



UNIVERSITÀ DI PARMA
DIPARTIMENTO DI INGEGNERIA E ARCHITETTURA

Dottorato di Ricerca in Tecnologie dell'Informazione
XXXVIII Ciclo

Rami O. M. Skaik

**Text-to-Image Generation from a Single Image with Pose
and Style Control**

DISSERTAZIONE PRESENTATA PER IL CONSEGUIMENTO
DEL TITOLO DI DOTTORE DI RICERCA

2022/2023 – 2024/2025

UNIVERSITÀ DI PARMA

Dottorato di Ricerca in Tecnologie dell'Informazione

XXXVIII Ciclo

**Text-to-Image Generation from a Single Image with Pose
and Style Control**

Coordinator:

Prof. Dr. Marco Locatelli

Tutor:

Chiar.mo Prof. Andrea Prati

Dottorando: *Rami O. M. Skaik*

2022/2023 – 2024/2025

*To my beloved parents, who shaped my path and live forever in
my heart.*

*To my close family: my wife Yasmeen, and my children Yara,
Sami, Sara, and Jouri,*

*And to my brothers, sisters, and their children for their
unwavering love and support.*

*To my supervisor, Prof. Andrea Prati, and co-authors Dr.
Leonardo Rossi and Dr. Tomaso Fontanini, for their guidance
and collaboration.*

*This work is dedicated to all who inspired and encouraged me
throughout this journey.*

Sommario

1	Background and Motivation	1
1.1	Evolution of Generative Models	1
1.2	Text-to-Image Generation	1
1.3	Challenges in Controllable Text-to-Image Generation	2
1.3.1	Ambiguity in Textual Prompts	2
1.3.2	Pose and Geometry Control	2
1.3.3	Style Transfer and Personalization	3
1.3.4	Data Efficiency and Scalability	3
1.4	Research Contributions of This Thesis	4
1.4.1	Mask Conditional Text-to-Image Generative Model (MCGM)	4
1.4.2	MCGM-Styler	5
1.5	Objectives of the Thesis	5
1.6	Structure of the Thesis	6
2	Related Work	7
2.1	Overview of Generative Models	8
2.2	Diffusion Models and Their Evolution	8
2.2.1	Diffusion Models	8
2.2.2	Forward and Reverse Diffusion Processes	9
2.2.3	Training Objective and Denoising Loss	9
2.2.4	Architectural Developments and Extensions	10
2.3	Text-to-Image Generation	10

2.4	Controllable Text-to-Image Synthesis	10
2.5	Personalization and Style Transfer in Diffusion Models	11
2.6	Few-Shot and Single-Image Learning	12
2.7	Summary and Positioning of This Thesis	12
3	Mask Conditional Text-to-Image Generative Model (MCGM)	15
3.1	Introduction	15
3.2	State of the Art	17
3.2.1	Fine-tuning diffusion models	17
3.2.2	Cross-attention	18
3.2.3	Personalization in diffusion models	18
3.3	Proposed Method	20
3.3.1	The MCGM Model Architecture:	20
3.3.2	Mask Injection	21
3.3.3	The Optimization Losses	22
3.4	Chapter Summary	23
3.5	Contributions of This Chapter	24
4	Free-Form styler for Mask Conditional Text-to-Image Generative Model (MCGM-Styler)	27
4.1	MCGM-Styler : Motivation and Overview	27
4.2	State of the art	29
4.3	MCGM-Styler Model Architecture	31
4.4	Style Representation and Conditioning	32
4.4.1	Preliminary: From MCGM to MCGM-Styler	32
4.4.2	Injecting Style into the Generation Process	33
4.5	Decoupling Content and Style	34
4.6	Inference	35
4.7	Chapter Summary	35
4.8	Contributions of This Chapter	36

5	Experimental Results	39
5.1	Results on MCGM	39
5.1.1	Experimental Setup	40
5.1.2	Training Details	40
5.1.3	Single-Subject Results	41
5.1.4	Multi-Subject Results	45
5.1.5	Discussion and Limitations	48
5.1.6	Ablation Study	48
5.1.7	User Study	49
5.2	Experimental Results: MCGM-Styler	51
5.2.1	Training Setup and Implementation Details	51
5.2.2	Qualitative Results	53
5.2.3	Comparison with State-of-the-Art Methods	57
5.2.4	Ablation Studies	58
5.2.5	Style-Guided Text-to-Image Generation	62
5.2.6	Quantitative Evaluation	62
5.2.7	User Study	63
6	Conclusion and Future Work	67
6.1	Thesis Summary	67
6.2	Main Contributions	68
6.3	Discussion of Findings	69
6.4	Limitations	69
6.5	Future Work	71
6.6	Final Remarks	72
	Bibliography	73
	Acknowledgment	83

List of Figures

2.1	The forward and backward processes of the diffusion model. Image credit: Arash Vahdat.	9
3.1	Overview of the proposed MCGM method. Starting from a single image with a set of subjects and the corresponding masks, it is possible to (a) generate any of those subjects using an arbitrary mask to set the pose, (b) generate multiple subjects at the same time using a single mask to control the pose and (c) generate multiple subjects at the same time using multiple masks to define each of their poses.	16
3.2	Model Architecture details: the left side is the training phase that is tasked to learn the subjects with the aid of text and mask encoders, and the right side represents the inference phase where the subjects are generated using text for expressing their context and mask to set their pose.	21

4.1	Overview of the MCGM-Styler capabilities. [On the left] starting from an image containing multiple subjects, the model is trained sequentially using an action pose mask, a target style, and a textual prompt. This enables users to generate scenes with arbitrary subject(s), pose(s), and style(s). The bottom row illustrates the generation of multiple subjects in a single style, the middle row shows a single-subject scenario, and the top row demonstrates the generation of images with mixed styles , where two or more artistic styles are combined within the same output. [On the right] the model generates paintings of a single subject rendered in various artistic styles. . . .	30
4.2	Pipeline of the MCGM-Styler: The left section illustrates the two-stage training process, with source and style images provided as user inputs. In the first stage, various concepts are learned using textual prompts and mask inputs. In the second stage, the style concept is trained. The token <code>< asset1 ></code> represents the cat in the source image, while <code>< s ></code> represents the style. Note that the (+) symbol indicates concatenation of the text with mask embeddings. The right section depicts the inference process, where a textual prompt description and a target position mask are used to generate the output scene.	32
5.1	Different generated images for the input source images in different scenes based on the text prompt and with the pose guided by the target masks.	42
5.2	Different images are generated for the same dog image using the same text prompts and different target mask poses; in (a), the dog is sitting on the ground based on the shape of two different masks. Next, in (b) the dog is running with two different poses based on the mask shape.	43
5.3	Different images generated by MCGM with shaped mask, MCGM with a blank mask, or the Break-a-scene model	44

5.4	The same prompt with different target masks was used to generate different positions for the dog playing football.	45
5.5	Some results for multi-subject source images used different text prompts associated with masks to generate scenes for a single subject.	46
5.6	Some results for multi-subject source images used different text prompts associated with mask support the action position in the scene, each column has a different case based on the number of subjects and the number of masks used in the scene.	47
5.7	Ablation results for two-subject images using different text prompts and masks. In particular, we tested MCGM with/without mask attention loss, compared with the Break-a-scene model	49
5.8	Survey results when asking which image, one generated with MCGM and the other with Break-a-scene, matches more with the pose indicated by the mask.	50
5.9	Survey results when asked to determine, with a scale from 1 to 5, how much a generated sample with our method matches the pose of a target mask. <i>T-resp 1con</i> and <i>T-resp 2con</i> are the total responses for single and multiple concept images, <i>w-sum 1con</i> and <i>w-sum 2con</i> are the weighted sum for single and multiple concept images and <i>Avg-rat 1con</i> and <i>Avg-rat 2con</i> are the average ratings for single and multiple concept images.	52
5.10	One Concept, One Style: Images generated from a single input image using the prompt "a painting of $\langle asset0 \rangle$ / $\langle asset1 \rangle$ in the style of $\langle s0 \rangle$." ,where $\langle asset0 \rangle$ refers to the left dog, $\langle asset1 \rangle$ represents the cat, and $\langle s0 \rangle$ denotes the target style. The figure highlights the effectiveness of style conditioning for both learned concepts.	54

5.11	One concept with varying masks and styles: Generated outputs are based on the same textual prompt, but guided by different styles and mask shapes. The first column uses only style guidance, while the second and third columns incorporate both style and mask guidance. The first row corresponds to $\langle asset1 \rangle$, representing the cat concept, and the second row corresponds to $\langle asset0 \rangle$, representing the left dog concept.	55
5.12	Multi-Concepts-Multi-Masks: (A) illustrates the case of using a single mask to guide only one of the concepts pose, specifically guiding the sitting pose for the cat. (B) demonstrates the case of using a single mask to guide actions for two concepts. In this example, the mask guides the sitting position for both the cat and the dog. (C) showcases the case of generating three concepts in one scene using one mask to guide their pose. Finally, (D) showcases the case of using multiple masks to guide multiple actions for multiple concepts, with one mask guiding a sitting dog and the other guiding a sleeping cat.	56
5.13	Results with multiple styles. The first two columns apply individual styles, while the third column demonstrates mixed-style generation.	57
5.14	Mixed-style synthesis with different trained style pairs, producing novel hybrid artistic renderings.	58
5.15	Comparison between our model and different combination of DreamStyler and DreamBooth. The same sentence was used in all the experiments: “ <i>a painting of $\langle asset0 \rangle$ in the style of $\langle s0 \rangle$.</i> ” Here, $\langle asset0 \rangle$ represents the dog in the source image, while $\langle s0 \rangle$ corresponds to the style image used in each experiment. (A) represents the results employing our full model for both learning the concept and the styles. (B) and (C) both employ DreamStyler for the style and MCGM and DreamBooth for the concept, respectively. In this case six style training stages are used. Finally, in (D) DreamStyler with a single style training stage in combination with DreamBooth is presented.	59

5.16	Effect of human description prompts in decoupling content and style. Different scenes without and with the human description technique. On the left, the source and style used are displayed. The top row shows the different prompts applied. The middle row presents three different concept scenes generated using various text prompts without the human description technique. The bottom row shows the results obtained by incorporating human description during training.	60
5.17	Contribution of the mask encoder: two experiments using the same source image, target pose mask, text prompt, and target style. The example to the right, without the mask encoder, fails to apply the target pose correctly.	61
5.18	Style training with different timesteps. For each experiments, three examples are shown. Optimal results are obtained at 800 steps, while larger values lead to overfitting.	62
5.19	Style-guided text-to-image generation: Comparison between our model and various configurations of DreamStyler model. The left column shows the input text prompt and target style, the right column displays results from our model, and the center presents DreamStyler outputs under different training setups, including one-step, six-step, and MCGM-pretrained variants.	66
6.1	Limitation in following the correct orientation of the mask. In this case, the direction of the output images (front view) is different from the mask direction (side view) even if the overall pose is correct. . .	70
6.2	Limitation when using more than two mask conditions. Indeed, the model struggles to apply the mask conditions for all subjects in the output scenes.	71

- 6.3 Decoupling limitation. Two examples of failure cases caused by complex multi-subject interactions in the style image. The left column shows the source image and prompt, the middle columns show the mask and style, and the right column shows the generated results , where the content is entangled with the source style image. 72

List of Tables

5.1	Quantitative Evaluation: Comparison of MCGM-Styler with Dream-Styler under different training conditions. Bold: best, <u>underline</u> : second best.	63
-----	---	----

Chapter 1

Background and Motivation

1.1 Evolution of Generative Models

Generative modeling has witnessed unprecedented advancements over the past decade. Early breakthroughs with Generative Adversarial Networks (GANs) demonstrated that neural networks could learn to synthesize high-quality data from noise distributions [1]. Although GANs were effective in producing sharp images, they suffered from training instability, mode collapse, and limited controllability [2].

A paradigm shift occurred with the introduction of diffusion probabilistic models (DPMs), which generate data by gradually denoising from a Gaussian noise distribution [2, 3]. These models, such as DDPM [4] and DDIM [5], significantly improved both stability and sample diversity. Diffusion models quickly became the backbone of modern text-to-image (T2I) frameworks [6, 7], demonstrating superior photorealism and scalability compared to GAN-based methods.

1.2 Text-to-Image Generation

Text-to-image (T2I) generation emerged as one of the most prominent applications of diffusion models. Systems such as DALL-E 2, Imagen, and Stable Diffusion showed that natural language could be directly translated into visually coherent images [7–

9]. These breakthroughs were enabled by coupling large-scale language encoders (e.g., CLIP, T5) with diffusion backbones, allowing semantically accurate alignment between text and generated images [10].

The societal and industrial impact of T2I models has been profound, spanning applications in *artistic creation, graphic design, architecture, film production, gaming, advertising, education, and healthcare visualization* [11, 12]. However, despite their promise, the need for more controllable, data-efficient, and customizable T2I generation remains a fundamental challenge.

1.3 Challenges in Controllable Text-to-Image Generation

1.3.1 Ambiguity in Textual Prompts

Although language provides a flexible interface, natural text is inherently ambiguous. Phrases such as “a person sitting” do not specify the exact body posture, orientation, or interaction with other subjects in the scene [13]. Similarly, describing style in words (e.g., “in impressionist style”) cannot capture fine-grained texture, brushwork, or composition details [14]. As a result, generated outputs often deviate from user expectations when precise control is required.

1.3.2 Pose and Geometry Control

Controlling human or object pose is especially challenging. Prior works such as ControlNet [15], Pose-Guided Diffusion [16], and Pose-to-Image frameworks [13] introduced skeletal or keypoint-based conditioning. These methods improve pose fidelity but typically require paired training datasets or extensive fine-tuning, making them less accessible for personalized use. Moreover, pose-guided synthesis often struggles with multi-subject consistency, where multiple characters must be aligned in realistic interactions.

1.3.3 Style Transfer and Personalization

Parallel research has focused on personalization and style transfer. Notable approaches include DreamBooth [17], which fine-tunes diffusion models on a few images to capture a subject identity, and Textual inversion [18], which learns custom word embeddings for new concepts. Although effective, these methods are computationally expensive and primarily address subject fidelity rather than artistic style.

Style-guided T2I generation has seen progress with methods like ControlStyle [14] and ArtAdapter [19], which learn to transfer visual appearance from reference artworks. However, these often require large curated datasets or fail to combine style transfer with pose fidelity, limiting their generalization.

1.3.4 Data Efficiency and Scalability

Another critical bottleneck is the reliance on large-scale datasets. State-of-the-art T2I diffusion models are typically trained on billions of image-text pairs [9], a requirement that hinders personalization. In contrast, real-world applications often demand generation based on only a handful of images, such as producing new images of a specific pet, object, or artistic style. Few-shot and single-image methods like Break-a-scene [20] and DreamBooth [17] attempt to address this, but they fall short in balancing pose, style, and semantic alignment simultaneously.

A closely related work that serves as the foundation of this research is *Break-a-scene* [21], which introduced an effective framework for extracting multiple visual concepts from a single image and recombining them into novel scenes. By optimizing concept-specific textual handles and subsequently fine-tuning a diffusion model, Break-a-scene enables personalized image generation from minimal data, making it particularly well suited for data-scarce scenarios. However, despite its effectiveness in concept disentanglement and recombination, the method provides limited control over the spatial configuration of generated subjects. In particular, the pose and precise spatial arrangement of individual objects remain difficult to specify, as control is primarily achieved through textual descriptions, which are inherently ambiguous.

This limitation motivates the first contribution of this thesis, the *Mask Conditional*

Generative Model (MCGM), which extends the Break-a-scene framework by introducing an explicit mask-based conditioning mechanism. By embedding spatial mask information alongside textual embeddings, MCGM enables direct and fine-grained control over the pose of one or multiple subjects during image generation, significantly improving structural consistency and user controllability. Building upon this framework, we further propose *MCGM-Styler*, an enhanced model that addresses an additional limitation of Break-a-scene; namely, the absence of explicit stylistic control. MCGM-Styler incorporates a dedicated training phase to learn style embeddings, allowing the model to simultaneously enforce pose constraints and apply either single or multiple artistic styles. Together, these extensions elevate Break-a-scene from a concept recomposition framework into a flexible and expressive system for pose- and style-controlled image generation from a single reference image.

1.4 Research Contributions of This Thesis

This thesis addresses the above challenges by proposing two interrelated frameworks: MCGM and MCGM-Styler. Both models are designed with an emphasis on **controllability, efficiency, and generalization**.

1.4.1 Mask Conditional Text-to-Image Generative Model (MCGM)

The Mask Conditional Generative Model (MCGM) extends the Break-a-scene framework [20] by introducing a novel **mask embedding injection** mechanism. Unlike conventional prompt-only conditioning, mask embeddings allow users to specify subject poses and spatial layouts with fine granularity. This enables:

- Pose-specific synthesis of single or multiple subjects.
- Multi-subject consistency in a shared scene.
- User-friendly control through intuitive mask inputs.

Through extensive experiments, MCGM demonstrates superior pose fidelity while retaining semantic alignment with textual prompts.

1.4.2 MCGM-Styler

MCGM-Styler builds on MCGM by incorporating **style transfer** capabilities. A novel training procedure enables the model to learn style embeddings from a single reference image. This allows:

- Simultaneous control over **pose and style**.
- Application of **single or multiple target styles**.
- Style mixing across different subjects in the same scene.

Compared to DreamStyler [14], MCGM-Styler achieves superior performance in both style fidelity and pose preservation. Importantly, it does so with significantly fewer data requirements, making it more practical for real-world applications.

1.5 Objectives of the Thesis

The primary objective of this thesis is to advance the field of controllable text-to-image generation by developing novel diffusion-based frameworks capable of learning from a single reference image while providing users with explicit control over the generated content. In particular, this research focuses on improving the controllability of personalized image generation through the integration of multiple conditioning signals.

To achieve this goal, the thesis pursues the following objectives. First, it proposes the Mask Conditional Generative Model (MCGM), a framework that introduces mask-based spatial conditioning to enable fine-grained pose control during image synthesis. Second, it extends this framework by introducing MCGM-Styler, which incorporates artistic style conditioning while preserving subject identity and pose, allowing flexible and controllable image generation. Third, the proposed models are extensively evaluated against state-of-the-art approaches through both quantitative and qualitative experiments, assessing their performance in terms of pose accuracy, style fidelity, and overall image quality. Finally, the thesis investigates the feasibility of generating high-quality personalized images from a single reference

image, demonstrating that controllable image synthesis can be achieved with minimal training data.

1.6 Structure of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 reviews the scientific background and related work, providing an overview of generative models, diffusion-based text-to-image synthesis, personalization techniques, pose control, and style transfer methods. Chapter 3 introduces the proposed Mask Conditional Generative Model (MCGM), describing its motivation, architecture, training procedure, and methodological contributions. Chapter 4 presents MCGM-Styler, an extension of MCGM that integrates artistic style conditioning into the generation process while maintaining explicit pose control. Chapter 5 reports the experimental evaluation of both proposed models, including quantitative and qualitative comparisons with state-of-the-art methods, ablation studies, and user evaluations. Finally, Chapter 6 summarizes the main contributions of the thesis, discusses the identified limitations, and outlines promising directions for future research in controllable generative artificial intelligence.

Chapter 2

Related Work

Generative models are among the most revolutionary elements of contemporary artificial intelligence, allowing the synthesis of realistic data, including text, audio, and images. These models seek to understand intricate data distributions and produce new samples that resemble observations from the real world. Over the past decade, significant progress has been achieved through various architectures such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models [3, 22–24].

In this chapter, we review the main generative architectures relevant to this thesis, with different levels of detail depending on their role in the proposed methods. Core models that directly inspire our work, namely diffusion models, Stable Diffusion, CLIP, and Break-a-scene, are described in greater depth, including their architectural design and training objectives. Other related approaches are discussed more concisely to provide contextual background while maintaining focus on the methods most relevant to MCGM and MCGM-Styler.

2.1 Overview of Generative Models

The earliest deep generative models, including Variational Autoencoders (VAEs) [22, 25], introduced probabilistic latent-variable frameworks for reconstructing and generating new samples. VAEs provide a principled approach to learning latent representations through approximate Bayesian inference but often suffer from producing blurry images due to the nature of their reconstruction loss [22, 26–28].

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [23], significantly advanced the field by formulating the generation process as a mini-max game between a generator and a discriminator. Numerous variants such as DC-GAN [29], StyleGAN [30], and BigGAN [31] improved stability, quality, and diversity of generated images. Despite their success, GANs exhibit challenges in training stability and mode collapse [32], [33], motivating exploration into alternative generative paradigms.

Autoregressive models such as PixelCNN [34] and Image Transformer [35] modeled pixel dependencies explicitly, achieving strong likelihood estimates at the cost of slow sampling speed. Methods based on diffusion were able to overcome these constraints and provide generative flexibility, stability, and excellent picture quality.

2.2 Diffusion Models and Their Evolution

2.2.1 Diffusion Models

Diffusion models have recently emerged as a dominant framework for high-quality image generation. Initially introduced by Sohl-Dickstein et al. [36], these models formulate generation as a gradual denoising process that transforms random noise into structured data. Due to their stable training dynamics and ability to produce high-fidelity samples, diffusion models have surpassed earlier generative approaches such as GANs in many image synthesis tasks.

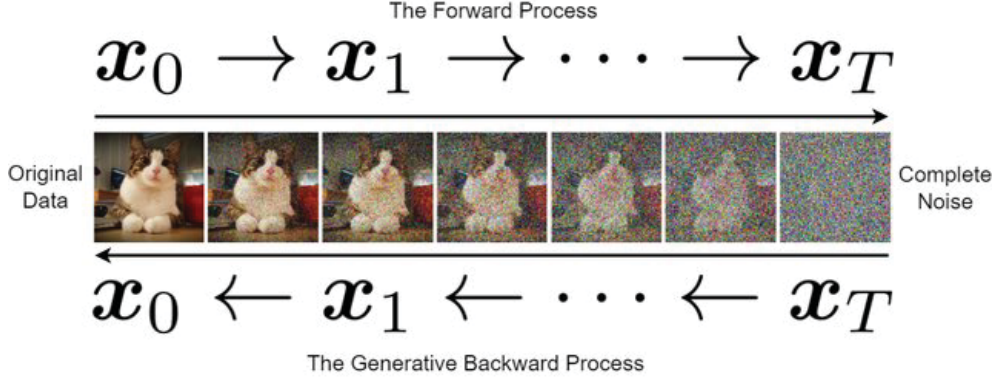


Figure 2.1: The forward and backward processes of the diffusion model. Image credit: Arash Vahdat.

2.2.2 Forward and Reverse Diffusion Processes

As shown in the Figure 2.1, A diffusion model consists of two complementary processes. The *forward process* progressively corrupts a data sample by adding Gaussian noise over a sequence of time steps until the sample approaches an isotropic noise distribution. The *reverse process* is learned by a neural network that gradually removes noise, reconstructing a sample from pure noise. This probabilistic formulation enables the model to approximate complex data distributions through iterative denoising [36].

2.2.3 Training Objective and Denoising Loss

The Denoising Diffusion Probabilistic Model (DDPM) [24] simplified training by directly predicting the noise added at each diffusion step using a mean squared error loss. This formulation significantly improved training stability and image quality. Subsequent extensions, such as DDIM [37], introduced deterministic sampling procedures that reduce inference time, while classifier-guided diffusion [3] enhanced sample fidelity through conditional guidance.

2.2.4 Architectural Developments and Extensions

Latent Diffusion Models (LDMs) [38] improved scalability by performing the diffusion process in a learned latent space rather than pixel space, greatly reducing computational requirements while maintaining visual quality. This design choice enabled large-scale models such as Stable Diffusion and facilitated widespread adoption of diffusion-based text-to-image systems.

More recent developments have explored alternative training objectives and architectures. Consistency Models [39] aim to generate samples in fewer steps by enforcing consistency across different noise levels, while Rectified Flow [40] reformulates diffusion as a continuous flow-matching problem. Additionally, Diffusion Transformers [41] extend diffusion modeling to video and multi-modal generation, demonstrating the flexibility and generality of the diffusion framework.

2.3 Text-to-Image Generation

The text-to-image generation task focuses on synthesizing coherent and semantically aligned images conditioned on textual descriptions. Early works like AttnGAN [42] used attention mechanisms to align words and image regions. The introduction of CLIP [10], a contrastively trained multimodal model, enabled large-scale alignment of text and vision, leading to the emergence of high-quality text-to-image diffusion models. DALL·E 2 [43], Imagen [44], and Stable Diffusion [38] demonstrated unprecedented fidelity by leveraging transformer-based text encoders and large-scale image-text datasets such as LAION-5B [45]. These systems have been surveyed comprehensively in [11], revealing that diffusion-based approaches outperform GAN-based methods in both realism and controllability.

2.4 Controllable Text-to-Image Synthesis

While text-to-image diffusion models can generate visually impressive content, they often lack explicit spatial or structural control. To address this, controllable diffusion

frameworks were proposed, allowing users to guide generation via additional conditions. ControlNet [46] introduced trainable adapters that accept control signals such as depth maps, segmentation masks, or human poses. Similarly, T2I-Adapter [47] enables lightweight plug-in control modules to extend pretrained diffusion models.

Cross-attention mechanisms play a fundamental role in modern text-to-image diffusion models by enabling the effective alignment of textual descriptions with visual representations. By injecting text embeddings into the denoising process, cross-attention allows the model to translate semantic concepts expressed in natural language into corresponding visual elements, substantially improving the fidelity and semantic consistency of generated images. This mechanism has become a core component of state-of-the-art systems, including DALL-E [48], Imagen [44], and Stable Diffusion [9], where cross-attention layers are applied throughout the generation process to guide image synthesis.

Beyond generation, cross-attention maps have been leveraged for image editing and interpretability. Prompt-to-Prompt [49] exploits cross-attention to modify generated images by selectively altering textual conditions, a technique later extended to real-image editing by Mokady *et al.* [50]. Attend-and-Excite [51] further demonstrates the interpretability of cross-attention by using attention maps to enforce object-level faithfulness during generation. In the context of personalized image synthesis, Break-a-scene [21] employs cross-attention maps to disentangle and recombine multiple concepts extracted from a single image. Building upon this framework, the proposed MCGM integrates an additional mask-based embedding into the cross-attention mechanism, enabling explicit control over subject pose. Furthermore, a dedicated cross-attention loss is introduced to strengthen the association between spatial masks and their corresponding visual concepts, improving controllability and structural consistency.

2.5 Personalization and Style Transfer in Diffusion Models

Personalization and style adaptation have become essential for customized generation. Textual inversion [52] learns new token embeddings representing a concept or

style, enabling personalized generation without full fine-tuning. DreamBooth [17] fine-tunes diffusion models with subject-specific images to preserve identity while generating novel contexts. LoRA [53] introduced parameter-efficient fine-tuning through low-rank adaptation, and subsequent works such as Custom Diffusion [54] and Style-Drop [55] enhanced multi-style generation and fidelity.

Recent approaches, including ArtAdapter [56], ControlStyle [57], and Style-A-Video [58], push this further by enabling dynamic style conditioning and temporal consistency across generated frames. These advances show the growing importance of personalization and controllable aesthetics in generative AI.

2.6 Few-Shot and Single-Image Learning

Despite the success of large-scale diffusion models, their dependency on massive datasets remains a challenge. Few-shot and single-image learning approaches attempt to generalize from limited data. Break-a-scene [59] extracts multiple concepts from a single image and recombines them in new contexts. This line of research inspires models like the Mask Conditional Generative Model (MCGM) and its extension, MCGM-Styler, which introduce mask-based and style-based conditioning to enable fine-grained pose and style control from minimal data.

2.7 Summary and Positioning of This Thesis

In summary, the evolution of generative models has transitioned from VAEs and GANs to highly flexible and controllable diffusion-based frameworks. Text-to-image synthesis has achieved remarkable realism, yet challenges persist in enabling user-driven control over pose, structure, and style, especially under data-scarce conditions. This thesis contributes to this evolving landscape through two complementary models: (1) **MCGM**, a Mask Conditional Text-to-Image Generative Model that integrates pose conditioning using spatial masks, and (2) **MCGM-Styler**, an extension that introduces style conditioning and multi-style synthesis. Together, these models advance controllable and efficient image generation by enabling both structural and stylistic

flexibility, even from a single training image.

Chapter 3

Mask Conditional Text-to-Image Generative Model (MCGM)

3.1 Introduction

Recent advances in diffusion-based text-to-image models have enabled the generation of highly realistic images from natural language descriptions. Beyond generic image synthesis, an increasingly important research direction is personalized generation, where a model is adapted to reproduce specific subjects from only a limited number of reference images. Several approaches have addressed single-concept personalization by fine-tuning pretrained diffusion models or by learning new textual embeddings that represent novel concepts. These techniques allow users to generate customized subjects in diverse contexts using textual prompts. However, while identity preservation has seen significant improvements, controllability over spatial attributes such as pose and structural configuration remains limited.

In particular, textual conditioning alone provides weak structural supervision. Since natural language descriptions are inherently ambiguous and underspecified with respect to spatial geometry, enforcing precise pose configurations becomes an ill-posed problem. As a result, personalized diffusion models may preserve subject identity while failing to satisfy explicit spatial constraints. For example, after person-

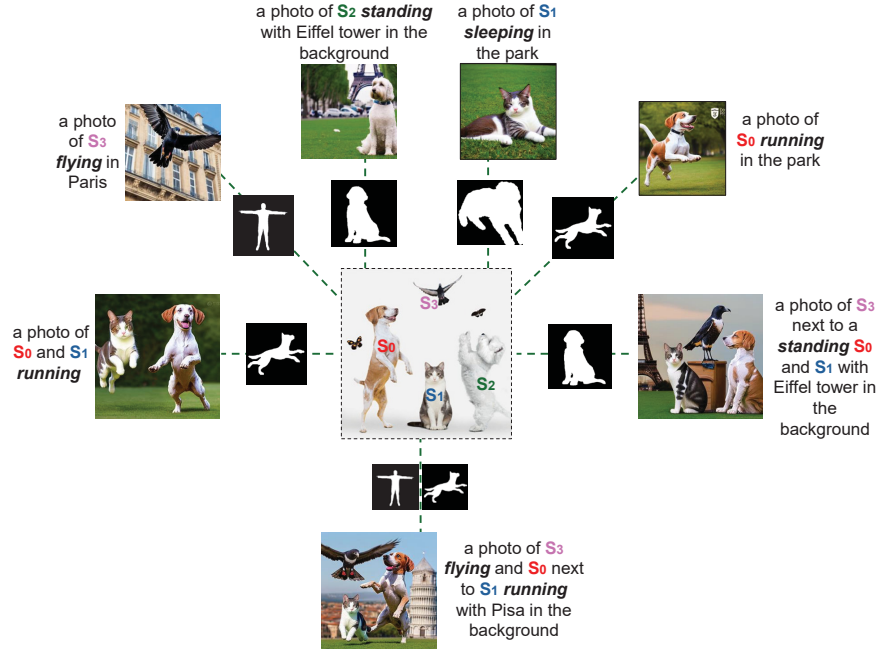


Figure 3.1: Overview of the proposed MCGM method. Starting from a single image with a set of subjects and the corresponding masks, it is possible to (a) generate any of those subjects using an arbitrary mask to set the pose, (b) generate multiple subjects at the same time using a single mask to control the pose and (c) generate multiple subjects at the same time using multiple masks to define each of their poses.

Generative models are one of the most revolutionary elements of contemporary artificial intelligence

alizing a diffusion model to generate a specific dog, it may still be difficult to enforce that the dog appears in a very specific sitting posture or orientation. This limitation reduces the practical applicability of existing personalization methods, especially in scenarios requiring fine-grained structural control.

To address this gap, this chapter introduces the Mask Conditional Generative Model (MCGM) (see Figure 3.1), a personalized diffusion framework that augments text-based conditioning with explicit spatial guidance. The key idea is to inject target

pose masks as an additional conditioning signal, allowing the model to interpret structural constraints during generation. By combining textual prompts with mask embeddings, The proposed model enables the synthesis of subjects that preserve identity while conforming to user-defined pose configurations.

Furthermore, we introduce a mask-guided cross-attention alignment loss that encourages the learned textual handles to focus on their corresponding spatial regions. This mechanism improves concept disentanglement and enhances pose consistency, particularly in multi-subject settings.

The remainder of this chapter is organized as follows. Section 3.2 reviews related work in personalized diffusion models and controllable generation. Section 3.3 presents the proposed MCGM architecture and training strategy. Experimental results and analysis are discussed in Chapter 5.

3.2 State of the Art

3.2.1 Fine-tuning diffusion models

Diffusion-based models [4, 5, 60, 61] emerged as the new state of the art for text-to-image generation [9, 49, 62, 63]. In this context, fine-tuning conditional diffusion models for text-to-image generation represents a critical area of research, aiming to enhance the performance, control, and application range of these models. Zhang *et al.* [64] highlighted that fine-tuning conditional diffusion models on specific datasets improves text-to-image generation performance. Then, Brown and Smith [65] provided a comparative analysis of fine-tuning strategies, offering insights into optimizing these processes, while Garcia *et al.* [66] emphasized the importance of hyperparameter tuning in enhancing the performance of fine-tuned diffusion models. In proposed model, to generate new images of the subject in different contexts and poses, it is based on the Break-a-scene [21] model, which has two phases: the first involves designating a collection of text tokens (handles), freezing the model weights, and optimizing the handles, while in the second one, the whole model weights are fine-tuned while continuing to optimize the handles too.

3.2.2 Cross-attention

Cross-attention mechanisms enable the model to effectively align and integrate information from textual descriptions with visual data, significantly enhancing the quality of generated images from textual inputs. In the domain of text-to-image generation, cross-attention layers have been effectively utilized in various state-of-the-art models. DALL-E [48] uses cross-attention to align text descriptions with image generation, ensuring the visual output accurately reflects the input text. Then, Imagen [44] employs a cascading pipeline of cross-attention mechanisms to integrate text embeddings into the image generation process, achieving high fidelity and resolution in the generated images. Additionally, Stable Diffusion [9] utilizes cross-attention layers throughout the denoising steps, allowing the model to generate highly-detailed and semantically-accurate images from textual descriptions. Cross-attention maps in text-to-image diffusion models are used by prompt-to-prompt [49] to edit generated images. This process was then extended to real images by Mokady *et al.* [50]. To match text-to-image generations, attend-and-excite [51] employs cross-attention maps as an explainability-based technique [67]. Furthermore, in Break-a-scene [21] cross-attention maps have the purpose of separating learned concepts. More in detail, as opposed to altering the input image, Break-a-scene concentrates on extracting textual handles from a scene and combining them into entirely new scenes. In MCGM, the cross-attention technique was utilized as in Break-a-scene but also incorporated the embedding of a mask image as extra information alongside text embeddings to help the model create a specific pose of the subject in the scene. Additionally, a novel cross-attention loss is designed to better connect the masks and the corresponding subjects.

3.2.3 Personalization in diffusion models

Recently, several approaches were proposed to reproduce in different contexts and edit real images using diffusion models. In both Textual inversion (TI) [68] and DreamBooth (DB) [17] the model learns how to generate a visual concept in different situations simply by a few images. A novel method is used by Textual inversion [68]

to embed concepts from the images into new terms in the text embedding space. By denoising supplied images, this technique increases the tokenizer dictionary’s size and maximizes more tokens. Conversely, DreamBooth [17] increases the diversity of the generated results by updating the UNet parameters with a class-specific prior preservation loss and representing concepts with low-frequency words (e.g., "V"). TI and DB are simple, flexible frameworks that provide the basis for many fine-tuning-based techniques. Moreover, weight deviations during fine-tuning are examined by Custom Diffusion [69], which highlights the critical role of cross-attention layer parameters. Furthermore, SVDiff [70] implements a successful personalization technique in the parameter space using the weight kernels’ singular-value decomposition. They also give a regularization term for mixing and unmixing, which enables the generation of two concepts that are close to each other. In addition, to enable quick personalization, ELITE [71] proposed a learning-based encoder that supports input masks and encodes a single image. Its global and local mapping networks enable precise and quick custom text-to-image generation. ELITE is still unable to produce more than one concept, though. Finally, Break-a-scene [21] can extract multiple concepts from a single image using a masking mechanism. Differently from SVDiff which requires several images for each of the concepts, Break-a-scene [21] operates on a single image containing multiple concepts. Our framework follows the Break-a-scene and ELITE idea by using only a single image for the concept, in contrast to other methods that require multiple images. In addition, unlike other recent methods, Our framework method uses a mask condition in line with the text to set the pose for one or more source subjects.

While Break-a-scene [20] focuses on extracting textual handles from masked regions, it does not provide explicit mechanisms to control target pose during generation. In contrast, MCGM introduces spatial conditioning at both architectural and optimization levels, enabling controllable pose synthesis beyond concept extraction.

3.3 Proposed Method

3.3.1 The MCGM Model Architecture:

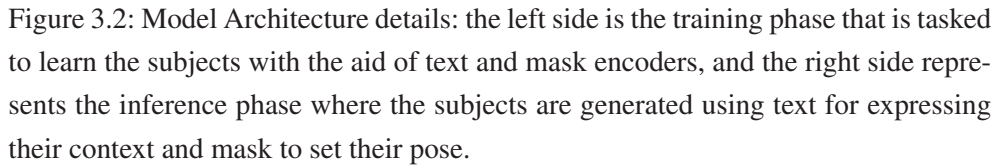
The MCGM architecture builds upon the Break-a-scene framework [21], extending it with additional pose control mechanisms. Break-a-scene goal is to extract N textual handles $\{v_i\}_{i=1}^N$, where the i th handle v_i represents the concept indicated by the mask M_i , given a single input image I and a set of N masks $\{M_i\}_{i=1}^N$ indicating the subjects in the image. Once the handles are acquired, they can be utilized to direct the synthesis of images containing the concepts in novel contexts and poses during inference by inserting them into textual prompts.

In order to train the MCGM to recognize multiple concepts from a single image, a small collection of image-text pairs with the format "A photo of $[v_x]$ and $[v_y]$..." is created, then the background is masked out based on the masks provided and adding a solid, arbitrary background. Each time, a random subset of tokens is chosen. During training, the model weights w_i and text embeddings are optimized using two distinct stages.

First, the model weights are frozen and the text embeddings are optimized to match the masked concepts. This enables a fast initial embedding while preserving the capability of the model, thereby offering a strong basis for the subsequent phase. Then, in the second stage, the model weights are unfrozen and optimized alongside the text tokens. This preserves editability while enabling a faithful replication of the extracted concepts. The concepts can be generated in a variety of contexts with minimal degradation in editability by using distinct text descriptions and masks.

In this work, the same scenario of Break-a-scene was used, but the two phases were optimized by adding a mask encoder whose purpose is encoding the mask to generate a mask embedding m_i for each token v_i and concatenating it with the text embedding t_i . The mask encoder projects the masks to embeddings having the same dimension as the text embeddings and then feeds both textual and mask embeddings in the cross-attention layers of the diffusion model.

By doing this, the mask embeddings will guide the model to specify the poses of the subjects.



3.3.2 Mask Injection

The mask encoder is employed in both training and inference processes. In the training process, the mask embeddings are generated for the input concepts in the source image, and these embeddings are used in the two optimization phases of the model. On the other hand, in the inference process, a set of arbitrary masks is selected and embedded to specify the target pose of the generated samples and are injected alongside the text embeddings in the model. By doing so, in the inference

process, the model uses the conditioning information provided by the mask encoder to generate outputs that are consistent with the mask. Loss functions described in the next sections are designed to ensure that the generated output matches the desired characteristics dictated by the mask.

Injecting mask embeddings into the cross-attention layers is a natural design choice, as cross-attention mechanisms are responsible for aligning conditioning tokens with spatial latent representations. By augmenting textual tokens with spatial mask embeddings, the model learns a joint representation that couples semantic identity with geometric structure.

3.3.3 The Optimization Losses

MCGM uses a masked version of the standard diffusion loss [72] to optimize the handles (and model weights, in the second optimization phase). This loss is calculated over the pixels of the subjects that are defined by the concepts masks:

$$L_{rec} = \mathbb{E}_{z_t, s, \epsilon \sim N(0,1), t} \left[\|\epsilon \odot M_s - \epsilon_\theta(z_t, t, c_s) \odot M_s\|_2^2 \right] \quad (3.1)$$

where z_t is the latent with added noise at time step t , s is the number of concepts, c_s is the set of embeddings obtained from text and masks, M_s is the set of masks, ϵ is the noise, and ϵ_θ is the denoising network.

Rebuilding the concepts faithfully is encouraged by using the *masked diffusion loss* [21] in pixel space. In addition to reconstructing the pixels of the learned subjects, the model needs to make sure that each handle only looks at the portion of the image where the related concept is located. To this end, another loss term—the cross-attention loss—has been introduced. This loss is defined as the *MSE* between the input masks and the attention calculated in the cross-attention layers of the model. The following cross-attention loss is used for each training phase:

$$L_{attn} = \mathbb{E}_{z_t, k, t} \left[\|\mathcal{C}\mathcal{A}_\theta(v_i, z_t) - M_{ik}\|_2^2 \right] \quad (3.2)$$

where $\mathcal{C}\mathcal{A}_\theta(v_i, z_t)$ is the attention map of the token v_i and the noisy latent z_t obtained from each cross-attention layer.

This formulation can be interpreted as an explicit spatial alignment regularizer. By minimizing the discrepancy between attention maps and ground truth masks, the model is encouraged to allocate attention mass to semantically relevant spatial regions. This reduces cross-concept interference and improves disentanglement in multi-subject scenarios.

In MCGM, mask embeddings are injected with the textual embeddings for each subject in order to define its pose. This additional conditioning may introduce ambiguity in associating each subject with its corresponding spatial region, potentially leading to concept mixing or uncontrolled compositions, as will be explained later in Chapter 5, the section of the experiments. Therefore, to control this process of distinguishing between the subjects in the output image, we added a mask cross-attention loss specifically designed for the mask tokens. The final cross-attention loss becomes as follows:

$$L_{Mattn} = L_{attn} + \lambda_m \mathbb{E}_{z,k,t} \left[\|\mathcal{CA}_\theta(m_i, z_t) - M_{ik}\|_2^2 \right] \quad (3.3)$$

where λ_m is the weight for the mask cross-attention loss, $\mathcal{CA}_\theta(m_i, z_t)$ is the cross-attention map between the mask token m_i and the noisy latent z_t .

Finally, the full training objective becomes:

$$L_{total} = L_{rec} + \lambda_{attn} L_{Mattn} \quad (3.4)$$

Adding L_{Mattn} to the loss ensures that concept handles focus on their respective regions. We set $\lambda_{attn} = 0.01$ for accurate spatial positions in the produced image.

3.4 Chapter Summary

In this chapter, the Mask Conditional Generative Model (MCGM) was presented, a diffusion-based framework designed to extend single-image personalized generation with explicit pose control. By integrating a mask-based conditioning mechanism into the Break-a-scene architecture, MCGM enables fine-grained spatial control over one or multiple subjects while preserving the semantic information provided by textual descriptions. The chapter detailed the motivation, architectural design, and training strategy of the proposed model.

The effectiveness of MCGM, including its ability to generate structurally consistent and pose-controlled images, is evaluated through quantitative and qualitative experiments in Chapter 5. That chapter presents a comprehensive analysis of the model’s performance and comparisons with existing state-of-the-art approaches.

3.5 Contributions of This Chapter

The main contributions introduced in this chapter can be summarized as follows: The main contributions introduced in this chapter can be summarized as follows:

- The proposal of the **Mask Conditional Generative Model (MCGM)**, a diffusion-based framework that enables pose-controlled personalized image generation from a single reference image.
- The integration of **mask-based spatial conditioning** within the cross-attention layers of the diffusion model, allowing explicit control over subject pose during image synthesis.
- The introduction of a **mask cross-attention loss** that improves the alignment between learned textual handles and spatial regions, particularly in multi-subject scenarios.
- The ability to generate images containing **multiple subjects with controllable poses** using a single training image, extending the capabilities of existing personalized diffusion models.

Transition to Style-Aware Generation

While MCGM enables effective pose-controlled personalized image generation, it does not explicitly address stylistic control in the generated images. In many practical scenarios, users may wish to generate subjects not only in specific poses but also in particular artistic styles. To address this limitation, the next chapter introduces

MCGM-Styler, an extension of MCGM that incorporates style conditioning into the generation process while preserving pose controllability.

Chapter 4

Free-Form styler for Mask Conditional Text-to-Image Generative Model (MCGM-Styler)

4.1 MCGM-Styler : Motivation and Overview

Text-to-image diffusion models have recently demonstrated remarkable capabilities in generating high-quality and semantically consistent images from natural language descriptions [9, 44, 62]. These models have enabled a wide range of applications, including artistic content creation, virtual environments, and design assistance. In addition to generic image synthesis, recent research has increasingly focused on *personalized generation*, where a model learns to reproduce specific subjects from a limited number of reference images while allowing them to be generated in new contexts using textual prompts [17, 68, 71]. Several approaches have addressed this problem by adapting pretrained diffusion models through fine-tuning strategies or by learning new textual embeddings representing novel concepts. Methods such as Textual inversion [68] and DreamBooth [17] allow a model to capture the identity of a subject and reproduce it in different scenes. However, despite their effectiveness in preserving subject identity, these approaches typically provide limited control over

structural attributes, such as the pose or spatial arrangement of objects within the generated image. To address this limitation, the Mask Conditional Generative Model (MCGM) [73] introduced a mask-based conditioning mechanism that enables explicit pose control during personalized image generation. By injecting pose masks into the diffusion process, MCGM allows users to guide the spatial configuration of subjects while maintaining their visual identity. This mechanism significantly improves the controllability of diffusion-based personalized generation, particularly in multi-subject scenarios. Nevertheless, while MCGM provides precise pose guidance, it does not explicitly address another important aspect of controllable image synthesis: artistic style adaptation. In many practical applications, users may wish not only to control the pose of the generated subjects but also to generate images that follow a particular artistic or visual style. Existing style-transfer approaches often require multiple training images or specialized training procedures, which limits their applicability in single-image personalization scenarios [74, 75]. To overcome these limitations, this chapter introduces **MCGM-Styler**, an extension of the Mask Conditional Generative Model designed to enable simultaneous control over object pose and visual style in text-to-image generation from a single reference image. The proposed approach integrates style conditioning into the MCGM framework through an additional training stage that allows the model to learn stylistic characteristics while preserving the pose control provided by mask conditioning. By combining textual prompts, pose masks, and style information, MCGM-Styler enables the generation of images that are consistent with the desired scene description, structural configuration, and artistic style.

The remainder of this chapter presents the architecture and training strategy of the proposed model. First, related work on style-aware generative models and controllable diffusion techniques is discussed. Then, the MCGM-Styler framework is introduced, including the integration of style conditioning into the mask-guided diffusion pipeline. Finally, implementation details and training procedures are described, while experimental evaluations are presented in Chapter 5.

MCGM-Styler builds upon the Mask Conditional Generative Model (MCGM) [73], introduced in Chapter 3, which enables pose-controlled image generation through

mask-based spatial conditioning. Figure 4.1 illustrates the capabilities of the proposed model. Starting from an image containing one or more subjects, MCGM-Styler is trained sequentially using an action pose mask, a target style image, and a textual prompt. This training strategy enables flexible scene generation in which users can synthesize images featuring any subject or combination of subjects in arbitrary poses and styles. As shown in the figure, the model supports multiple usage scenarios, including single-subject generation, multi-subject scenes generated in a unified style, and scenes composed using mixed artistic styles. Additionally, MCGM-Styler can generate multiple stylistic variations of a single subject, highlighting its ability to disentangle content, pose, and style.

Overall, MCGM-Styler extends the controllability of personalized diffusion models by jointly modeling three complementary factors: subject identity, spatial pose, and artistic style.

4.2 State of the art

Recent works have explored the integration of artistic style into diffusion-based image generation. Several approaches extend pretrained text-to-image diffusion models in order to learn stylistic representations from a small number of reference images. For instance, StyleDrop [76] adapts large diffusion models to mimic artistic styles through style-specific fine-tuning, while DreamStyler [77] introduces a training strategy for learning style tokens that guide the generation process toward specific artistic patterns.

Stylized text-to-image generation has also been investigated in applications related to art and design. Methods such as InstantStyle [78], DreamStyler [77], and UnZipLoRA [79] focus on improving the disentanglement between style and content to produce visually coherent outputs while preserving semantic consistency with textual prompts. More recently, Mamba-ST [80] explored the use of state-space models for style transfer, demonstrating that alternative generative architectures can also effectively capture stylistic representations.

Despite these advances, most existing style-personalization methods primarily

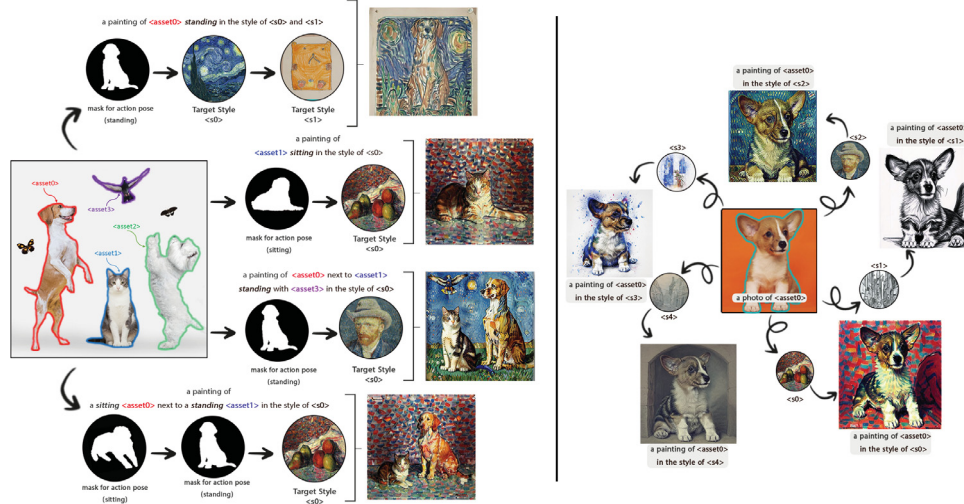


Figure 4.1: **Overview of the MCGM-Styler capabilities.** [On the left] starting from an image containing multiple subjects, the model is trained sequentially using an action pose mask, a target style, and a textual prompt. This enables users to generate scenes with arbitrary subject(s), pose(s), and style(s). The bottom row illustrates the generation of multiple subjects in a single style, the middle row shows a single-subject scenario, and the top row demonstrates the generation of images with **mixed styles**, where two or more artistic styles are combined within the same output. [On the right] the model generates paintings of a single subject rendered in various artistic styles.

focus on stylistic adaptation and do not provide explicit spatial control over the generated subjects. In particular, they lack mechanisms to guide pose or structural configuration, which limits their applicability in scenarios requiring precise scene composition or multi-object arrangements.

MCGM-Styler addresses this limitation by extending the Mask Conditional Generative Model (MCGM) [73] with an additional style-learning stage. By combining mask-based spatial conditioning with style embeddings extracted from reference im-

ages, the proposed architecture enables simultaneous control over subject identity, pose, and artistic style during the generation process. Furthermore, the model supports the simultaneous composition of multiple artistic styles within a single generated scene, allowing users to synthesize scenes where different subjects can share a unified style or exhibit mixed stylistic characteristics derived from multiple reference styles.

4.3 MCGM-Styler Model Architecture

The architecture of MCGM-Styler builds directly upon the MCGM framework [73] and is structured into two sequential training stages. In both stages, the model relies on a latent diffusion backbone equipped with cross-attention layers that integrate textual and auxiliary conditioning signals.

In the first stage, the model learns object concepts and their spatial configuration using a text encoder for concept-specific tokens, a mask encoder that embeds spatial pose information, and a diffusion model that integrates these embeddings through cross-attention.

In the second stage, the model is fine-tuned to learn a style representation from a single reference style image, encoded through a dedicated style token. During inference, the trained model generates an image from random noise, guided jointly by textual descriptions specifying content and style and spatial masks defining the target pose of the subjects.

The overall architecture of the proposed MCGM-Styler is illustrated in Figure 4.2. The model follows a two-stage training strategy designed to disentangle concept, pose, and style representations while maintaining controllable image synthesis from minimal data. The first stage focuses on learning subject-specific concepts and their associated poses using textual and mask-based conditioning, while the second stage specializes the model to capture stylistic attributes from a single style reference image. During inference, these learned representations are jointly leveraged to generate images guided by a textual prompt, a target pose mask, and one or more style tokens.

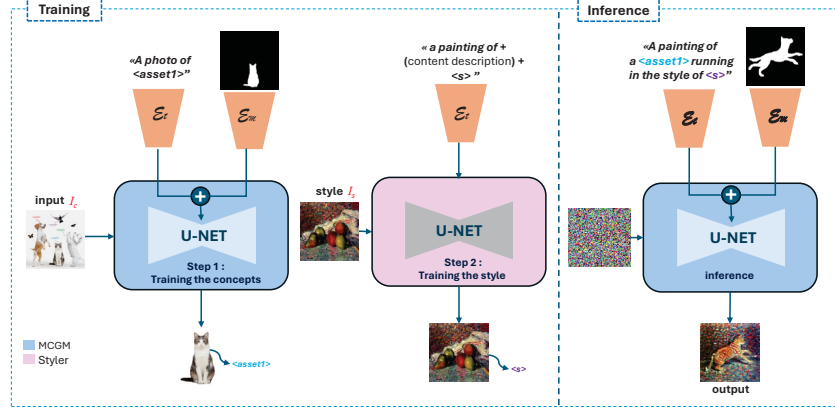


Figure 4.2: Pipeline of the MCGM-Styler: The left section illustrates the two-stage training process, with source and style images provided as user inputs. In the first stage, various concepts are learned using textual prompts and mask inputs. In the second stage, the style concept is trained. The token $\langle asset1 \rangle$ represents the cat in the source image, while $\langle s \rangle$ represents the style. Note that the (+) symbol indicates concatenation of the text with mask embeddings. The right section depicts the inference process, where a textual prompt description and a target position mask are used to generate the output scene.

4.4 Style Representation and Conditioning

4.4.1 Preliminary: From MCGM to MCGM-Styler

The objective of MCGM-Styler is to generate images of arbitrary subjects in arbitrary poses and styles, using only a single reference image per subject and a single style image. The proposed approach builds upon the Mask Conditional Generative Model (MCGM), which itself extends the Break-a-scene framework by introducing spatial mask conditioning for pose control.

Given a source image I_c containing one or more subjects, MCGM introduces a set

of N learnable textual handles $\{v_i\}_{i=1}^N$, where each handle corresponds to a specific subject in the image. Each subject is associated with a spatial mask M_i describing its target pose. A mask encoder $\mathcal{E}_m$ maps each mask to a mask embedding m_i , which is concatenated with the corresponding textual embedding t_i . The combined embeddings are injected into the cross-attention layers of the diffusion model, enabling explicit and fine-grained pose control during generation.

While MCGM provides effective control over subject identity and pose, it lacks an explicit mechanism for controlling visual style. MCGM-Style addresses this limitation by introducing an additional training stage dedicated to learning stylistic representations, thereby extending the controllability of the generation process.

4.4.2 Injecting Style into the Generation Process

To enable style control, MCGM-Style introduces a second training stage in which the model learns a style representation from a single reference image. This is achieved by learning a new style token s_* that can be included in textual prompts at inference time.

The training process is divided into two stages. The first stage focuses on learning subject concepts and their poses, while the second stage focuses on learning the style.

Stage One: Learning Concepts and Poses. In the first stage, the model is trained to extract and disentangle subject concepts from a single image I_c using textual prompts and spatial masks. A small set of image–text pairs is constructed using prompts such as "A photo of $[v_x]$ and $[v_y]$ ", where each learnable token v_i represents a subject. The background is masked out to constrain learning to the subjects of interest.

Textual inversion is used to optimize the token embeddings, and a masked diffusion reconstruction loss [72] is applied to restrict the reconstruction error to the masked regions:

$$L_{rec1} = \mathbb{E}_{z,s,\varepsilon \sim N(0,1),t} \left[\|\varepsilon \odot M_s - \varepsilon_\theta(z_t, t, c_s) \odot M_s\|_2^2 \right] \quad (4.1)$$

where z_t denotes the latent variable with noise added at time step t , s represents the number of concepts, c_s is the set of embeddings obtained from text and masks, M_s corresponds to the set of masks, ε indicates the noise, and ε_θ refers to the de-noising network.

To ensure that each textual handle attends exclusively to its corresponding subject, a cross-attention loss is introduced:

$$L_{attn} = \mathbb{E}_{z,k,t} \left[\|\mathcal{C} \mathcal{A}_\theta(v_i, z_t) - M_{ik}\|_2^2 \right]. \quad (4.2)$$

Since pose mask embeddings are injected alongside textual embeddings, an additional mask-based cross-attention loss is included:

$$L_{Mattn} = L_{attn} + \lambda_m \mathbb{E}_{z,k,t} \left[\|\mathcal{C} \mathcal{A}_\theta(m_i, z_t) - M_{ik}\|_2^2 \right]. \quad (4.3)$$

The total loss for the first training stage is therefore:

$$L_{T1} = L_{rec1} + \lambda_{t1} L_{Mattn}. \quad (4.4)$$

Stage Two: Learning the Style. In the second stage, the model is fine-tuned to learn a style representation from a single style image I_s . A learnable style token s_* is introduced using textual inversion, and the model is trained using prompts of the form "A painting of s_* ". The standard diffusion loss [72] is applied:

$$L_{T2} = \mathbb{E}_{z,\varepsilon,t} \left[\|\varepsilon - \varepsilon_\theta(z_t, t, s_*)\|_2^2 \right]. \quad (4.5)$$

After completing both training stages, the model is capable of generating images conditioned on subject identity, pose masks, and one or more learned style tokens.

4.5 Decoupling Content and Style

A fundamental challenge in learning style from a single image is the entanglement of content and style. Without additional constraints, the model may encode object identity, composition, or background information into the style token, leading to unintended content transfer during inference.

To mitigate this issue, MCGM-Styler introduces a lightweight content–style decoupling strategy based on contextual descriptions. During style training, a contextual descriptor p_c describing non-style attributes of the style image such as objects, layout, and background is inserted into the training prompt.

Let the standard prompt be defined as $P = [p_0, s_*]$. Instead, the proposed approach uses:

$$C = [p_0, p_c, s_*]. \quad (4.6)$$

An example training prompt is "A painting of a pencil, pears, and apples on a cloth, in the style of s_* ." The contextual description typically consists of 5-15 words and can be prepared by a non-expert user in less than one minute. This strategy encourages the model to encode only stylistic characteristics in the style token while preserving content and pose control through textual prompts and masks.

4.6 Inference

At inference time, MCGM-Styler generates images starting from random noise using a pre-trained diffusion model conditioned on multiple inputs. The user provides a textual prompt describing the scene, a target pose mask specifying the spatial configuration of the subject(s), and one or more learned style tokens. The pose mask embeddings and textual embeddings are injected into the cross-attention layers, enabling simultaneous control over subject identity, pose, and style.

This design allows flexible generation scenarios, including single-subject or multi-subject scenes, single-style or mixed-style compositions, and fine-grained pose manipulation. The effectiveness of the proposed approach is quantitatively and qualitatively evaluated in Chapter 5, where MCGM-Styler is analyzed jointly with MCGM.

4.7 Chapter Summary

This chapter presented *MCGM-Styler*, a style-aware extension of the Mask Conditional Generative Model designed to enable controllable text-to-image generation

conditioned on subject identity, pose, and artistic style. Building upon the pose-controlled generation capabilities of MCGM, the proposed approach introduces an additional training stage that allows stylistic attributes to be learned from a single reference image.

The two-stage training strategy enables an effective disentanglement of subject concepts, spatial configuration, and visual style while maintaining minimal data requirements. Spatial mask conditioning preserves fine-grained pose control, whereas the learned style tokens allow flexible appearance customization, supporting single-style, multi-style, and mixed-style generation scenarios. In addition, a lightweight content-style decoupling strategy based on contextual descriptions was introduced to reduce entanglement between stylistic and semantic attributes, improving robustness during inference.

Overall, MCGM-Styler extends the original MCGM [73] framework into a unified system capable of generating diverse and personalized images with precise structural and stylistic control from limited supervision.

4.8 Contributions of This Chapter

The main contributions introduced in this chapter can be summarized as follows:

- A style-aware extension of the Mask Conditional Generative Model that enables simultaneous control over subject identity, pose, and artistic style using a single reference image.
- A two-stage training strategy that separates concept and pose learning from style learning, improving modularity and controllability in diffusion-based image generation.
- A lightweight content-style decoupling mechanism based on contextual descriptions, reducing unintended content transfer while requiring minimal user effort.

- Support for flexible generation scenarios, including single-subject and multi-subject scenes, as well as single-style, multi-style, and mixed-style compositions.

The effectiveness of the proposed MCGM-Styler framework is evaluated through extensive qualitative and quantitative experiments. In the next chapter, we present a comprehensive experimental analysis of both MCGM and MCGM-Styler, including comparisons with state-of-the-art methods, ablation studies, and evaluations of pose accuracy, style fidelity, and visual quality. These results demonstrate the advantages and limitations of the proposed approaches in a variety of generation scenarios.

Chapter 5

Experimental Results

This chapter presents a comprehensive experimental evaluation of the proposed models introduced in this thesis. We begin by analyzing the Mask Conditional Generative Model (MCGM), which extends Break-a-scene [20] by enabling explicit pose control through mask-based conditioning. Subsequently, we evaluate the extended MCGM-Styler model, which further introduces style control while preserving pose consistency.

The experiments are designed to answer the following research questions: (i) Can pose information be reliably injected into text-to-image diffusion models trained on a single image? (ii) Does mask-based conditioning improve controllability without degrading image quality? (iii) Can multiple subjects and actions be handled simultaneously? (iv) Can pose control and style transfer be combined effectively?

5.1 Results on MCGM

This section focuses on the experimental evaluation of MCGM, while results for MCGM-Styler are presented later in this chapter.

5.1.1 Experimental Setup

The objective of these experiments is to evaluate the ability of MCGM to generate pose-controlled images from a single reference image. In particular, we analyze how the proposed mask encoder and mask cross-attention loss affect pose accuracy, subject disentanglement, and overall image quality.

MCGM incorporates an additional mask encoder and a mask-based conditioning mechanism injected into the diffusion cross-attention layers. This allows the generation process to be guided not only by textual prompts but also by explicit spatial constraints.

The experimental evaluation is conducted using images collected from multiple sources. A subset of the reference images is selected from the COCO dataset [81], which provides high-quality instance segmentation masks. To increase the diversity of the evaluation, additional images depicting different subjects, including both humans and animals performing various actions, are manually collected. For these images, segmentation masks are automatically generated using the STEGO unsupervised segmentation framework [82]. All images and masks are resized to a fixed resolution of 512×512 pixels to ensure consistency during training. Since MCGM is designed for single-image personalization, each experiment is trained using a single reference image containing one or more target subjects together with their corresponding segmentation masks. The same experimental protocol is adopted across all experiments, while different reference images are used to evaluate the robustness of the proposed framework across diverse subjects, scenes, and pose configurations.

5.1.2 Training Details

The training procedure follows the two-stage optimization strategy proposed in Break-a-scene, which has proven effective for single-image personalization. The proposed MCGM model is implemented in PyTorch as an extension of the official Break-a-Scene framework and all experiments are conducted on an NVIDIA A100 GPU. Unless otherwise specified, the remaining implementation settings follow those of the original Break-a-Scene model to ensure a fair comparison with the baseline. In

the first stage, only the textual embeddings corresponding to the learned concepts are optimized. A relatively high learning rate of 5×10^{-4} is used to encourage rapid adaptation of the token embeddings to the target concepts.

In the second stage, the diffusion UNet and text encoder weights are fine-tuned using a lower learning rate of 2×10^{-6} . This conservative learning rate is necessary to preserve the generalization capabilities of the pretrained model while refining it to incorporate pose information.

Both stages use the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 1×10^{-8} . The total number of optimization steps is set to 800, with an equal split between the two stages. This configuration represents a trade-off between training efficiency and generation quality, as confirmed empirically.

5.1.3 Single-Subject Results

We first analyze the behavior of MCGM in single-subject scenario. We decided to test this scenario using different configurations:

Textual prompts with coherent masks: Figure 5.1 presents examples where textual prompts are paired with coherent pose masks. The generated images demonstrate that MCGM successfully integrates semantic information from the text prompt with spatial constraints from the mask. Although the generated subject does not exactly replicate the mask shape, the overall pose and orientation are preserved. This behavior is desirable, as it allows the model to maintain visual realism while respecting the intended action.

Same textual prompt with different mask shapes. To further analyze the contribution of the mask encoder, we generate images using identical textual prompts but different pose masks. As illustrated in Figure 5.2, the resulting images exhibit significantly different poses while maintaining consistent appearance and semantics. This confirms that pose control is primarily governed by the mask input rather than the text description; for instance, Figs. 5.2(a) and 5.2(b) show that the dog in the input source

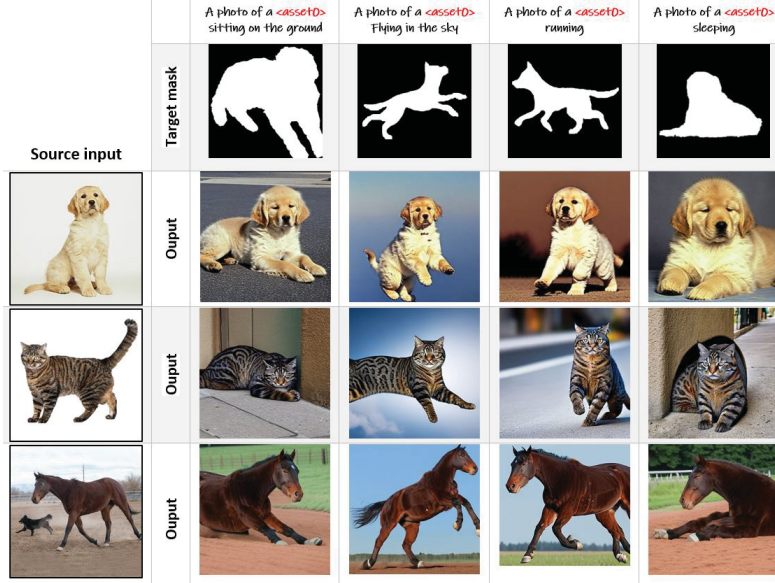


Figure 5.1: Different generated images for the input source images in different scenes based on the text prompt and with the pose guided by the target masks.

is (a)sitting/(b)running in different positions based on the mask, although using the same text prompt.

Comparison with blank masks and Break-a-scene. Figure 5.3 compares three configurations: MCGM with shaped masks, MCGM with blank masks, and the original Break-a-scene model which does not employ a mask during inference. When shaped masks are used, the generated subjects closely follow the desired pose. In contrast, Blank Masks and Break-a-scene rely solely on textual descriptions, resulting in unconstrained poses. This experiment highlights that MCGM generalizes Break-a-scene by enabling pose control while retaining compatibility with the original framework. Practically, the generated samples in Figure 5.3 are obtained with six pairs of prompts (three for the creature and three for the dog). It is straightforward to notice that scenes generated with the shaped masks follow the textual prompt and the mask poses, whereas, by injecting the blank masks or using the Break-a-scene model, the

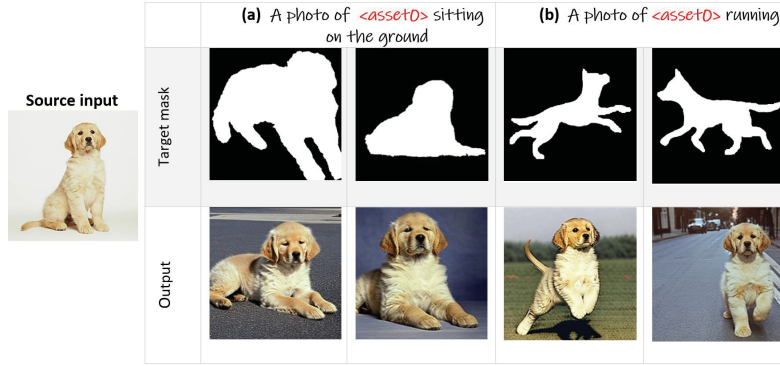


Figure 5.2: Different images are generated for the same dog image using the same text prompts and different target mask poses; in (a), the dog is sitting on the ground based on the shape of two different masks. Next, in (b) the dog is running with two different poses based on the mask shape.

scene follows only the textual text by applying a random shape. Indeed, our model can be seen as a more powerful version of Break-a-scene: when using a blank mask our model is conceptually the same, but when injecting a shaped mask it can control the pose of the generated samples.

General-Subject Generation with Mask Conditioning Beyond reproducing specific concepts learned from a reference image, an important question is whether MCGM can generalize pose control to broader semantic categories. To investigate this capability, we evaluate the model in a setting where no specific subject instance from the source image is used. Instead, the model is guided solely by a generic textual description of the subject class (e.g., *a dog*) together with different target pose masks.

In this experiment, the same textual prompt is used across all generations, while varying only the injected pose mask. Figure 5.4 illustrates representative results for the prompt "a dog playing football" where each generated image corresponds to a different target mask. Despite the absence of a specific subject instance from the training image, the model successfully generates plausible dogs that adopt poses consistent

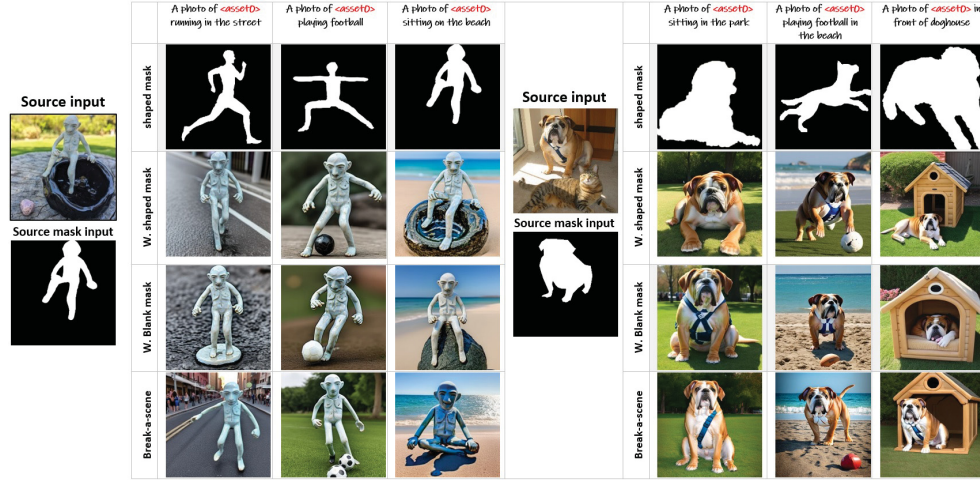


Figure 5.3: Different images generated by MCGM with shaped mask, MCGM with a blank mask, or the Break-a-scene model

with the provided masks.

These results demonstrate that the pose-conditioning mechanism in MCGM is not limited to memorized subject instances. Instead, the model learns a more abstract representation of pose that can be transferred across object categories. Notably, the generated poses do not rigidly replicate the mask shape; rather, the mask acts as a structural guide that the model adapts to maintain visual realism.

An especially illustrative example is shown in the rightmost image of Figure 5.4, where a human-shaped mask is used to guide the generation of a dog. Despite the anatomical differences between humans and dogs, the model produces a coherent image of a dog adopting a pose that clearly reflects the intended action. This behavior highlights the flexibility of the proposed mask-conditioning strategy and indicates that pose is learned as a high-level structural prior rather than as a strict pixel-level constraint.

Overall, this experiment confirms that mask injection provides an additional and independent degree of freedom beyond textual prompting. By decoupling pose control from subject identity, MCGM enables users to specify what should appear in the

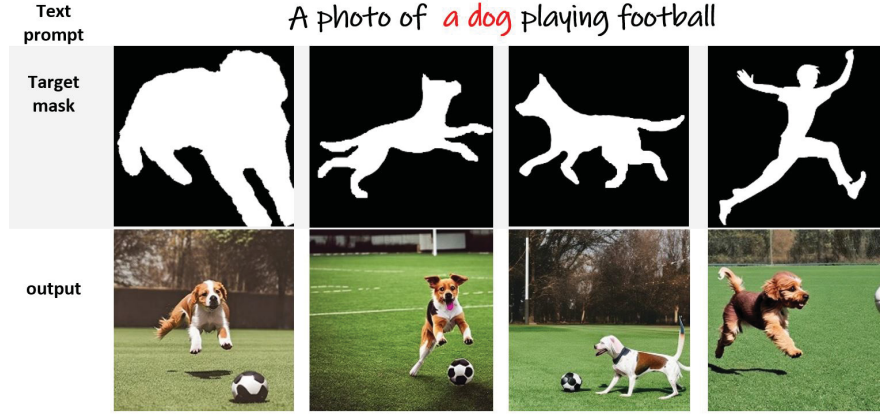


Figure 5.4: The same prompt with different target masks was used to generate different positions for the dog playing football.

image through text and how it should be positioned or posed through masks, significantly enhancing controllability in single-image text-to-image generation.

5.1.4 Multi-Subject Results

Generating scenes containing multiple subjects presents additional challenges compared to single-subject generation. The model must not only preserve the identity of each subject, but also correctly disentangle their representations, associate each subject with the appropriate textual description, and, when applicable, enforce distinct pose constraints. In this section, we evaluate the ability of MCGM to handle these challenges when learning from a single image containing multiple concepts.

Learning Multiple Concepts from a Single Image

The first requirement for multi-subject generation is the ability to learn and separate multiple visual concepts from a single reference image. Unlike traditional personalization approaches that rely on multiple images per subject, MCGM inherits from Break-a-scene the capability to extract several concepts simultaneously using concept-specific textual handles.

During training, each subject in the source image is associated with a distinct textual handle and a corresponding segmentation mask. This design encourages the diffusion model to learn disjoint representations for each concept while sharing a common background. The effectiveness of this disentanglement is evaluated qualitatively through downstream generation tasks.

Single-Subject Generation from Multi-Subject Sources

We first assess whether MCGM can selectively generate a single subject from a multi-subject source image. Figure 5.5 and Figure 5.6(b) illustrates results obtained by conditioning the model on a single textual handle and its associated mask, while ignoring the remaining subjects present in the source image.

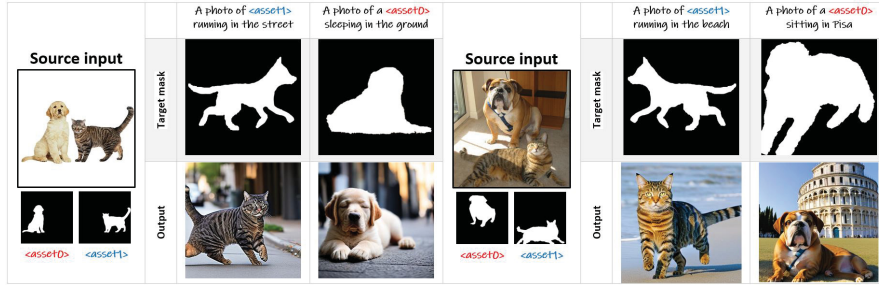


Figure 5.5: Some results for multi-subject source images used different text prompts associated with masks to generate scenes for a single subject.

The results show that the model successfully isolates the requested subject and generates coherent scenes that follow both the textual description and the injected pose mask. Importantly, other subjects from the source image are not inadvertently reproduced, indicating that the learned representations remain well-separated. This behavior demonstrates that MCGM maintains precise control over subject selection even when multiple concepts are learned jointly.

		(a) One object – No action/mask A photo of <asset1> in the beach	(b) One object – One action/mask A photo of <assetD> sleeping in the park	(c) Two objects – One mask/action A photo of <assetD> and <asset1> sitting in the street	(d) Three objects – One mask/action A photo of <assetD> and <asset1> and standing next to <asset2> with swimming pool in the background	(e) Two object – Two masks/actions a photo of a sleeping <asset1> looking for a flying <asset3> in the street	(f) Three object – Two masks/actions a photo of <asset3> flying and <assetD> next to <asset1> running with Pisa in the background
Source input	Prompt						
	Mask condition	No mask					
Source masks input	Output						

Figure 5.6: Some results for multi-subject source images used different text prompts associated with mask support the action position in the scene, each column has a different case based on the number of subjects and the number of masks used in the scene.

Multi-Subject Generation with Shared Pose Constraints

We next evaluate scenarios in which multiple subjects are generated simultaneously while sharing a common pose constraint. In this setting, a single mask is provided to guide the action or posture of multiple subjects in the scene.

As shown in Figure 5.6(c) and (d), MCGM is capable of generating scenes containing two or three subjects performing similar actions, such as standing or running, guided by a single mask. Despite sharing the same pose constraint, the generated subjects maintain distinct appearances and identities, and the overall scene remains visually coherent.

This experiment highlights the ability of the model to interpret the mask as a high-level structural guide rather than a rigid template, allowing it to adapt the pose constraint appropriately for different subjects.

Multi-Subject Generation with Independent Pose Constraints

A more challenging scenario arises when different subjects within the same scene are required to adopt distinct poses. To evaluate this capability, we provide multiple

masks, each associated with a specific textual handle.

Figure 5.6(e) and (f) demonstrate that MCGM can successfully generate scenes with two or three subjects performing different actions simultaneously, each following its corresponding pose mask based on its generated action. The model correctly associates each mask with the intended subject, avoiding pose confusion or concept mixing.

This result confirms the effectiveness of the proposed mask cross-attention mechanism, which enforces a strong correspondence between subject tokens, mask embeddings, and spatial regions in the generated image.

5.1.5 Discussion and Limitations

Overall, the multi-subject experiments demonstrate that MCGM extends pose-controlled generation beyond isolated subjects to complex scenes involving multiple interacting entities. The model consistently preserves subject identity, respects pose constraints, and avoids common failure modes such as subject blending or incorrect mask assignment.

Nevertheless, we observe that as the number of subjects and masks increases, the generation task becomes more challenging, occasionally leading to minor spatial inconsistencies or reduced visual fidelity. These limitations are primarily attributed to the increased complexity of the conditioning signals and the reliance on a single reference image for learning all concepts.

Despite these challenges, MCGM represents a significant advancement over existing single-image personalization methods by enabling flexible and controllable multi-subject scene generation with explicit pose control.

5.1.6 Ablation Study

To quantify the contribution of the proposed mask cross-attention loss, we conduct an ablation study comparing three configurations: MCGM with mask attention loss, MCGM without mask attention loss, and the baseline Break-a-scene model.

As shown in Figure 5.7, removing the mask attention loss leads to incomplete

concept separation, resulting in mixed or ambiguous subjects. This behavior occurs because the diffusion loss alone does not explicitly enforce spatial correspondence between learned handles and image regions. By contrast, incorporating the mask attention loss significantly improves both subject disentanglement and pose accuracy.



Figure 5.7: Ablation results for two-subject images using different text prompts and masks. In particular, we tested MCGM with/without mask attention loss, compared with the Break-a-scene model

These results confirm that the mask cross-attention loss plays a critical role in stabilizing training and enabling reliable pose control, especially in multi-subject scenarios.

5.1.7 User Study

Due to the lack of standardized quantitative metrics for evaluating pose-conditioned single-image generation, we conducted a user study to assess both the perceptual

quality of the generated images and the accuracy of pose alignment. The user study involved 82 voluntary participants recruited through an online Google Forms questionnaire. The survey was distributed through academic and professional networks to individuals from diverse backgrounds, including undergraduate and graduate students, university professors, researchers, and professionals from different disciplines, representing a broad range of ages and levels of expertise. The objective was to obtain feedback from both technical and non-technical users, thereby providing a more comprehensive assessment of the generated images. To ensure a fair comparison between the evaluated methods, all participants evaluated the same set of generated images, randomly selected from the complete set of experimental results, using an identical questionnaire and evaluation criteria. Their responses were automatically collected through Google Forms, exported for analysis, and used to compute the quantitative results presented in this chapter. The questionnaire consisted of two types of questions: (i) multiple-choice comparisons between MCGM and Break-a-scene, and (ii) Likert-scale ratings assessing the alignment between the generated images, the textual prompts, and the target pose masks. In total, 1,886 individual responses were collected and analyzed.

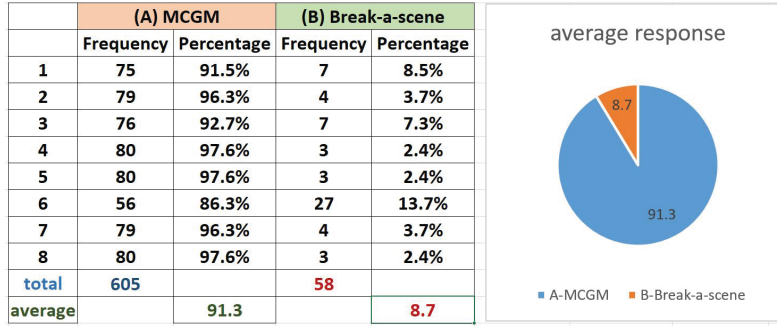


Figure 5.8: Survey results when asking which image, one generated with MCGM and the other with Break-a-scene, matches more with the pose indicated by the mask.

As shown in Figure 5.8, 91.3% of participants preferred images generated by MCGM when asked to identify the output that best matched the target pose. Given the mask and two synthesized images (MCGM v.s. Break-a-scene), the evaluators

were asked to select one image: "Please choose the image that is closest to the mask pose".

Furthermore, as shown in Figure 5.9, in the second part of the survey, The Likert scale (scale from 1 to 5) questions are employed to rate Text-Mask alignment: "Select the degree to which the photo result matches the input text and mask." Fifteen examples were provided, 11 for single-subject images and 4 for two-subject images. Given the original input image, the text prompt, the mask used to generate the output, and the generated images, the evaluators were asked to choose the degree of matching between the output and the text and the mask pose used. the average rating on the Likert scale is **3.8** for both one-concept and multi-concept images.

These findings provide perceptual validation that MCGM improves pose controllability without compromising image quality.

5.2 Experimental Results: MCGM-Styler

This section presents the experimental evaluation of *MCGM-Styler*, the second main contribution of this thesis. While the previous chapter sections demonstrated that MCGM enables precise pose control through mask conditioning, MCGM-Styler extends this framework by introducing an additional level of controllability: **artistic style transfer**.

The objective of MCGM-Styler is therefore not to re-establish pose conditioning only, but rather to show that the model can simultaneously preserve (i) the identity of learned concepts, (ii) control the target pose of the concepts, and (iii) the fidelity of one or multiple target artistic styles.

5.2.1 Training Setup and Implementation Details

As described in Section 3.3, the training procedure of MCGM-Styler is composed of two sequential stages. Both stages follow the same two-phase optimization strategy introduced in Break-a-scene and inherited in MCGM.

	1	2	3	4	5
Q.no	Frequency	Frequency	Frequency	Frequency	Frequency
1	4	4	5	22	40
2	4	6	10	17	37
3	3	7	18	22	25
4	2	9	11	24	29
5	11	10	15	17	21
6	3	9	12	20	31
7	8	7	8	27	25
8	10	9	8	18	30
9	2	8	6	24	35
10	9	13	10	16	27
11	12	10	13	18	22
one-concept	68	92	116	225	322
12	2	7	17	25	24
13	5	7	13	15	35
14	5	15	13	19	23
15	5	8	8	22	31
two-concepts	17	37	51	81	113
total	85	129	167	306	435
T-resp 1con	823	w-sum 1con	3110	Avg-rat 1con	3.8
T-resp 2con	299	w-sum 2con	1133	Avg-rat 2con	3.8

Figure 5.9: Survey results when asked to determine, with a scale from 1 to 5, how much a generated sample with our method matches the pose of a target mask. *T-resp 1con* and *T-resp 2con* are the total responses for single and multiple concept images, *w-sum 1con* and *w-sum 2con* are the weighted sum for single and multiple concept images and *Avg-rat 1con* and *Avg-rat 2con* are the average ratings for single and multiple concept images.

Stage 1: Concept and Pose Learning. In the first stage, the model learns the subject concept(s) from a single reference image while grounding their spatial configuration through mask conditioning.

Stage 2: Style Injection. In the second stage, the model is further trained to capture stylistic characteristics from one or more style reference images, enabling controllable stylization while preserving the learned concept identity.

Each stage is trained with two optimization phases:

- **Embedding optimization:** text embeddings are optimized with a learning rate

of 5×10^{-4} .

- **Model fine-tuning:** UNet and text encoder weights are fine-tuned with a lower learning rate of 2×10^{-6} .

The Adam optimizer is used throughout, with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay 1×10^{-8} . Each phase runs for 400 optimization steps, resulting in a total of 800 steps per training stage.

Hyperparameter Selection. The learning rates (5×10^{-4} for embedding optimization and 2×10^{-6} for fine-tuning) were chosen after a grid search over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ for the first phase and $\{1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}\}$ for the second phase. The selected rates provided the best trade-off between convergence speed and preservation of pretrained weights.

We also evaluated different timestep budgets $\{400, 600, 800, 1000\}$, and found that 800 steps yield optimal style fidelity without overfitting.

Computational Requirements. The model requires approximately 8GB of GPU memory for inference at 512×512 resolution. The complete training procedure (concept learning + style learning) takes roughly 32 minutes on a single NVIDIA A100 GPU, making the approach efficient even in low-data personalization settings.

5.2.2 Qualitative Results

Single-Concept Generation Across Diverse Styles

We first evaluate the ability of MCGM-Styler to extract a single visual concept from a reference image and render it in multiple artistic styles. The objective is to verify that style conditioning successfully modifies the appearance of the generated output while preserving the identity of the learned subject.

Figure 5.10 illustrates several examples where a single concept is generated under five distinct artistic styles. The results confirm that MCGM-Styler effectively captures stylistic features such as brush texture, color palette, and painting structure, while maintaining subject consistency.

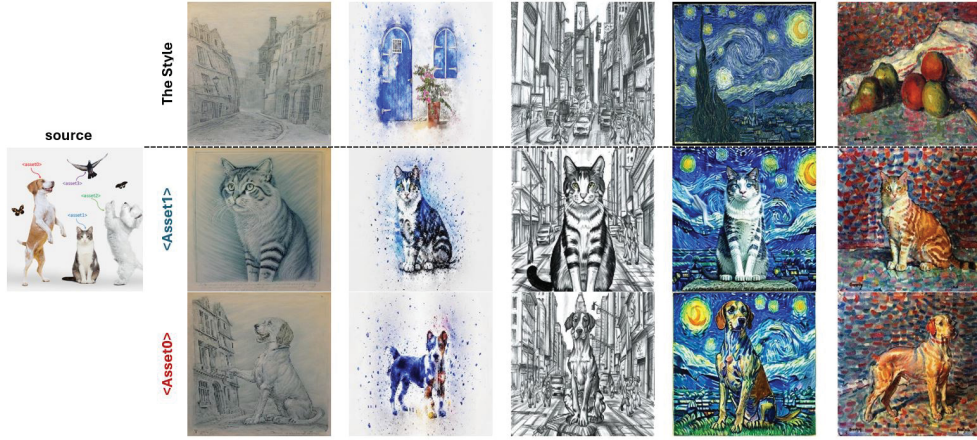


Figure 5.10: One Concept, One Style: Images generated from a single input image using the prompt "a painting of $\langle asset0 \rangle$ / $\langle asset1 \rangle$ in the style of $\langle s0 \rangle$." ,where $\langle asset0 \rangle$ refers to the left dog, $\langle asset1 \rangle$ represents the cat, and $\langle s0 \rangle$ denotes the target style. The figure highlights the effectiveness of style conditioning for both learned concepts.

Joint Pose and Style Conditioning

The main strength of MCGM-Styler lies in its ability to combine multiple conditioning signals: textual prompts, pose masks, and artistic styles. This enables users to generate subjects not only in a desired appearance but also in a spatial pose configuration.

Figure 5.11 presents two cases:

- Style-only conditioning, where the subject adopts a generic pose.
- Style + mask conditioning, where pose guidance directly shapes the generated action.

These results demonstrate that stylistic transfer does not compromise pose controllability, confirming that structural constraints and artistic rendering remain complementary within the proposed framework.

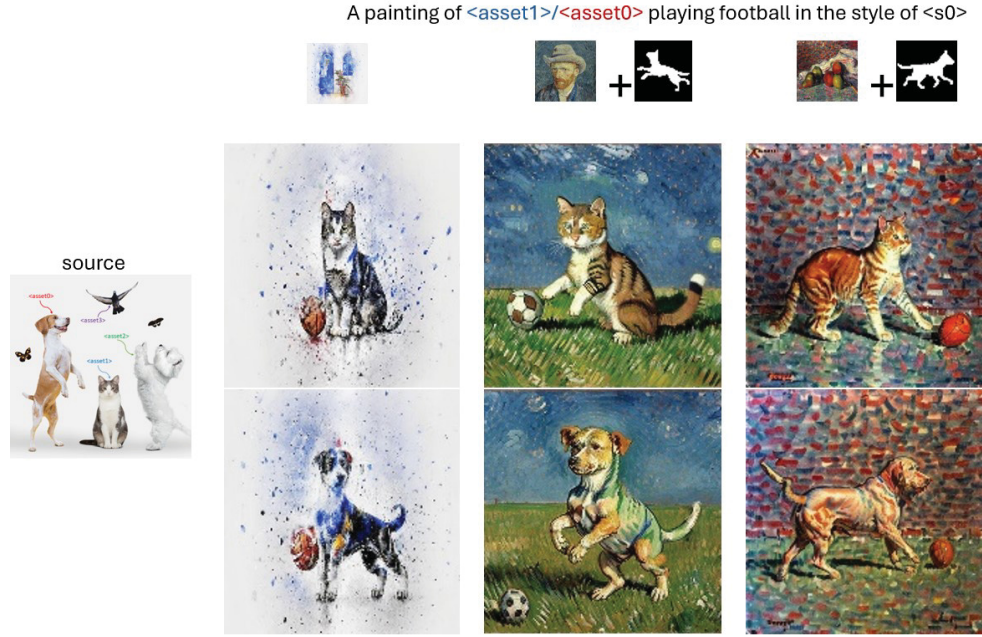


Figure 5.11: One concept with varying masks and styles: Generated outputs are based on the same textual prompt, but guided by different styles and mask shapes. The first column uses only style guidance, while the second and third columns incorporate both style and mask guidance. The first row corresponds to `<asset1>`, representing the cat concept, and the second row corresponds to `<asset0>`, representing the left dog concept.

Multi-Concept Stylized Scene Generation

Beyond single-subject synthesis, MCGM-Styler can generate coherent scenes involving multiple learned concepts, while still applying style and pose constraints.

Figure 5.12 illustrates several scenarios:

- A single mask guiding one concept.
- A single mask shared across multiple concepts.
- Multiple masks guiding multiple actions simultaneously.

These examples highlight the flexibility of the proposed multi-conditional framework for complex controllable scene generation.



Figure 5.12: Multi-Concepts-Multi-Masks: (A) illustrates the case of using a single mask to guide only one of the concepts pose, specifically guiding the sitting pose for the cat. (B) demonstrates the case of using a single mask to guide actions for two concepts. In this example, the mask guides the sitting position for both the cat and the dog. (C) showcases the case of generating three concepts in one scene using one mask to guide their pose. Finally, (D) showcases the case of using multiple masks to guide multiple actions for multiple concepts, with one mask guiding a sitting dog and the other guiding a sleeping cat.

Mixed-Style Generation

An additional capability of MCGM-Styler is the ability to learn multiple artistic styles and blend them during inference. This enables the generation of novel hybrid artistic outputs that combine characteristics from more than one target style.

Figure 5.13 shows results obtained after training with two styles. The model can generate outputs using either style independently or a mixture of both, depending on the style tokens used at inference. For instance, the prompt "a photo of $\langle \text{asset0} \rangle$ in style of $\langle s0 \rangle$ ", this case will generate one concept in one style, but when we use the prompt "photo of $\langle \text{asset0} \rangle$ / $\langle \text{asset1} \rangle$ in the style of $\langle s0 \rangle$ and $\langle s1 \rangle$ ", that

means that the concepts will generated with mixed styles style1 and style2 together in singel scene.

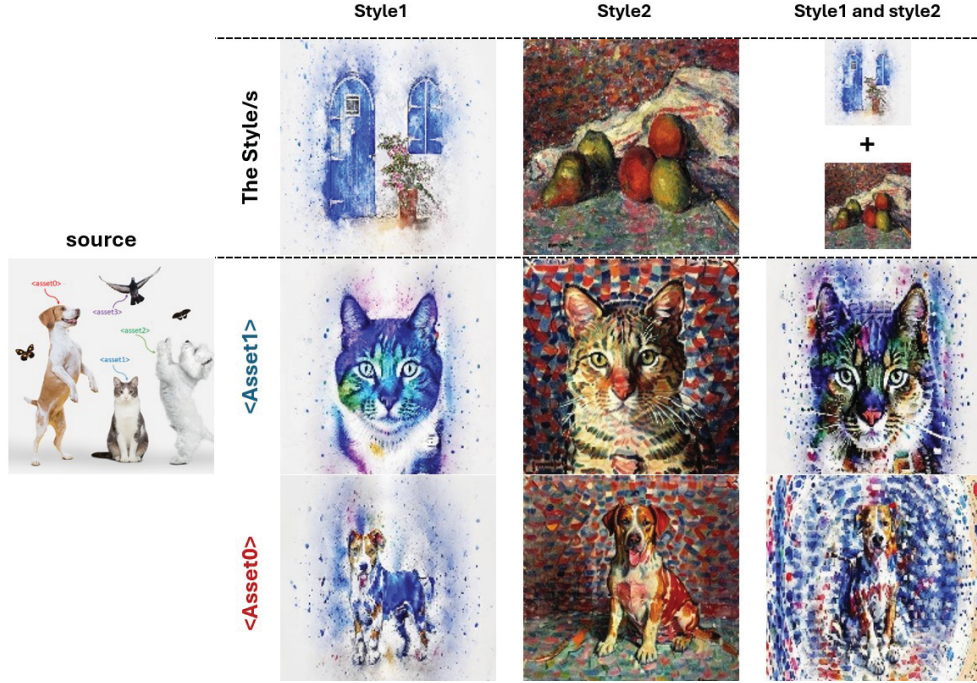


Figure 5.13: Results with multiple styles. The first two columns apply individual styles, while the third column demonstrates mixed-style generation.

Further examples with different style pairs are shown in Figure 5.14, illustrating that each combination produces a unique blended artistic appearance.

5.2.3 Comparison with State-of-the-Art Methods

To evaluate the capability of our model to transfer a style to a particular object, we performed experiments to test our model’s style transfer performance and compared it in Figure 5.15 with the DreamStyler model [77] in different scenarios. Firstly, in (B), DreamStyler was trained using MCGM [73] to learn the concept to which the style needs to be applied. In this case, the results were superior to the original



Figure 5.14: Mixed-style synthesis with different trained style pairs, producing novel hybrid artistic renderings.

DreamStyler model, which utilized DreamBooth as a pre-trained model and required multiple training stages for optimal results. This configuration is shown in column (C). Unlike DreamStyler, which often requires multi-stage training to achieve strong results, MCGM-Styler achieves superior stylization quality within a single-stage style adaptation process, while better preserving concept identity. To ensure a fair comparison, we also evaluated our single-stage model against DreamStyler trained in a single stage. The results presented in column (A) show that our model outperforms DreamStyler in all scenarios.

5.2.4 Ablation Studies

Decoupling Style and Content

A key challenge in style transfer is avoiding the leakage of semantic content from the style reference image into the generated scene. To address this, MCGM-Styler introduces a lightweight human description prompt that explicitly separates non-style attributes from stylistic cues.

Figure 5.16 shows that without this description, the model tends to copy objects present in the style image. and tends to copy objects present in the style image.

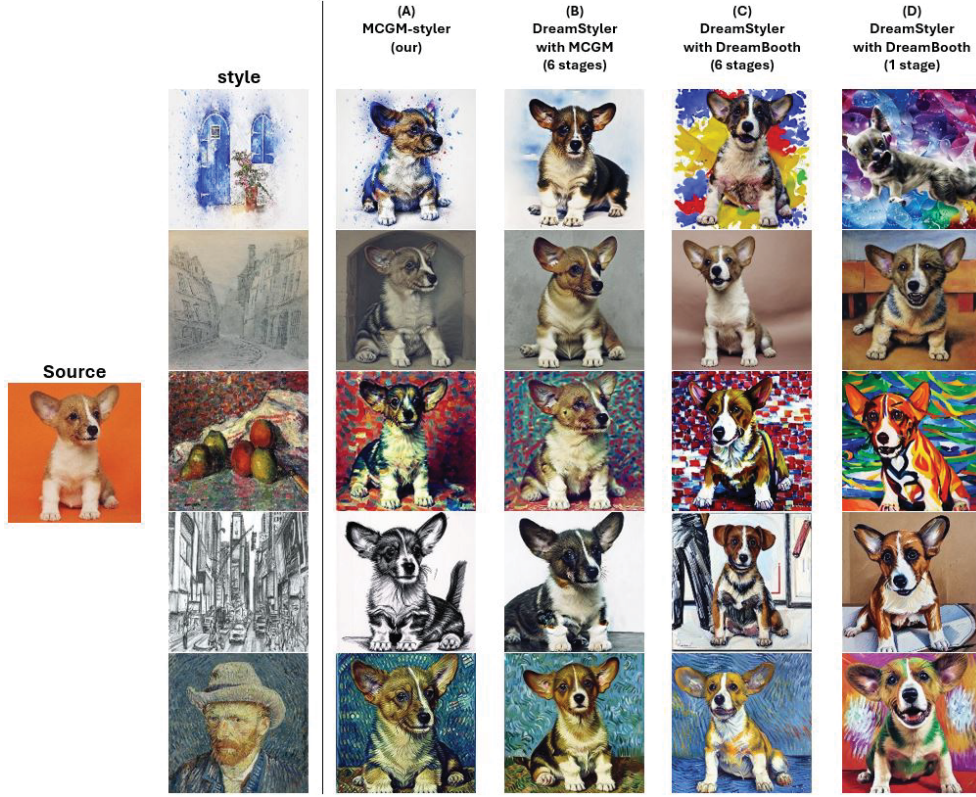


Figure 5.15: Comparison between our model and different combination of DreamStyler and DreamBooth. The same sentence was used in all the experiments: “*a painting of $\langle asset0 \rangle$ in the style of $\langle s0 \rangle$.*” Here, $\langle asset0 \rangle$ represents the dog in the source image, while $\langle s0 \rangle$ corresponds to the style image used in each experiment. (A) represents the results employing our full model for both learning the concept and the styles. (B) and (C) both employ DreamStyler for the style and MCGM and DreamBooth for the concept, respectively. In this case six style training stages are used. Finally, in (D) DreamStyler with a single style training stage in combination with DreamBooth is presented.

More in detail, our system integrates a new human description prompt in the training process with the objective of decoupling style and content from the style im-

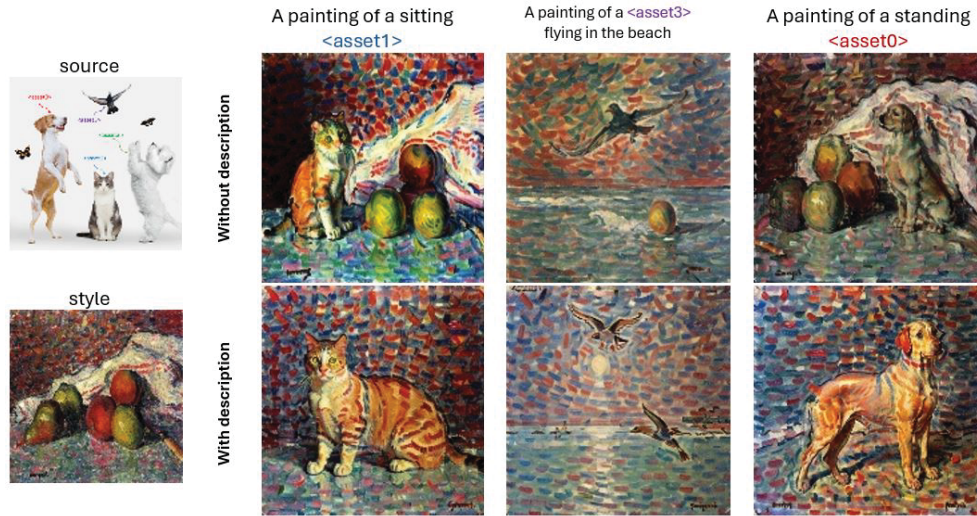


Figure 5.16: Effect of human description prompts in decoupling content and style. Different scenes without and with the human description technique. On the left, the source and style used are displayed. The top row shows the different prompts applied. The middle row presents three different concept scenes generated using various text prompts without the human description technique. The bottom row shows the results obtained by incorporating human description during training.

age. This description consists of a concise list of non-style attributes (5–15 words) describing the objects, composition, and background in the style image. Short but specific descriptions (e.g., ‘pencil, pears, and apples on a cloth’) are sufficient for effective decoupling. Preparing the description requires less than one minute from a non-expert user. An example training prompt is ‘A painting of a pencil, pears, and apples on a cloth, in the style of s*.’ This lightweight annotation process offers a practical balance between user effort and model performance. From the figure it is evident how, without this description, the model struggles to generate the concepts and instead copies elements that are present in the style image. So as shown in Figure 5.16, with the description, the system achieves more effective style-content decoupling, improving generalization.

Contribution of the Mask Encoder

We also evaluate the importance of the mask encoder component. As shown in Figure 5.17, removing the mask encoder significantly reduces pose accuracy, even when the correct style is applied. For example, when using the prompt “a painting of <asset1> playing football in the style of <s0>”, the generated object (cat) fails to align its action (playing football) with the target mask pose,

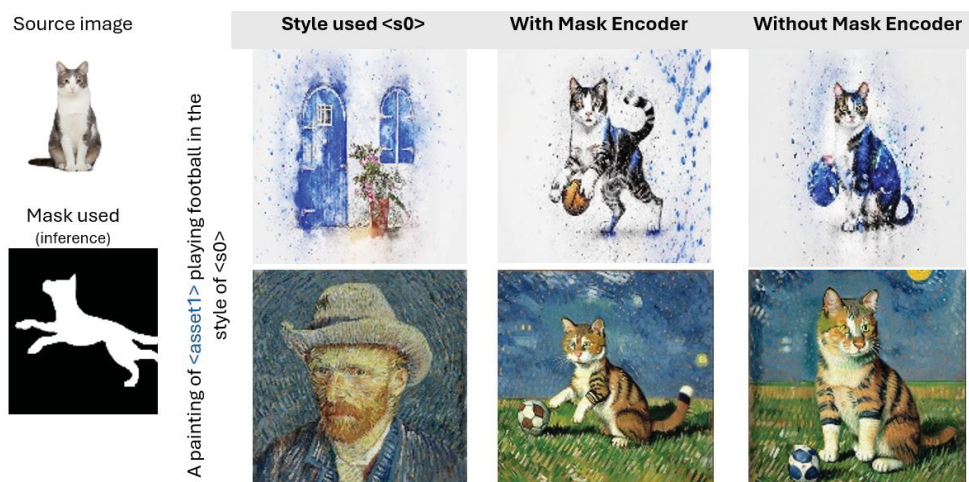


Figure 5.17: Contribution of the mask encoder: two experiments using the same source image, target pose mask, text prompt, and target style. The example to the right, without the mask encoder, fails to apply the target pose correctly.

Effect of Training Timesteps

Finally, we analyze the impact of different timestep budgets during style learning. during the style training phase, we trained the model using different number of timesteps. As shown in Figure 5.18, before reaching 800 timesteps, the model fails to accurately apply the style, resulting in generated images that do not fully align with the reference style. At 800 timesteps, the generated images achieve an optimal painting style that closely matches the reference. However, beyond 800 timesteps, the model begins to

overfit, prioritizing the style over the trained concept. This overfitting is influenced by the initial concept training phase. Therefore, 800 timesteps provide the best balance between style fidelity and concept preservation.

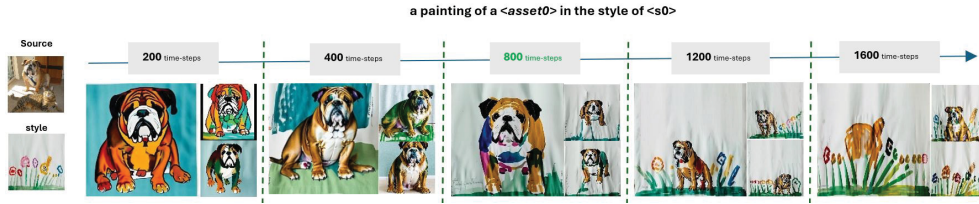


Figure 5.18: Style training with different timesteps. For each experiment, three examples are shown. Optimal results are obtained at 800 steps, while larger values lead to overfitting.

5.2.5 Style-Guided Text-to-Image Generation

MCGM-Styler can also be applied beyond concept personalization, enabling style-guided generation of generic objects from text prompts alone (a bridge, a house, etc.) rather than specific concepts from the source image to demonstrate the effectiveness of our system in injecting the desired style into the generated images. This extends the applicability of the model to broader text-to-image synthesis tasks.

Figure 5.19 presents examples of stylized text-to-image generation and comparisons with DreamStyler configurations, confirming the strong style controllability of our approach.

5.2.6 Quantitative Evaluation

We quantitatively compare MCGM-Styler with DreamStyler under the same training scenarios discussed above. For each method, we generated 125 images (25 samples per style across five styles).

We compute:

- DINO similarity matrix [83]

- CLIP-I score [84]

for each image and calculated the *average similarity* for each set of 25 images per style. Finally, we derived the *overall average similarity score* for each scenario. Table 5.1 shows that MCGM-Styler achieves the highest similarity scores, confirming its superior performance in style transfer quality.

Table 5.1: **Quantitative Evaluation:** Comparison of MCGM-Styler with DreamStyler under different training conditions. **Bold:** best, underline: second best.

Method	DINO $\uparrow$ (Cosine)	DINO $\downarrow$ (Euclidean)	CLIP-I $\uparrow$ (Cosine)
DreamStyler (1 stage)	0.075	38.5	0.591
DreamStyler (6 stage)	0.203	36.1	0.609
DreamStyler (with MCGM)	<u>0.207</u>	<u>35.4</u>	<u>0.643</u>
MCGM-Styler	0.436	29.6	0.708

5.2.7 User Study

To complement quantitative evaluation, we conducted a user study with 25 participants (age 21–38, mixed academic and non-academic backgrounds). The study included 25 questions divided into three parts:

1. Style transfer fidelity and realism (10 questions)
2. Preference comparisons against DreamStyler baselines (10 questions)
3. Consistency of joint pose + style transfer in multi-concept scenes (5 questions).

In Part 1 participants rated the realism and fidelity of stylized outputs given a specific target style. in Part 2, comprising object-level and scene-level evaluations across different styles, in which our method was compared to DreamStyler under various training setups as detailed in the table, and in Part 3, participants assessed the consistency of pose and style across scenes generated from a source image with multiple objects, a target position mask, a target style, and a guiding text prompt.

Participants reported strong satisfaction, with 77.5% of positive responses (scale 4: 33.33%, and scale 5: 44.16%) in Part 1 and 68.8%(scale 4: 36.45%, and scale 5: 32.33%) in Part 3, indicating high satisfaction with both the fidelity of the style and the integrated transfer of the pose-style. Furthermore, our method was preferred over three baseline methods approaches in 65% of object-level comparisons and 61.68% of full-scene evaluations.

These results confirm that MCGM-Styler provides an effective and flexible solution for controllable pose- and style-guided image generation from minimal data.

Chapter Summary

This chapter presented the experimental evaluation of the proposed frameworks, MCGM and its extension MCGM-Styler.

First, the results of MCGM demonstrated that incorporating mask-based conditioning into concept learning enables controllable pose-guided scene generation from a single input image. The qualitative experiments confirmed that the model can generate both single-subject and multi-subject scenes while following user-defined spatial layouts. Furthermore, the ablation study validated the importance of the proposed mask cross-attention loss in improving concept disentanglement and pose accuracy. User study findings also supported the effectiveness of mask guidance, showing strong alignment between generated subjects and target pose conditions.

Second, the evaluation of MCGM-Styler showed that the framework can successfully extend pose control toward style-conditioned generation. The model was able to synthesize paintings of personalized concepts across diverse artistic styles, including mixed-style compositions and multi-concept stylized scenes. Comparisons with state-of-the-art approaches such as DreamStyler highlighted the advantages of the proposed two-stage training strategy, achieving stronger style fidelity while requiring fewer optimization stages. Quantitative metrics and user study results further confirmed the superiority of MCGM-Styler in transferring both pose and style in a unified generation process.

Overall, the experiments demonstrate that the proposed multi-conditional gen-

erative framework provides a flexible and efficient solution for controllable text-to-image synthesis, supporting user-defined subjects, poses, and artistic styles within a single personalized model.

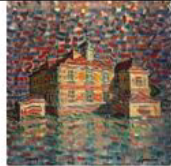
<i>prompt</i>	<i>Style image</i>	DreamStyler 1stages	DreamStyler 6stages	DreamStyler With MCGM	MCGM-Styler (Ours)
"A painting of a house "					
"A painting of a lighthouse on a cliff "					
"A painting of a knight "					
"A painting of a sailboat on the sea "					
"A painting of a bridge "					

Figure 5.19: Style-guided text-to-image generation: Comparison between our model and various configurations of DreamStyler model. The left column shows the input text prompt and target style, the right column displays results from our model, and the center presents DreamStyler outputs under different training setups, including one-step, six-step, and MCGM-pretrained variants.

Chapter 6

Conclusion and Future Work

6.1 Thesis Summary

In recent years, diffusion-based generative models have achieved remarkable success in text-to-image synthesis, enabling the creation of highly realistic and diverse visual content from natural language descriptions. Despite these advances, personalized generation from a single reference image remains a challenging task, particularly when additional controllability is required. Most existing personalization approaches focus primarily on learning the identity of a subject from limited examples. However, they often lack explicit mechanisms to control important structural attributes such as object pose, spatial placement, and artistic style. This limitation reduces the usability of such models in practical applications, where users typically require more flexible and precise control over the generated scenes.

This thesis addressed these challenges by introducing a multi-conditional personalized diffusion framework that enables controllable image generation from a single input image. In particular, we proposed two complementary models:

- **MCGM**, which extends concept personalization by introducing explicit pose control through mask conditioning.
- **MCGM-Styler**, which further extends the framework to support style-conditioned generation while preserving pose controllability.

Together, these contributions provide a unified approach for generating scenes that are simultaneously controllable in terms of subject identity, pose structure, and artistic style.

6.2 Main Contributions

The primary contributions of this thesis can be summarized as follows:

- We proposed the **Mask Conditional Generative Model (MCGM)**, a multi-conditional personalized diffusion framework that extends single-image concept learning with explicit spatial controllability. By introducing a dedicated mask encoder and injecting mask embeddings into the cross-attention layers, the model enables pose-guided generation while preserving subject identity.
- We introduced a **mask cross-attention alignment loss** that explicitly regularizes the association between textual handles and their corresponding spatial mask regions. This loss improves concept disentanglement and enhances pose consistency in multi-subject generation scenarios.
- We proposed **MCGM-Styler**, an extension of MCGM that incorporates artistic style conditioning through a two-stage training strategy, enabling users to generate personalized subjects in arbitrary styles while maintaining pose control.
- We introduced a lightweight **human description-based decoupling** technique to reduce content-style entanglement during style learning, improving generalization across complex style references.
- We validated both frameworks through extensive qualitative results, ablation studies, quantitative evaluations, and user studies, demonstrating consistent improvements over state-of-the-art baselines such as Break-a-scene and Dream-Styler.

6.3 Discussion of Findings

The experimental results presented in Chapter 5 confirmed the effectiveness of the proposed multi-conditional diffusion framework.

For **MCGM**, the integration of pose masks significantly enhanced the level of controllability compared to standard text-only personalization methods. The model was able to generate subjects performing diverse actions and adapting to various spatial configurations, even when the target masks were only approximate representations of the desired pose. Furthermore, multi-subject experiments demonstrated that the proposed attention-based losses play a crucial role in preventing concept mixing and ensuring that each learned handle focuses on the correct image region.

For **MCGM-Styler**, the two-stage training approach successfully enabled the transfer of artistic styles while preserving both identity and pose structure. Unlike existing methods that require multi-stage optimization to achieve high style fidelity, MCGM-Styler achieved competitive and often superior results with only a single style training stage. Additionally, the model supports multi-style blending, offering new creative possibilities beyond traditional style transfer.

Quantitative evaluations using DINO similarity and CLIP-based metrics confirmed that MCGM-Styler produces outputs that are both stylistically consistent and semantically aligned. User studies further reinforced these findings, showing a clear preference for the proposed framework over competing baselines in terms of pose adherence, style quality, and overall realism.

Overall, the results demonstrate that adding explicit spatial and stylistic conditioning to personalized diffusion models is a promising direction for controllable image synthesis.

6.4 Limitations

Despite the strong performance of the proposed models, several limitations remain.

In the case of **MCGM**, the model sometimes struggles to correctly interpret the orientation of target masks. While the generated subject typically matches the desired

action and position, the viewpoint or directional alignment may differ from the mask specification. for instance As shown in Fig. 6.1, when using a mask condition for the subject with a side view, the model generates images for the subjects using the correct mask action position but with a front view.

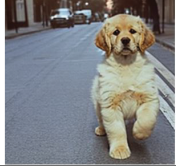
Text prompt	A photo of <code><asset0></code> running	A photo of <code><asset0></code> playing football in the beach	A photo of <code><asset0></code> running in the street
mask orientation			
output			

Figure 6.1: Limitation in following the correct orientation of the mask. In this case, the direction of the output images (front view) is different from the mask direction (side view) even if the overall pose is correct.

Furthermore, when more than two distinct pose masks are provided simultaneously, the model may fail to accurately satisfy all constraints, leading to partial misalignment in complex multi-subject scenes (see Fig 6.2. The text prompt suggests three actions for three subjects: sleeping `asset1`, flying `asset3`, and running `asset0` . knowing that: `asset0` represents the dog, `asset1` the cat, and `asset3` the bird). The mask conditions consist of three masks, one mask for each action; the results show that the positions of the mask conditions did not apply to the generated subjects.

For **MCGM-Style**, although the model effectively combines pose and style control, disentangling content from highly abstract or compositionally complex style references remains challenging. In particular, as shown in fig.6.3 style images containing multiple interacting objects or strong structural patterns may introduce unwanted content leakage into the generated output. Additionally, both frameworks rely on a

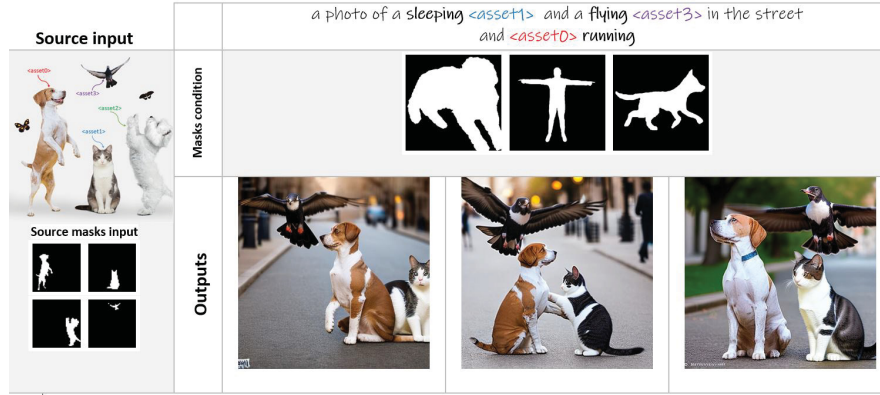


Figure 6.2: Limitation when using more than two mask conditions. Indeed, the model struggles to apply the mask conditions for all subjects in the output scenes.

fine-tuning procedure, which may still present scalability challenges when extending to very large numbers of concepts, longer prompts, or complex structural synthesis.

6.5 Future Work

The limitations discussed above suggest several directions for future research.

First, future extensions may focus on improving pose accuracy in complex scenes by supporting a larger number of pose masks. This would enable reliable generation in scenarios involving multiple subjects performing independent actions.

Second, investigate strategies to enhance the model’s capacity to interpret longer prompts, thereby enabling the generation of more detailed and coherent scenes

Third, automated style captioning and context extraction could significantly improve the usability of MCGM-Styler by removing the need for manual human descriptions. Integrating vision language models to generate contextual prompts automatically represents a promising direction.

Fourth, future work could explore more advanced disentanglement strategies to decouple the content and style from the style image.

Finally, extending the framework beyond static image synthesis toward video

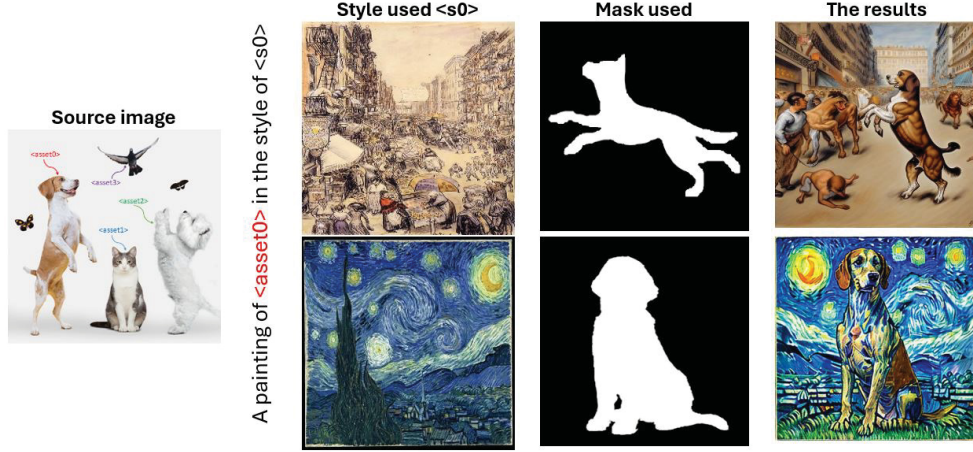


Figure 6.3: Decoupling limitation. Two examples of failure cases caused by complex multi-subject interactions in the style image. The left column shows the source image and prompt, the middle columns show the mask and style, and the right column shows the generated results, where the content is entangled with the source style image.

generation, 3D-aware diffusion, or interactive editing applications could further broaden the impact of the proposed approach.

6.6 Final Remarks

This thesis introduced a novel multi-conditional diffusion framework for personalized image synthesis from a single reference image. By combining concept learning, pose control through mask conditioning, and style injection through two-stage training, the proposed models provide a flexible and effective solution for controllable scene generation. The contributions of MCGM and MCGM-Styler highlight the potential of diffusion models not only as powerful generative tools but also as controllable and user-adaptive frameworks for creative applications. We believe that this work represents a meaningful step toward more expressive, efficient, and controllable generative modeling systems.

Bibliography

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [6] Alex Nichol et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [7] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.
- [8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, et al. Hierarchical text-conditional image generation with clip latents. In *ICML*, 2022.

- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [10] Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [11] Han Zhang et al. Text-to-image generation: A survey. *Foundations and Trends in Computer Graphics and Vision*, 2023.
- [12] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.
- [13] Clément Bonnet, Ariel N. Lee, Franck Wertel, Antoine Tamano, Tanguy Cizain, and Pablo Ducru. From text to pose to image: Improving diffusion model control and quality. *arXiv preprint arXiv:2411.12872*, 2024.
- [14] Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. *arXiv preprint arXiv:2311.05463*, 2023.
- [15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.
- [16] Hsu-Yung Cheng, Chia-Cheng Su, Chi-Lun Jiang, and Chih-Chang Yu. Pose transfer with multi-scale features combined with latent diffusion model and controlnet. *Electronics*, 14(6):1179, 2025.
- [17] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*, 2023.
- [18] Rinon Gal et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

- [19] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. *arXiv preprint arXiv:2405.15287*, 2024.
- [20] Omri Avrahami, Ohad Fried, et al. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2210.03103*, 2022.
- [21] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.
- [22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.
- [23] Ian Goodfellow et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [26] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014. URL: <https://arxiv.org/abs/1401.4082>.
- [27] Lucas Theis, Aaron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2016. URL: <https://arxiv.org/abs/1511.01844>.
- [28] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric.

- In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016. URL: <https://arxiv.org/abs/1512.09300>.
- [29] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.
- [30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [31] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.
- [32] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [34] Aaron van den Oord et al. Pixel recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2016.
- [35] Niki Parmar et al. Image transformer. In *International Conference on Machine Learning (ICML)*, 2018.
- [36] Jascha Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.
- [37] Jiaming Song et al. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

- [38] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [39] Yang Song et al. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.
- [40] Lingchen Liu et al. Rectified flow: A unified framework for flow and diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.
- [41] Ziyu Wang et al. Video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [42] Tao Xu et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [43] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [44] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [45] Christoph Schuhmann et al. Laion-5b: Large-scale open dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks*, 2022.
- [46] Lvmin Zhang et al. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.
- [47] Menghan Mou et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

- [48] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [49] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [50] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.
- [51] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.
- [52] Rinon Gal et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [53] Edward J Hu et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- [54] Chitwan Kumari et al. Custom diffusion: Customized text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [55] David Hinton et al. Styledrop: Text-to-image generation in any style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [56] Xin Wang et al. Artadapter: Efficient style transfer via adapter-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

- [57] Hyun Lee et al. Controlstyle: Style-conditioned diffusion models for artistic image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [58] Tian Zhang et al. Style-a-video: Consistent stylization of video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [59] Omri Avrahami et al. Break-a-scene: Extracting multiple concepts from a single image. In *ACM SIGGRAPH*, 2022.
- [60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [61] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [62] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.
- [63] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.
- [64] L. Zhang et al. Fine-tuning conditional diffusion models for text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- [65] Benjamin Ilse and Frederick Blackwood. Comparative analysis of finetuning strategies and automated evaluation metrics for large language models in customer service chatbots. 2024.

- [66] J. Garcia et al. Hyperparameter optimization in fine-tuned diffusion models for textual image generation. *Neural Information Processing Systems (NeurIPS)*, 2022.
- [67] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.
- [68] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [69] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, Wenbo Li, Renjing Pei, Fan Li, and Wangmeng Zuo. Mc²: Multi-concept guidance for customized multi-concept generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2802–2812, 2025.
- [70] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7323–7334, 2023.
- [71] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15943–15953, 2023.
- [72] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023.
- [73] Rami Skaik, Leonardo Rossi, Tomaso Fontanini, and Andrea Prati. Mcgm: Mask conditional text-to-image generative model. *arXiv preprint arXiv:2410.00483*, 2024.

- [74] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [75] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.
- [76] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, David Fleet, and Mohammad Norouzi. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.
- [77] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681, 2024.
- [78] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.
- [79] Chang Liu, Viraj Shah, Aiyu Cui, and Svetlana Lazebnik. Unziplora: Separating content and style from a single image. *arXiv preprint arXiv:2412.04465*, 2024.
- [80] Filippo Botti, Alex Ergasti, Leonardo Rossi, Tomaso Fontanini, Claudio Ferrari, Massimo Bertozzi, and Andrea Prati. Mamba-st: State space model for efficient style transfer. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 7786–7795, February 2025.
- [81] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

-
- [82] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.
 - [83] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.
 - [84] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18062–18071, 2022.

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Andrea Prati, for his continuous guidance, support, and encouragement throughout my PhD journey. His expertise, insightful comments, and constructive feedback have been invaluable in shaping this research.

I would also like to thank my co-authors, Dr. Leonardo Rossi and Dr. Tomaso Fontanini, for the fruitful collaboration, scientific discussions, and their valuable contributions to the research presented in this thesis.

My sincere appreciation goes to the Department of Engineering and Architecture at the University of Parma for providing a stimulating academic environment that enabled me to conduct this research. I am also deeply grateful to Qatar University for the PhD scholarship that supported my doctoral studies.

I would like to extend my gratitude to the Islamic University of Gaza (IUG), my home university, and especially to my colleagues in the College of Information Technology for their support, encouragement, and inspiration throughout my academic journey.

On a personal note, I dedicate special thanks to my family. I am deeply grateful to my mother for her endless love, sacrifices, and encouragement throughout my life. I also remember my father, who is no longer with us, but whose values, support, and belief in education continue to inspire me.

My deepest gratitude goes to my wife, Yasmeen, for her unwavering support, patience, and understanding throughout the years of my PhD journey. I am also grateful to my beloved children, Yara, Sami, Sara, and Jouri, whose love and joy have been a

constant source of motivation and strength.

I also wish to acknowledge my brothers and sisters, and their children, for their constant support, encouragement, and inspiration throughout my life. Their care and presence have enriched my journey in countless ways.

Finally, I would like to acknowledge my homeland, Gaza, whose resilience and spirit have always inspired me to pursue knowledge and contribute to the advancement of science and education.