Reg. No: 2025/57/N/ST6/05041; Principal Investigator: mgr Artur Kasymov

Democratizing Diffusion Models: Guidance-Driven Optimization of Existing Models for Enhanced Quality and Diversity with Compliance-Preserving Unlearning

Diffusion models have become foundational to generative AI, enabling high-quality synthesis of images, audio, and text. However, their widespread adoption is hindered by two critical issues: (1) existing models often produce outputs with limited diversity or suboptimal quality, restricting their utility for creative or specialized tasks, and (2) concerns about data privacy, intellectual property, and regulatory compliance (e.g., under the EU AI Act) limit their deployment in safety-critical or public domains. These challenges disproportionately affect non-expert users, small businesses, and underserved communities, who lack resources to train models from scratch or address compliance risks.

This project advances two complementary subgoals to democratize access to diffusion models while addressing technical and societal challenges. First, we will develop guidance-driven optimization techniques to enhance the quality and diversity of outputs from existing pretrained diffusion models (e.g., Stable Diffusion, FLUX). These methods will enable non-expert users and resource-constrained organizations to refine model behavior post-training—improving resolution, coherence, and creative diversity without full retraining or architectural changes. Second, we will pioneer compliance-preserving unlearning strategies tailored to diffusion models, which are classified as General Purpose AI (GPAI) systems under the EU AI Act. This work will address risks such as embedded biases in open-source models (e.g., stereotypical representations in image generation) and ensure alignment with GPAI obligations, including transparency, risk mitigation, and data privacy. By enabling targeted removal of harmful or regulated content (e.g., copyrighted material, sensitive personal data), our unlearning framework will make diffusion models safer for public deployment while preserving their generative capabilities.

For Guidance-Driven Optimization, we will explore novel diffusion guidance techniques that refine pretrained models (e.g., Stable Diffusion) post-training to enhance output quality and diversity. This includes modifying classifier-free guidance through adaptive scaling strategies and integrating auxiliary tools like LoRA, ControlNet, or IP-Adapter to steer generation toward user-defined goals (e.g., artistic fidelity, domain-specific accuracy). While minimizing retraining requirements to lower computational barriers, our focus is on improving general output quality (e.g., resolution, coherence) and diversity (e.g., multimodal creativity). This will enable non-expert users to customize models for niche applications—such as personalized art tools or low-resource creative workflows—while maintaining efficiency.

For Compliance-Preserving Unlearning, we will develop strategies to enforce policy compliance in diffusion models, which are classified as General Purpose AI (GPAI) under the EU AI Act. This includes mitigating biases (e.g., stereotypical representations) and preventing generation of content violating intellectual property or branding policies (e.g., unauthorized corporate logos, regulated symbols). By targeting specific data patterns linked to non-compliant outputs, our unlearning framework will suppress harmful or policy-violating generations while preserving overall model performance. This work directly addresses gaps in open-source models (OSS) that inadvertently propagate biases or violate GPAI obligations, such as transparency and risk mitigation for high-impact systems.

By bridging the gap between technical innovation and societal needs, this work will:

- Democratize access: enable users to enhance and customize pretrained models locally, reducing reliance on resource-intensive training.
- Foster trust: ensure generative AI systems respect privacy, comply with regulations, and avoid harmful biases or content.
- Boost creativity: provide tools for artists, educators, and small businesses to generate high-quality, diverse outputs tailored to niche needs.

By uniting technical innovation with ethical deployment, this work will position diffusion models as accessible, compliant tools for equitable AI advancement.