Overview - Generative and Multimodal Learning for Vision and Language

The fields of multimodal learning and generative models have undergone significant advancements in recent years, driven by the increasing availability of large datasets and the development of innovative architectures capable of handling complex, real-world tasks. As data from various modalities become more accessible, the need for models that can effectively integrate and generate such data has become increasingly important. This dissertation focuses on advancing these areas through a series of contributions that explore the interplay between different modalities and the enhancement of generative modeling, particularly for multimodal data.

1 Multimodal Learning

Multimodal learning involves the integration and processing of information from multiple sensory modalities, such as visual, auditory, and textual data. By combining diverse sources of information, multimodal learning systems aim to achieve a more holistic understanding of the world, enhancing performance in areas like scene understanding, object recognition, natural language processing, and content generation. The ability to leverage multiple modalities allows these systems to compensate for the limitations of individual modalities, leading to more robust and accurate models. The field has gained significant attention due to the increased abundance of multimodal data and the demand for systems that can effectively interpret and utilize this wealth of information. Challenges in multimodal learning include effective data fusion, handling modality-specific noise, dealing with heterogeneous data distributions, and overcoming data scarcity in less common modalities.

In this dissertation, we contribute to the field of multimodal learning by developing new methods for cross-modal integration, self-supervised learning, and joint training of large autoregressive models. Our work aims to advance the state-of-the-art by addressing key challenges in multimodal fusion and by proposing efficient strategies for leveraging auxiliary modalities to enhance learning outcomes. These contributions are detailed in Chapters 2, 4, and 5.

2 Generative Models

Generative models have become a cornerstone of modern machine learning, enabling the sampling on new data from a learned data distribution. Among the various generative modeling techniques, diffusion models

have recently emerged as a powerful approach for generating high-quality data across multiple domains, particularly in image synthesis.

Denoising Diffusion Models (DDMs) have shown remarkable success in producing realistic images through an iterative denoising process. However, despite their high-quality outputs, diffusion models often face challenges related to inference efficiency due to the need for numerous denoising steps. This has led to a surge of interest in developing methods to accelerate diffusion models without compromising the quality of the generated samples. Additionally, personalized content generation remains a challenging aspect, requiring models that can adapt to specific user-defined concepts efficiently.

In this dissertation, we address these challenges by introducing novel methods to improve the efficiency and personalization capabilities of generative models. We develop techniques for fast sampling in diffusion models and propose a finetuning-free approach for personalized text-to-image generation. Through these contributions, detailed in Chapters 3 and 6, we aim to enhance the practicality and versatility of generative models in various applications.