

From Data to Meaning: Multimodal Semantic Search for Rococo-Era Prints

Thomas Hudcovic¹, Ines Röckl², Julian Jachmann², Gabriel Zachmann¹

¹University of Bremen, Institute of Computer Graphics & Virtual Reality, Germany

²University of Regensburg, Institute of Art History, Germany

Background & Motivation

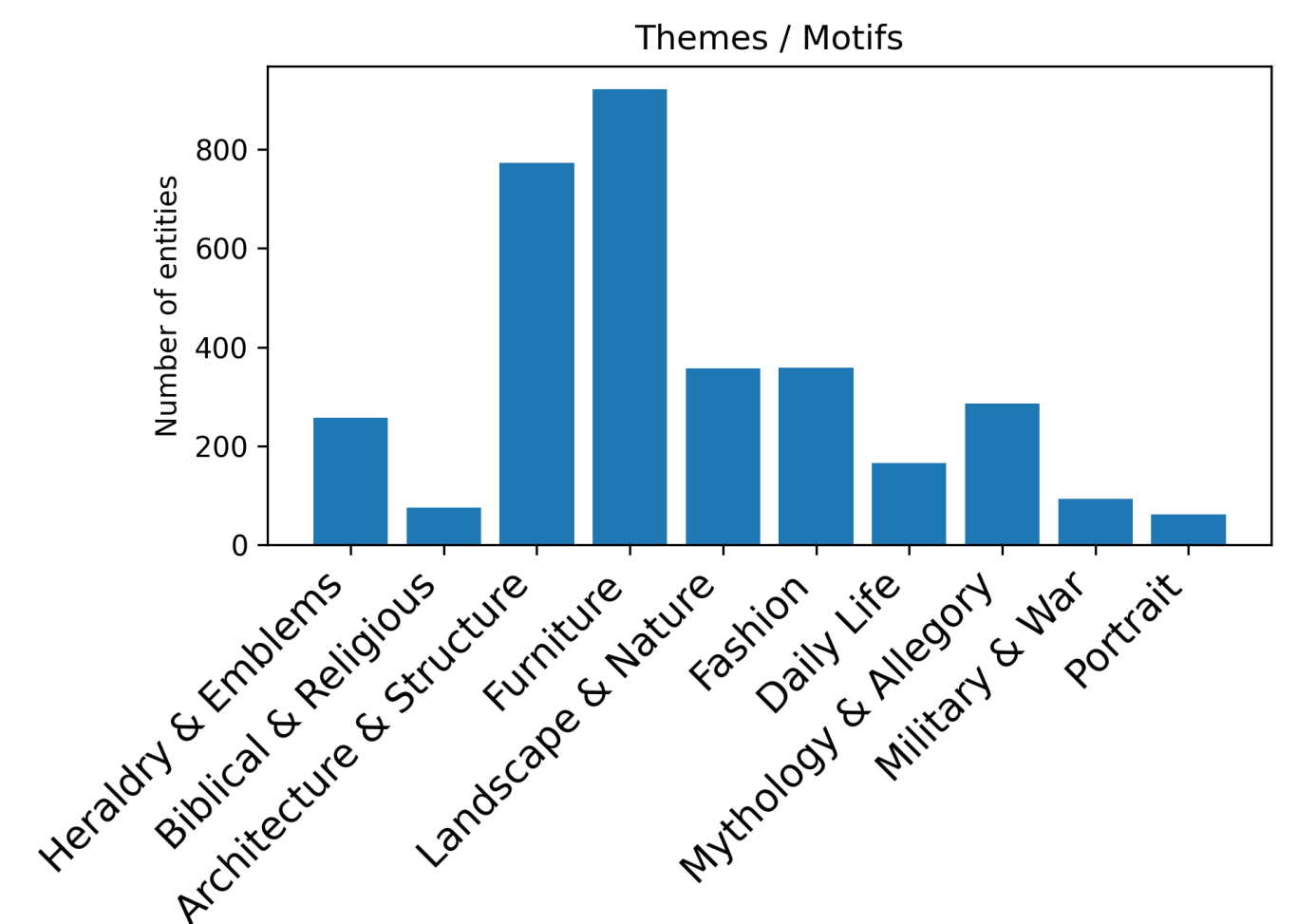
- The defining ornament of the Rococo period (ca. 1720s - 1780s)
- Rocaille are morphologically hard to capture, classify and define



Our Dataset

- 1605 high-resolution images
- Extensive metadata & expert text description & segmentation masks
- Metadata → Knowledge graph

```
{
  :
  "anmerkungen": "zwei Putti mit ausgeprägter Gestik in einem durchbrochenen Rocailleaufbau;
  feinerige Rocaille, mittelmäßiger Anteil an Voluten, dünn, recht großmaßstäblich, Sockel,
  Architektur",
  "anmerkungen_en": "two putti with pronounced gestures in a broken Rocaille structure;
  Fibrous rocaille, mediocre proportion of volute, thin, quite large, base, architecture",
  :
}
```

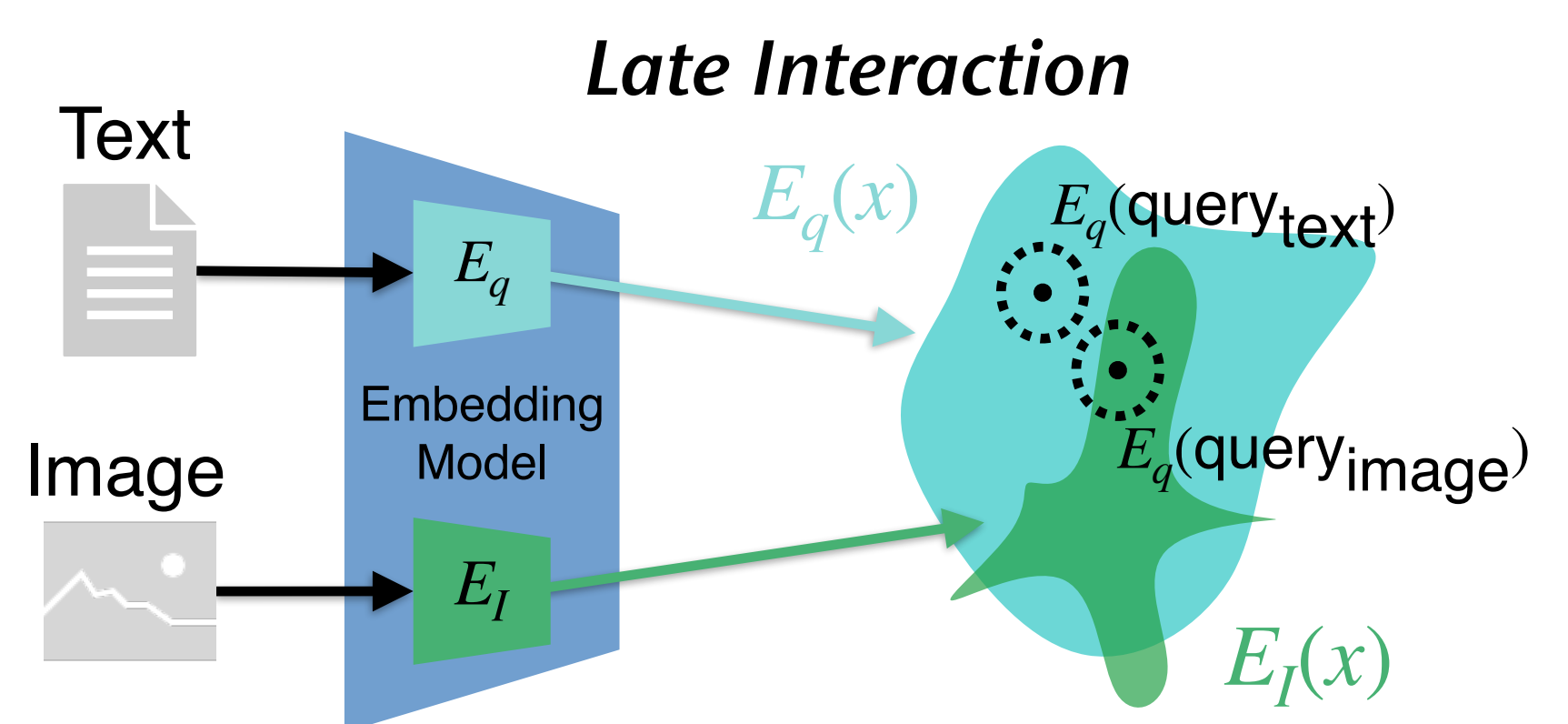


Multimodal Semantic Search

- Multimodal embedding by one encoder per modality (feature spaces are aligned)

1. Common practice: *Late Interaction*

- One query per modality (text, image), results are combined afterwards



2. Our novel approach: *Mid Interaction*

- Fusion of multimodal embeddings into *one* feature space (using gated linear attention)

$$\text{fuse}(q, T, \tau) = \text{norm}\left((1 - g)q + g \cdot \text{norm}\left(T^T \text{softmax}\left(\frac{Tq}{\tau}\right)\right)\right),$$

$$g = \sigma\left(10(\max_i T_i^T q - 0.5)\right), \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

- **Advantage: no new embeddings models necessary!**
no heavy Vision Language Models necessary.

