AI Factory Austria AI:AT

Shaping the Future of AI

Michela Vignoli
Scientist

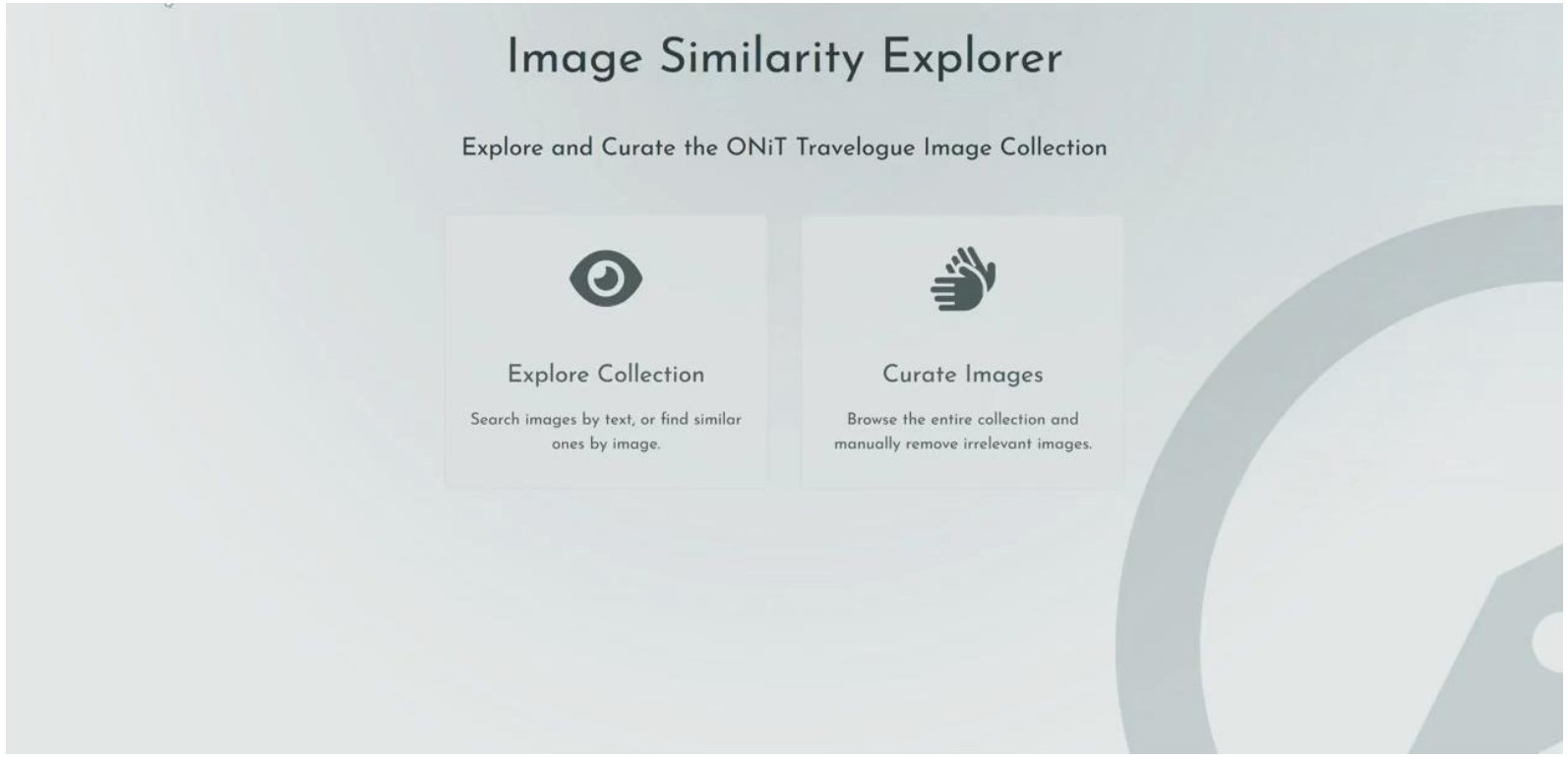
10.06.2026

Search Images with AI – Hands-On Introduction to CLIP

Agenda

- **Introduction to Vision–Language Models** (45 min.)
  Overview of core concepts behind models like CLIP and how they connect images and text
- **Live Demo: ONiT Explorer in Action** (20 min.)
  Explore a real-world application and see similarity search with image–text embeddings
- Short break (15 min.)
- **Hands-On Session:** Build Your Own Similarity Pipeline (55 min.)
  Implement image–text matching and test text query functionality (own images or ONiT dataset)
- Wrap-Up & Discussion (10 min.)
- 18:30 End of course

Search Images with AI – Hands-On Introduction to CLIP



Introduction: Vision-Language Models

- In recent years, **multimodal AI models** have become increasingly important.
- These models are designed to process and link information from different modalities
  → e.g., **text, images, audio, or video**.
- Traditional Machine Learning models focus on one modality (e.g., image recognition, text processing). Multimodal models learn to recognize **relationships between different types of data**.

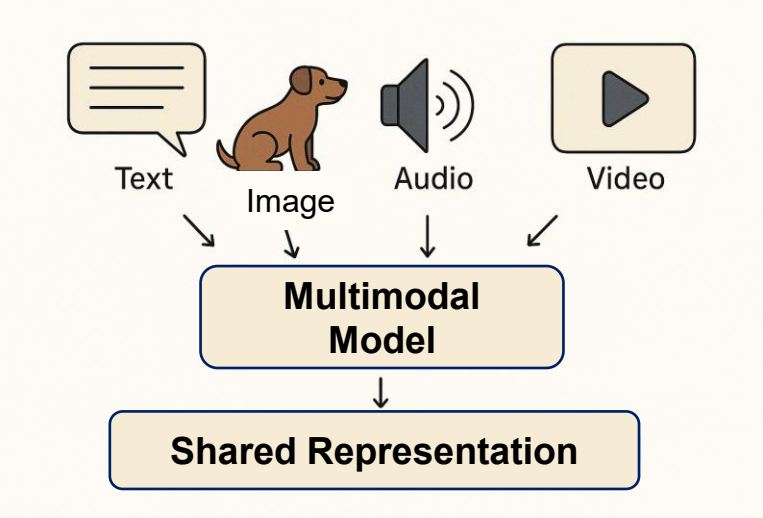


Image generated with ChatGPT-4.
Prompt: A diagram illustrating the basic function of multi-modal models.

Introduction: Vision-Language Models

- An example is **CLIP** (Contrastive Language–Image Pretraining), which was developed by OpenAI in 2021.
- CLIP is trained to **match text descriptions with images**.
- E.g., it can attribute the sentence “a dog jumps over a fence” to a similar image – even without the image having been manually labeled beforehand.

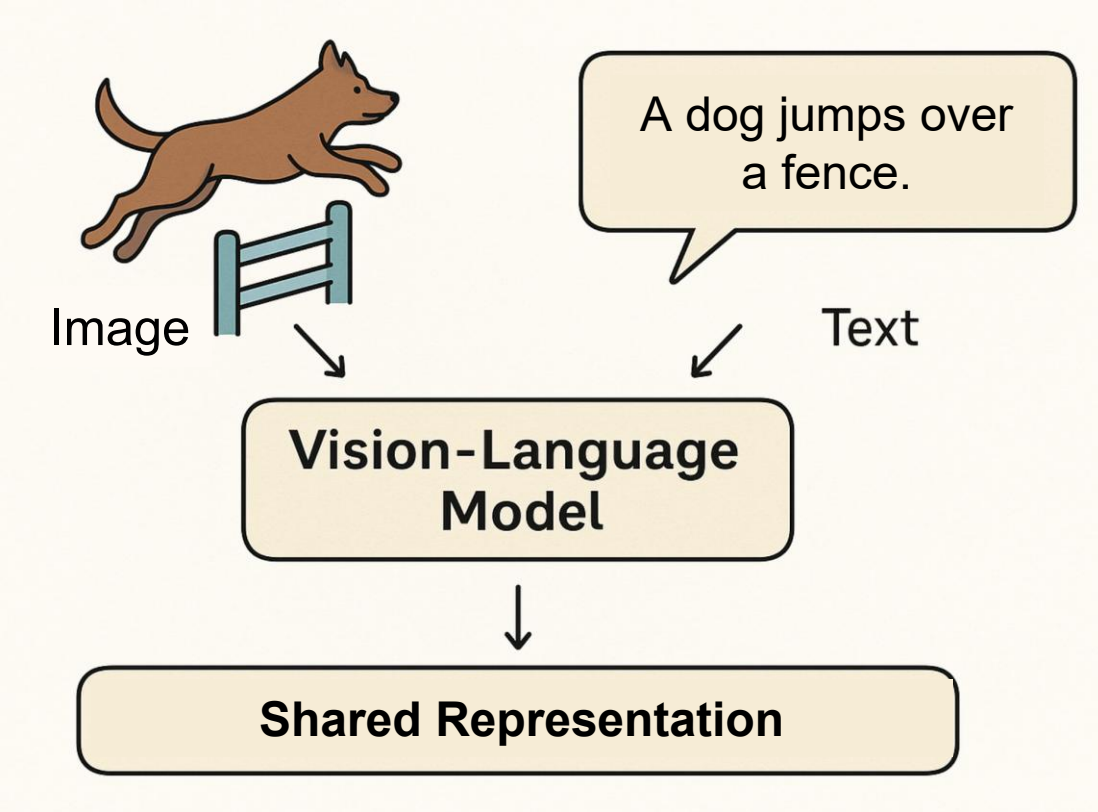


Image generated with ChatGPT-4. Prompt: A diagram illustrating the basic function of vision-language models.

Introduction: Vision-Language Models

- The strength of CLIP and similar systems lies in their ability to **generalize semantically**.
- This opens up a wide range of potential applications, including **image search without the need to label your data first**.
- Multimodal models represent an important step toward **more general-purpose, flexible AI systems** capable of processing complex information.

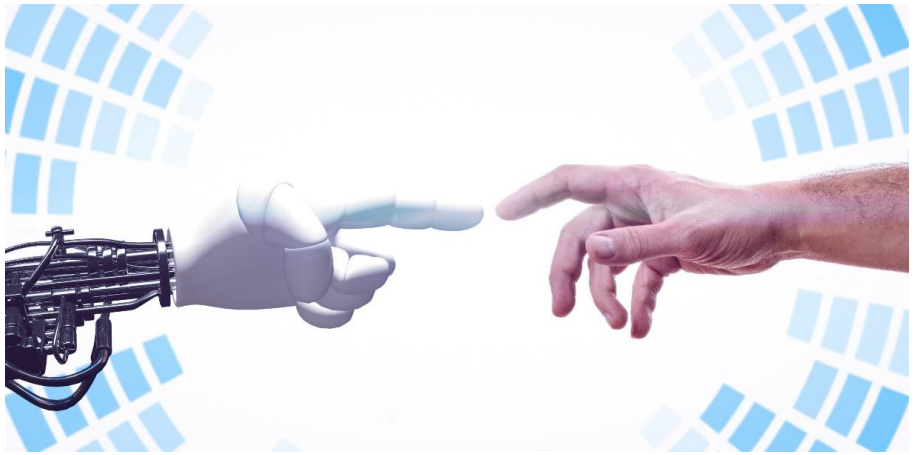


Image Source: https://pxhere.com/en/photo/1638452
License: CC0

Introduction: Vision-Language Models

- **Multimodal Linking**
  - Shared understanding of images and text
  - Recognition of semantic relationships between modalities
- **Zero-shot and few-shot capabilities**
  - Models can solve new tasks without explicit training
  - No specific classifiers or labelling required
- **Better generalization**
  - Robust against new objects, concepts, and domains
  - Wide applicability (heterogeneous datasets, real-world scenarios)

Image generated with Stable Diffusion (DreamStudio). Prompt: Human computer interaction.

Introduction: Vision-Language Models

- **Flexible Applications**
  - Text-based image search
  - Automated image captions (captioning)
  - Visual question answering
  - Support for creative processes (e.g., prompt-based art)
- **Elimination of fixed classification rules**
  - No limited, predefined set of labels as with traditional image classifiers
  - Natural language as an input interface

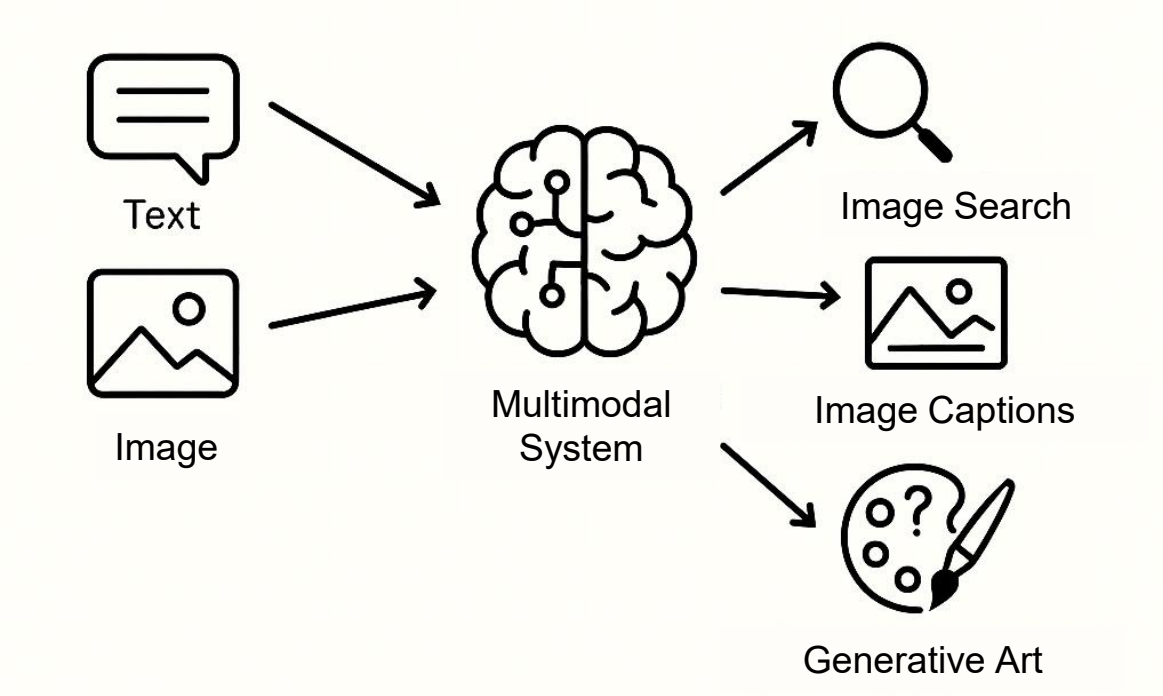
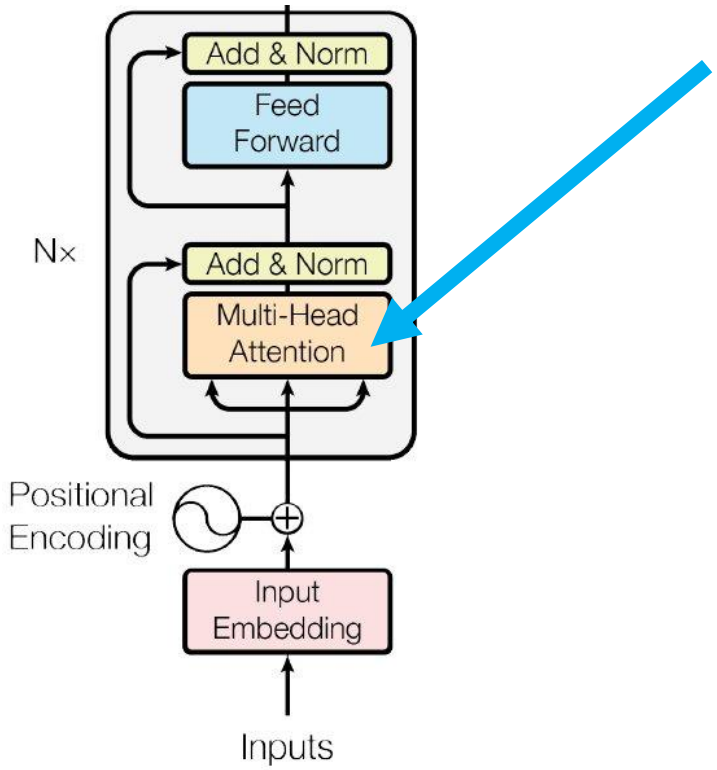


Image generated with ChatGPT-4.
Prompt: Create a graphic: The strength of CLIP and similar multimodal systems lies in their ability to perform semantic generalization. […]

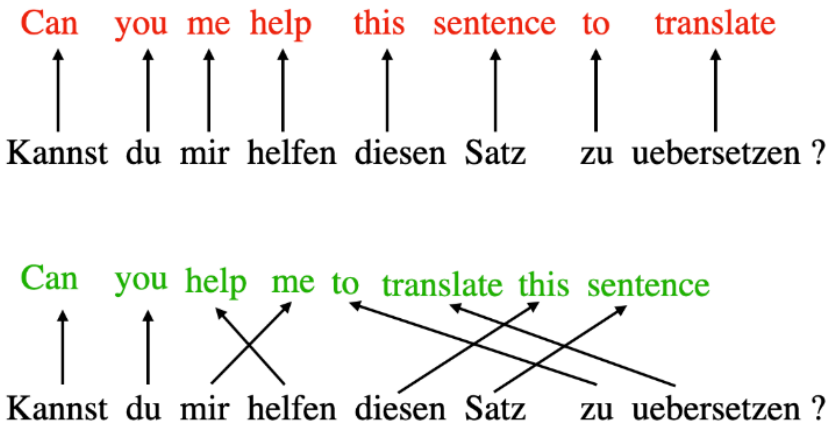
Introduction: Transformers

- **Transformer Encoder**

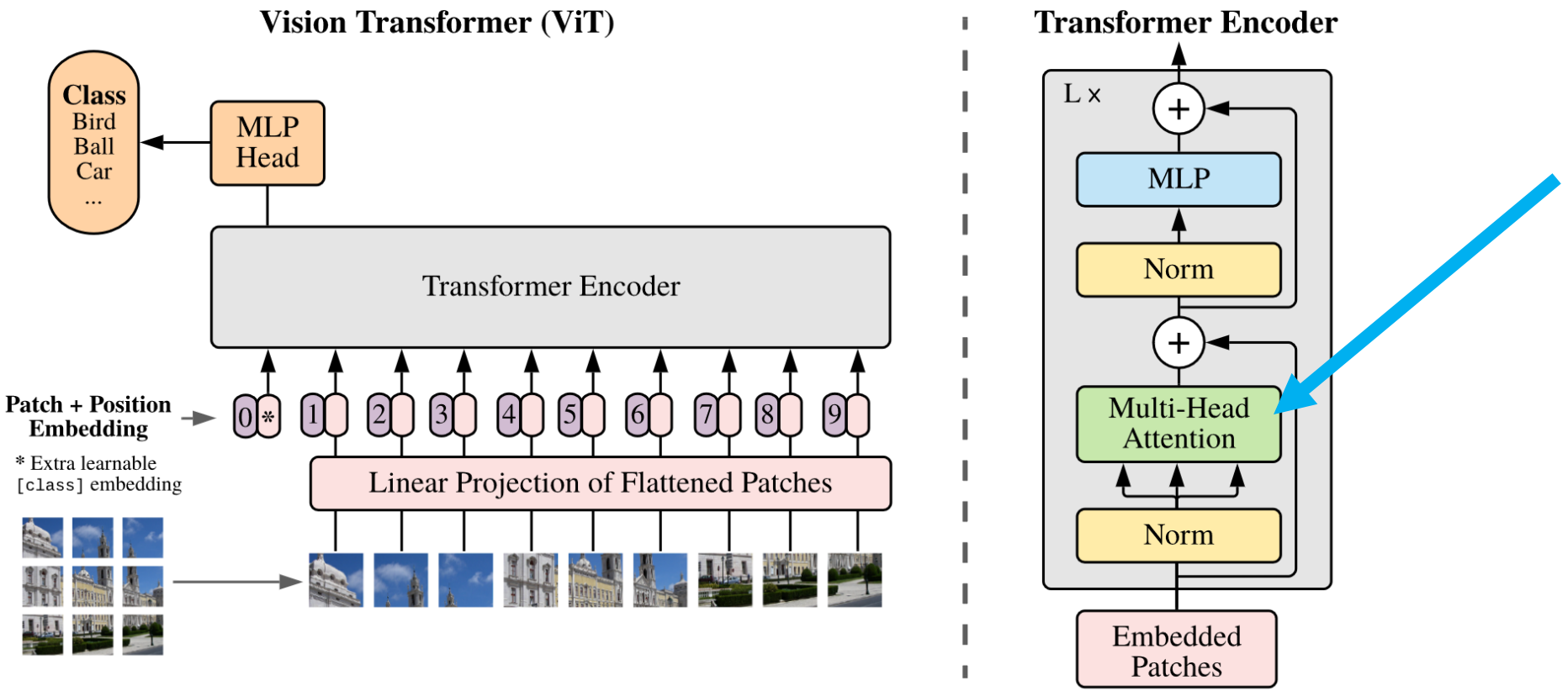


Source: Vaswani et al. 2017, p. 3

- **Self-attention Layer**



Introduction: Transformers



Source: Dosovitskiy et al. 2020, https://arxiv.org/abs/2010.11929, p. 3

Introduction: Transformers

Source: https://theaisummer.com/xai/

- **Self-attention Layer**
  - Allows the neural network to focus on the necessary features/objects
  - Using so-called class activation maps (CAM), we can visualize the regions in the image that the self-attention layer has amplified as most relevant for classification (heatmaps)

Concept: Multi-Modal Embedding

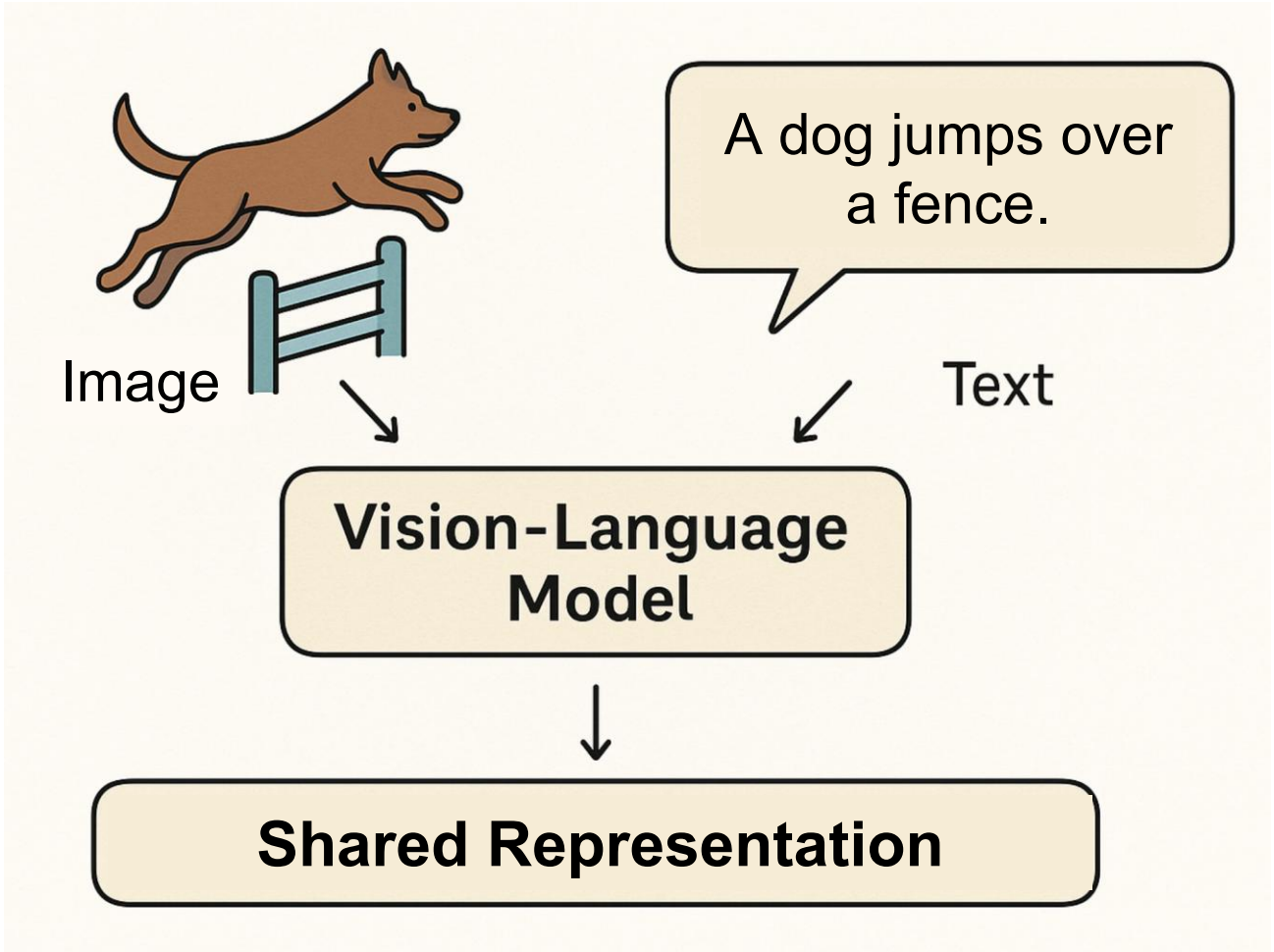
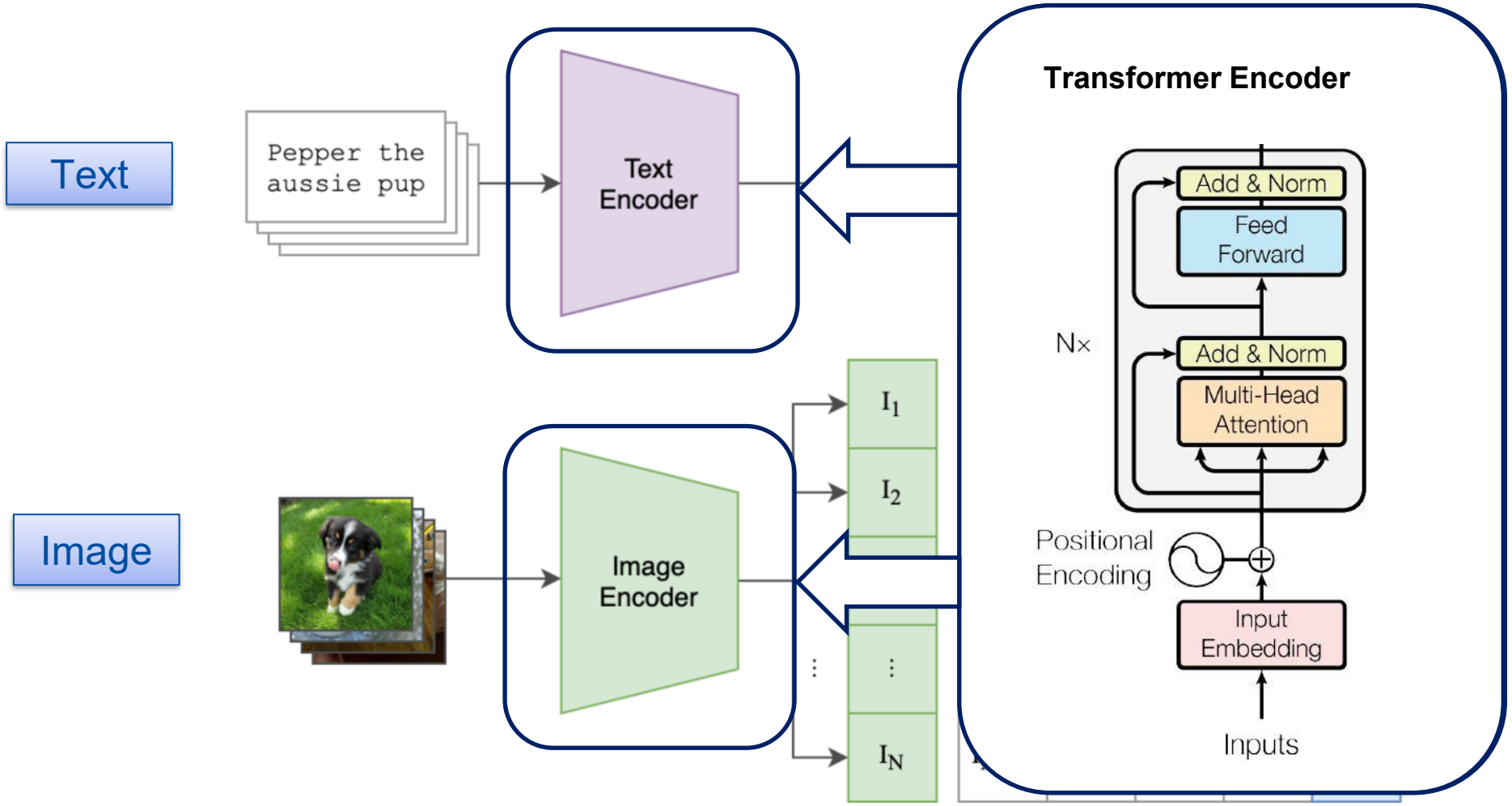


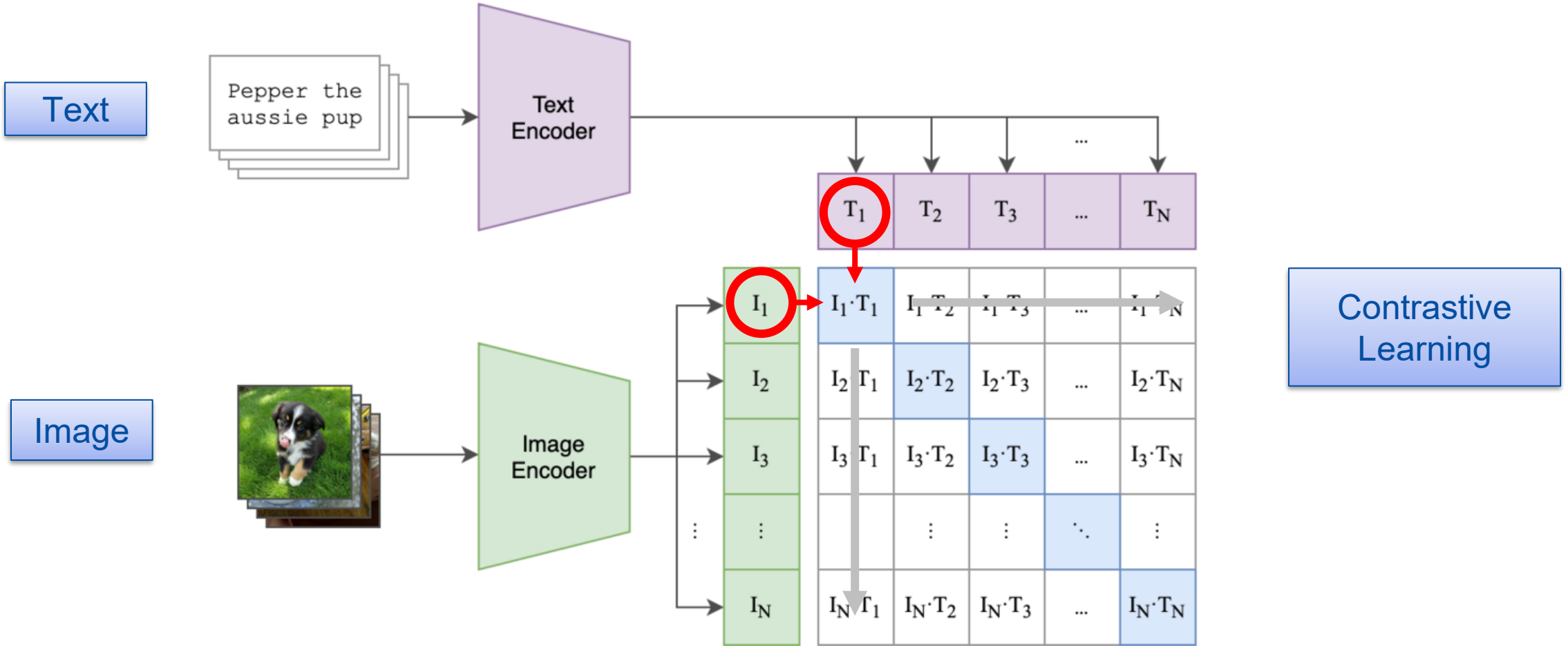
Image generated with ChatGPT-4. Prompt: A diagram illustrating the basic function of vision-language models.

Concept: Multi-Modal Embedding



Source: Radford et al. 2021, https://arxiv.org/abs/2103.00020, p. 2)

Concept: Multi-Modal Embedding



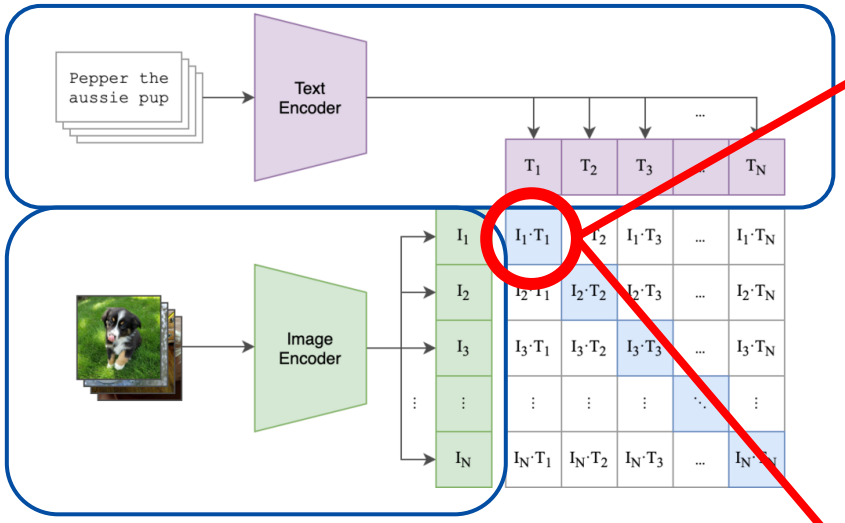
Source: Radford et al. 2021, https://arxiv.org/abs/2103.00020, p. 2)

Concept: Multi-Modal Embedding

AUSTRIAN INSTITUTE OF TECHNOLOGY

Embeddings / Vectors

Text



Exotic fish

Text Encoder

$$\begin{matrix} 1 & 2 & \cdots & n \end{matrix}$$
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}$$
`torch.Size([1, 512])`

High similarity

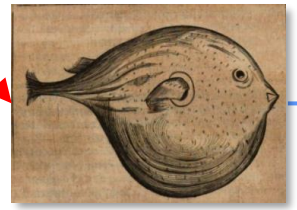


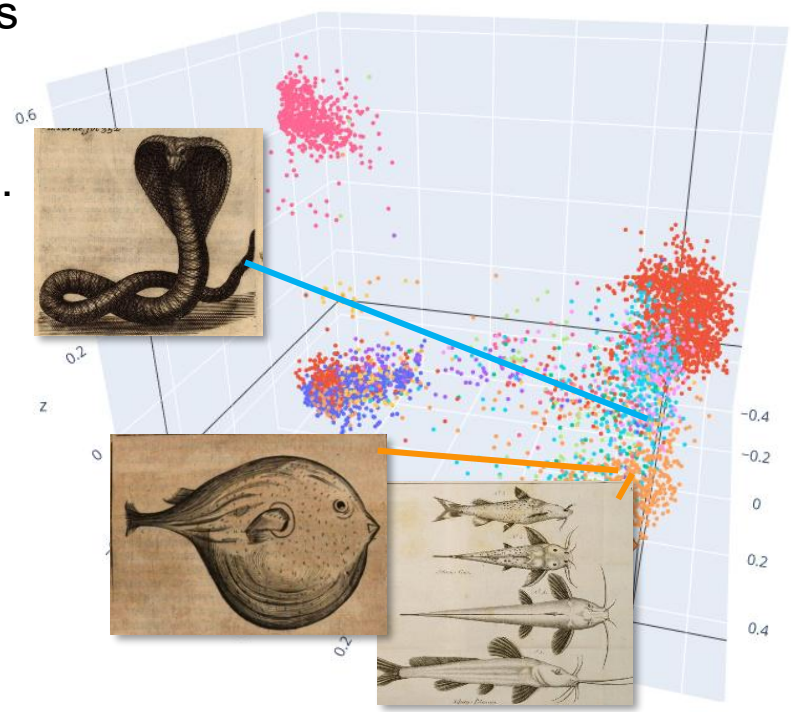
Image Encoder

$$\begin{matrix} 1 & 2 & \cdots & n \end{matrix}$$
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}$$
`torch.Size([1, 512])`

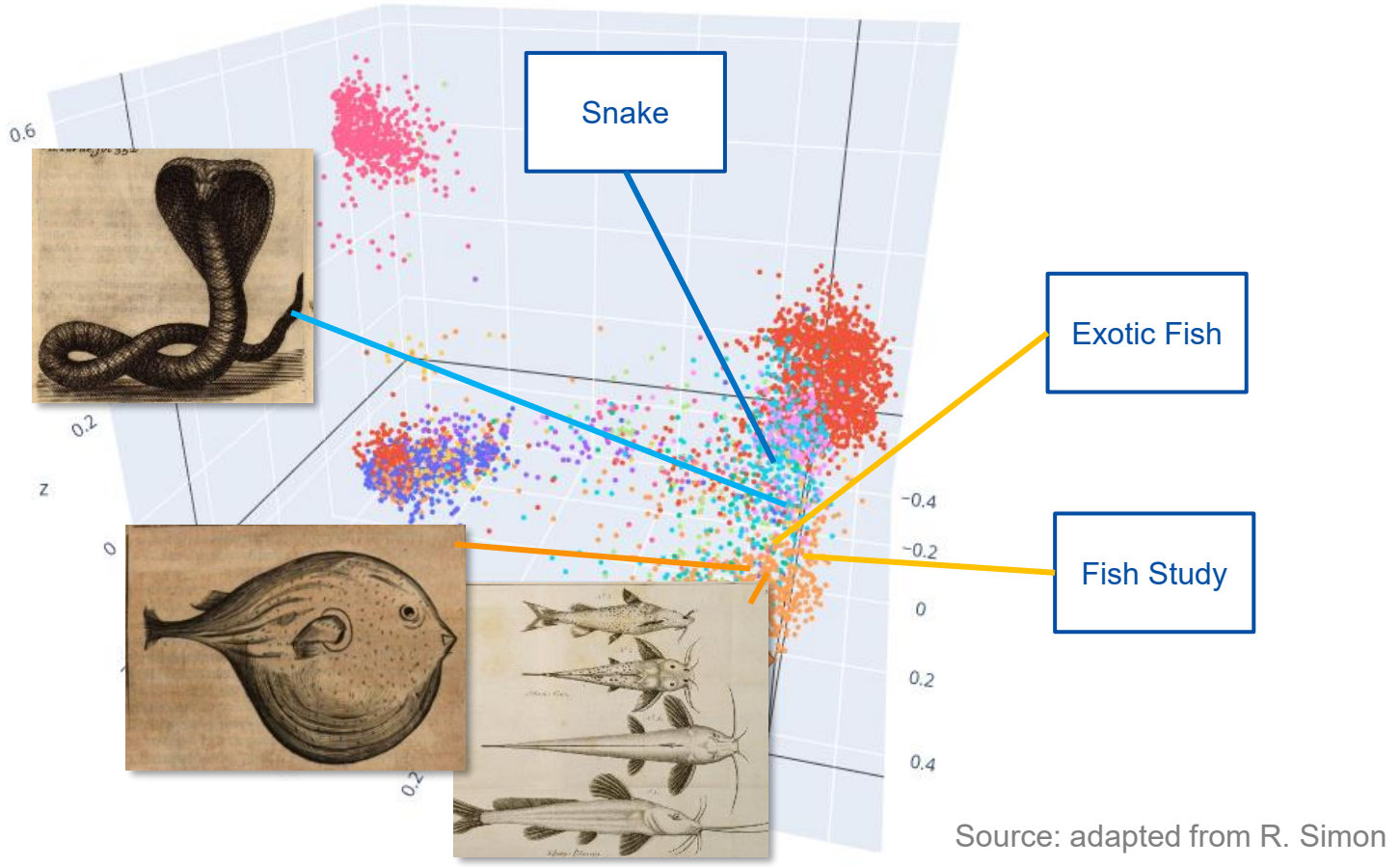
Image

Concept: Semantic Vector Space

- **Basic Concept**
  - Text, images, etc. are converted into **vectors** (sequences of numbers).
  - These vectors are located in a **high-dimensional space**.
  - **Semantically similar content** is close together, while dissimilar content is farther apart.
- **Properties**
  - **Distances ≈ semantic similarity**
    → e.g., “dog” and “cat” are closer than “dog” and “car.”
  - Forms the basis for **semantic search** and **similarity comparisons**.



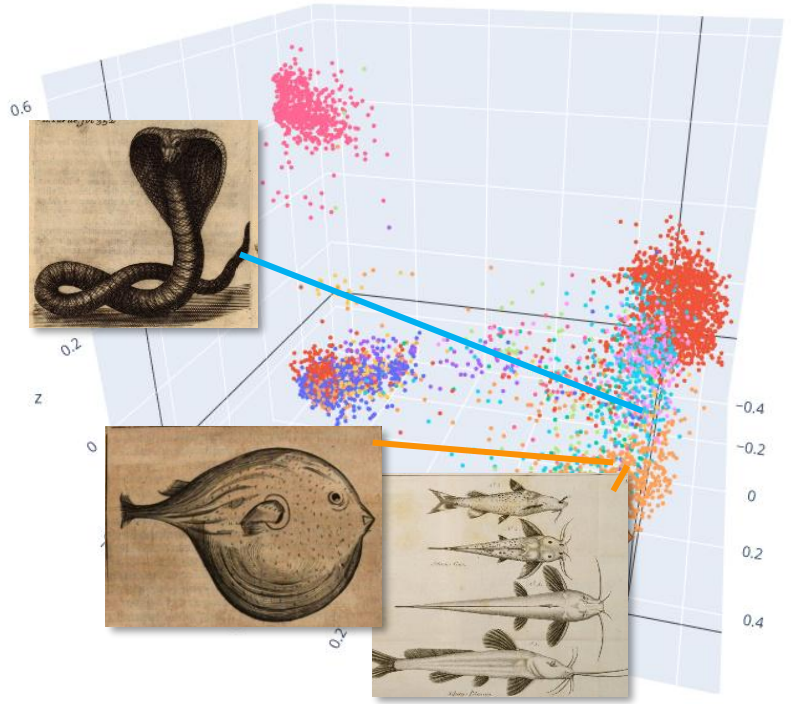
Concept: Semantic Vector Space



Source: adapted from R. Simon

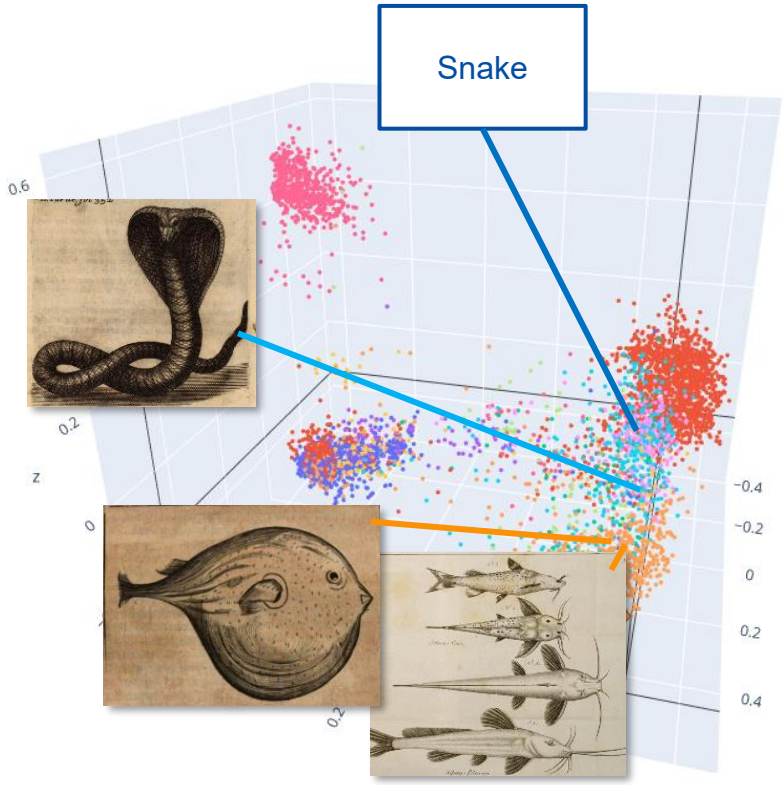
Concept: Semantic Search

- **Semantic Search**
  - Uses **vector representations (embeddings)** of text, images, etc.
  - Calculates **similarity in vector space** (e.g., cosine similarity)
  - Recognizes **meaning and context**, not just keywords
- **Advantages**
  - Can take synonyms, paraphrases, and context into account
  - Applicable to **multimodal data** (CLIP: text ↔ image)



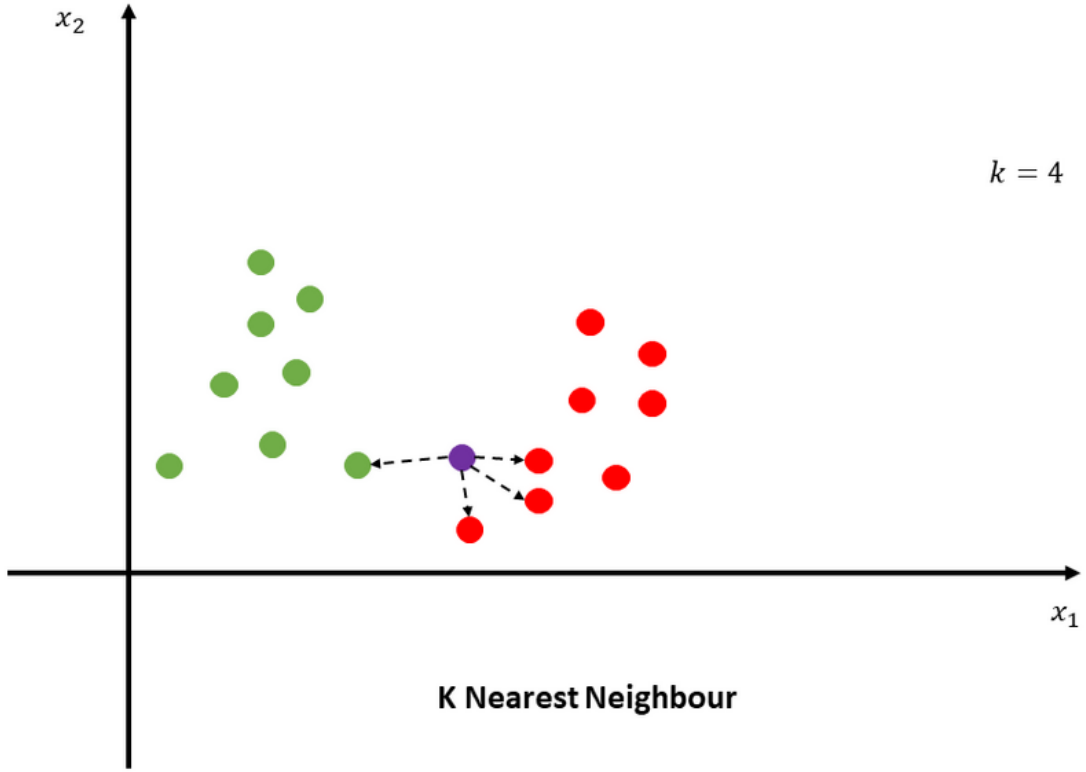
Concept: Semantic Search

- **How it works:**
  - All objects are represented in a **shared vector space**.
  - A new query is also embedded as a vector.
  - The **nearest neighbors** are determined using a distance metric (e.g., cosine similarity, Euclidean distance).
- **Applications:**
  - Image or text search based on content similarity
  - Recommendation systems
  - Clustering and anomaly detection



Concept: Semantic Search

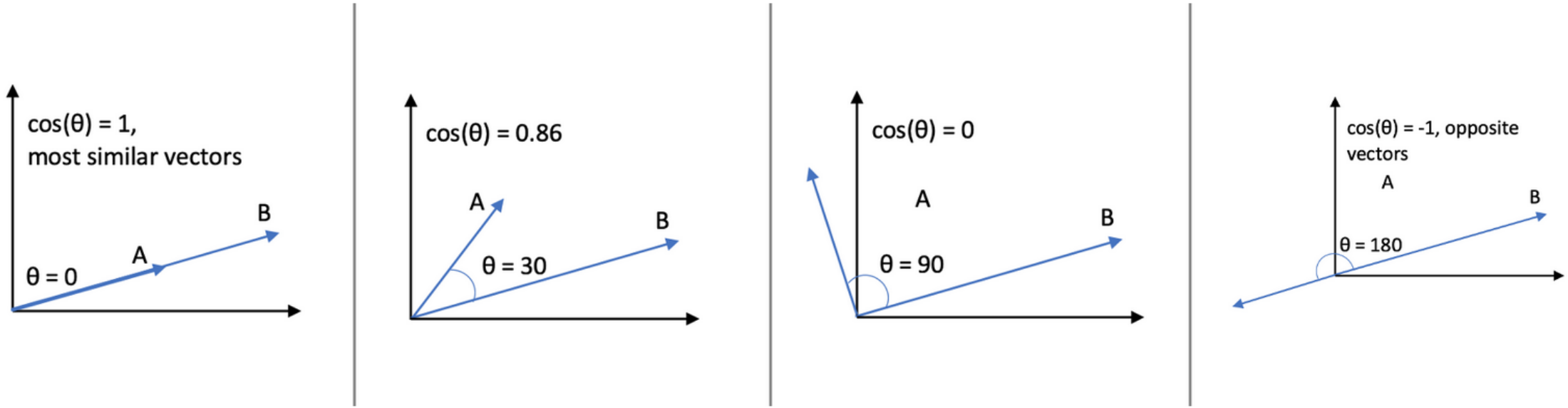
- **Nearest Neighbor**
  - **Basic concept:** Find the data points that are most similar to a given query (e.g., image, text, vector).
  - **Advantage:** Flexible search that does not require an exact match but is based on semantic proximity.



Source: Wikimedia 2021 (CC BY-SA)

Concept: Semantic Search

- **Similarity Metrics: Cosine Similarity**
  - Measures the angle between two vectors
  - Widely used in text and multimodal embeddings (BERT, CLIP, etc.)



Source: Towards Data Science, 2020

Questions?

Search Images with AI – Hands-On Introduction to CLIP

Agenda

- **Introduction to Vision–Language Models** (45 min.)
  Overview of core concepts behind models like CLIP and how they connect images and text
- **Live Demo: ONiT Explorer in Action** (20 min.)
  Explore a real-world application and see similarity search with image–text embeddings
- Short break (15 min.)
- **Hands-On Session:** Build Your Own Similarity Pipeline (55 min.)
  Implement image–text matching and test text query functionality (own images or ONiT dataset)
- Wrap-Up & Discussion (10 min.)
- 18:30 End of course

Live Demo:
ONiT Explorer in Action

Ottoman Nature in Travelogues (ONiT)

- Semantic search in historical books with rich illustratinos
- Interdisciplinary workflow to extract and analyse images and texts in historical travelogues (16th-19th century)

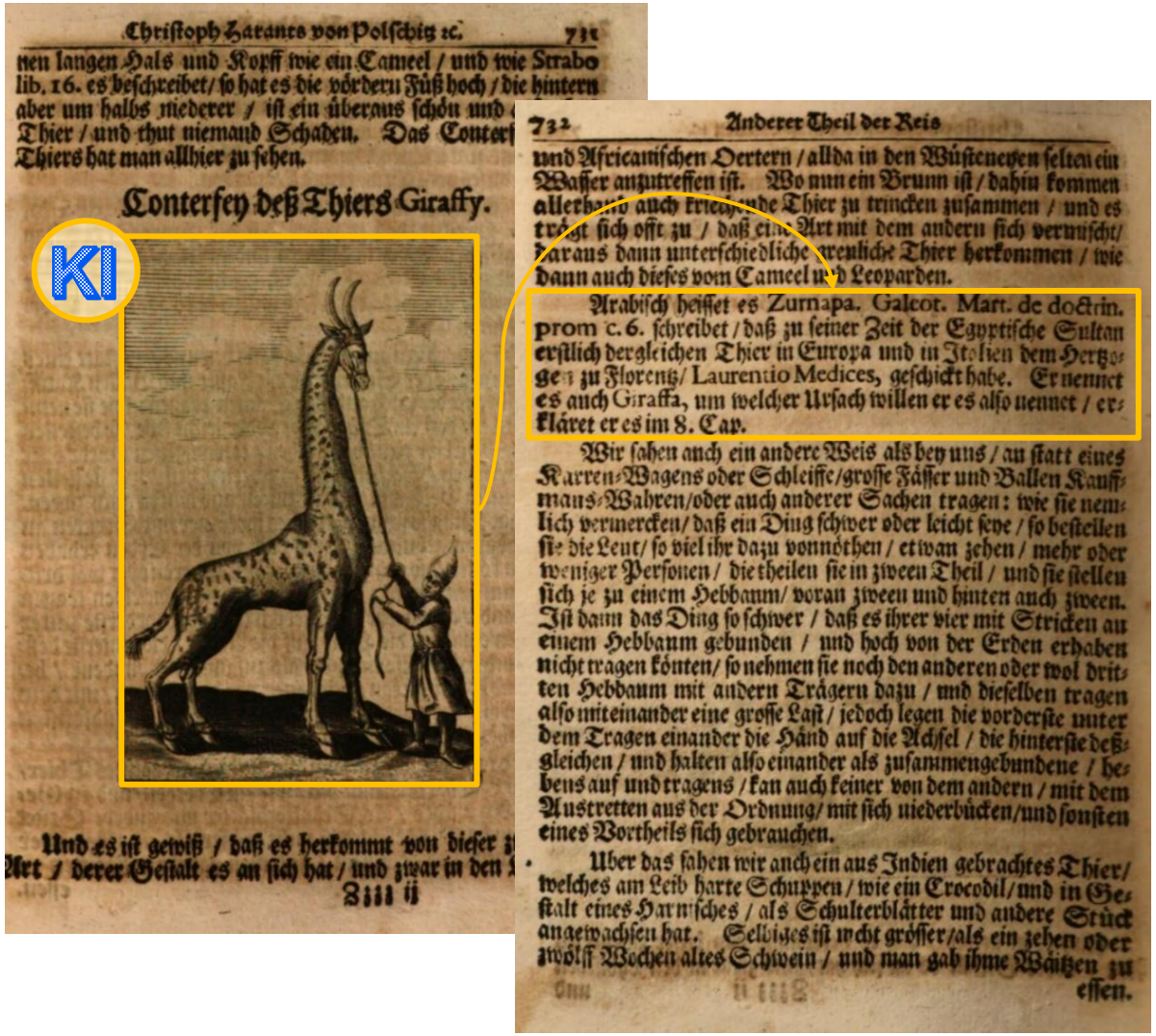
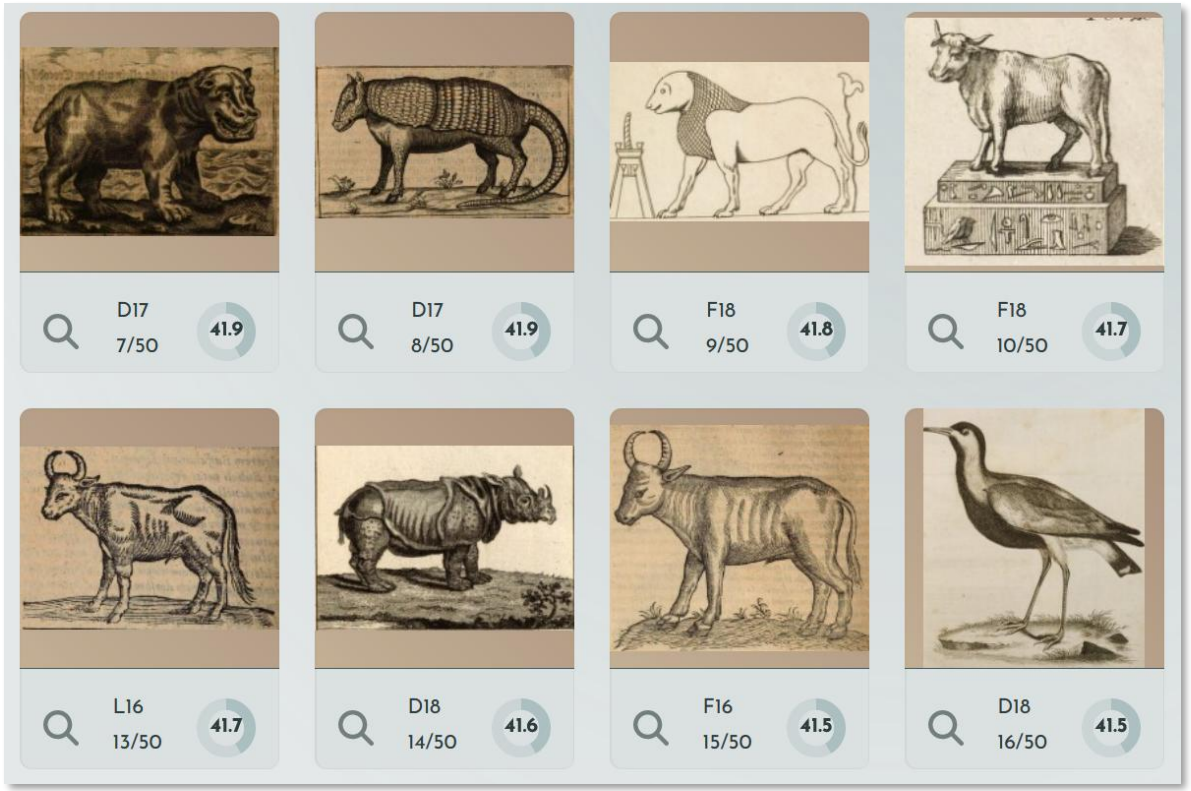


Image source: http://data.onb.ac.at/rep/10A8CA5D, pp. 771-772

Live Demo: ONiT Explorer in Action

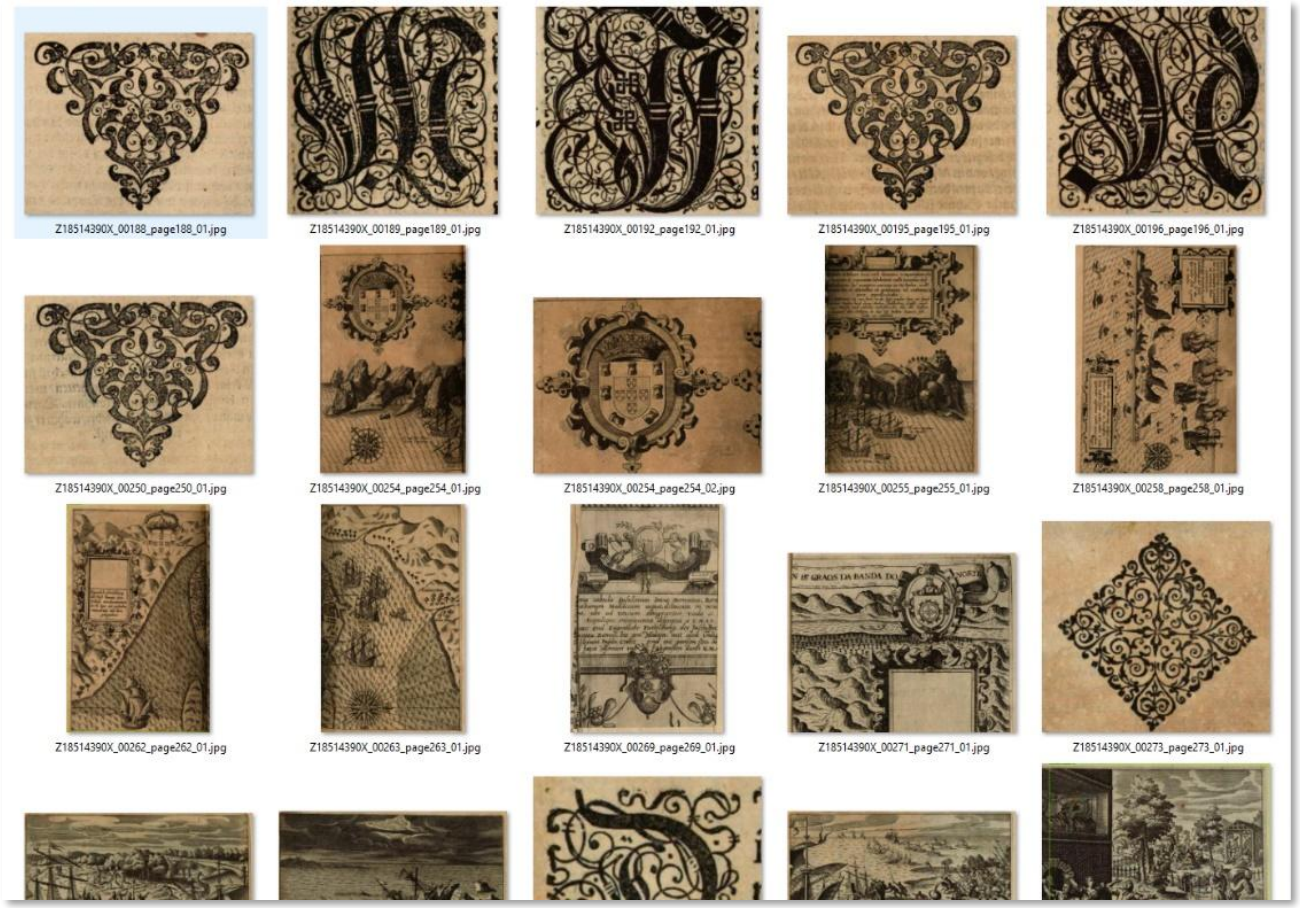


- **Focus on nature representations** (*flora, fauna, landscapes, maps*)
  - **Step 1: Extraction**
    - Automatic with pretrained tool
  - **Step 2: Curation**
    - Semi-automatic
  - **Step 3: Classification**
    - Manual annotation
    - Fine-Tuning and retrieval with CLIP

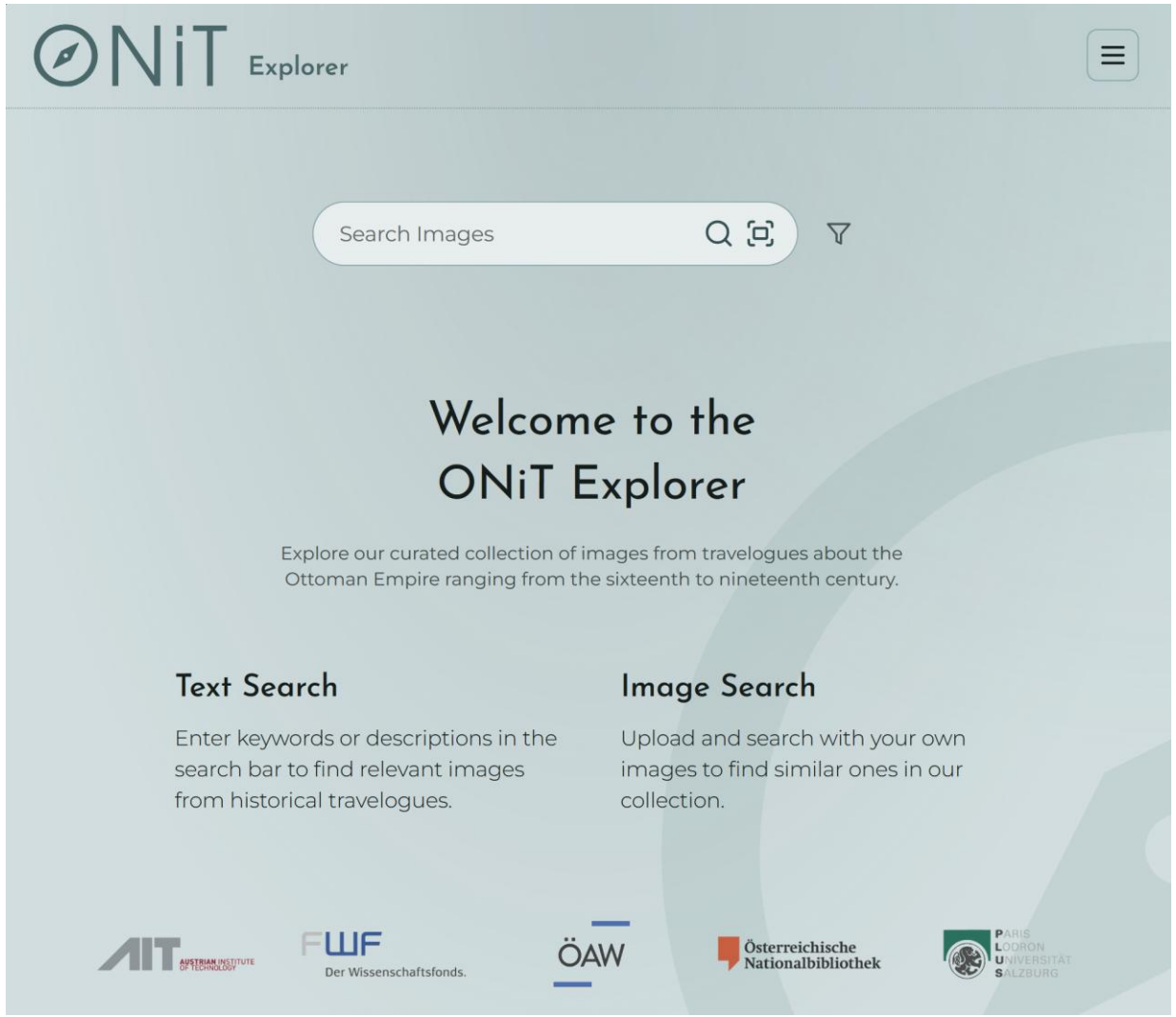
Live Demo: ONiT Explorer in Action

Data preparation

- Automatic extraction:
  - 22‘300 images from over 1‘500 digitized books
- Curation:
  - Sort out irrelevant material (erroneous detections, book stamps, ornaments, illuminated initials, other images)
  - Result: 8‘700 images with nature representations



Live Demo: ONiT Explorer in Action



ONiT Explorer

- Web application to explore image corpus
- Based on CLIP
- Semantic search of image vectors
- Search with text- or image prompts
- Advantage: no metadata needed
- Link: https://labs.onb.ac.at/de/tool/onit-explorer/

Live Demo: ONiT Explorer in Action

Fine-tuned

Class (ICONCLASS)	Total Examples	Recall@K (k=200)	Precision @K (k=200)	Recall@K (k=1,000)	Precision@ K (k=1,000)	R-Precision
plants/vegetation 25G	5,708	0.02	0.71	0.1	**0.62**	**0.51**
landscapes 25H	4,234	0.05	**0.98**	0.22	**0.95**	**0.7**
animals 25F	3,506	0.05	**0.92**	0.17	**0.69**	**0.46**
maps/atlases 25A	1,081	0.16	0.89	0.68	0.72	0.69

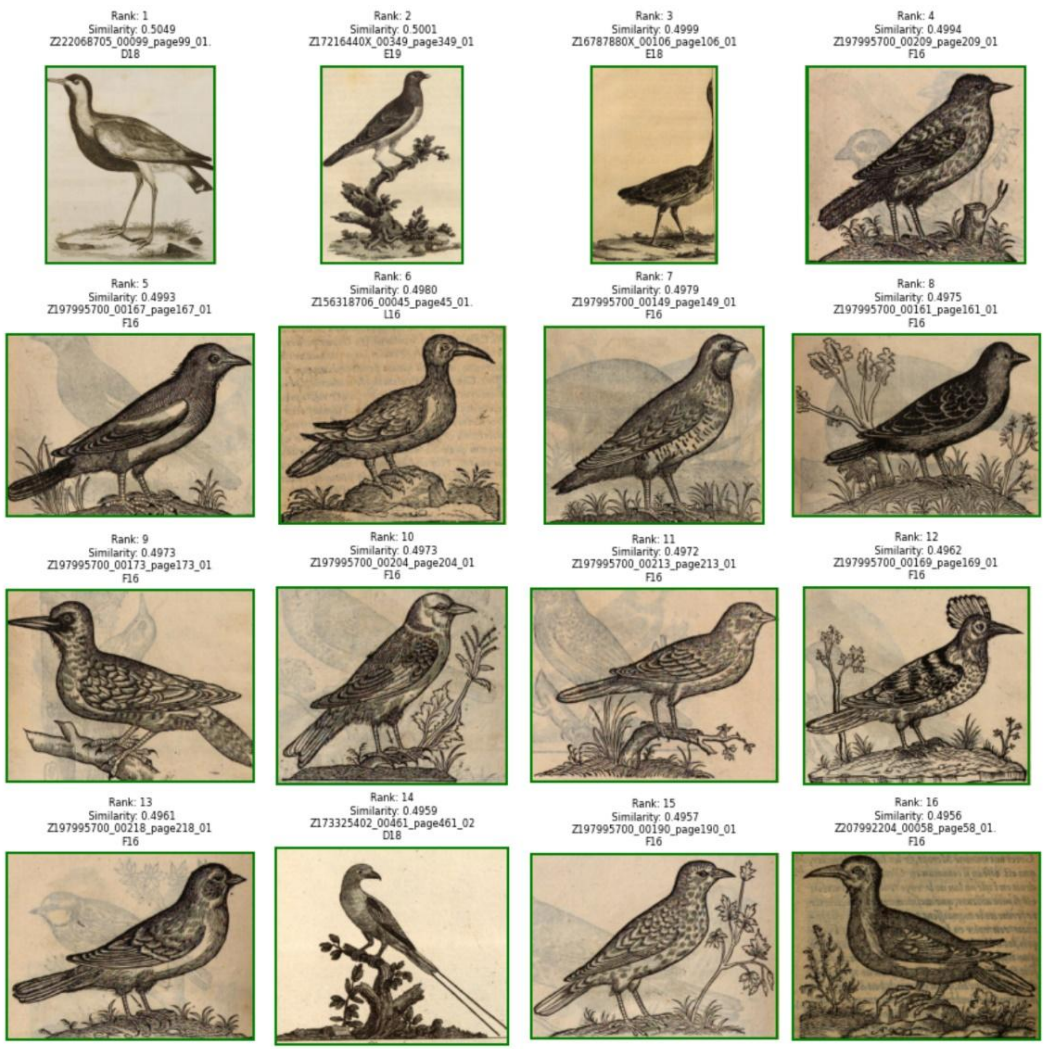
Base model

Pretrained CLIP Model						
plants/vegetation 25G	5,708	0.03	**0.86**	0.08	0.45	0.17
landscapes 25H	4,234	0.05	**0.98**	0.22	0.9	0.67
animals 25F	3,506	0.05	**0.93**	0.17	0.59	0.28
maps/atlases 25A	1,081	0.17	**0.94**	0.71	**0.77**	**0.75**

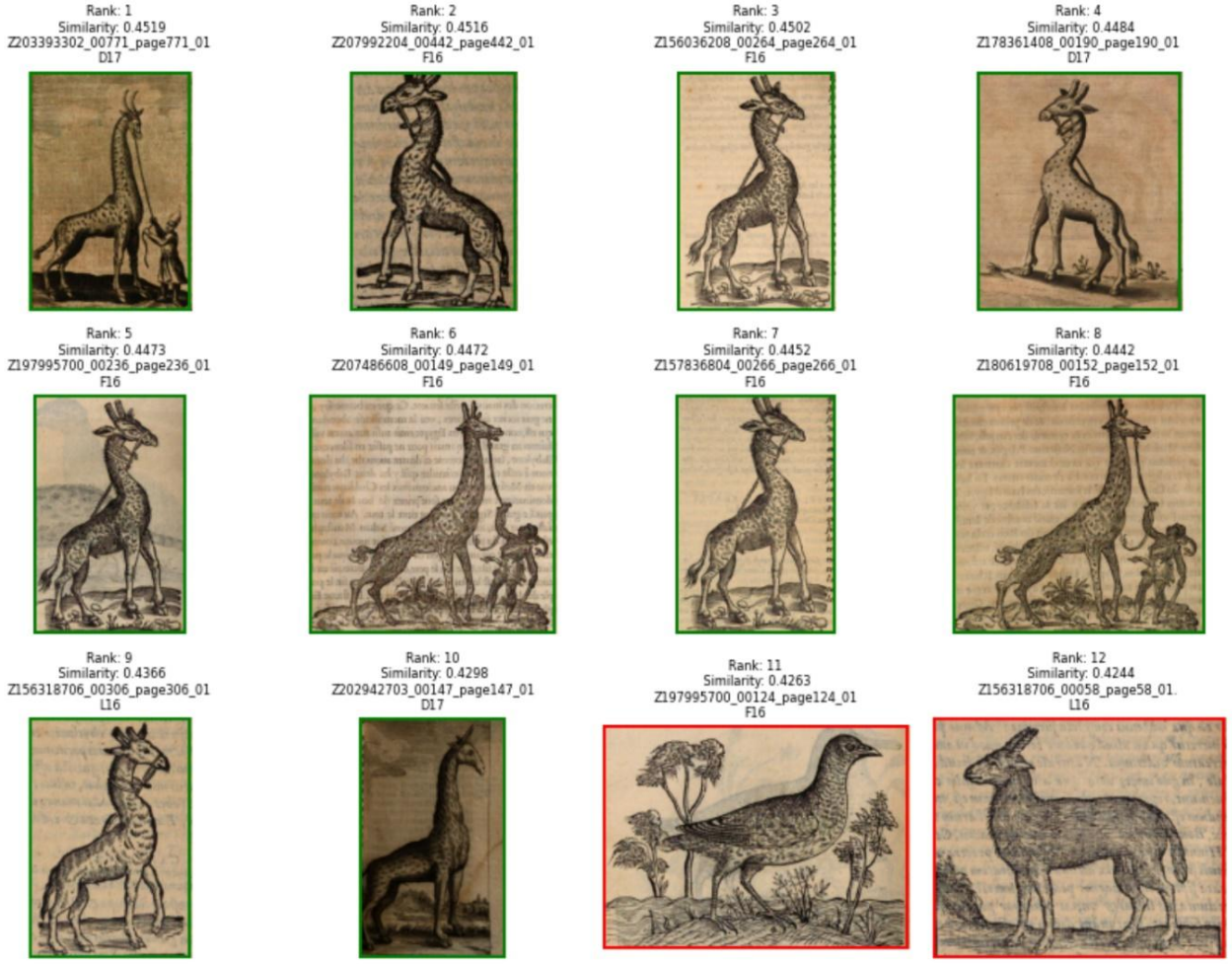
Live Demo: ONiT Explorer in Action

AUSTRIAN INSTITUTE OF TECHNOLOGY

birds

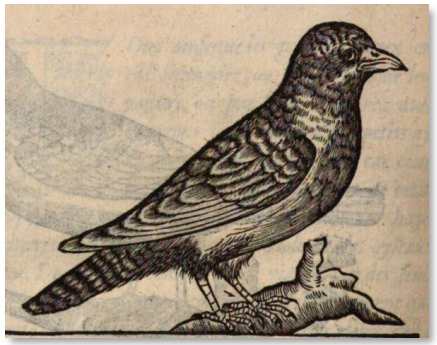


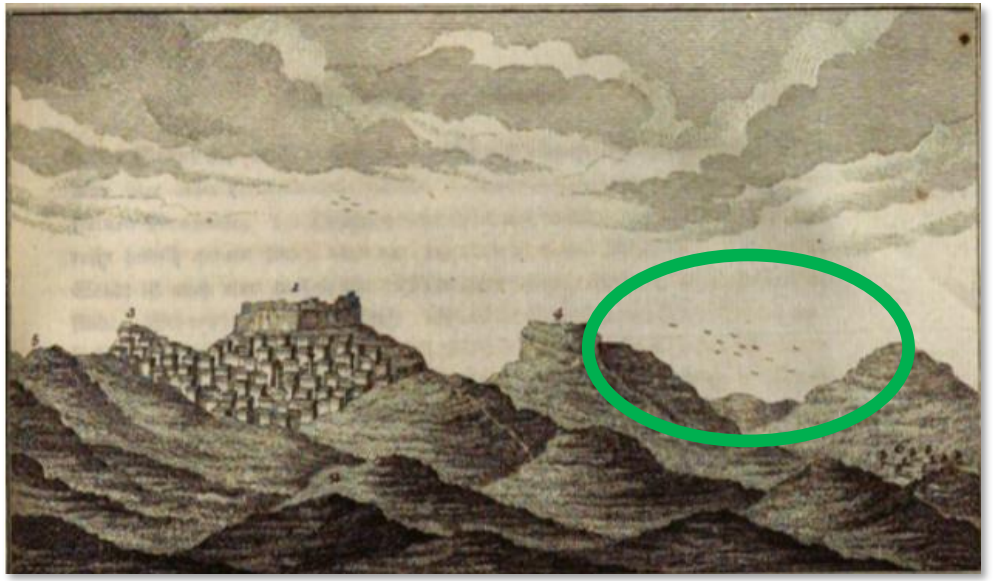
giraffe



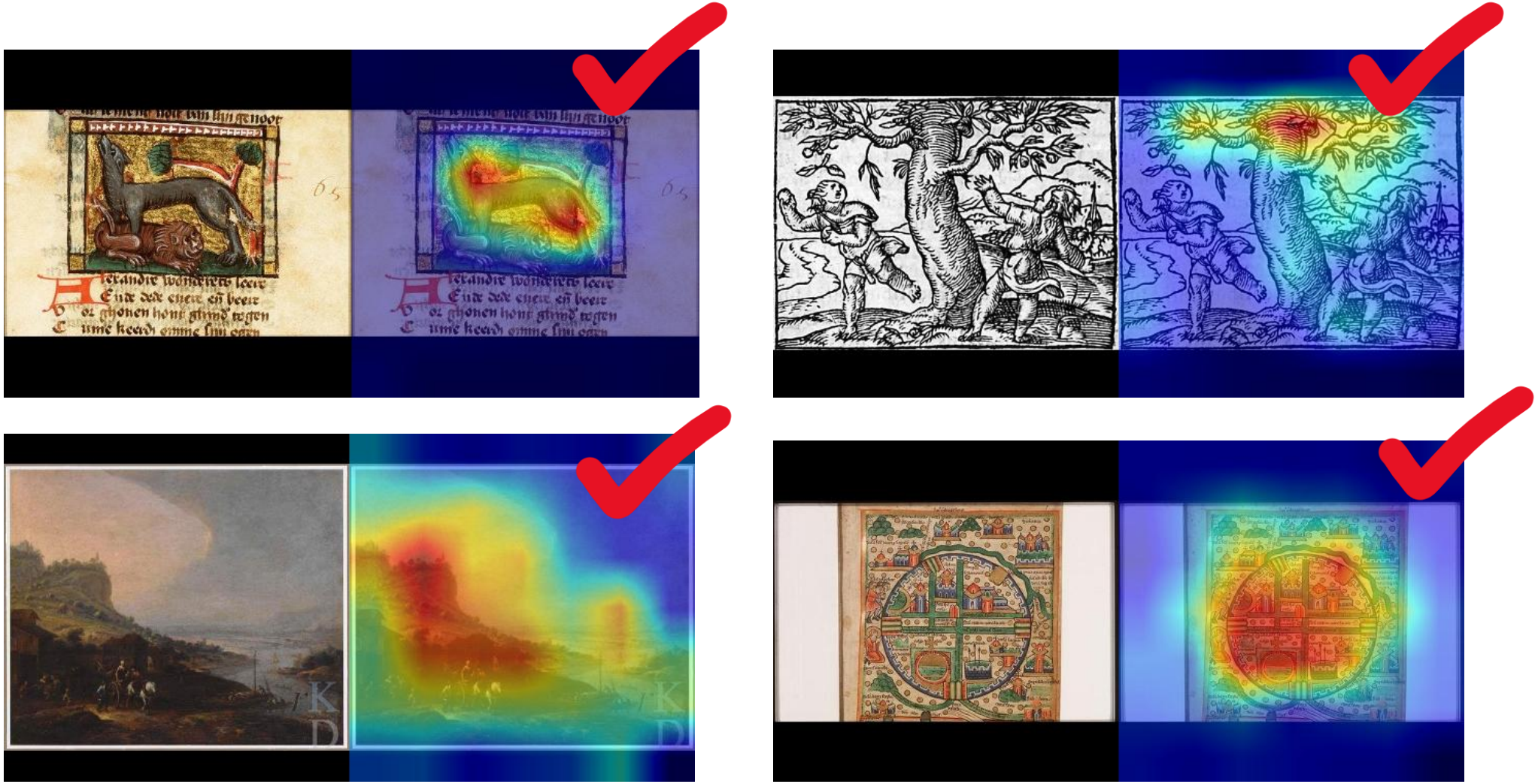
Live Demo: ONiT Explorer in Action

- Birds





Live Demo: ONiT Explorer in Action



Source: Vignoli, Michela et al. (2023) *Impact of AI: Gamechanger for Image Classification in Historical Research?* In Konferenzbeiträge der Digital History 2023, Berlin. <https://doi.org/10.5281/zenodo.8322398>.

Live Demo: ONiT Explorer in Action

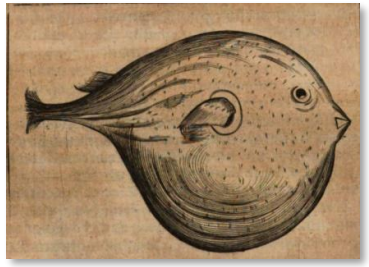
GS: “an image of animals”

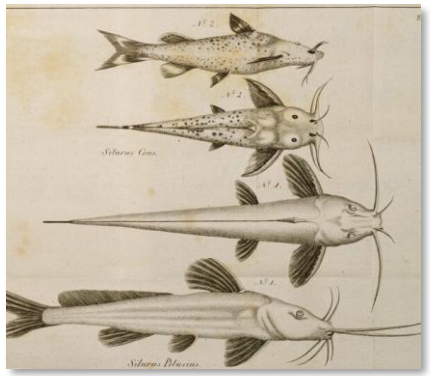
Most probable class: landscapes

Source: Vignoli, Michela et al. (2023) *Impact of AI: Gamechanger for Image Classification in Historical Research?* In Konferenzbeiträge der Digital History 2023, Berlin. <https://doi.org/10.5281/zenodo.8322398>.

Live Demo: ONiT Explorer in Action

- **Conclusion**
  - Similarity ranking gives a first semantic structuring
    → however, it is incomplete and imprecise
  - Helpful for discovering contents
    → hovewer, manual revision and quality control required
  - Bias:
    - Images resembling CV-training datasets are ranked higher
    - Concepts encoded through learning of formal visualisations and semantics from training data (*dataset bias*, *conceptual bias*)





Questions?

Short break – 15 min.

Search Images with AI – Hands-On Introduction to CLIP

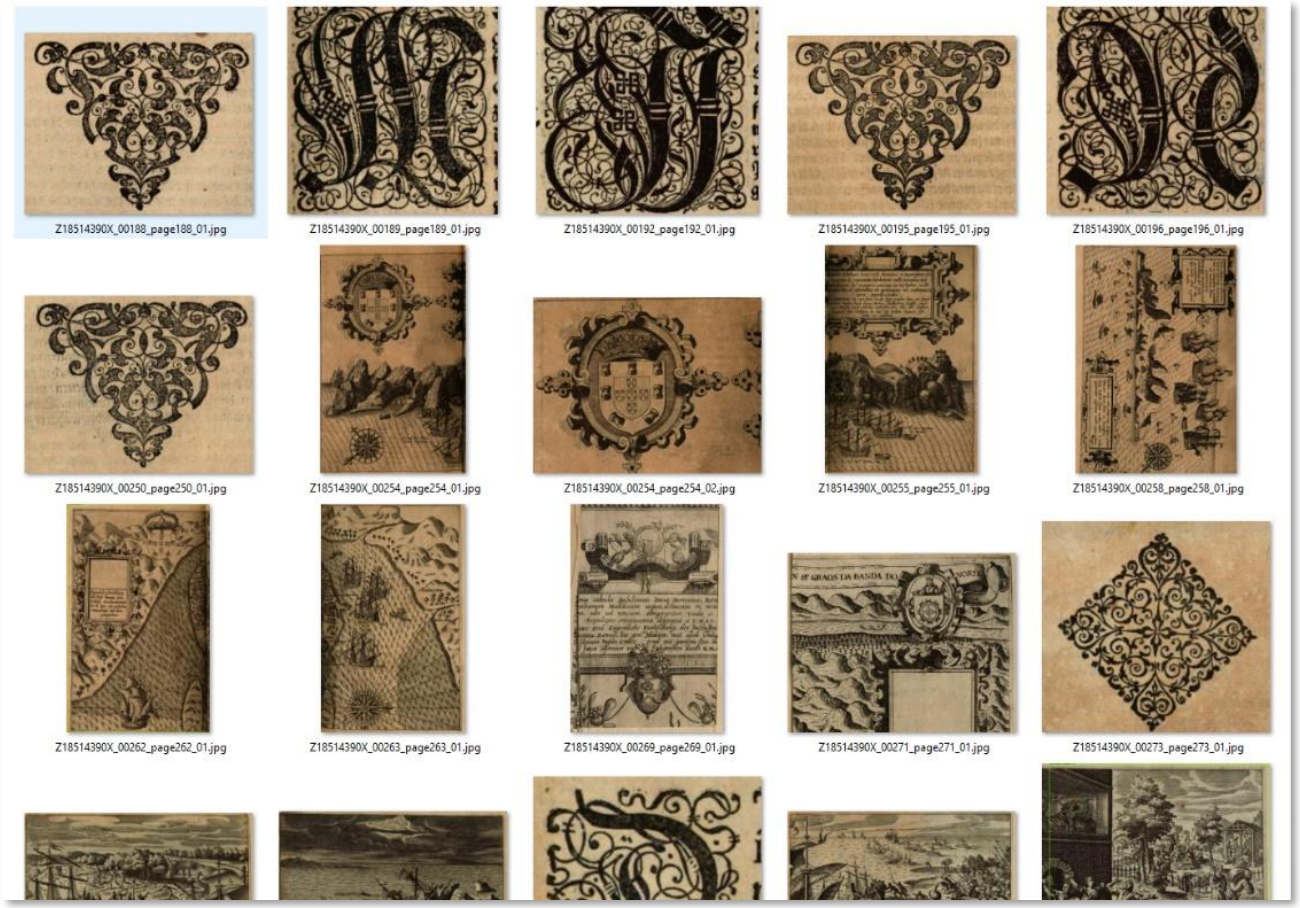
Agenda

- **Introduction to Vision–Language Models** (45 min.)
  Overview of core concepts behind models like CLIP and how they connect images and text
- **Live Demo: ONiT Explorer in Action** (20 min.)
  Explore a real-world application and see similarity search with image–text embeddings
- Short break (15 min.)
- **Hands-On Session:** Build Your Own Similarity Pipeline (55 min.)
  Implement image–text matching and test text query functionality (own images or ONiT dataset)
- Wrap-Up & Discussion (10 min.)
- 18:30 End of course

Hands-On Session: Build Your Own Similarity Pipeline

Login to VSC

- Link: https://jupyterhub.vsc.ac.at/
- Login with credentials provided to you



Search Images with AI – Hands-On Introduction to CLIP

Agenda

- **Introduction to Vision–Language Models** (45 min.)
  Overview of core concepts behind models like CLIP and how they connect images and text
- **Live Demo: ONiT Explorer in Action** (20 min.)
  Explore a real-world application and see similarity search with image–text embeddings
- Short break (15 min.)
- **Hands-On Session:** Build Your Own Similarity Pipeline (55 min.)
  Implement image–text matching and test text query functionality (own images or ONiT dataset)
- Wrap-Up & Discussion (10 min.)
- 18:30 End of course

Thank you!

Contact

Michela Vignoli

Scientist
AIT Austrian Institute of Technology
AI Factory Austria AI:AT

+43 664 88390661
michela.vignoli@ait.ac.at
michela.vignoli@ai-at.eu

AI Factory Austria AI:AT
Karl-Farkas-Gasse 22
1030 Vienna, Austria

training@ai-at.eu
info@ai-at.eu

ai-at.eu

@ai-factory-austria

Funded by

under discussion with

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).

Training & Skills Development

• AI Workshops • Courses • Webinars

ai-at.eu/trainings

Scan to explore
our events

AI Factory Austria AI:AT Consortium