DeMeTeR: Interpreting Diffusion Models Through Representations

Diffusion models are the latest revolution of generative modelling in computer vision. However, we still lack an in-depth understanding of their inner workings from both an empirical and theoretical standpoint. Improving this understanding is not only crucial to their continued development. It also allows for enhancing their safety and provides opportunities to develop new methods for explainability of predictive models in computer vision. Considering that, the main goals of the DeMeTeR project are:

1. to broaden the practical and theoretical understanding of diffusion-specific *latent representations* and architecture-specific *internal representations* of diffusion models,
2. to develop novel methods of manipulating these representations that allow for enhancing *safety* and *explainability* of deep learning models.

Questions and hypotheses: We aim to address the following research questions and hypotheses. How to improve the understanding of state-of-the-art diffusion models trained on images through their *latent* and *internal representations*? What are the critical *limitations* of their generative process and how to overcome them? *Latent* and *internal representations* of diffusion models trained on images allow precise *control* of the generative process. How to transfer this understanding to improve the *safety* of diffusion-based *foundation models*? How to exploit these representations for *explaining* visual predictive models? Understanding the representations of diffusion models can improve the *safety* and *explainability* of other deep learning models.

Impact: Developing methods to better understand and interpret diffusion models has a direct impact on their capabilities and potential for improvement. Approaching this problem from the perspective of their *internal* and *latent representations* also greatly benefits other important research domains. With the emergence of diffusion-based *foundation models*, enhancing their *safety* is crucial in ensuring their *safe* and *responsible* use. The development of new methods for explainability of predictive models with the use of diffusion models is increasingly noticeable in the area of explainable artificial intelligence. In both cases, proper understanding and manipulation of the *representations* of diffusion models is key. Progress in this line of research is, therefore, beneficial to both the research community and its further practical applications.

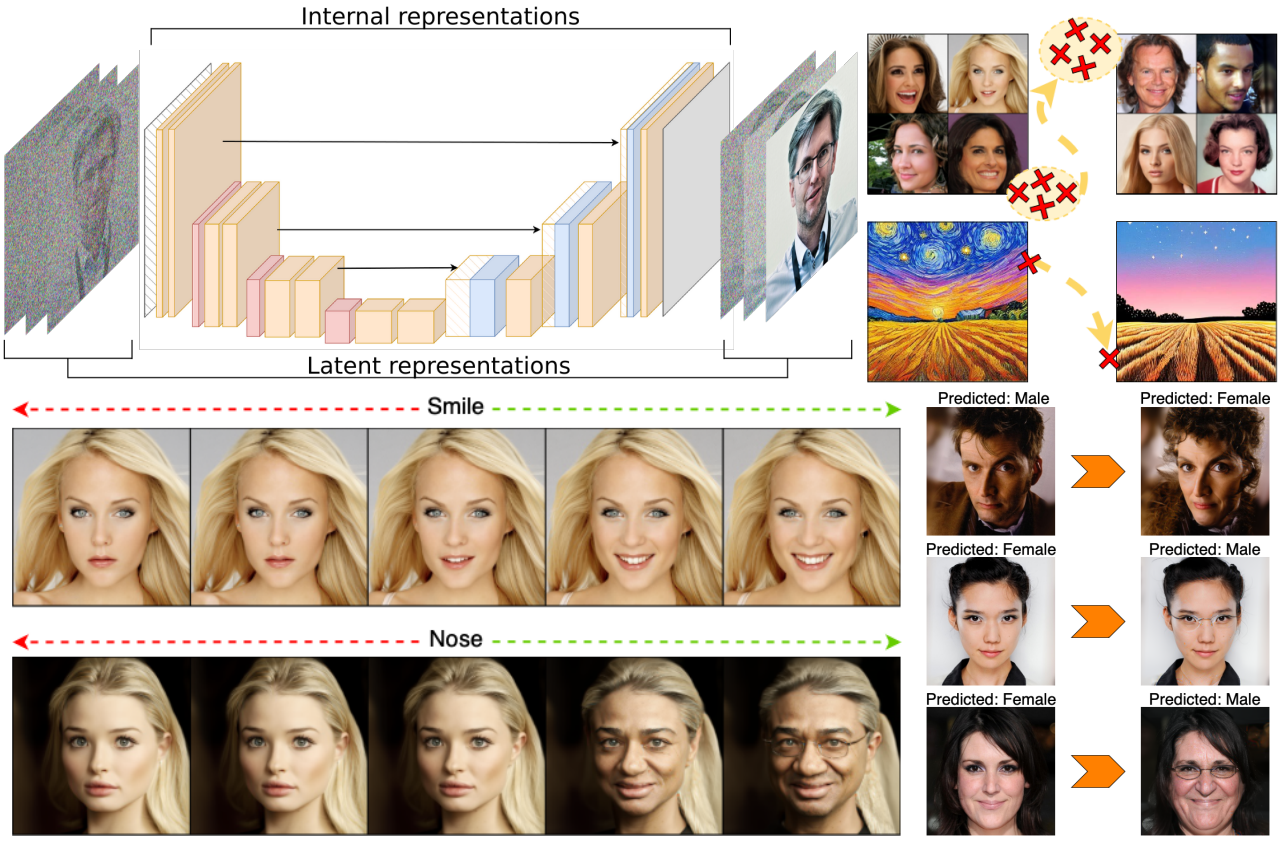

Figure 1: Interpreting diffusion models can be done from the perspective of diffusion-specific *latent representations* and architecture-specific *internal representations* **(top left)**. A better understanding of these representations allows for eliminating biases inherent to the training data, erasing undesirable concepts from the generated content **(top right)**, improving the generative capabilities such as attribute disentanglement **(bottom left)** and developing new methods for explainability of other predictive models in computer vision, e.g. generation of counterfactual explanations, a result of our preliminary research **(bottom right)**. Source: see full project description.