Multimodal Machine Learning Lab

Winter Semester 2024/2025

Niklas Deckers and Martin Potthast

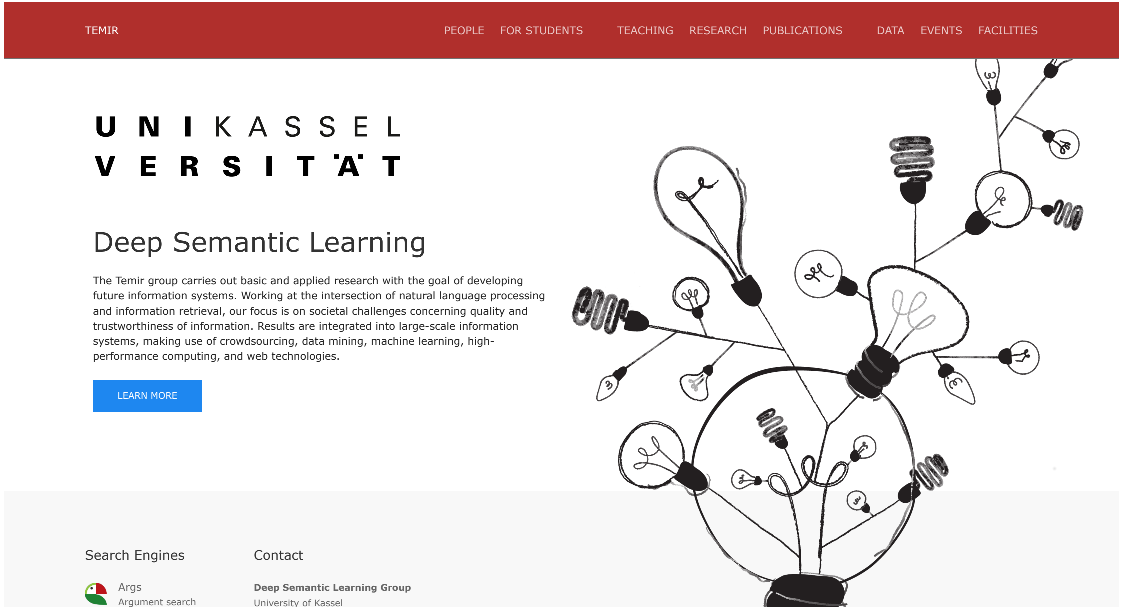
Deep Semantic Learning
University of Kassel and hessian.AI

About us

Martin Potthast

Niklas Deckers

You can say “you” to us



Deep Semantic Learning

The Temir group carries out basic and applied research with the goal of developing future information systems. Working at the intersection of natural language processing and information retrieval, our focus is on societal challenges concerning quality and trustworthiness of information. Results are integrated into large-scale information systems, making use of crowdsourcing, data mining, machine learning, high-performance computing, and web technologies.

Search Engines

Args
Argument search

Contact

Deep Semantic Learning Group
University of Kassel

[kassel.webis.de]

Agenda

❑ Motivation

❑ Getting to Know Each Other

❑ Lab Organization

❑ Foundations

❑ Research Objectives for This Lab

❑ Initial Task

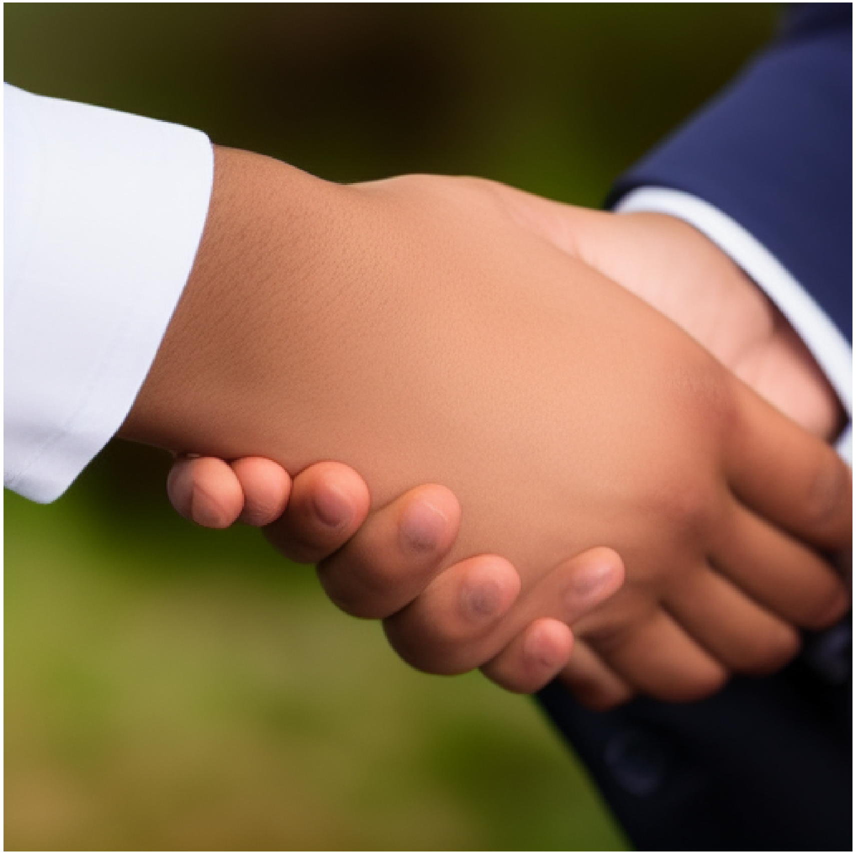
❑ Cluster Onboarding

Introduction

- Generative models are widely used
- LLMs like ChatGPT can be used to generate texts on a high level
- Models like GANs introduced high-quality image generation
- With *multimodal* text-to-image generation models like Stable Diffusion, controlling the images using prompts is possible

Introduction

- Some quality issues in the generated images
- Main focus of current research seems to be improving this quality



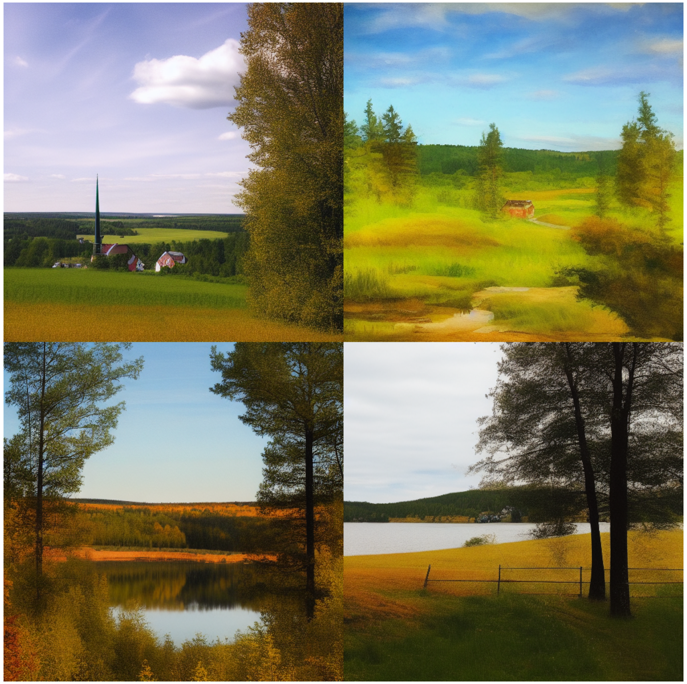
Introduction

- Some improvements with model increments

- Prompt engineering will be the next most important bottleneck

Interactive Motivation

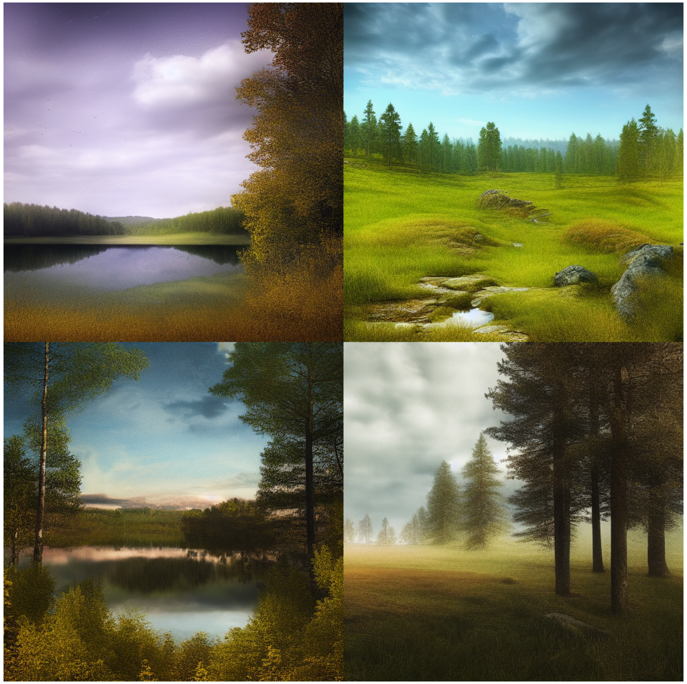
How can these images be improved?



swedish landscape

Interactive Motivation

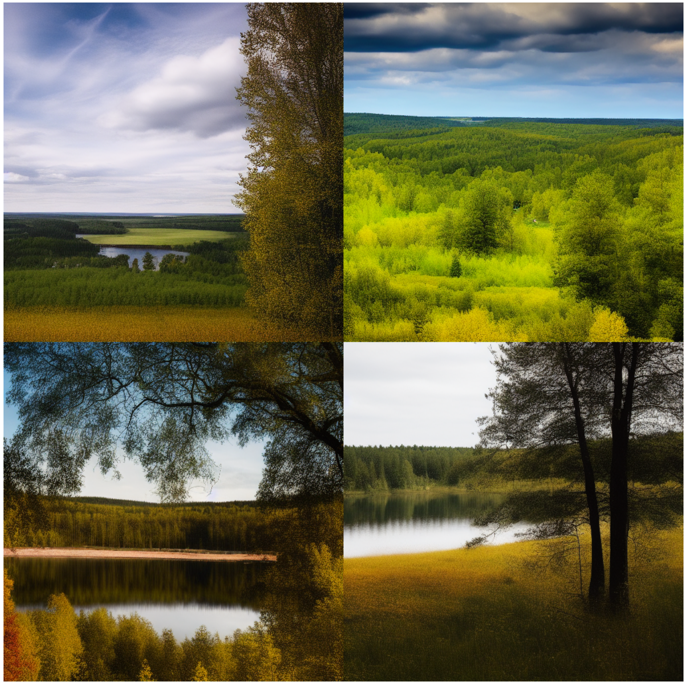
How can these images be improved?



`swedish landscape photorealistic`

Interactive Motivation

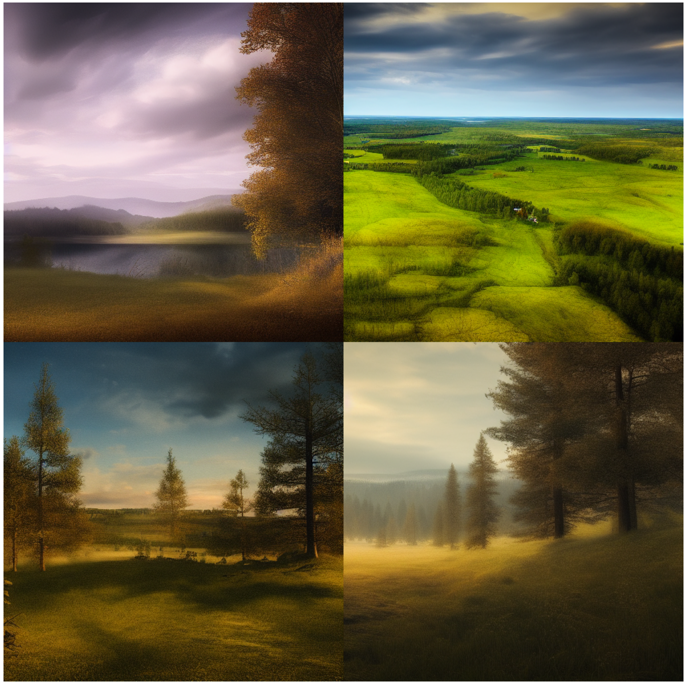
How can these images be improved?



`swedish landscape made by a nikon`

Interactive Motivation

How can these images be improved?



swedish landscape 4k high resolution award winning image trending on artstation

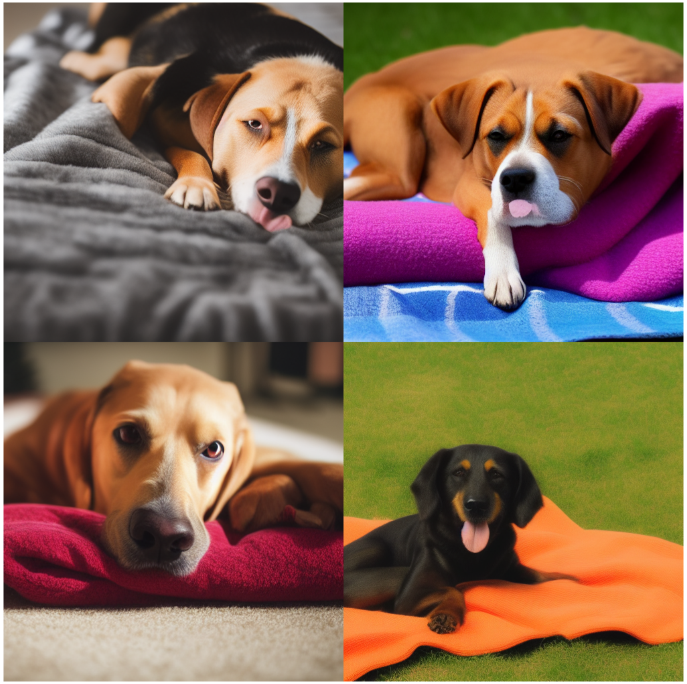
Interactive Motivation

How can we generate an image of this specific dog in a different context?

`person with a dog`

Interactive Motivation

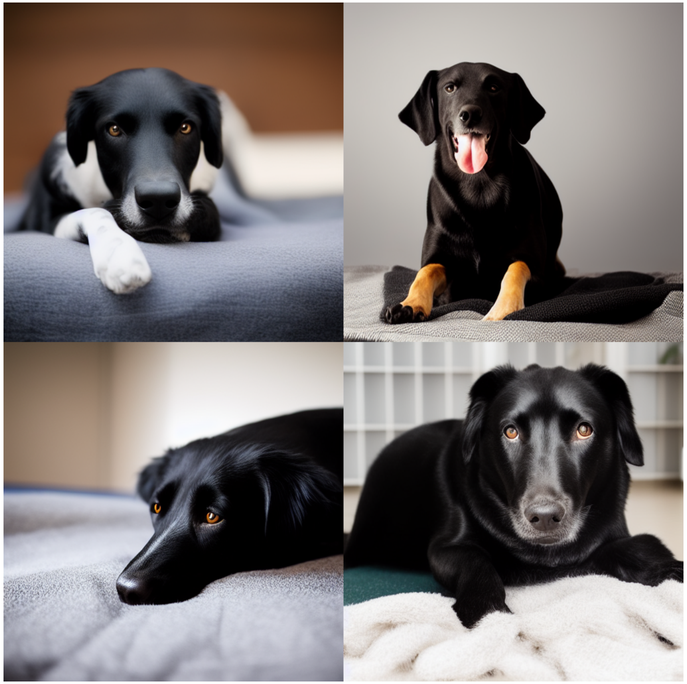
How can we generate an image of this specific dog in a different context?



`dog sitting on a blanket`

Interactive Motivation

How can we generate an image of this specific dog in a different context?



`black dog with a gray nose sitting on a blanket`

Interactive Motivation

How can we generate an image of this specific dog in a different context?

Prompt to Niklas: `Create an image of precisely this dog in a different context.`

Problems?

Interactive Motivation

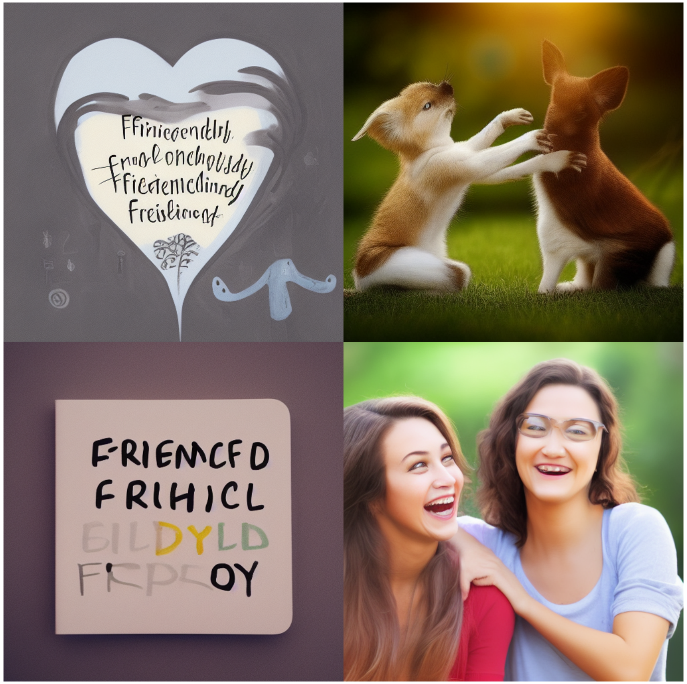
How can the following concepts be represented?

`friendship`

`diligence`

Interactive Motivation

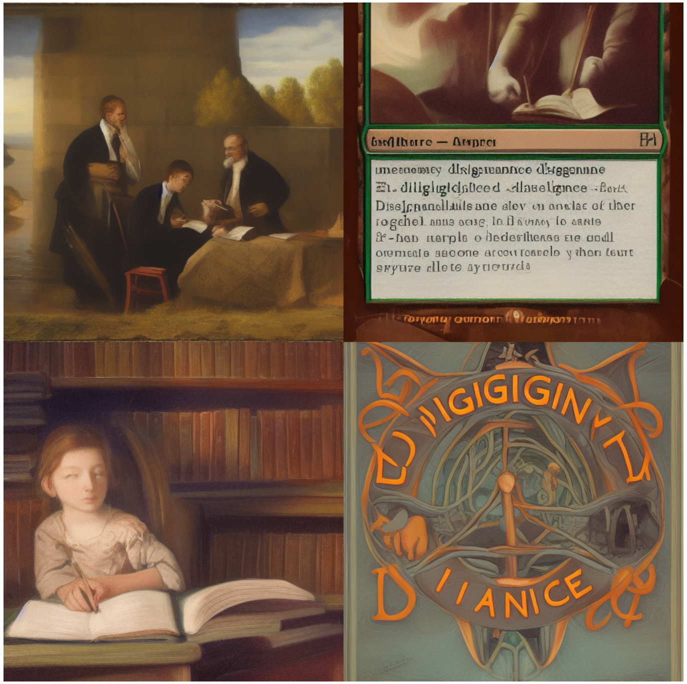
How can the following concepts be represented?



`friendship`

Interactive Motivation

How can the following concepts be represented?



`diligence`

Is this really the best visual representation of these concepts?

Interactive Motivation

Using the prompt to control the generated images is nice, but insufficient.

Getting to Know Each Other

What experience do you have in the following subjects?

- Text-to-image models like Stable Diffusion

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering

Getting to Know Each Other

What experience do you have in the following subjects?

❑ Text-to-image models like Stable Diffusion

❑ Prompt engineering

❑ Machine learning, deep learning

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering
- ❑ Machine learning, deep learning
- ❑ Language models like BERT

Getting to Know Each Other

What experience do you have in the following subjects?

❑ Text-to-image models like Stable Diffusion

❑ Prompt engineering

❑ Machine learning, deep learning

❑ Language models like BERT

❑ Vision models, Unets, vision transformers

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering
- ❑ Machine learning, deep learning
- ❑ Language models like BERT
- ❑ Vision models, Unets, vision transformers
- ❑ CLIP

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering
- ❑ Machine learning, deep learning
- ❑ Language models like BERT
- ❑ Vision models, Unets, vision transformers
- ❑ CLIP
- ❑ Programming in Python

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering
- ❑ Machine learning, deep learning
- ❑ Language models like BERT
- ❑ Vision models, Unets, vision transformers
- ❑ CLIP
- ❑ Programming in Python
- ❑ PyTorch

Getting to Know Each Other

What experience do you have in the following subjects?

- ❑ Text-to-image models like Stable Diffusion
- ❑ Prompt engineering
- ❑ Machine learning, deep learning
- ❑ Language models like BERT
- ❑ Vision models, Unets, vision transformers
- ❑ CLIP
- ❑ Programming in Python
- ❑ PyTorch
- ❑ Git, SSH, Slurm

Lab Organization

- Weekly consultations
- At the end of the semester: Written report (in groups) and presentation

Learning Objectives

- Work in a structured and self-supervised manner
- Work on a project of a larger scope
- Deal with open-ended tasks
- Groupwork and communication skills
- Apply and extend current research and tools in the field of generative models
- Develop and carry out experiments
- Scientific writing
- Apply machine learning to a real life problem

We would like to do real research with you!

Foundations

- ❑ IR terminology [webis.de]
- ❑ Discriminative vs. generative models [webis.de]
- ❑ Deep learning basics and backpropagation [webis.de]
- ❑ Embedding models [mlvu.github.io]
- ❑ Quick intro to GANs, diffusion networks, Stable Diffusion and CLIP [webis.de]
- ❑ Descriptive vs. creative approach, infinite index, interpolation [webis.de]
- ❑ Iterative prompt engineering, optimization and navigation in the prompt embedding space [webis.de]
- ❑ Integrating different modalities from pairwise datasets [arxiv.org]

Research Objectives for This Lab

- When people use IR systems, they do not need information that they already have
- Similarly, if text-to-image models are used for inspiration: Users want creative input
- However, it is unclear what to prompt

Research Objectives for This Lab

- There are two probability distributions:
  - The (a-priori) probabilities of the generative system
  - Probabilities defined by user surprise (low probability $=$ high surprise)
- *Give the user a maximum surprise with the condition of correct generation, i.e., low model surprise*
- How can this user probability be modeled?

Research Objectives for This Lab

- This gives multiple steps for research:
  - Find the probability distribution of the user surprise
  - Optimize w.r.t. this probability
  - Effectiveness evaluation (create datasets and scenarios to verify success)
    stock images might be a powerful tool to get abstract objectives for the analysis

- Important requirements:
  - The result must be on-topic (to avoid user confusion)
  - Thus, build a definition of neighborhood in the prompt embedding space to find relevant information
    which might be more complicated than just taking a fixed distance in the prompt embedding space since it results in varying distances in the image space
  - We assume that factuality is already part of the generative model (which is still utopic)

Research Objectives for This Lab

- The research projects will use different input modalities:
  Automated user interviews, navigation via examples, gestures, eyetracking, watch time evaluation, . . .

Initial Task

- Given two text prompts and an interpolation coefficient $0 \leq \lambda \leq 1$ (and a single fixed seed),
  generate the interpolated image between these two prompts.
- Try different interpolation methods (LERP, NLERP, SLERP)
- Use the `diffusers` library
  https://huggingface.co/blog/stable_diffusion
- https://github.com/huggingface/diffusers/blob/main/src/diffusers/pipelines/stable_diffusion/pipeline_stable_diffusion.py

Cluster Onboarding

- Until next week: Complete onboarding
- Complete the initial task (in groups)
- Papers & useful resources will be published to temir.org