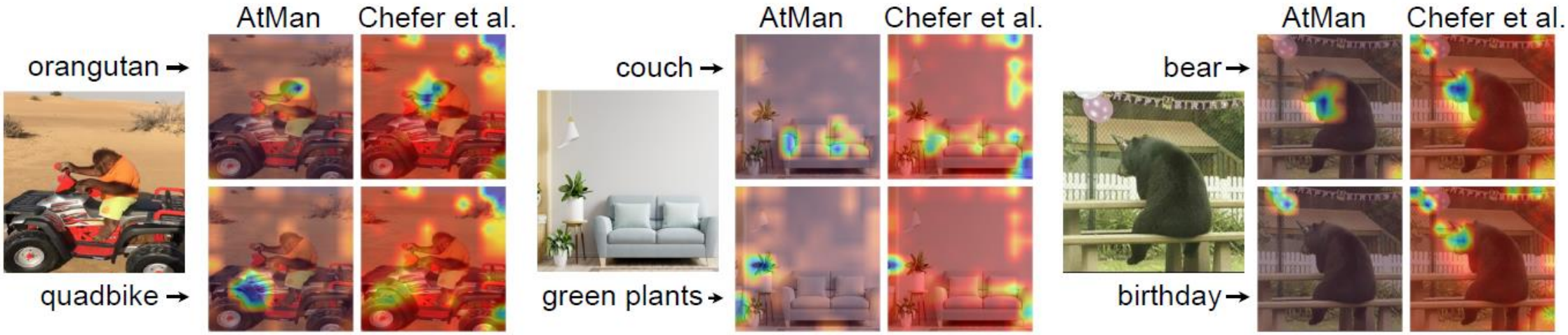
AtMan –XAI on generative Models



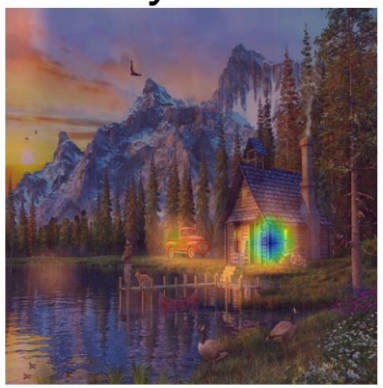
> Motivation

Multimodal prompt

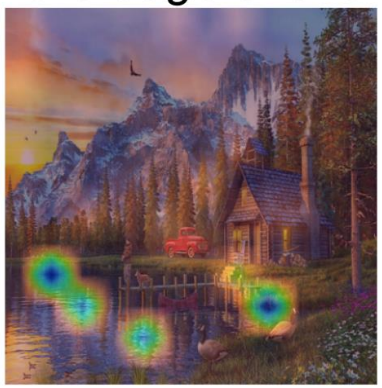
This is a painting of

Completion and AtMan Expl.

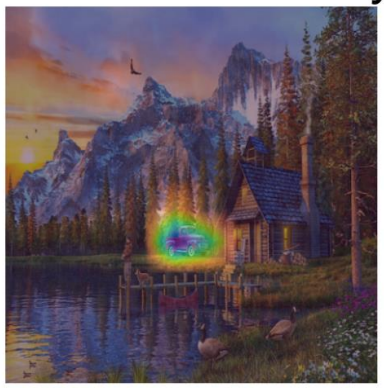
a lonely cabin



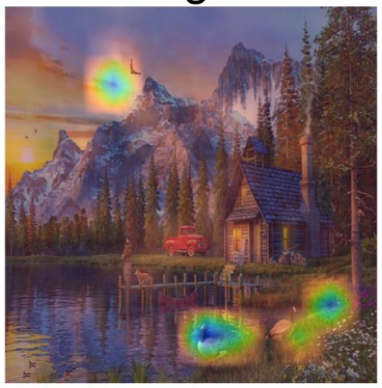
on the edge of a lake



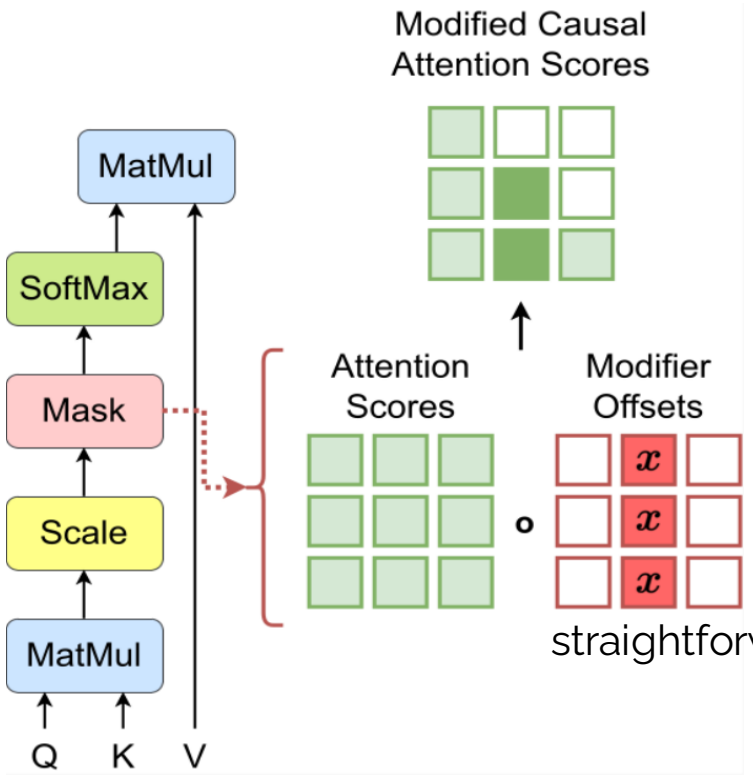
with a truck nearby



and the geese ...

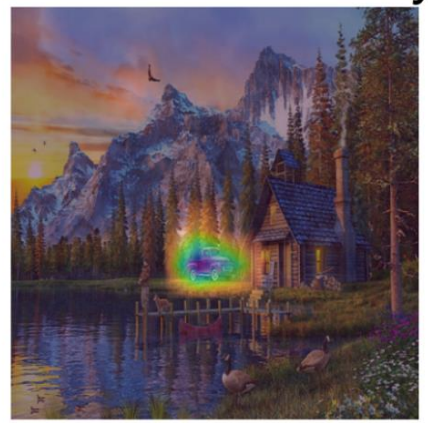


> XAI Method – Perturbating the Input

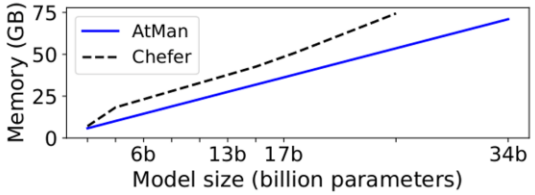


straightforward in hidden-space

with a truck nearby

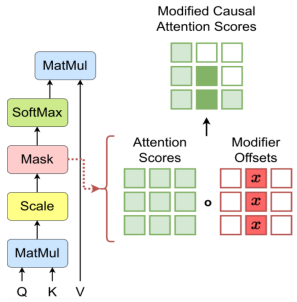


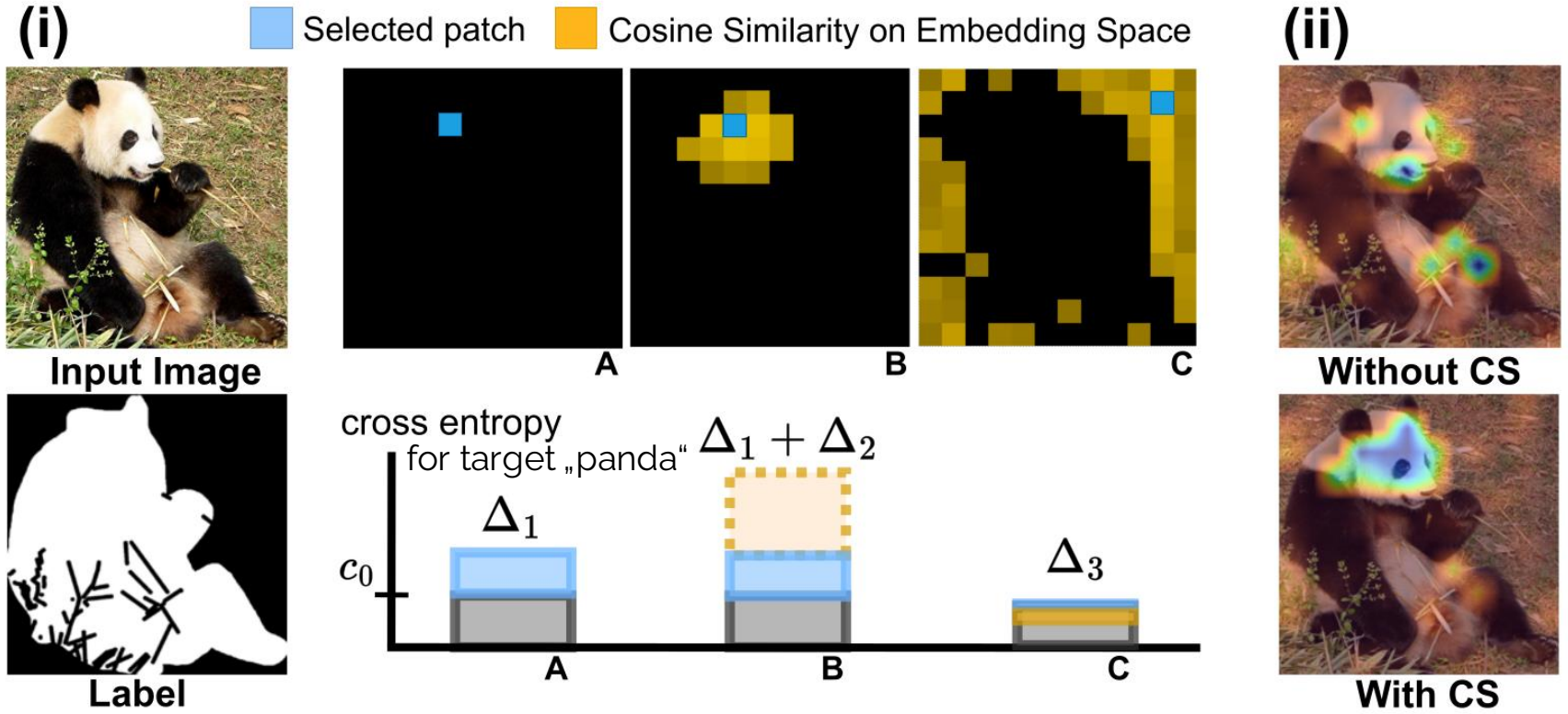
tricky in input-space



tricky with gradients

> Method – Cosine Similarity





> Results – Text/ Squad

Context with AtMan Explanation

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia.

Question	Label
From which countries did the Norse originate?	Denmark, Iceland and Norway
When were the Normans in Normandy?	10th and 11th centuries
In what country is Normandy located?	France

> Results – GQA

c.1) Q: Does the man

c.2) Q: Does the man ride a horse? A: *No.*

Negative Positive

> Results – Multiple Concepts

b) Q: How many animals are in this picture? A: *two*.

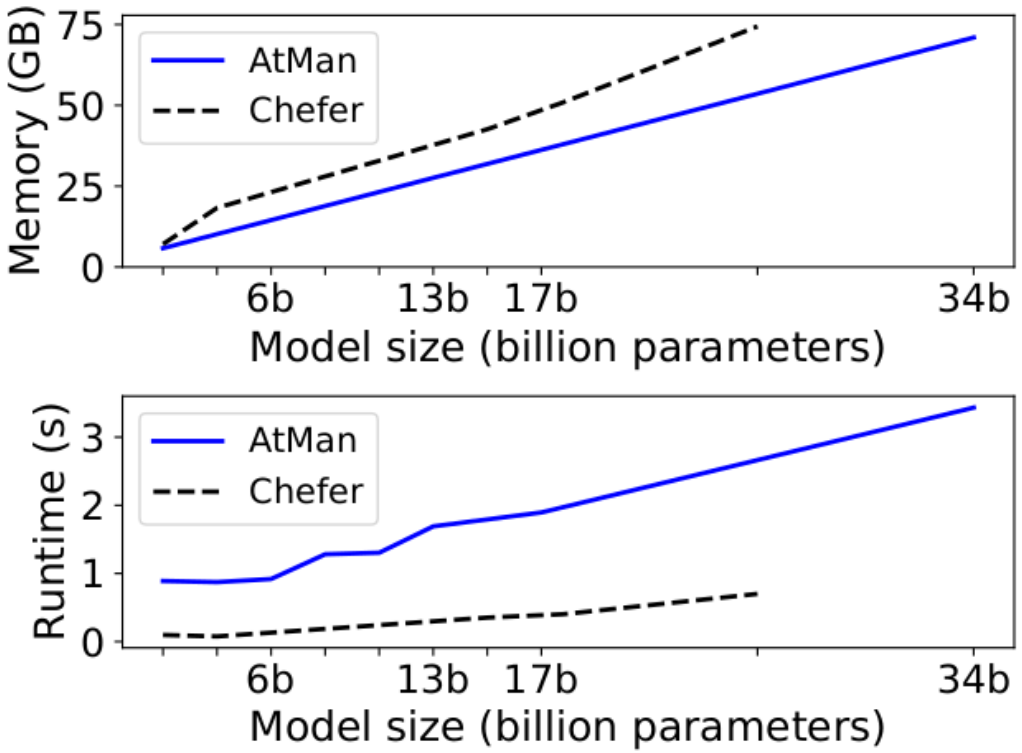
> Results – True Multimodality

> Results – Performance

- Only relying on forward passes

- No memory overhead

- No additional XAI-workers required



ATMAN: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation

Björn Deiseroth[1,2,3*] **Mayukh Deb**[1*] **Samuel Weinbach**[1*] **Manuel Brack**[2,4]
Patrick Schramowski[2,3,4,5] **Kristian Kersting**[2,3,4]
[1]Aleph Alpha [2]Technical University Darmstadt
[3]Hessian Center for Artificial Intelligence (hessian.AI)
[4]German Center for Artificial Intelligence (DFKI) [5]LAION

19 Jan 2023

https://github.com/Aleph-Alpha/AtMan

AtMan –XAI on generative Models

Björn Deiseroth
PhD Student

bjoern.deiseroth@aleph-alpha.com

www.aleph-alpha.com
Aleph Alpha GmbH
Grenzhöfer Weg 36
69123 Heidelberg
Germany