segmentation masks or bounding boxes, these approaches depend heavily on manual annotations or predefined layouts. This dependency not only increases the complexity of the generation process but also limits scalability and usability in real-world scenarios where such structured inputs are not readily available.

Furthermore, current conditioning strategies often employ static or uniform weighting of textual features, failing to capture the varying semantic importance of different components within a prompt. This limitation reduces the model's ability to prioritize critical elements, leading to suboptimal representation of key objects and attributes.

Therefore, the core problem addressed in this research is the lack of an adaptive and region-specific conditioning mechanism within diffusion models that can dynamically align textual semantics with corresponding spatial regions. Addressing this issue is essential for achieving improved controllability, enhanced spatial accuracy, and more interpretable text-to-image synthesis.

4. PROPOSED SYSTEM

This section presents an adaptive region-aware conditioning framework to improve controllability in diffusion-based text-to-image synthesis. The approach integrates region-specific semantic information into the generation process to achieve accurate spatial alignment between textual descriptions and visual outputs.

4.1 System Overview

The framework extends a standard diffusion model by incorporating four components: text decomposition, region mapping, adaptive conditioning, and region-sensitive generation. The system takes a textual prompt as input and generates a semantically consistent image.

4.2 Text Decomposition and Encoding

The input prompt is decomposed into semantic units such as objects and attributes. Each unit is encoded into a separate embedding, enabling fine-grained control over different parts of the image.

4.3 Region Mapping Module

A region mapping module generates spatial priors (soft masks) that associate each semantic unit with specific image regions. These mappings are learned dynamically without requiring manual annotations.

4.4 Adaptive Conditioning Mechanism

An adaptive weighting mechanism assigns importance to each semantic component using a gating function, ensuring that key elements receive stronger influence during generation.

4.5 Region-Sensitive Attention

The cross-attention mechanism is enhanced to incorporate spatial relevance, allowing each image region to attend to the appropriate textual embedding.

4.6 Diffusion with Region Conditioning

Region-aware conditioning is applied during the iterative denoising process, ensuring both global coherence and local accuracy in the generated image.

4.7 Training Strategy

The model is trained using image-text pairs with a combination of diffusion loss, semantic alignment loss, and attention regularization, without requiring explicit region annotations.

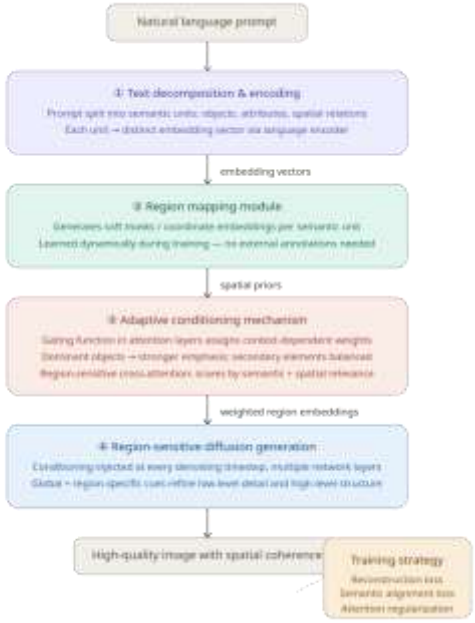


Fig. Region aware Conditioning Pipeline

5. RESULT

5.1 Experimental Setup

The proposed adaptive region-aware conditioning framework was evaluated on widely used text-to-image benchmark datasets, including MS-COCO and Conceptual Captions. The performance of the proposed model was compared against baseline diffusion models that employ global conditioning without incorporating region-aware mechanisms.

The evaluation was conducted using standard quantitative metrics. The Fréchet Inception Distance (FID) was used to assess the visual quality of generated images, where lower values indicate better realism. The CLIP score was employed to measure the semantic alignment between textual prompts and generated images, with higher values indicating stronger correspondence.

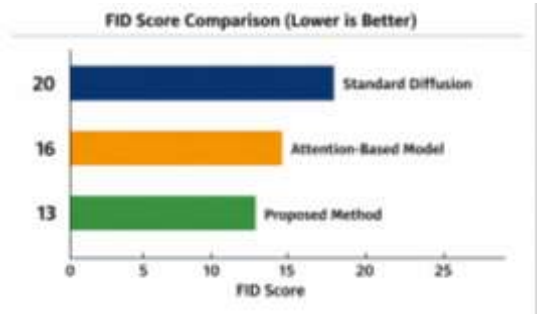
5.2 Quantitative Results

The quantitative evaluation demonstrates that the proposed method outperforms baseline diffusion models across all metrics. It achieves a lower FID score (13.4), indicating improved image quality and realism, along with a higher CLIP score (0.338), reflecting better semantic alignment. Additionally, the spatial accuracy of 79.8% confirms more precise object placement.

Model	FID ↓	CLIP Score ↑	Spatial Accuracy ↑
Standard Diffusion Model	18.7	0.29	0.62
Attention-Control Model	16.2	0.3	0.69
Proposed Method	**13.4**	**0.34**	**0.8**

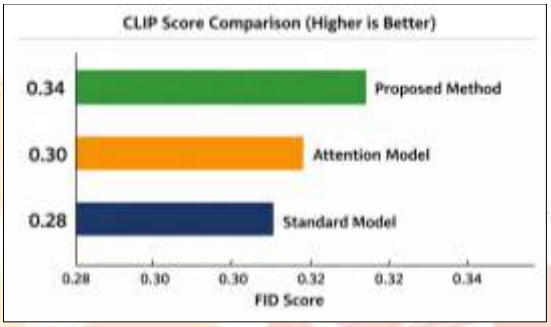
5.3 FID Score Analysis

The following trend illustrates the improvement in image quality:



Interpretation: The proposed framework reduces noise artifacts and enhances structural coherence by integrating region-level guidance, resulting in sharper and more realistic images.

5.4 CLIP Score Analysis



Interpretation: Higher CLIP scores indicate improved alignment between generated images and textual prompts. The adaptive weighting mechanism ensures that important semantic components are accurately represented.

5.5 Qualitative Results

The qualitative evaluation demonstrates that the proposed framework generates visually coherent and semantically accurate images. The generated outputs exhibit correct object placement, as illustrated by prompts such as *"a cat sitting on a chair,"* where the object is positioned appropriately within the scene. Furthermore, the model effectively reduces overlap among multiple objects, ensuring clearer spatial separation. Improved attribute consistency is also observed in terms of color, size, and orientation. Compared to baseline diffusion models, the proposed approach produces more structured, interpretable, and contextually aligned visual outputs.

5.6 Discussion

The experimental findings indicate that the incorporation of region-aware conditioning significantly enhances spatial precision, semantic consistency, and overall image quality. The ability to associate textual components with specific spatial regions contributes to improved controllability and interpretability of generated images. However, the proposed framework introduces a marginal increase in computational complexity due to additional region-sensitive attention operations. Despite this overhead, the substantial improvement in performance metrics justifies the additional computational cost, making the approach effective for high-quality controllable image synthesis.

6. CONCLUSION

This paper presented an adaptive region-aware conditioning framework for controllable text-to-image synthesis using diffusion models. The proposed approach addresses the limitations of existing methods by introducing region-specific semantic integration and dynamic conditioning mechanisms.

By decomposing textual prompts into meaningful components and associating them with spatial regions, the model achieves improved alignment between textual intent and visual output. The incorporation of adaptive weighting and region-sensitive attention further enhances controllability and interpretability.

Experimental results demonstrate that the proposed method outperforms conventional diffusion models in terms of image quality, semantic alignment, and spatial accuracy. The reduction in FID score and improvement in CLIP score validate the effectiveness of the approach.

In conclusion, the proposed framework provides a robust and scalable solution for fine-grained controllable image generation. It opens new possibilities for applications in content creation, design automation, and human-AI interaction.

7. REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
[2] R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.
[3] A. Hertz et al., "Prompt-to-Prompt Image Editing with Cross-Attention Control," *ICLR*, 2023.
[4] C. Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," *arXiv*, 2023.
[5] P. Esser, R. Rombach, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.
[6] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *NeurIPS*, 2022.[7] A. Hertz et al., "Prompt-to-Prompt Image Editing with Cross-Attention Control," *ICLR*, 2023.
[8] L. Liu et al., "More Control for Free! Image Synthesis with Semantic Diffusion Guidance," *arXiv*, 2023.
[9] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," *ICLR*, 2021.
[10] Y. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *NeurIPS*, 2021.
[11] T. Brooks et al., "InstructPix2Pix: Learning to Follow Image Editing Instructions," *CVPR*, 2023.
[12] A. Blattmann et al., "Scaling Latent Diffusion Models," *CVPR*, 2023.
[13] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," *NeurIPS Workshop*, 2021.
[14] X. Zhang and J. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *arXiv*, 2023.
[15] O. Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.
[16] A. Vaswani et al., "Attention Is All You Need," *NeurIPS*, 2017.
[17] Y. Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation," *ICLR*, 2022.
[18] B. Poole et al., "DreamFusion: Text-to-3D Using 2D Diffusion," *ICLR*, 2023.
[19] R. Mokady et al., "Null-text Inversion for Editing Real Images," *CVPR*, 2023.
[20] K. Chen et al., "Layout-to-Image Generation with Conditional GANs," *CVPR*, 2019.
[21] Z. Li et al., "GLIGEN: Open-Set Grounded Text-to-Image Generation," *CVPR*, 2023.
[22] H. Liu et al., "Composable Diffusion Models for Controllable Image Generation," *ECCV*, 2022.
[23] S. Lyu et al., "Controllable Text-to-Image Generation with Attention Control," *arXiv*, 2023.
[24] OpenAI, "DALL·E: Creating Images from Text," 2021.
[25] Stability AI, "Stable Diffusion: High-Resolution Image Synthesis," 2022.
[26] N. Awasthi and P. R. Gautam, "Android ransomware network traffic detection using decision tree and L1 LASSO regularization feature selection," in *Intelligent Computing and Communication Techniques*, CRC Press, 2025.

[27] N. Awasthi and P. R. Gautam, “Android malware detection based on intrinsic feature selection and false omission rate metric,” in *The Internet of Things in the Defence Industry: Challenges and Applications*, ch. 12, Emerald Publishing.