

École polytechnique de Louvain

Graph-Based retrieval of artworks from narrative inputs for memory-triggering applications

Author: **Safiya NOUIDI**
Supervisor: **Benoît MACQ, Sebastien LUGAN**
Reader: **Eric PIETTE**
Academic year 2025–2026
Master [120] in Mathematical Engineering

Abstract

Many deep-learning approaches to artwork retrieval from narrative inputs offer little interpretability: the reasoning that links a story to a retrieved artwork remains opaque. This master’s thesis proposes a hybrid, graph-based alternative that performs retrieval as an explicit two phase pipeline. The method relies on pre-trained neural representations for encoding, but carries out ranking and reasoning inside an explicit heterogeneous graph, using shortest paths (Dijkstra) and cosine similarity. In the first phase, an artwork is retrieved from the semantic content of the story (narrative input), where the emotion inferred from each artwork already contributes to the score; in the second phase, the best artwork from the pool is retrieved, according to a target emotion specified by the user, after being linked to a pre-defined set of nine emotions.

In a small exploratory study with ten participants, the method reached an 80% success rate on a controlled image retrieval task and a mean opinion score of 3.1/5 on emotional alignment, while satisfying the five-second inference-time set as a reasonable target. We also evaluated the algorithm with quantitative and qualitative benchmarks against CLIP. These preliminary results suggest that an interpretable, graph-based approach is a promising direction for artwork retrieval from narrative input, offering an explainable alternative to purely deep-learning pipelines.

Acknowledgements

I would like to express my gratitude to everyone who has supported me during this master's thesis and, more broadly, throughout my academic years.

I thank my supervisors, Benoît Macq and Sébastien Lugan, for the opportunity to work on such a meaningful project and for their feedback. I would also like to thank my friends and family for their support throughout my academic and personal journey.

Contents

List of Abbreviations	7
1 Introduction	8
2 State of the Art	11
2.1 Dementia and Alzheimer’s disease	11
2.2 Use of AI for Dementia	13
2.3 AI for Cultural Heritage and Art Retrieval	14
2.4 CLIP and Image-Text Representation	14
2.5 Embeddings and Vector Search	16
2.6 Previous Master’s Thesis	17
2.7 Graph Theory	17
3 Dataset	21
3.1 RMFAB Dataset	21
3.1.1 Dataset statistics	22
3.2 Hierarchical label extraction	22
3.2.1 Semantic similarity	23
3.2.2 Depth-weighted labels	23
3.3 Correction algorithm	29
4 Story-to-Art Association	31
4.1 Mission statement	31
4.2 Phase One: Semantic Retrieval	32
4.3 Formulas used and visualisation	35
4.4 Phase Two: Emotional Alignment	45
4.5 Illustrative examples of algorithm outputs	47
5 Interpretability and Bias Analysis	50
5.1 Distinguishing the relevant notions	50
5.2 Interpretability for whom?	51

5.3	The explanation-reality gap	51
5.4	Bias in cultural heritage AI	52
5.5	Auditability as an advantage of the graph	53
6	Evaluation Protocol and Benchmarks	54
6.1	Participants and procedure	54
6.2	Recovering the target image	54
6.3	Comparison with baseline and emotional alignment	56
6.3.1	Baseline algorithm	57
6.3.2	Mean Opinion Score (MOS) scale	57
6.4	Benchmark comparison with CLIP	58
6.5	Qualitative comparison with CLIP	62
7	Limitations and Future Perspectives	67
7.1	Limitations	67
7.1.1	Loss of syntactic structure through tokenisation	67
7.1.2	Nature of the iconographicInterpretation field	67
7.1.3	Variable emotional performance	68
7.1.4	Small testing samples	68
7.1.5	Heavy dependence on heuristic thresholds	69
7.1.6	Risk of triggering distressing memories	69
7.1.7	Ethical considerations	70
7.2	Future Perspectives	70
7.2.1	Higher-level and application directions	71
7.2.2	Technical improvements	72
8	Conclusion	74
	Appendix A	81
	Appendix B	88

List of Figures

2.1	Visualization of an example of Dijkstra’s Algorithm	19
3.1	Jacques d’Arthois, <i>Paysage d’hiver</i>	24
3.2	Extracted subjectTerms hierarchy of <i>Paysage d’hiver</i>	25
3.3	Peter Paul Rubens et atelier, <i>Le martyre de saint Liévin</i>	26
3.4	Extracted subjectTerms hierarchy of <i>Le martyre de saint Liévin</i>	27
3.5	Extracted iconographicTerms hierarchy of <i>Le martyre de saint Liévin</i>	28
4.1	Step 1: Text preprocessing of the artwork descriptions.	39
4.2	Step 2: Caption-label extraction and filtering.	40
4.3	Step 3: Hierarchical-label extraction from the SubjectMatter fields.	41
4.4	Step 4: Counting of label co-occurrences and edge weighting.	42
4.5	Step 5: Construction of the multimodal graph.	43
4.6	Step 6: Online processing of a query memory and retrieval.	44
4.7	Result of Phase 1 for the input story of Example 1.	47
4.8	Result of Phase 2 for the input story of Example 1.	48
4.9	Result of Phase 1 for the input story of Example 2.	49
4.10	Result of Phase 2 for the input story of Example 2.	49
6.1	The five images used for the controlled retrieval test.	55
6.2	Indexing-time comparison between the proposed graph-based algorithm and CLIP.	60
6.3	Inference time per query (emotion filter excluded) for the proposed algorithm and CLIP, on a logarithmic scale.	61
6.4	Memory Δ RSS during indexing for the proposed algorithm and CLIP.	62
6.5	Qualitative comparison for Query 1 (<i>un portrait d’une femme assise</i>): top-five artworks from the proposed method (top row) and CLIP (bottom row).	63
6.6	Qualitative comparison for Query 2 (<i>une scène de bataille historique</i>): top-five artworks from the proposed method (top row) and CLIP (bottom row).	64

6.7	Qualitative comparison for Query 3 (<i>un paysage avec des montagnes et un lac</i>): top-five artworks from the proposed method (top row) and CLIP (bottom row).	65
6.8	Qualitative comparison for Query 4 (<i>une nature morte avec des fruits</i>): top-five artworks from the proposed method (top row) and CLIP (bottom row).	66

List of Tables

1	List of Abbreviations	7
3.1	Main characteristics of the filtered RMFAB dataset	22
6.1	Results of the first (controlled retrieval) protocol	56

List of Abbreviations

Abbreviation	Definition
AD	Alzheimer’s Disease
AI	Artificial Intelligence
CLIP	Contrastive Language-Image Pre-training
CST	Cognitive Stimulation Therapy
DL	Deep Learning
FAISS	Facebook AI Similarity Search
GNN	Graph Neural Network
IDF	Inverse Document Frequency
RMFAB	Royal Museums of Fine Arts of Belgium
RT	Reminiscence Therapy
SG	Serious Game
STS	Semantic Textual Similarity
ViT	Vision Transformer
YOLO	You Only Look Once

Table 1: List of Abbreviations

Chapter 1

Introduction

To motivate graph-based retrieval of artworks from narrative inputs, consider memory-triggering applications. One application of the tool developed is reminiscence-based interaction for people with neurodegenerative diseases and dementia, of which Alzheimer’s disease (AD) is the most common form, as discussed further in Chapter 2. AD and other forms of dementia represent one of the major public health challenges of our time, affecting millions of people worldwide and placing a heavy burden on patients, families, and caregivers. In the absence of curative treatments, non-pharmacological interventions such as reminiscence and art therapy have gained increasing attention for their ability to stimulate autobiographical memory and support emotional well-being. Visual artworks, in particular, can act as powerful memory triggers, evoking personal recollections and initiating meaningful interaction between patients and their relatives.

In this context, this thesis explores how artificial intelligence can help connect personal narratives to evocative artworks. Given a short story describing a personal memory, the goal is to retrieve an artwork that resonates with it. While recent deep-learning (DL) methods, CLIP in particular, have shown strong performance on such retrieval tasks, they have some limitations, as discussed in Section 2.4. DL models are often regarded as *black boxes* and typically require large amounts of data.

The present work aims to propose an interpretable and efficient alternative to the DL strategies used for artwork retrieval from narrative inputs. It relies on explicit relations between the different modalities, using a graph-based method.

The main contributions of this master’s thesis are the following:

- the construction of a heterogeneous, multilayer graph built from the data of the *Royal Museum of Fine Arts of Belgium* (RMFAB), whose nodes include artwork images, their captions, textual labels extracted from the narrative, hierarchical labels from the museum’s iconographic classification

(see Chapter 3), and inferred emotions;

- a method for mapping the words of a narrative onto the nodes of this graph, based on cosine similarity between pre-trained sentence embeddings;
- a retrieval score combining shortest paths (Dijkstra), inverse document frequency (IDF), and hierarchical depth, in which the emotion inferred from each artwork already contributes to the ranking;
- a second, emotion based selection step that chooses the best candidate in the pool according to a target emotion specified by the user.

Together, these components form a two-phase retrieval procedure that jointly accounts for semantics and emotions.

It should be emphasised that the proposed method is not independent of deep learning: pre-trained models play an essential role in it. Sentence-CamemBERT provides the textual embeddings, ResNet50 provides visual features for artworks without a textual description, and FAISS performs efficient nearest-neighbour search over these learned representations. The novelty of our approach lies not in discarding these models, but in using their representations while performing the *reasoning and ranking* inside an explicit graph structure.

Interpretability is central to this work, as discussed in Chapter 5. Being able to explain why a given artwork is recommended, by following the path from the input story to the output image, is a key asset of the approach and a property that purely deep-learning pipelines rarely provide.

As said before, an application of the tool developed is reminiscence-based interaction for people with neurodegenerative diseases. The framework is therefore presented as a research prototype motivated by possible future applications in which patients and their relatives could share personal memories in written form. It should be noted that the system has not yet been evaluated with patients, caregivers or therapists, and that any clinical use would require further validation.

This work is organised as follows:

- the second chapter presents the state of the art;
- the third chapter describes in detail the structure of the RMFAB dataset used throughout the project;
- the fourth chapter presents the main algorithm, in which a relative of a person with neurodegenerative diseases submits a textual story describing a personal memory and the system retrieves an evocative artwork;
- chapter five is a small analysis of interpretability and the different types of biases;

- an evaluation of the algorithm with qualitative and quantitative benchmarks is described in chapter six;
- this thesis ends with a discussion on limitations and future perspectives and finally, the conclusion.

Chapter 2

State of the Art

This chapter presents the background necessary to situate this master’s thesis within existing research. We successively reviews the medical context, the relevant artificial-intelligence methods (with a focus on CLIP), and the principles of graph theory on which the proposed framework relies.

2.1 Dementia and Alzheimer’s disease

The most common form of dementia is Alzheimer’s disease (AD). It represents a major and growing global health challenge. Due to population aging, the number of new diagnoses is expected to rise significantly in the coming years, highlighting the urgent need for innovative tools to support both prevention and care. According to recent estimates, AD accounts for approximately 60-80% of dementia cases worldwide [1].

Clinically, AD is characterised by progressive cognitive decline, including memory loss, language impairments, perceptual deficits, and reduced executive function. These impairments are not isolated: because cognitive processes rely on interconnected neural networks, deterioration in one domain often propagates to others. In addition to cognitive symptoms, patients frequently experience behavioural and psychological changes such as apathy, irritability, depression, and mood fluctuations. Emotional states play a role in modulating brain activity and cognition, and may further influence disease progression [2].

The burden of AD extends beyond patients to caregivers, who often experience high levels of stress, depression, and physical health issues. This makes AD not only a neurological condition but also a major social and public health concern [3]. This is one reason explaining the chose to include the patient’s relatives in the loop.

From a biological perspective, AD is associated with multiple pathological

mechanisms, including the accumulation of amyloid- β plaques and tau protein abnormalities. Its development is influenced by a combination of genetic and environmental factors, such as hypertension, obesity, diabetes, and depression [4, 5]. Despite advances in pharmacological treatments, current drugs remain limited in their ability to stop or reverse disease progression [6]. As a result, non-pharmacological interventions have gained increasing attention.

Lifestyle factors such as physical activity and cognitive engagement have been shown to positively impact brain plasticity and cognitive function [7]. In particular, Cognitive Stimulation Therapy (CST) has been demonstrated to improve cognitive functioning, communication, and quality of life in people with dementia [8]. These approaches highlight the importance of care strategies that go beyond purely medical treatment.

Reminiscence Therapy

Among non-pharmacological interventions, Reminiscence Therapy (RT) plays a central role. It consists of stimulating autobiographical memory through exposure to familiar cues such as images, music, or personal objects. RT is particularly suited to AD because long-term memories are often preserved longer than recent ones.

Clinical studies have shown that RT can improve cognitive performance, reduce depressive symptoms, and enhance daily functioning in patients with moderate AD. Beyond cognitive benefits, RT also promotes social interaction and emotional well-being, making it a valuable tool for both patients and caregivers [9].

Art Therapy and Visual Stimulation

Art therapy represents a natural extension of reminiscence-based approaches. Visual artworks can act as powerful memory triggers, engaging both cognitive and emotional processes. Unlike language, which is primarily processed in specific brain regions, visual art has been reported to engage distributed neural networks that may remain relatively preserved in the presence of neurodegeneration [10].

Recent studies have suggested that interventions based on visual arts can improve cognitive functioning and may even influence biological markers associated with aging [11]. These findings suggest that artworks constitute a particularly rich resource for therapeutic applications.

Serious Games

Another promising direction is the use of Serious Games (SGs), which combine entertainment with therapeutic or cognitive objectives. Early work already showed

that computer-based games could improve self-esteem and social interaction in elderly populations. More recent systems integrate interactive technologies to stimulate cognitive functions and improve emotional well-being. Serious games designed for dementia patients can target specific cognitive abilities such as memory, attention, and problem-solving [12]. Usability remains a key challenge, requiring simple interfaces, adaptive difficulty, and accessibility considerations tailored to elderly users [13, 14].

However, the topic of serious games goes beyond the scope of this thesis; it would be an interesting direction for future work, as described in section 7.

2.2 Use of AI for Dementia

Artificial Intelligence (AI) has emerged as a powerful tool across multiple aspects of dementia (AD in particular) research and care. At the diagnostic level, it supports early detection through neuroimaging analysis, risk prediction by identifying modifiable factors, and the development of personalised treatment strategies.

Beyond diagnosis, AI systems have increasingly been explored for direct patient interaction and caregiver support. Large Language Models (LLMs), in particular, have shown promising results in facilitating meaningful conversations, providing cognitive stimulation, and assisting caregivers in daily tasks. Large Multimodal Models (LMMs) extend this capability further by combining textual and visual inputs within unified architectures [15], making them promising candidates for applications at the intersection of healthcare and cultural heritage [16].

These systems reflect a broader shift toward more interactive, technology-assisted care. However, most existing AI approaches, including LLMs and LMMs, rely heavily on deep learning (DL), which comes with important limitations in this context. DL models typically require large annotated datasets and often lack interpretability, making it difficult to explain their outputs.

Graph-based approaches offer a complementary perspective that addresses these shortcomings. While they have been used to model brain connectivity and disease progression, their potential for therapeutic applications remains largely underexplored, as discussed in Section 2.7. By explicitly modelling relationships between entities, graph-based methods provide greater transparency and can integrate multiple data sources within a unified, interpretable framework, motivating the approach proposed in this work.

2.3 AI for Cultural Heritage and Art Retrieval

AI in Museums

Cultural heritage institutions have increasingly adopted AI techniques to manage, annotate, and explore their collections. Automatic image analysis methods, such as object detection models like YOLO, enable the extraction of visual elements from artworks [17, 18]. These methods can enrich data and facilitate large-scale indexing of collections.

Despite these advances, applying AI to artworks remains challenging. Models are typically trained on natural images, leading to performance degradation when applied to artistic representations, which exhibit high stylistic variability. This limitation motivates the exploration of alternative approaches that rely less on purely visual recognition.

Exploratory Search

In museums, users may not have a precise target but instead seek inspiration or emotional resonance. Traditional search systems assume a well-defined query and a single correct answer. In contrast, exploratory search addresses open-ended information needs, where users refine their goals through interaction with the system.

Previous work emphasised the importance of supporting discovery and evolving information needs [19, 20]. These ideas are closely aligned with the objectives of art-therapy applications, where the goal is not only retrieval accuracy but also meaningful engagement.

2.4 CLIP and Image-Text Representation

Definition and motivation

Recent advances in multimodal learning have enabled joint representations of images and text. CLIP (Contrastive Language-Image Pre-training) learns aligned embeddings by maximising similarity between matching image-text pairs [21]. The pre-training task, which predicts which caption corresponds to which image, is an efficient and scalable way of learning high-level visual representations. This approach leverages the internet, which is full of images paired with captions [21], which removes the need to annotate them manually. Moreover, using natural language for supervision allows the model to connect text and visual representations.

How it works

CLIP trains a pair of models, one dedicated to understanding images and the other to understanding text, using a contrastive objective. This method has enabled applications in many fields, including cross-modal retrieval, text-to-image generation and aesthetic classification. In practice, the two encoders (image and text) project their inputs into a shared embedding space. Within this space, the vector of an image and the vector of its associated text must be close to one another, while being distant from the representations of other, unpaired images and texts. The similarity between two vectors is computed as the cosine similarity of their normalised versions.

On the text encoder side, the models used are typically Transformers. In the original OpenAI report, a Transformer with 63 million parameters (12 layers, 512-wide, 8 attention heads) is used with lowercase BPE (Byte Pair Encoding) on a vocabulary of 49,152 tokens, with a context length capped at 76 for efficiency reasons. For the image encoder, several architectures were explored, ranging from variants of ResNet to Vision Transformers (ViT). The ViT-L/14 model, trained and then fine-tuned at 336×336 resolution via FixRes, proved to be the best-performing model. The training data is a key factor in CLIP's success: the dataset used, known as WebImageText (WIT), contains 400 million image-caption pairs collected from the internet [21].

Zero-shot transfer capabilities

One of CLIP's strengths lies in its zero-shot transfer capabilities: the model can be applied to new tasks without any specific retraining. By not optimising directly for a given benchmark, the system is more representative and robust, reducing the distribution robustness gap by up to 75% compared to supervised models [21]. This generalisation capability is particularly valuable in fields where annotated data is scarce, such as cultural heritage [22].

Application to artwork retrieval

In addition to zero-shot classification, the approach can be used for content-based image retrieval, outperforming many competing approaches. One advantage of the method is that it can be applied to small datasets, thanks to CLIP's extensive pre-training, enabling its deployment on relatively small collections such as those found in small to medium-sized museums [22].

CLIP’s limitations

Despite its strong performance, CLIP exhibits several important limitations that motivate the exploration of alternative approaches, such as the one proposed in this thesis.

First, several studies have shown that CLIP lacks satisfactory interpretability. This shortcoming stems partly from the inherent nature of Vision Transformers (ViTs) and partly from the non-smoothness introduced by false negatives during contrastive training [23]. This opacity is especially problematic in sensitive domains such as art therapy for individuals with Alzheimer’s disease, where the ability to explain recommendations is arguably both an ethical and a therapeutic requirement. In contrast, the graph-based framework proposed in this work makes it possible to trace the full path from the input narrative to the recommended artwork.

Second, in practice, CLIP handles only very short text inputs (around 20 tokens effectively). This constraint not only limits the usefulness of its text encoder but also affects the image encoder’s ability to capture fine-grained details and relationships. During training, the input text typically consists of brief summaries containing only primary attributes, which encourages the model to focus on the most salient elements while ignoring secondary details [24]. In the context of this thesis, where inputs consist of potentially long and nuanced personal narratives, this limitation becomes non-negligible.

CLIP also relies primarily on global semantic alignment between text and image, without any explicit mechanism for modelling emotional content. As a result, it cannot inherently distinguish between descriptions that are semantically similar but emotionally opposite. For art-therapy applications, where the emotional valence of the selected artwork may plausibly influence the user’s experience, this is a potentially important limitation.

In this work, CLIP is used as a reference benchmark, both in terms of computational efficiency and qualitative performance, to evaluate the proposed framework. This comparison is appropriate, as CLIP represents one of the most directly applicable multimodal deep-learning approaches to text-to-image retrieval. The limitations outlined above are precisely the gaps that the graph-based system introduced in this thesis aims to address.

2.5 Embeddings and Vector Search

Efficient similarity search in high-dimensional spaces is enabled by tools such as FAISS, which supports scalable indexing and retrieval [25, 26].

FAISS (Facebook AI Similarity Search) is a library that allows fast nearest-neighbour search over very large collections of vectors. The idea is the following:

if each item (image or text) has been turned into a vector (an embedding) that captures its meaning, then finding the items most similar to a new query by comparing it one by one to every single item becomes prohibitively slow when the collection is large. FAISS addresses this by organising the vectors so that searching through millions of them becomes fast, using specialised indexing techniques that group similar vectors together and allow the search to skip items that are clearly poor matches. In the context of this thesis, FAISS is used to quickly match the words from a user’s story to the closest related concepts already stored in the graph, making the retrieval process fast enough for interactive use.

2.6 Previous Master’s Thesis

This master’s thesis builds on two previous ones.

On the one hand, Victor Rijks [27] investigated the use of multimodal retrieval models to automatically select artworks from museum collections, addressing challenges related to data scarcity through synthetic data generation.

On the other hand, Silva and Bejjani [28] proposed ArtEvoke, a system that links personal narratives to artworks using LLMs, multimodal models, and vector search. Their platform enables both memory reconstruction and exploratory interaction.

While these approaches demonstrate the potential of AI in this domain, they rely primarily on deep learning and embedding-based similarity. The present work extends this line of research by introducing a graph-based framework that explicitly models relationships between narratives, emotions, and artworks, providing improved interpretability and reducing dependence on large annotated datasets.

2.7 Graph Theory

Heterogeneous Graphs for Multimodal Representation

A graph is formally defined as $G = (V, E)$, where V is a set of nodes and E a set of weighted edges encoding relationships between them. Each node $v \in V$ may carry a feature vector (of dimension d) $\mathbf{x}_v \in \mathbb{R}^d$, and each edge e_{ij} a weight $w_{ij} \in \mathbb{R}$. When nodes and edges belong to different types, the structure is referred to as a *heterogeneous graph*: a representation particularly suited to multimodal data, where entities of distinct nature (images, texts, labels) coexist and interact through diverse semantic relations [29].

This property is directly relevant to the domain of artwork analysis. Jiang et al. [30] showed, in the context of painting caption generation, that a heterogeneous knowledge graph integrating nodes for images, authors, titles, techniques, and historical periods can capture stylistic and contextual relationships that a purely

visual model cannot represent. Their KALE system, which constructs such a multilayer graph from the metadata of the SemArt dataset, significantly improves the quality of generated descriptions compared with existing approaches. This work illustrates the added value of an explicitly structured representation of museum collections, precisely the principle adopted here using the Royal Museums of Fine Arts of Belgium dataset.

Propagation and Similarity through Shortest Paths

In Graph Neural Networks (GNNs), the core mechanism is *message passing*: each node iteratively aggregates representations from its neighbours, allowing it to benefit from information located several hops away. Wu et al. [29] distinguish two major families of convolutional GNNs: *spectral* methods, which operate in the frequency domain through the graph Laplacian, and *spatial* methods, which directly aggregate representations from a node’s topological neighbourhood. In both cases, the foundational principle remains the same: two nodes connected by a short path in the graph are considered semantically close.

The system proposed in this work does not train a parametric GNN; instead, we adopt this same principle in an explicit manner. Rather than learning representations through backpropagation, we use Dijkstra’s algorithm to compute shortest-path distances between the tokens of an input narrative and the image nodes in the graph. An image node receives a high score when it is reachable through short paths from many query tokens (formally analogous to the notion of higher-order neighbourhoods in GNNs).

Dijkstra’s algorithm

Dijkstra’s algorithm finds the shortest path from a single source node to all other nodes in a graph with non-negative edge weights. It maintains a set of unmarked nodes and repeatedly selects the closest unmarked node, marks it as visited, and then updates the distances to its neighbours if a shorter path is found through it:

$$w_j \leftarrow \min\{w_j, w_k + d_{kj}\},$$

where w_j is the tentative path weight assigned to node j and d_{kj} is the direct edge weight (distance) between node k and node j . This process repeats until all nodes are marked, at which point the shortest paths from the source to every other node are known, and each path can be reconstructed by backtracking from a destination node to the origin [31].

To make it clearer, on figure 2.1, you can find a visualization of an example of Dijkstra’s Algorithm coming from [32].

Execution of Dijkstra's algorithm

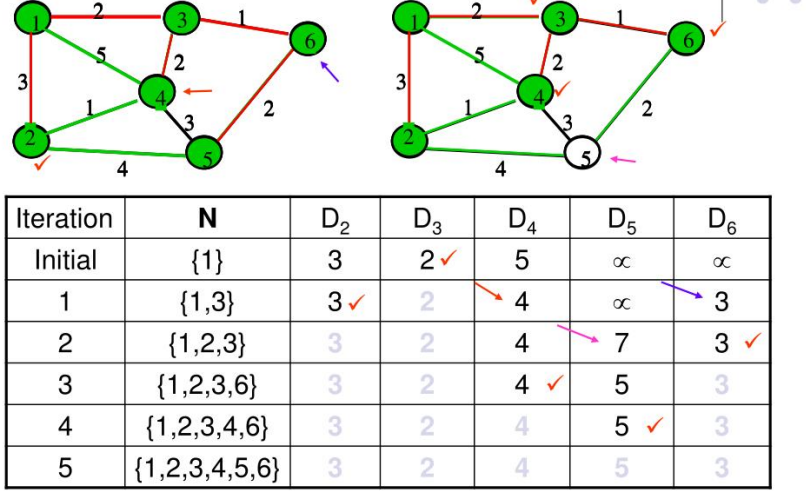


Figure 2.1: Visualization of an example of Dijkstra's Algorithm

Co-occurrence Weighting and Hierarchical Structure

Gao et al. [33], in their survey of GNN-based recommender systems, identify three main motivations for adopting graph-based approaches. The first is **high-order connectivity**: GNNs capture relationships between entities that are not directly connected, revealing latent similarities that pairwise comparisons cannot detect. The second is the exploitation of **structural properties** of data whose topology is itself semantically meaningful. The third is the ability to provide **enriched supervision signals**, even in sparse-data scenarios.

These three motivations apply directly to our framework. In our graph, edge weights come from three distinct sources, which is kept conceptually separate. *Structural* edges have fixed weights and encode the composition of an artwork (for instance, the link between an image and its caption, or between an image and a hierarchical label). *Co-occurrence* edges are the only ones derived from the data during the offline construction of the graph: two labels that frequently appear together in artwork descriptions are linked by a stronger edge, i.e. a shorter path. *Visual-similarity* edges connect artworks whose ResNet50 features are close, and are used to attach artworks that have no textual description. As detailed in Chapter 3, the hierarchical structure of our data is also exploited by assigning higher weights to more specific labels (greater depth d), following the idea that discriminative terms should contribute more strongly to the similarity score.

Interpretability as a Central Requirement

Finally, Gao et al. [33] emphasise that explainability has become a central objective of modern recommender systems: a system must not only produce accurate results, but also justify them. The graph structure makes it possible to trace, step by step, the path from a token in the input narrative to the retrieved artwork, through intermediate nodes such as captions, labels, and emotions. Interpretability, explainability and biases are further discussed in Chapter 5.

As noted above, this transparency constitutes a decisive advantage over deep-learning approaches such as CLIP, where the path from input to output is encoded within millions of parameters without any interpretable intermediate structure, a limitation discussed in Section 2.4.

Chapter 3

Dataset

This chapter provides an overview of the dataset on which the project is based: the online library of the *Royal Museums of Fine Arts of Belgium* (RMFAB), which we were fortunate to have access to and whose data format follows the *Thesaurus Garnier* [34]. What follows describes only the components of the dataset relevant to our work.

3.1 RMFAB Dataset

The main entry considered for each artwork is **SubjectMatter**, which is divided into four fields:

- **subjectTerms**: objects present in the artwork.
- **iconographicTerms**: iconographic elements present in the artwork.
- **conceptualTerms**: abstract concepts represented in the artwork.
- **iconographicInterpretation**: a continuous textual description of the artwork.

However, not all artworks possess all four fields. Some are described only through **subjectTerms**, for example.

Each sub-entry, except *iconographicInterpretation*, follows a structured notation defined by the following markers:

- (start of a group
-) end of a group
- ; sibling object or new parent object

- : child of the previous object

This structure allows us to exploit the hierarchical nature of the **SubjectMatter** field and integrate it into our graph-based algorithm. We implemented a correction algorithm, as the RMFAB data were manually entered and may contain errors. This correction relies on lexical similarity and will be explained in Section 3.3.

3.1.1 Dataset statistics

Before describing the hierarchy construction, the main characteristics of the RMFAB dataset are presented in Table 3.1. These figures describe the full filtered collection as characterised in the preceding work of Rijks [27], on which the present dataset is based.

Characteristic	Value
Raw artwork records provided	6 869
Records with at least one usable image	5 301
Records with <code>iconographicInterpretation</code>	516 (9.7%)
Records with <code>subjectTerms</code>	3 610 (68.1%)
Records with <code>iconographicTerms</code>	2 817 (53.1%)
Records with <code>conceptualTerms</code>	670 (12.6%)
Unique labels across all SubjectMatter fields	3 860
Language of the descriptions	French

Table 3.1: Main characteristics of the filtered RMFAB dataset

Subset used in this work. For practical reasons, the experiments reported in this thesis do not operate on the entire collection at once. From the filtered dataset, it samples the artworks that possess an `iconographicInterpretation` (the text-eligible records), together with a smaller pool of artworks that have no textual description but do have an image (up to one fifth of the former). The latter are attached to the graph through visual similarity edges, as described in Chapter 4. Sampling is performed with a fixed random seed so that the constructed graph is reproducible across runs.

3.2 Hierarchical label extraction

A program was implemented to exploit the hierarchical nature of the labels in the main algorithm, described in Chapter 4. The logic of the implementation is as follows:

- it identifies the main groups as the text appearing before the opening parenthesis;
- a main group ends with the closing parenthesis;
- it splits the content inside the parentheses on the semicolon separator ; to obtain individual tokens;
- for each token:
 - if the token contains a colon :, the part before the colon is a parent node and the part to the right of the colon is a child node, with the possibility of having grandchildren (children of children);
 - if the token contains no colon :, it uses semantic similarity, computed with *sentence-camembert-large*, to decide whether it is a child of the most recent parent or a new parent.

3.2.1 Semantic similarity

According to previous work [35], a strong model for assessing semantic similarity in French is *sentence-camembert-large*. In Table 12 of that paper, *sentence-camembert-large* reaches a score of 0.86 on the Semantic Textual Similarity (STS) task. It should be emphasised that this figure is a *correlation* between the model’s similarity scores and human similarity judgments, and not a proportion of correct decisions: it does not mean that the model "agrees with humans 86% of the time". A value close to 1 nonetheless indicates a strong alignment between the model’s similarity ordering and human perception, which is what our hierarchy construction step relies on.

When no colon is present, a token is attached to the most recent parent if its cosine similarity to that parent exceeds a fixed threshold; otherwise it is treated as a new parent. This threshold was set empirically to 0.45. A manual inspection of a sample of these corrections confirmed that the resulting parent-child assignments were plausible.

3.2.2 Depth-weighted labels

The depth of a label is a number corresponding to its position in the hierarchy:

- depth 0 is for the main groups;
- each direct child of a main group, i.e. each parent, is associated with depth 1;
- depths 2 and 3 are for children and grandchildren respectively.

As described in Chapter 4, this depth is used to weight the different edges of the graph.

Example 1

Figure 3.1 shows a painting by Jacques d'Arthois (figures attributed to Pieter Bout) named "Paysage d'hiver". Figure 3.2 shows the corresponding hierarchy for **subjectTerms**. There are two main groups (*paysage* and *scène*). In *paysage* there are seven parents but no children, while in *scène* there are seven parents and four children, distributed between *animal* and *outil*. The associated **subjectTerms** field was the following:

paysage (forêt ; arbre ; chemin ; étang ; village ; maison ; église) ; scène (homme ; femme ; enfant ; jeu ; charette ; animal : cheval ; chien ; oiseau ; outil : hache)



Figure 3.1: Jacques d'Arthois, *Paysage d'hiver*.

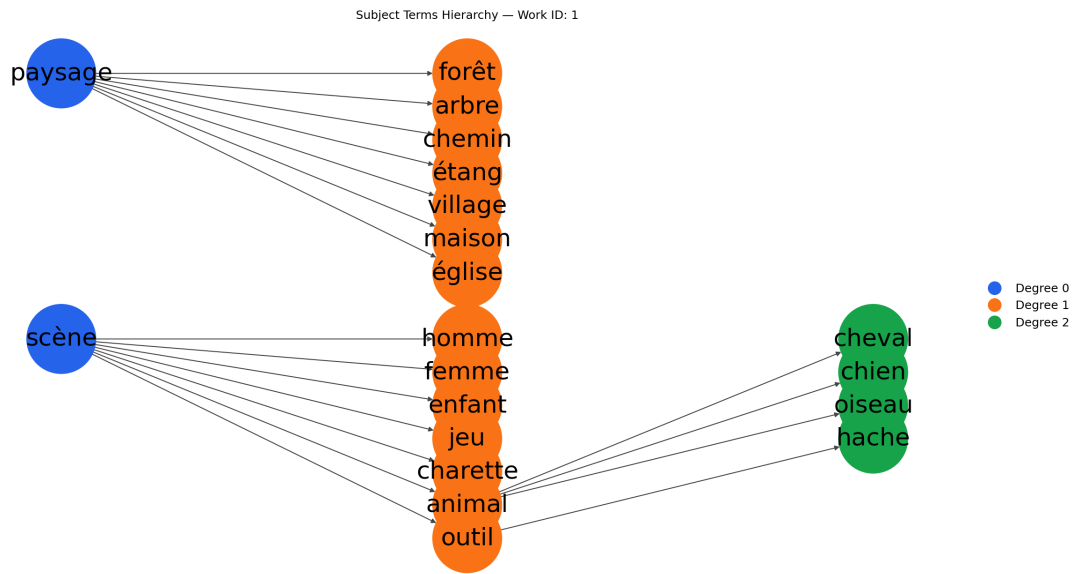


Figure 3.2: Extracted *subjectTerms* hierarchy of *Paysage d'hiver*.

The output hierarchy was:

```

'paysage': ['forêt', 'arbre', 'chemin', 'étang', 'village', 'maison', 'église'], 'forêt': [],
'arbre': [], 'chemin': [], 'étang': [], 'village': [], 'maison': [], 'église': [], 'scène':
['homme', 'femme', 'enfant', 'jeu', 'charette', 'animal', 'outil'], 'homme': [],
'femme': [], 'enfant': [], 'jeu': [], 'charette': [], 'animal': ['cheval', 'chien', 'oiseau'],
'cheval': [], 'chien': [], 'oiseau': [], 'outil': ['hache'], 'hache': []

```

Example 2

Figure 3.3 shows a painting by Peter Paul Rubens and workshop (Peter Paul Rubens et atelier) named "Le martyre de saint Liévin". Unlike Example 1, this record also contains **iconographicTerms** and **conceptualTerms**, which allows us to illustrate the hierarchy construction for all three fields.



Figure 3.3: Peter Paul Rubens et atelier, *Le martyre de saint Liévin*.

Figure 3.4 shows the corresponding hierarchy for **subjectTerms**. There is a single main group, *groupe de figures*, with fifteen direct children (*homme*, *vêtement liturgique*, *crosse*, *soldat*, *animal*, *être imaginaire*, *nu*, *palme*, *couronne*, *arme*, *lance*, *pince*, *langue*, *nuage*, *foudre*). Among these, four have their own children: *vêtement liturgique* (*chape*, *étole*, *mitre*), *animal* (*cheval*, *chien*), *être imaginaire* (*aile*), and *arme* (*épée*). The associated **subjectTerms** field was the following:

groupe de figures (homme ; vêtement liturgique : chape ; étole ; mitre ; crosse ;
soldat ; animal : cheval ; chien ; être imaginaire : aile ; nu ; palme ; couronne ;
arme : épée ; lance ; pince ; langue ; nuage ; foudre)

The output hierarchy was:

{'groupe de figures': ['homme', 'vêtement liturgique', 'crosse', 'soldat', 'animal',
'être imaginaire', 'nu', 'palme', 'couronne', 'arme', 'lance', 'pince', 'langue', 'nuage',

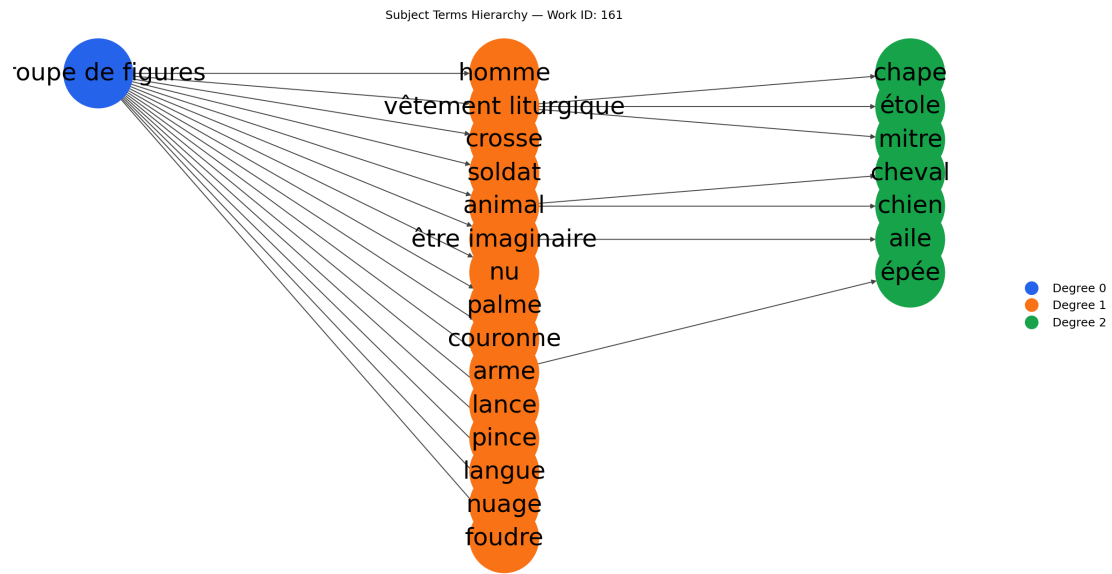


Figure 3.4: Extracted **subjectTerms** hierarchy of *Le martyre de saint Liévin*.

'foudre'], 'homme': [], 'vêtement liturgique': ['chape', 'étole', 'mitre'], 'chape': [], 'étole': [], 'mitre': [], 'crosse': [], 'soldat': [], 'animal': ['cheval', 'chien'], 'cheval': [], 'chien': [], 'être imaginaire': ['aile'], 'aile': [], 'nu': [], 'palme': [], 'couronne': [], 'arme': ['épée'], 'épée': [], 'lance': [], 'pince': [], 'langue': [], 'nuage': [], 'foudre': []}

Figure 3.5 shows the corresponding hierarchy for **iconographicTerms**. Here again there is a single main group, *scène religieuse*, with six children and no grandchildren: *St. Liévin*, *martyre*, *évêque*, *Gand*, *ange*, *bourreau*. The associated **iconographicTerms** field was the following:

scène religieuse (St. Liévin ; martyre ; évêque ; Gand ; ange ; bourreau)

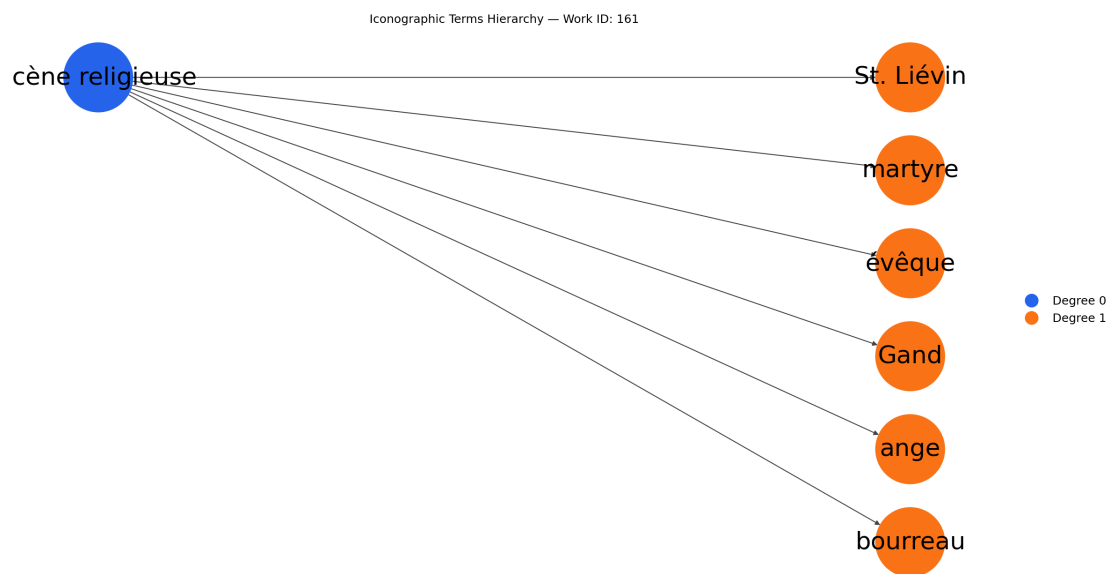


Figure 3.5: Extracted *iconographicTerms* hierarchy of *Le martyre de saint Liévin*.

The output hierarchy was:

```
{ 'scène religieuse': [ 'St. Liévin', 'martyre', 'évêque', 'Gand', 'ange', 'bourreau' ], 'St.
  Liévin': [], 'martyre': [], 'évêque': [], 'Gand': [], 'ange': [], 'bourreau': [] }
```

Finally, the **conceptualTerms** field for this record was the following:

jésuite ; Contre-Réforme

Since this string contains no parentheses, it was first handled by the program as a single bare node rather than being split on the ";" separator. As a result, the output hierarchy collapsed both concepts into a single node, { 'jésuite ; Contre-Réforme': [] }, instead of producing two independent roots *jésuite* and *Contre-Réforme*. This issue was fixed, and it now yields two separate parent nodes. The output hierarchy is:

```
{ 'jésuite': [], 'Contre-Réforme': [] }
```

3.3 Correction algorithm

As mentionned in Section 3.1, we implemented a correction algorithm that detects any error in the **iconographicTerms**, **conceptualTerms** or in the **iconographicInterpretation** fields. You can find the simplified pseudo code in Algorithme 1.

Algorithm 1: Semantic Hierarchy Correction

Input: Hierarchy H , threshold τ

Output: Corrected hierarchy H

```

1 Compute embeddings and similarities between all nodes;
2 repeat
3   foreach child node  $c$  do
4     Find the node  $p^*$  most similar to  $c$ ;
5     if  $p^*$  fits better than  $c$ 's current parent (by  $\tau$ ) then
6       Remember this as a candidate move;
7     end
8   end
9   Apply the single best move: re-attach  $c$  under  $p^*$ ;
10 until no good move is found;
11 return  $H$ ;
```

A "move" is taking one child node and detaching it from its current parent, then attaching it under a better-fitting parent.

Example

For the following **subjectTerms**, corresponding to "Le retour de la kermesse" by Jacques d'Arthois:

paysage (forêt ; arbre ; fleur ; chemin ; étang ; maison) ; scène (homme ; femme ; danse ; instrument de musique : cornemuse ; animal : chien ; vache ; oiseau)

The *correction algorithm* implemented gives this output:

```

- Original subjectTerms -
paysage -> ['forêt', 'arbre', 'fleur', 'chemin', 'étang', 'maison']
          forêt -> []
          arbre -> []
          fleur -> []
          chemin -> []
          étang -> []
          maison -> []
```

```

scène -> ['homme', 'femme', 'danse', 'instrument de musique', 'animal']
      homme -> []
      femme -> []
      danse -> []
instrument de musique -> ['cornemuse']
      cornemuse -> []
animal -> ['chien', 'vache', 'oiseau']
      chien -> []
      vache -> []
      oiseau -> []

```

- Corrected subjectTerms -

```

paysage -> ['forêt', 'fleur', 'chemin', 'étang', 'maison']
      forêt -> ['arbre']
      arbre -> []
      fleur -> []
      chemin -> []
      étang -> []
      maison -> []
scène -> ['homme', 'femme', 'danse', 'instrument de musique', 'animal']
      homme -> []
      femme -> []
      danse -> []
instrument de musique -> ['cornemuse']
      cornemuse -> []
animal -> ['chien', 'vache', 'oiseau']
      chien -> []
      vache -> []
      oiseau -> []

```

As we can see, the correction algorithm detects that "arbre" (tree) is a child of "forêt" (forest), even though this was not the case in the original **subjectTerms** field.

Chapter 4

Story-to-Art Association

This chapter presents the core of the master's thesis. It defines our mission, describes the two-phase algorithm, and shows some results.

4.1 Mission statement

The goal was to create a tool that satisfies the following requirements:

- involves the relatives of people with neurodegenerative diseases and dementia in the process,
- triggers memories,
- triggers emotions,
- is easy to use, with a simple interface,
- retrieves an artwork relatively fast (less than 5 seconds per request),
- is interpretable (based on graph theory in our case).

Global idea

The core idea of this system is the following: the relatives of a patient submit textual stories describing personal memories involving the patient. The system then retrieves artworks from the museum database described in Chapter 3 that are semantically and emotionally aligned with these narratives.

The entire algorithm was implemented in Python. When building the graph offline, only the artworks containing an *iconographicInterpretation* field were used as text-eligible records, treating this field as the artwork-related "story"; artworks

without such a description are attached to the graph through visual similarity, as described in section 4.2. There are some limitations due to the use of *iconographicInterpretation* as the artwork-related "story"; it will be discussed in Chapter 7.

The proposed system is divided into two phases. The first phase retrieves the artworks that are most semantically similar to the input memory, relying on the textual content of both the memory and the artworks; the emotion inferred from each artwork already contributes to this first score, as detailed in section 4.2. The second phase then re-ranks the pool of top artworks retrieved during the first phase and selects the one that best matches a desired emotion provided by the user. This ensures that the final result is both thematically coherent and emotionally appropriate.

The general pseudo code of the main algorithm is reported in Appendix B.

4.2 Phase One: Semantic Retrieval

During this first phase, the emphasis is on semantic similarity. The goal is to build a structured graph of the artwork collection and to efficiently compare it against an input memory.

Textual Analysis

Stories (whether from the input memory or from artwork descriptions) are preprocessed as follows: the text is lowercased, punctuation is removed, and the result is tokenised into individual words. Tokenising into individual words causes us to lose part of the syntactic structure and word order of the sentence, and therefore some of its overall meaning. This is a limitation of the graph that is acknowledged and that could be addressed in future work, as discussed in Chapter 7.

To extract only the most meaningful tokens, three filters are applied:

- **Stopwords:** grammatical function words (articles, prepositions, pronouns, etc.) that carry little semantic information are discarded. The stopwords list is defined manually and covers the most common French function words.
- **Short tokens:** words of two characters or fewer are systematically removed, as they are rarely semantically meaningful in isolation.
- **Duplicate tokens:** within a single story, the same word is counted only once to avoid artificially inflating its importance.

In addition, tokens are filtered using a hard frequency threshold: any token appearing more than 50 times across the full raw corpus is treated as too generic and discarded, alongside standard stopwords and tokens of length ≤ 2 . Now are computed two complementary statistics that are used later, at different stages of the pipeline:

- **Global frequency:** the total number of occurrences of a word across all documents. This is used to build the exclusion set for the caption-label filter described above.
- **Document frequency:** the number of documents containing a given word at least once. This is used later, during Phase One retrieval scoring, for IDF weighting of query labels (see Section 4.3).

Both statistics are computed on the full raw corpus rather than on the working subset, so that frequency thresholds and IDF values are correctly calibrated.

Artwork Labels

Beyond the textual *iconographicInterpretation*, artworks in the museum dataset are annotated with structured data fields: *subjectTerms*, *iconographicTerms*, and *conceptualTerms*. As described in Chapter 3, these fields carry an intrinsic hierarchical structure. For example, a term may have parent, child, and grandchild concepts within a controlled vocabulary. Each term is extracted and assigned a depth value d corresponding to its level in this hierarchy, computed using the `creation_hierarchy` and `compute_levels` functions. When the same term appears across multiple fields, the maximum depth across all fields is retained.

Exploiting this hierarchy by assigning higher edge weights allows deeper (more specific) labels in the graph. The intuition is that a very specific term is more discriminative and informative than a broader term, and should therefore contribute more strongly to the similarity score.

Image Analysis

To complement the textual information, visual features are extracted directly from the artwork images using a convolutional neural network. For each image, a 2048-dimensional feature vector is derived using ResNet50, pre-trained on ImageNet.

These visual features serve a specific role in the graph construction: for artworks that lack a textual *iconographicInterpretation*, the L2-normalised feature vectors are used to compute pairwise cosine similarities against all artworks that *do* possess a description. An artwork without a description is linked in the graph to its

most visually similar annotated neighbours, retaining at most three such links, each subject to a cosine similarity threshold of 0.5. The resulting edges carry a weight of $0.7 \times \text{sim}$ and are flagged as `visual_similarity` edges, allowing the undescribed artwork to benefit from the textual labels of related artworks through graph propagation.

Emotion Extraction

Each artwork is assigned a single dominant emotion by comparing the embedding of its *iconographicInterpretation* with the embeddings of a predefined set of emotion labels (see Section 4.4 for the list). The closest emotion label, measured by cosine similarity, is selected. This assignment is computed once during preprocessing and stored as an emotion node linked to the artwork. The inferred emotion is used in *both* phases: in Phase One it participates in the retrieval score through its emotion node (see Section 4.3), and in Phase Two it is compared against the emotion desired by the user. Artworks without a textual description are assigned the neutral label *autre* by default, as no textual signal is available to infer their emotional content.

Multimodal Graph Construction

Inspired by Zhang’s work on tripartite graphs [36], the entire dataset is represented as a weighted undirected graph G . The nodes of this graph belong to several categories:

- **Image nodes:** one node per artwork, identified by its unique `workID` (unique number associated with each artwork in the dataset to recognise it).
- **Caption nodes:** one node per *iconographicInterpretation*.
- **Caption label nodes:** individual tokens extracted from artwork captions after preprocessing. Each caption label is connected to its caption node with an edge of weight 0.8.
- **Hierarchical label nodes:** terms from *subjectTerms*, *iconographicTerms*, and *conceptualTerms*, connected directly to their image node. Their edge weight is $w = 0.5 + 0.5 \cdot \log(1 + d)$, where d is the depth of the term in the hierarchy, so that more specific terms have stronger connections.
- **Emotion nodes:** one node per emotion label, linked to image nodes with an edge of weight 0.8, based on the dominant emotion assigned during preprocessing.

The edges of the graph therefore come from three conceptually distinct sources. *Structural* edges have *fixed* weights and encode the composition of an artwork: the caption-caption label edge (0.8), the image-hierarchical label edge ($0.5 + 0.5 \log(1 + d)$), and the image-emotion edge (0.8). *Co-occurrence* edges connect two labels and are the only edges whose weights are *derived from the data* during the offline construction of the graph, as described in Section 4.3. *Visual-similarity* edges ($0.7 \times \text{sim}$) connect undescribed artworks to visually similar annotated ones. Weighted shortest-path distances derived from all these edges are the mechanism used to measure similarity between an input memory and the artworks in the database.

Mapping a Query into the Graph

When a user submits a memory, the input text is preprocessed in the same way as the artwork descriptions: tokenised and filtered. The resulting query tokens, however, may not appear directly as existing nodes in the graph.

To bridge this gap, *sentence-camembert-large* is re-used, a French sentence embedding model, to map each query token to its nearest counterpart in the graph. All graph nodes of type `caption_label`, `hier_label`, and `mood` are pre-embedded at indexing time and stored in a FAISS index for efficient retrieval. At query time, each query token is encoded as a dense vector and compared against this index using cosine similarity. Up to five matching graph nodes are retained per query token, subject to a similarity threshold of 0.45 (found heuristically). If no match exceeds this threshold, a guaranteed fallback mechanism selects the single closest graph node with a reduced similarity weight of $\times 0.6$.

Emotion tokens are handled separately from textual tokens: they are looked up directly by name in the graph rather than through the embedding search, since their representation is not semantic.

4.3 Formulas used and visualisation

As noted above, images, captions, and their associated labels are represented as nodes in a graph G . Among these, the label-label edges are weighted according to co-occurrence frequencies observed during the offline construction of the graph. The key idea is that two labels that tend to appear together in the same artwork or memory are likely to be semantically related, and this relationship is encoded as a strong connection (a short path) in the graph.

Co-occurrence Weight

For each pair of labels (i, j) , the co-occurrence weight is defined as:

$$w_{ij} = \frac{\text{count}(i, j)}{\max_{(a, b)} \text{count}(a, b)}$$

where $\text{count}(i, j)$ is the number of times labels i and j co-occur, and the denominator normalises all weights to the range $[0, 1]$. Distinguish three types of co-occurrence:

- co-occurrence of two caption labels within the same artwork description;
- co-occurrence of two hierarchical artwork labels within the same artwork;
- cross-modal co-occurrence between a caption label and a hierarchical artwork label for the same artwork.

Hierarchical co-occurrences are further weighted by the depth of the labels involved: deeper (more specific) labels receive a bonus proportional to $\log(1 + d)$, where d is the depth in the hierarchy. This ensures that specific co-occurrences are considered more informative than generic ones.

Once raw co-occurrence weights are computed and normalised, they are transformed by a square-root function before being added as graph edges: $w_{\text{edge}} = \max(\sqrt{w_{ij}}, 0.1)$. This keeps very frequently co-occurring pairs as strong edges while compressing the gap between high and moderate co-occurrence, and the minimum of 0.1 ensures that even rarely co-occurring pairs contribute a non-negligible signal during graph propagation.

Weight-Distance Conversion

Dijkstra’s algorithm [37] requires edge *distances*, not *weights*: a high weight should correspond to a short distance. Therefore, weights are converted to distances as follows:

$$\text{dist}(i, j) = \frac{1}{\max(w_{ij}, \varepsilon)}$$

with a small ε preventing division by zero. This inversion ensures that a strongly co-occurring pair of labels (high w) is represented by a short path in the graph, while weakly related pairs (low w) are far apart.

Shortest-Path Similarity

For each query label that has been mapped into the graph, the shortest-path distance to all image nodes is computed using Dijkstra’s algorithm, with a maximum propagation cutoff of 4.0. Each image accumulates a score over all query labels:

$$\text{score}(\text{image}) = \sum_{l \in \mathcal{L}_{\text{query}}} \frac{\text{IDF}(l) \cdot \text{sim}(l) \cdot \text{boost}(l)}{1 + d^*(\text{image}, l)}$$

where:

- $\mathcal{L}_{\text{query}}$ is the set of query labels, comprising the textual tokens *and* the detected emotion label,
- $\text{IDF}(l)$ is the inverse document frequency of label l (see below),
- $\text{sim}(l) \in [0, 1]$ is the cosine similarity between the query token and its matched graph node,
- $\text{boost}(l)$ is a node-type multiplier: 1.2 for **caption_label** nodes, $1.1 \times (1 + 0.1d)$ for **hier_label** nodes at depth d , and 1.0 for emotion nodes,
- $d^*(\text{image}, l)$ is the shortest-path distance from label node l to the image node.

More informally, an artwork scores highly if it is reachable via short paths from many relevant query labels, weighted by how distinctive those labels are and by the specificity of the matched graph node. Note that, because the emotion label is part of $\mathcal{L}_{\text{query}}$, the emotion inferred for each artwork already influences the Phase One ranking, and is not confined to Phase Two.

IDF Weighting

Not all query labels are equally informative. A label that appears in nearly every artwork should contribute less to the score than a rare, discriminative label. Using Inverse Document Frequency (IDF) to capture this:

$$\text{IDF}(l) = \log\left(\frac{N}{\text{df}(l) + 1}\right) + 1$$

where N is the total number of documents in the corpus and $\text{df}(l)$ is the number of documents containing label l . The $+1$ smoothing prevents division by zero, and the outer $+1$ ensures all IDF values remain strictly positive. This same formula is applied uniformly to all query labels.

Node Degree Regularisation

Images that are richly annotated (many labels, many edges) naturally accumulate more score simply because they have more paths converging toward them. To correct for this structural bias, scores are normalised by a function of the image node’s weighted degree:

$$\text{score}_{\text{norm}}(\text{image}) = \frac{\text{score}(\text{image})}{1 + \sqrt{\text{deg}(\text{image})}}$$

where $\text{deg}(\text{image})$ is the weighted degree of the image node in the graph. The square-root function provides a softer penalty than $\log(1 + \cdot)$: it meaningfully corrects for the structural bias while still allowing well-connected (richly described) artworks to rank highly when they are genuinely relevant.

Final Retrieval

The top-5 candidates (regarding to $\text{score}_{\text{norm}}$) returned by Phase One are displayed to the user and simultaneously passed to Phase Two for emotional re-ranking. Phase Two then selects, among these five candidates, the artwork whose inferred emotion best matches the user’s desired emotional state.

Visualisation of the algorithm

The algorithm is illustrated in Figures 4.1 to 4.6. Steps 1-3 correspond to pre-processing and label extraction, Step 4 to the counting of co-occurrence weights, Step 5 to graph construction, and Step 6 to the shortest path retrieval.

Step 1 — Raw multimodal input

A story (text) and the candidate artworks, still isolated

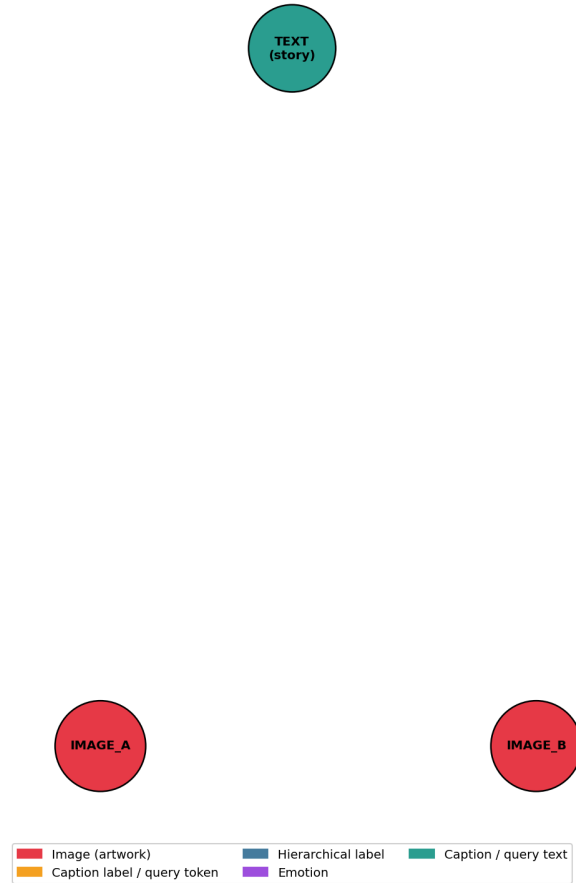


Figure 4.1: Step 1: Text preprocessing of the artwork descriptions.

Step 2 — Tokenization & caption labels

Text and descriptions split into tokens (lowercased, stopwords removed)

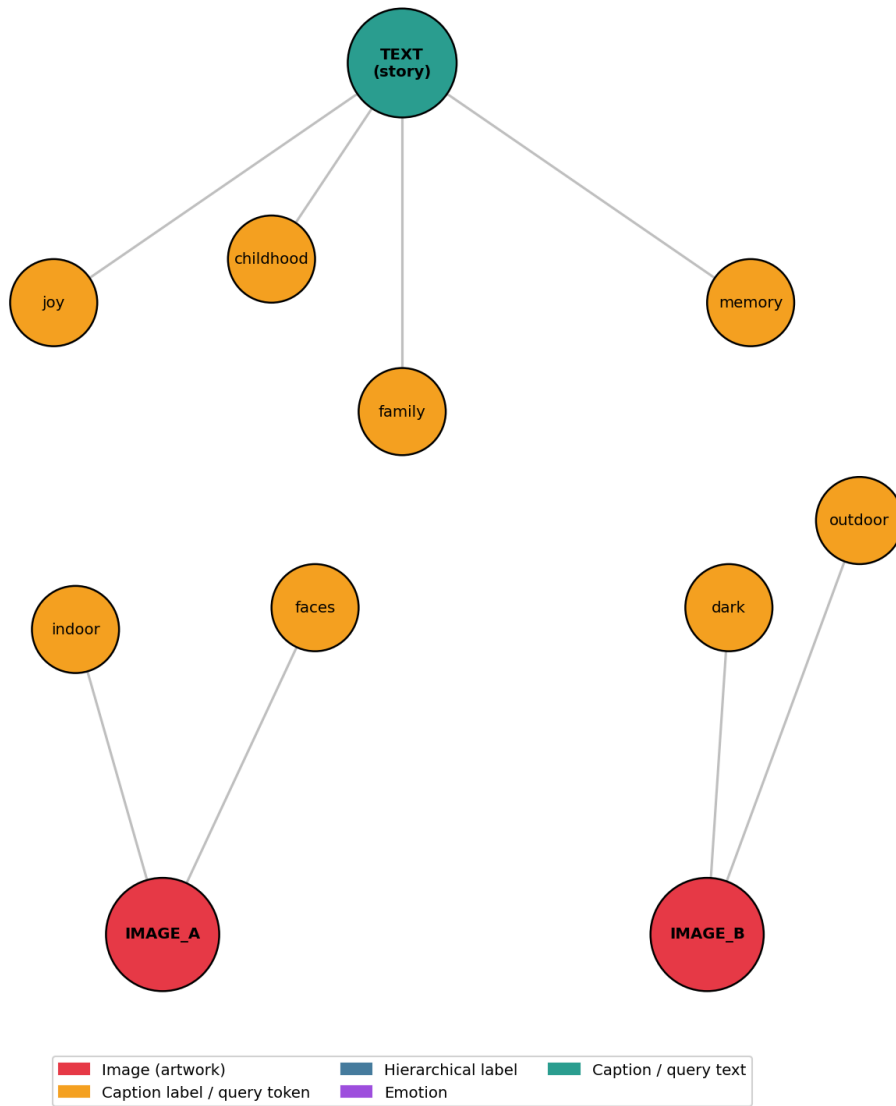


Figure 4.2: Step 2: Caption-label extraction and filtering.

Step 3 — Hierarchical labels & emotions

Adding hierarchical labels (depth) and one dominant emotion per artwork

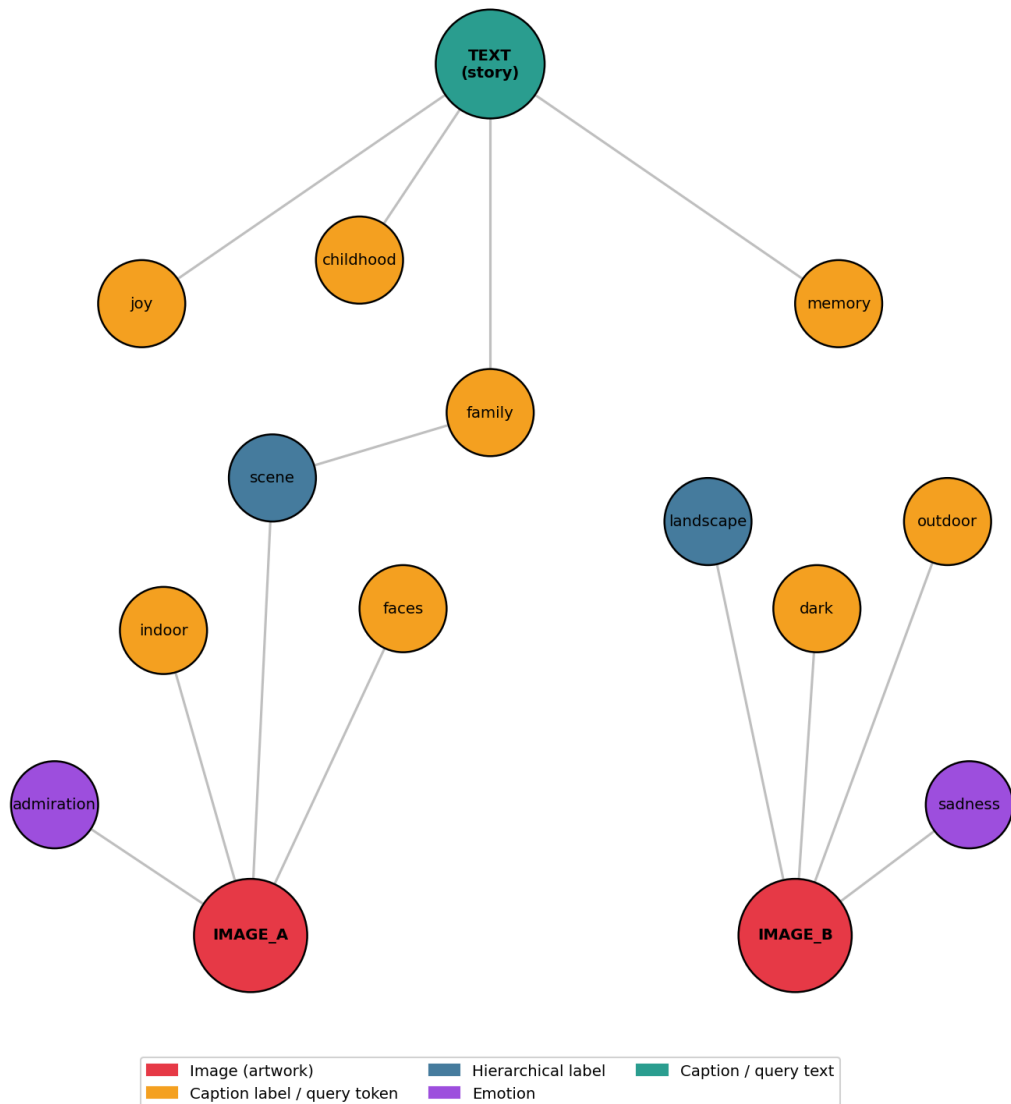


Figure 4.3: Step 3: Hierarchical-label extraction from the SubjectMatter fields.

Step 4 — Learning semantic co-occurrences

Co-occurring label pairs are linked (weighted blue edges)

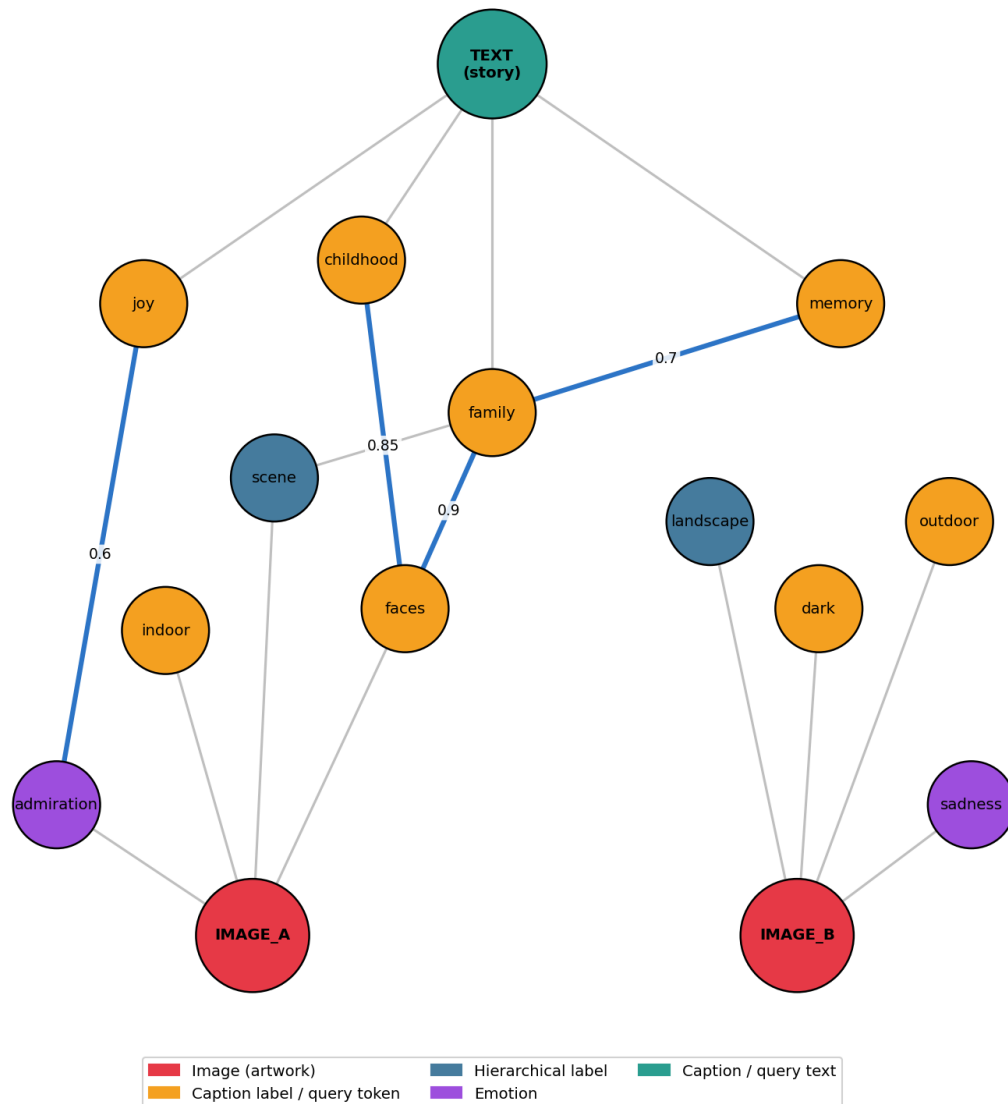


Figure 4.4: Step 4: Counting of label co-occurrences and edge weighting.

Step 5 — Full multimodal graph

Full heterogeneous graph: images, captions, labels, hierarchy, emotions

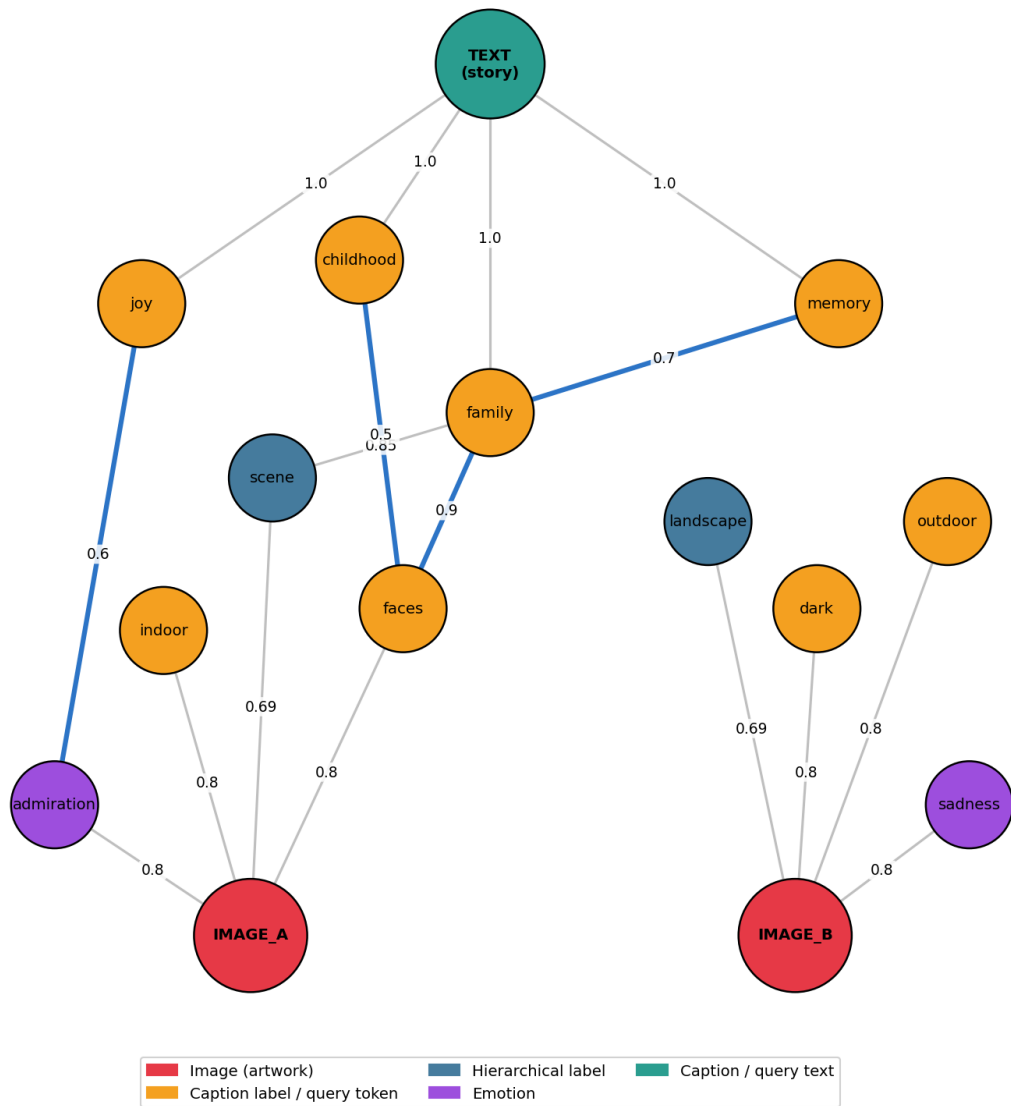


Figure 4.5: Step 5: Construction of the multimodal graph.

Step 6 — Shortest-path retrieval (Dijkstra)

The weighted shortest path links the query to IMAGE_A: the returned artwork

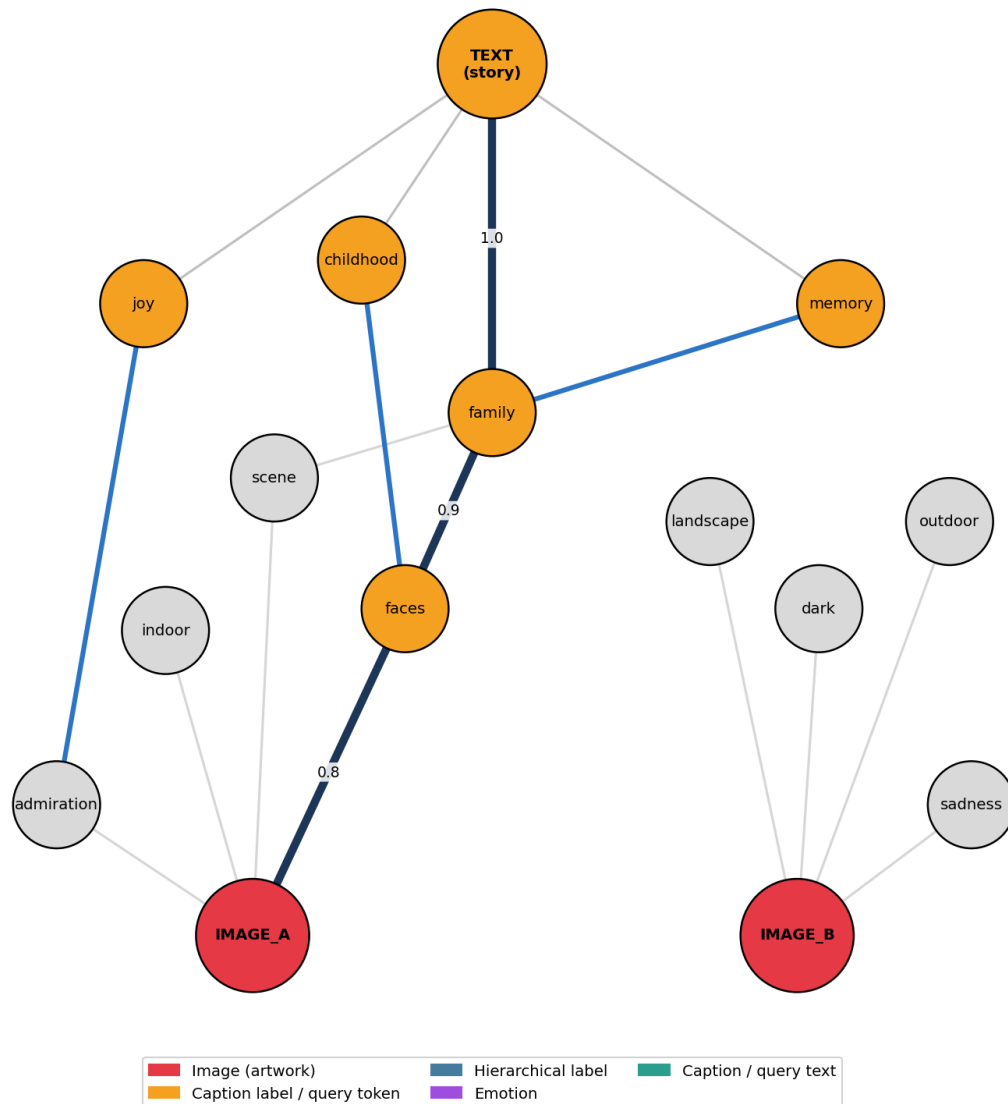


Figure 4.6: Step 6: Online processing of a query memory and retrieval.

4.4 Phase Two: Emotional Alignment

Phase One returns a set of semantically relevant artworks, but semantic relevance alone does not guarantee emotional coherence with the context. In an art-therapy context, the emotional tone of the selected artwork may matter: presenting a grieving person with a joyful or chaotic artwork, even if thematically related to their memory, could plausibly be not appropriate. Matching a desired emotional state (whether to validate a feeling or to gently shift a mood) is often described as an important aspect of art-therapy practice.

To address this, Phase Two selects, from the pool of top-5 semantically relevant artworks retrieved in Phase One, the artwork whose inferred emotion is most aligned with a desired emotion specified by the user (for example a therapist or a caregiver).

Each artwork was assigned a dominant emotion during preprocessing: the *iconographicInterpretation* was encoded into a vector using the same sentence embedding model (*sentence-camembert-large*) and compared against the embeddings of the predefined emotion labels. The emotion with the highest cosine similarity was retained as the artwork’s dominant emotion. This assignment is performed in batches for efficiency: captions are processed in chunks of 200 (on GPU) or 50 (on CPU), and the resulting emotion assignments are cached for the duration of the session.

At query time, the desired emotion entered by the user is similarly embedded and compared against the dominant emotions of the five candidate artworks. The artwork whose emotion embedding is closest to the desired emotion is selected.

Emotion categories

The set of nine emotion categories used in this work is adapted from the ArtEmis dataset and model [38]. ArtEmis is a large-scale dataset that pairs visual artworks with emotion annotations and free-form explanations written by human annotators. Its central premise is that the emotional impact of art goes beyond what is literally depicted, encompassing metaphor, atmosphere, and symbolic content.

ArtEmis treats emotion as a primary axis for understanding art, motivated by the observation that, for many viewers, a central function of art is to communicate feelings. Building on its nine-way scheme, the following nine categories are used, labelled in French to match the language of the museum descriptions:

1. Amusement (*amusement*)
2. Awe / Admiration (*admiration*)
3. Contentment / Satisfaction (*satisfaction*)

4. Excitement (*excitation*)
5. Anger (*colère*)
6. Disgust (*dégoût*)
7. Fear (*peur*)
8. Sadness (*tristesse*)
9. Something else / Other (*autre*)

The first four are positive emotions, the next four are negative, and the last is a neutral category. This scheme is rich enough to capture meaningful emotional distinctions, yet compact enough to remain tractable for embedding-based matching.

Valence Constraint

Emotions that are semantically related but of opposite polarity (such as "joy" and "sadness") may have a moderately high cosine similarity simply because they share the emotional domain. A pure cosine-similarity search could therefore accidentally select an artwork of opposite emotional valence.

To prevent this, a **valence constraint** is introduced. Each emotion is assigned a polarity:

- **Positive valence:** amusement, admiration, satisfaction, excitement.
- **Negative valence:** anger, disgust, fear, sadness.
- **Neutral:** *something else*.

When the desired emotion and a candidate artwork's emotion have opposite valences, the candidate's score is strongly penalised by a factor of 0.05, making it effectively ineligible unless no better-valenced option is available. Neutral emotions are not penalised in either direction.

For emotions entered freely by the user (not from the predefined list), the valence is inferred by finding the closest predefined emotion in the embedding space and using its valence.

Final Selection

The final artwork returned by the system is the one that jointly satisfies two criteria:

- it is among the most semantically relevant artworks with respect to the input memory (Phase One), as determined by the graph-based shortest-path score;

- among these semantically relevant candidates, it is the most emotionally aligned with the desired emotion (Phase Two), subject to the valence constraint.

4.5 Illustrative examples of algorithm outputs

Below are two examples of input and output of the algorithm. They are qualitative illustrations of the intended behaviour; a rigorous evaluation of performance is carried out in Chapter 6.

Example 1

Input story: Ma soeur aimait boire du thé au petit déjeuner. Elle s’installait à table et profitait de sa boisson. C’était son rituel du matin. (My sister liked to drink tea for breakfast. She would sit down at the table and enjoy her drink. It was her morning ritual.)

Phase 1 output:



Figure 4.7: Result of Phase 1 for the input story of Example 1.

The five images are ordered from the closest match to the least semantically relevant. In this example, the first three images include a table, the first two include a woman, and only the first shows a woman at a table drinking her tea.

Figure 4.7 shows, above each image, its ID, the title of the artwork, its score, and the main emotion attached to it.

Input emotion: Joie/Joy.

Phase 2 output:



Figure 4.8: Result of Phase 2 for the input story of Example 1.

As shown in Figure 4.8, in this example the best image semantically is also the best emotionally, within the set found in Phase One.

Example 2

Input story: Mon oncle était passionné de navigation. Il possédait même un bateau. Un jour, alors qu'il était en mer, il s'est retrouvé dans une tempête et a eu très peur. (My uncle was passionate about sailing. He even owned a boat. One day, while he was out at sea, he got caught in a storm and was very scared.)

Phase 1 output: As shown in Figure 4.9, the first three images are quite similar and portray the input story well.



Figure 4.9: Result of Phase 1 for the input story of Example 2.

Input emotion: Colère/Anger.

Phase 2 output:



Figure 4.10: Result of Phase 2 for the input story of Example 2.

In this example, the best image of Phase One is not the best emotionally. The algorithm therefore returns the second image of Phase One as the result of Phase Two.

Chapter 5

Interpretability and Bias Analysis

A central motivation of this thesis is that transparency matters as much as retrieval accuracy in the intended application. This chapter examines what our graph-based approach does or does not offer in terms of interpretability, and uses the explicit structure of the graph to analyse potential biases. Following the terminological concerns raised across the interpretability literature [39, 40, 41], we distinguish carefully between several notions that are often conflated: graph transparency, explainability, faithfulness, and the ability to support meaningful reminiscence.

5.1 Distinguishing the relevant notions

The literature repeatedly warns that terms such as "interpretability", "explainability", "transparency" and "faithfulness" are used interchangeably even though they denote different properties, and that this conflation is harmful [41]. We therefore separate the concepts we rely on.

- **Transparency:** the ability to inspect the explicit structure that produced a result. In our system, every edge, weight, and shortest path in the graph can be examined directly. This corresponds to what [39] associate with *completeness*, an explanation that describes the operation of a system accurately, up to the limit of revealing all of its internal operations. In our case, revealing the full graph is a genuinely complete description of the retrieval mechanism.
- **Explainability / interpretability:** the ability to give an account of *why* a given artwork was retrieved that is understandable to a human [39]. For us, this means following the path from a query token, through caption labels, hierarchical labels, and emotion nodes, to the retrieved image. Interpretability is tied to the cognition and vocabulary of the user, so it is not an absolute property of the system.

- **Faithfulness:** the degree to which the explanation accurately reflects the true reasoning process behind the decision [41].
- **Plausibility:** how convincing an interpretation is to humans [41]. Plausibility and faithfulness are independent, and a plausible but unfaithful explanation can be a bad outcome in sensitive settings.
- **Meaningful reminiscence:** whether the retrieved artwork actually triggers autobiographical memory or emotional engagement in a user. This is a therapeutic outcome, and it is not guaranteed by any of the previous properties.

A system can be transparent, faithful and explainable while still failing to support meaningful reminiscence, and conversely a black-box system could, in principle, trigger memories effectively. Our contribution concerns the first notions; the last would require clinical validation, as discussed in Chapter 7.

5.2 Interpretability for whom?

Interpretability is not an absolute property; it is relative to an audience, because it depends on the knowledge, vocabulary and expectations of the person receiving the explanation [39]. The explanations our system can produce serve different people in different ways:

- for a **caregiver or therapist**, the traced path can justify why an artwork was suggested, supporting an informed decision about whether to show it to a patient;
- for a **developer**, the explicit graph allows inspection of the mechanisms and detection of biased associations;
- for a **patient**, the raw path is unlikely to be directly useful, and the explanation would need to be mediated by a human.

Being explicit about the intended audience avoids overstating what the system offers: our explanations are primarily useful to the humans in the loop rather than to the end user directly.

5.3 The explanation-reality gap

A crucial point concerns what the graph paths actually represent. A link in the graph reflects a co-occurrence, a hierarchical relation, or a visual similarity observed

in the data; it does not necessarily correspond to a real semantic or causal relation in the world. Two labels frequently appearing together in museum descriptions are connected by a short path, but this statistical association may encode a cultural convention or a historical bias rather than a genuine conceptual link.

This distinction, explained in [40], suggests that "explanation" is misleading terminology: an account of how a *model* works is not an account of how the *world* works. Our path is a faithful account of the system's own reasoning but not of the underlying reality. Our system explains its own retrieval process; it does not explain human memory formation, nor does it claim that the associations it exploits are objectively correct. Recognising this gap is what makes the transparency of the system responsible rather than misleading, and guards against the false confidence [40].

5.4 Bias in cultural heritage AI

Discussions of AI applied to cultural heritage collections increasingly revolve around several intertwined forms of bias [42]. These are directly relevant to our system, since it inherits both the biases of its pre-trained components and those of the museum data itself:

- **Representational bias:** certain groups, themes, or periods may be over- or under-represented in the collection.
- **Cataloguing bias:** the manual annotations reflect the assumptions, vocabulary, and priorities of the cataloguers.
- **Language bias:** the descriptions are in French and follow a specific iconographic vocabulary, which constrains what can be expressed and matched.
- **Embedding bias:** the pre-trained models (sentence-camembert-large, ResNet50) carry biases learned from their own training data, independent of the museum collection.

A concrete example that could arise in our own results is the **woman-kitchen association**: for the still-life query, the graph method promotes a kitchen scene (*La cuisinière*) featuring a woman to the top of its ranking (see Chapter 6), reflecting how the hierarchical label structure links food, kitchen, and domestic roles. Such an association may encode a gendered cultural assumption present in the descriptions rather than an unbiased semantic relation.

5.5 Auditability as an advantage of the graph

A key argument of this chapter is that the explicit graph is not merely transparent for its own sake: it actively facilitates the analysis of the biases listed above. Concretely, the architecture supports:

- the identification of biased associations, by inspecting which label-label edges carry strong co-occurrence weights;
- the analysis of cultural assumptions, by tracing recurring patterns in the retrieved paths;
- the auditing of emotional pathways, by examining which labels drive a given emotion assignment;
- the transparency of retrieval decisions, by exposing the full path for any result.

This connects to the central thesis of [40]: for high-stakes or sensitive decisions, an inherently transparent model is easier to debate, troubleshoot and correct than a black box, and one should prefer it whenever it is available at comparable performance. Our approach yields two properties that a purely deep-learning pipeline such as CLIP does not provide as readily. The first is **auditability**: because every decision is traceable to responsible edges, a biased association can in principle be located and corrected at its source, rather than approximated by a second model whose own faithfulness would itself need to be verified [40]. The second is the **absence of hallucination**: since the system can only retrieve artworks that are actually present in the graph and reachable through real edges, it cannot fabricate non-existent relations or content, in contrast to generative models.

We nonetheless temper these claims in line with [41], who argue that "inherent interpretability" is a claim that must be verified rather than assumed, and that faithfulness is better treated as a graded property than as a binary one. Our graph is faithful to its own retrieval computation by construction, but this does not certify that the associations it encodes are meaningful, unbiased, or clinically effective. Together, transparency and auditability improve **trust**: a caregiver or auditor can, at least in principle, verify and challenge the system's behaviour, while remaining aware of what the explanations do and do not establish.

Chapter 6

Evaluation Protocol and Benchmarks

To evaluate our algorithm quantitatively and qualitatively, three distinct protocols were established. Those are presented in this chapter. The first two (6.2 and 6.3) aim to probe the algorithm’s functional behaviour and its precision relative to a *baseline* algorithm, detailed in Section 6.3. The third one (6.4) compares the speed and resource usage of our algorithm against CLIP (described in Chapter 2). Qualitative comparison between our graph-based algorithm and CLIP was also carried out.

Before presenting the protocols, it should be noted that all of the studies reported in this chapter are small and exploratory. They involve ten participants recruited from my relatives and acquaintances, a convenience sample, and each participant contributed a single query per test. The results should therefore be read as preliminary indications of the system’s behaviour, not as statistically robust evidence of superiority over any other method.

6.1 Participants and procedure

The ten participants were relatives and acquaintances who were not compensated and took part on a voluntary basis. None of them had prior knowledge of which retrieval system produced which result. The details of the tests are reported in Appendix A.

6.2 Recovering the target image

The first test evaluates the retrieval accuracy of the system in a *controlled* setting. Importantly, in this protocol participants describe an artwork they are looking at,

rather than recalling an autobiographical memory; it therefore measures whether the system can retrieve a known target from a literal description, and does *not* evaluate the memory stimulation use case. The procedure was:

- Participants were shown a randomly chosen artwork and asked to describe it in their own words;
- this description was used as the input query for the algorithm;
- a trial was marked as **successful** if the original image appeared within the top-5 retrieved artworks, and a **failure** otherwise.

This protocol was conducted with ten participants and five different artworks, so that each artwork was tested twice. The artworks were chosen to be diverse in style, artist, and content (objects, landscape, and/or people); they are shown in Figure 6.1.



(a) *La veillée. Jeanne la folle.*



(b) *Le déjeuner*



(c) *Un port en orient*



(d) *Vase de fleurs*



(e) *Louis Faider à trois ans*

Figure 6.1: The five images used for the controlled retrieval test.

Results

The results of this test are given in Table 6.1. There are eight successes and two failures, i.e. 80% of trials succeeded in this small sample. The two failures both involved artworks depicting specific, historical people: because participants did not know who was portrayed, their descriptions were more generic, which plausibly explains the lower retrieval accuracy for those works.

	La veillée	Le déjeuner	Un port en orient	Vase de fleurs	Louis Faider
Test 1	Success	Success	Success	Success	Failure
Test 2	Failure	Success	Success	Success	Success

Table 6.1: Results of the first (controlled retrieval) protocol

We also note that whenever retrieval succeeded, the target image was ranked first among the top five, except for a single case in which it was ranked second.

6.3 Comparison with baseline and emotional alignment

For the comparison test, the two sets of five artworks (ours and the baseline) were displayed side by side; the left/right placement of the two systems was randomised across participants so that position could not be confused with method. Each participant contributed a single story and a single desired emotion, so that every test corresponds to one independent query.

Because the predefined emotion categories may not match the words participants naturally use, freely entered emotions (for example *nostalgie*, *fatigue*, *calme*, *bonheur*) were projected onto the nine categories of Section 4.4 by taking the closest category in the sentence-embedding space, using the same procedure as the valence inference described in Chapter 4.

In other words, the second test consisted of two phases, mirroring the two-stage architecture of our algorithm described in Chapter 4:

- **Phase 1 (Narrative retrieval):** participants were asked to share a personal story of a few sentences, used as the input. They then performed a side-by-side comparison to decide which set of five retrieved artworks was more relevant: those from our algorithm or those from the *baseline* described in Subsection 6.3.1, without knowing which was which. They could also state that they had no preference.

- **Phase 2 (Emotional alignment):** participants specified a target emotion; our algorithm then retrieved an artwork based on this emotional input. Participants rated the emotional relevance of the resulting artwork on a five-point scale, as defined in Subsection 6.3.2.

6.3.1 Baseline algorithm

The baseline is a simple keyword-matching retrieval system built on set overlap rather than on any semantic understanding. When a query is entered, it is tokenised and filtered in the same way as every artwork caption, leaving each caption, and the query itself, as a set of meaningful keywords. To score how well an artwork matches the query, the algorithm compares the two token sets using two classic set-similarity measures: the Jaccard similarity, which relates the size of the intersection to the union of both sets, and the overlap coefficient, which relates the intersection to the size of the query alone. These two scores are blended with equal weight ($\alpha = 0.5$) into a single combined score. Every artwork is scored this way, the results are sorted from highest to lowest, and the top five are returned. There is thus no notion of synonymy or learning, which is exactly what makes it a useful point of comparison against our more elaborate algorithm.

6.3.2 Mean Opinion Score (MOS) scale

The Mean Opinion Score is a simple subjective rating, on an ordinal scale usually ranging from 1 to 5, commonly used to quantify perceived quality [43].

In our case, the scale was:

- 1: opposite emotion to the desired one;
- 2: not emotionally relevant;
- 3: slightly emotionally relevant;
- 4: coherent emotion;
- 5: fully emotionally relevant.

Results and discussions

For the first phase (narrative retrieval), out of ten participants:

- 7 preferred the results of our algorithm;
- 2 preferred the results of the baseline;

- 1 had no preference.

In this small exploratory study, participants more frequently preferred the results of our proposed method over those of the lexical baseline. Given the sample size, this should be read as an encouraging indication rather than as evidence of a statistically significant improvement.

For the second phase (emotional alignment), the ratings varied substantially between participants:

- 1 participant rated the result 5/5;
- 3 participants rated it 4/5;
- 3 participants rated it 3/5;
- 2 participants rated it 2/5;
- 1 participant rated it 1/5.

These ten ratings have a mean of 3.1/5 and a median of 3. The spread of the ratings (from 1 to 5) indicates that emotional alignment is perceived very differently across users and leaves clear room for improvement.

6.4 Benchmark comparison with CLIP

Settings

The benchmark was conducted on a randomly sampled subset of 1,000 artworks from the RMFAB dataset, restricted to items with a non-empty *iconographicInterpretation* field. The sample is fixed by a random seed (42) for reproducibility. Four test queries were used, chosen to cover subjects representative of the museum collection:

- *un portrait d'une femme assise* / a portrait of a seated woman
- *une scène de bataille historique* / a scene of a historical battle
- *un paysage avec des montagnes et un lac* / a landscape with mountains and a lake
- *une nature morte avec des fruits* / a still life with fruits

Two phases were evaluated: indexing and inference. Given the caching effects discussed below, a *cold-cache* scenario (first run, no cached embeddings) is distinguished from a *warm-cache* scenario (subsequent runs). The inference phase was run five times per query, preceded by one warm-up run. The average time and standard deviation are reported for each phase and method. Memory consumption was measured as the increase in *Resident Set Size* (RSS) of the Python process, captured via *psutil* before and after each phase, as it reflects the extra memory each operation adds rather than memory already in use.

Method

For our graph-based algorithm, indexing included corpus frequency computation, caption-label extraction, hierarchical-label extraction, emotion assignment, graph construction, co-occurrence counting, and building the FAISS vector index over all graph nodes. Inference included query preprocessing, graph-based shortest-path scoring via Dijkstra’s algorithm, dense retrieval over the full dataset, and score fusion. The emotion filter (Phase Two) was excluded from the inference timing of both methods.

For CLIP, indexing consisted of loading the model into memory and encoding all artwork descriptions (and images when available) into a matrix of 512-dimensional vectors. Inference consisted of encoding the query with the text encoder and computing the dot product between the query vector and the pre-indexed matrix.

To force a full rebuild on every measurement, the graph index cache was explicitly invalidated before each indexing run.

It should also be noted that CLIP’s text encoder is limited to 77 tokens (roughly 300 characters in French). Longer inputs are truncated, so in practice CLIP only processes the beginning of a query. For the personal narratives of three to five sentences targeted here, this truncation is non-negligible. The proposed system, whose dense-retrieval branch uses sentence-camembert-large with a 512-token context window, encodes the full narrative without truncation. This asymmetry should be kept in mind when interpreting the inference-time comparison: part of CLIP’s speed advantage stems from the smaller amount of text it actually processes.

Results

Indexing time. As shown in Figure 6.2, the graph-based algorithm indexes the subset in approximately 5.3 seconds on average, against approximately 13.8 seconds for CLIP. However, this mean is not a reliable basis for concluding that the proposed system is faster to index. Its relative variability is about 118% (an absolute standard deviation of roughly 6 seconds for a 5.3-second mean), with individual runs ranging from a small fraction of the mean to well above it, whereas

CLIP’s standard deviation is about 1%. The graph-based figure is dominated by whether the CamemBERT embeddings are already cached: the first cold run is expensive and unpredictable, while warm runs are much cheaper. Because the standard deviation exceeds the mean, this comparison should be treated as exploratory and refrain from claiming a firm indexing speed advantage.

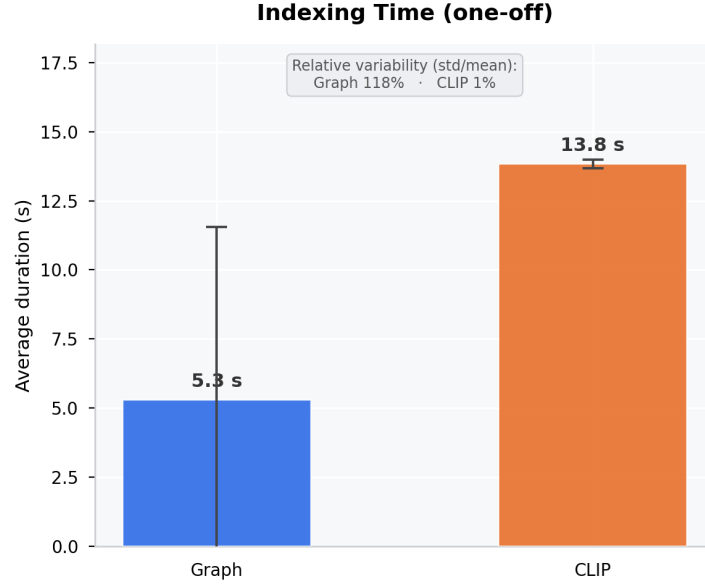


Figure 6.2: Indexing-time comparison between the proposed graph-based algorithm and CLIP.

Inference time. Figure 6.3 presents inference times per query on a logarithmic scale. CLIP’s inference time is consistently around 0.011 seconds across all four queries, since it always performs the same fixed operations (one forward pass through the text encoder followed by one dot product). Query content does not affect its cost.

The graph-based algorithm shows clear variation between queries, from about 0.233 seconds for the landscape query to about 0.548 seconds for the portrait query.

Despite CLIP’s speed advantage, the inference times of the proposed system remain well within acceptable bounds for the intended application: all four queries are processed in well under one second, far below the 5-second design requirement stated in Chapter 4, and compatible with interactive use.

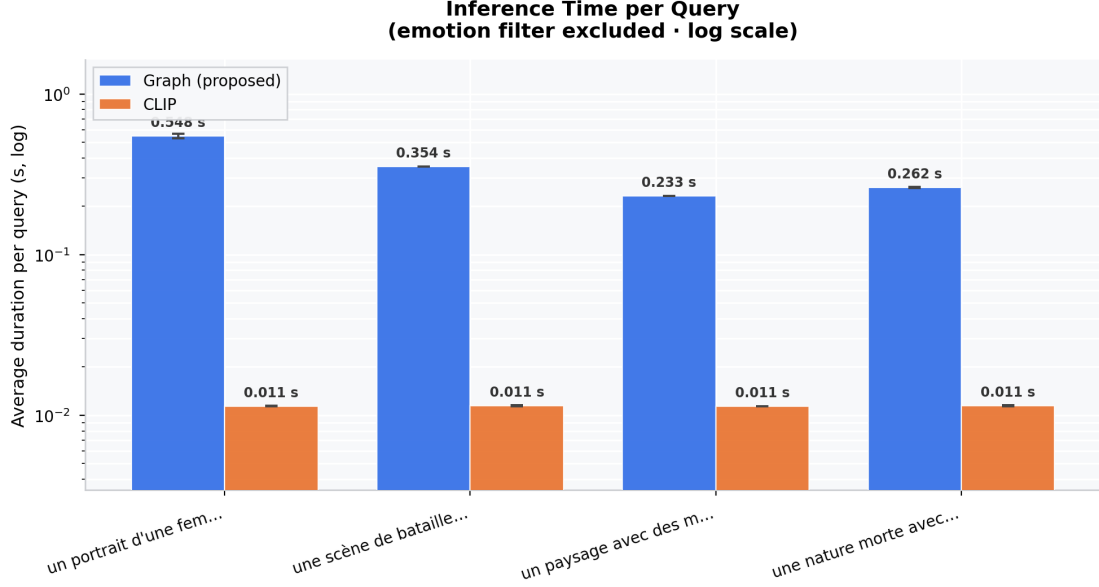


Figure 6.3: Inference time per query (emotion filter excluded) for the proposed algorithm and CLIP, on a logarithmic scale.

Memory consumption. Figure 6.4 presents the Δ RSS memory comparison during indexing. The graph-based system consumes approximately 1,416 MB, while CLIP requires approximately 594 MB.

The higher footprint of the proposed system follows directly from its rich internal representation. Every unique token across the 1,000 descriptions is embedded by CamemBERT and stored as a 1,024-dimensional float32 vector in a session cache that is not evicted. In addition, the NetworkX graph (node attributes, edge weights, structural metadata) and the FAISS index over all graph nodes contribute to the total. The 1,416 MB therefore represents the cost of keeping the full semantic structure that enables efficient Dijkstra-based lookups at inference time.

CLIP’s 594 MB reflects loading the ViT-B/32 weights and encoding 1,000 descriptions and the available images into 512-dimensional vectors, a thinner representation that stores no graph and no hierarchical structure.

Both systems show non-negligible standard deviations in their memory measurements, and the error bar on the graph-based system is particularly large. CLIP’s smaller but still non-negligible error bar suggests that even repeated model loading and corpus encoding are subject to similar caching variability.

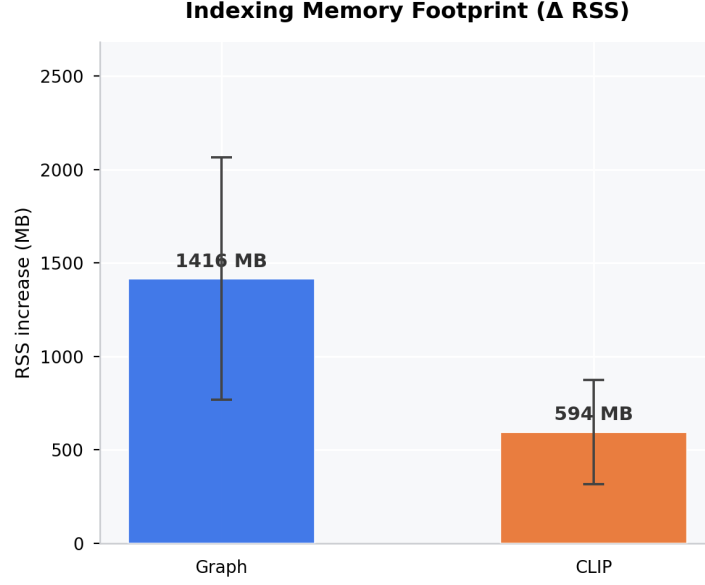


Figure 6.4: Memory ΔRSS during indexing for the proposed algorithm and CLIP.

6.5 Qualitative comparison with CLIP

Beyond quantitative performance, Figures 6.5–6.8 compare the top-five artworks returned by each method for the four test queries, with the proposed graph-based method on the top row and CLIP on the bottom row. For each method, the five highest-ranked artworks that have a locatable image file are shown.

For the portrait query (Figure 6.5, *un portrait d’une femme assise*), the proposed method returns a coherent set of seated portraits and interior scenes, led by *La dame au collier jaune*. CLIP also surfaces relevant portraits, including the same *La dame au collier jaune* at rank three, but its ranking mixes in looser matches such as a standing religious figure, suggesting that it keys on the visual notion of a single female figure more than on the specific notion of a seated pose. The two methods overlap on at least one artwork, which is reassuring given that they rank over entirely different representations.

Qualitative Comparison — Query 1: "un portrait d'une femme assise"



Figure 6.5: Qualitative comparison for Query 1 (un portrait d'une femme assise): top-five artworks from the proposed method (top row) and CLIP (bottom row).

For the battle query (Figure 6.6, *une scène de bataille historique*), both methods show crowded, multi-figure compositions. The graph method returns dense crowd scenes such as *Le combat de Carnaval*, while CLIP retrieves explicitly violent, crowded scenes such as *Épisode des Journées de...* and *La furie espagnole*. Here CLIP's results are arguably more accurate for the literal notion of a battle, whereas the graph method generalises towards large historical gatherings. Notably, *Les Trentains de Berthe...* appears in both rankings, confirming partial agreement.

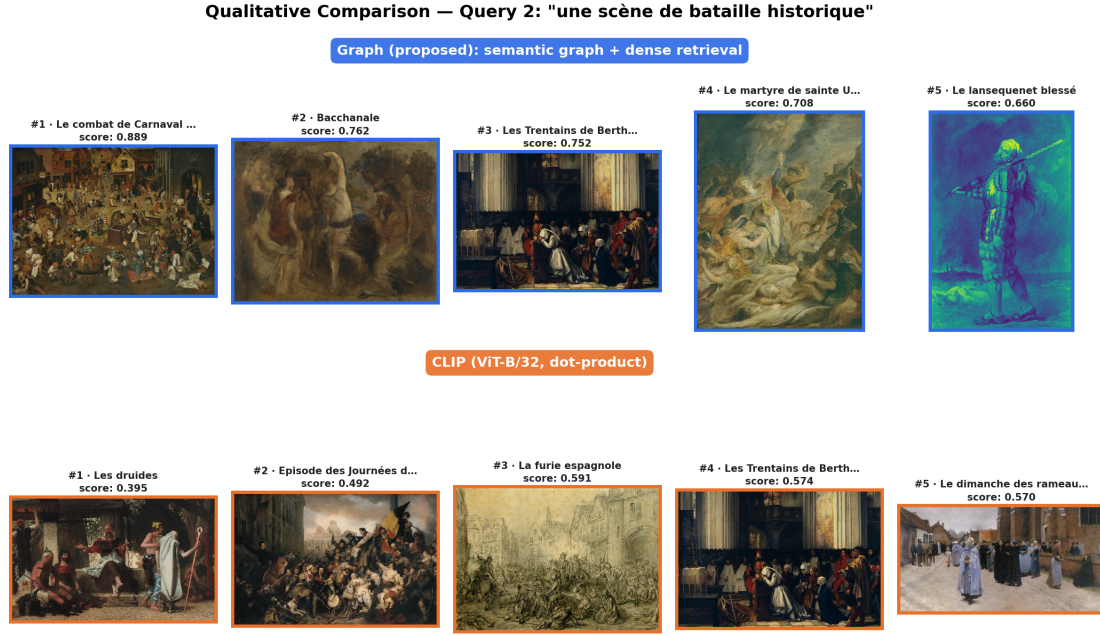


Figure 6.6: Qualitative comparison for Query 2 (*une scène de bataille historique*): top-five artworks from the proposed method (top row) and CLIP (bottom row).

For the landscape query (Figure 6.7, *un paysage avec des montagnes et un lac*), agreement between the two methods is strongest. Both retrieve *Le Mont Rose* and *Le passage du gué*, two landscapes with mountains and water that match the query almost literally, and both fill the remaining slots with further landscapes. This is the query for which the visual and semantic signals align most closely. A few graph results (such as a *Calvaire* and a *Triptyque*) are religious scenes set against a mountainous background, showing that the label-based approach can be drawn in by the landscape *setting* of an otherwise non-landscape work.

Qualitative Comparison — Query 3: "un paysage avec des montagnes et un lac"

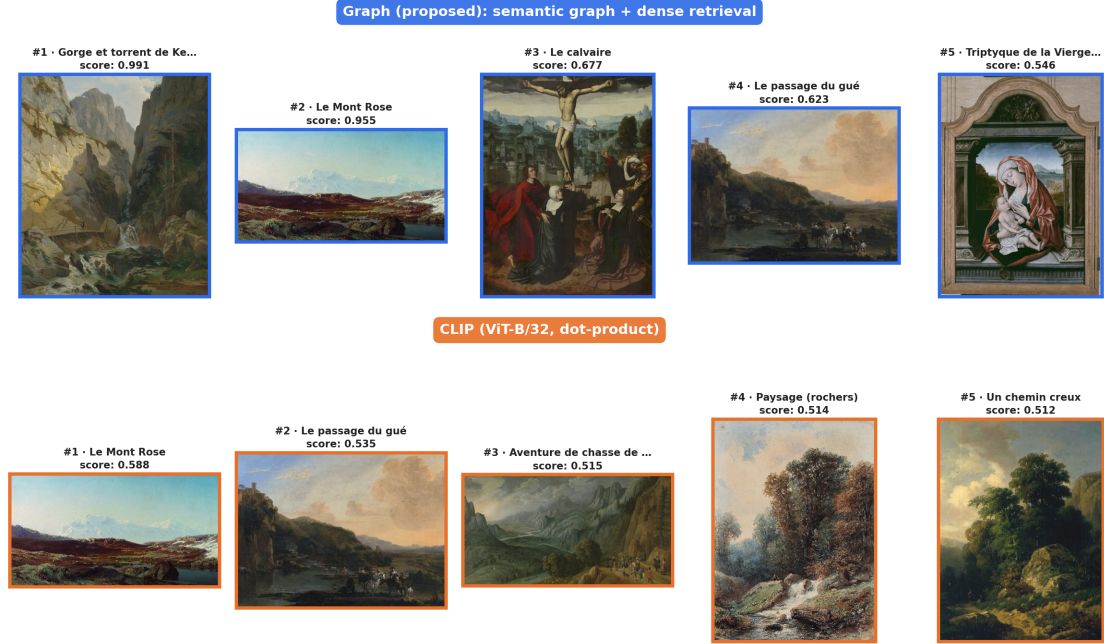


Figure 6.7: Qualitative comparison for Query 3 (un paysage avec des montagnes et un lac): top-five artworks from the proposed method (top row) and CLIP (bottom row).

For the still-life query (Figure 6.8, *une nature morte avec des fruits*), both methods converge on still lifes and pantry scenes: *Le garde-manger*, *Nature morte d'hiver*, and *Nature morte avec guirlande* all appear in both rankings, again showing substantial overlap. The graph method promotes a thematically similar kitchen scene (*La cuisinière*) to the top of its list, illustrating how the hierarchical label structure links food, kitchen, and still-life concepts.

Qualitative Comparison — Query 4: "une nature morte avec des fruits"

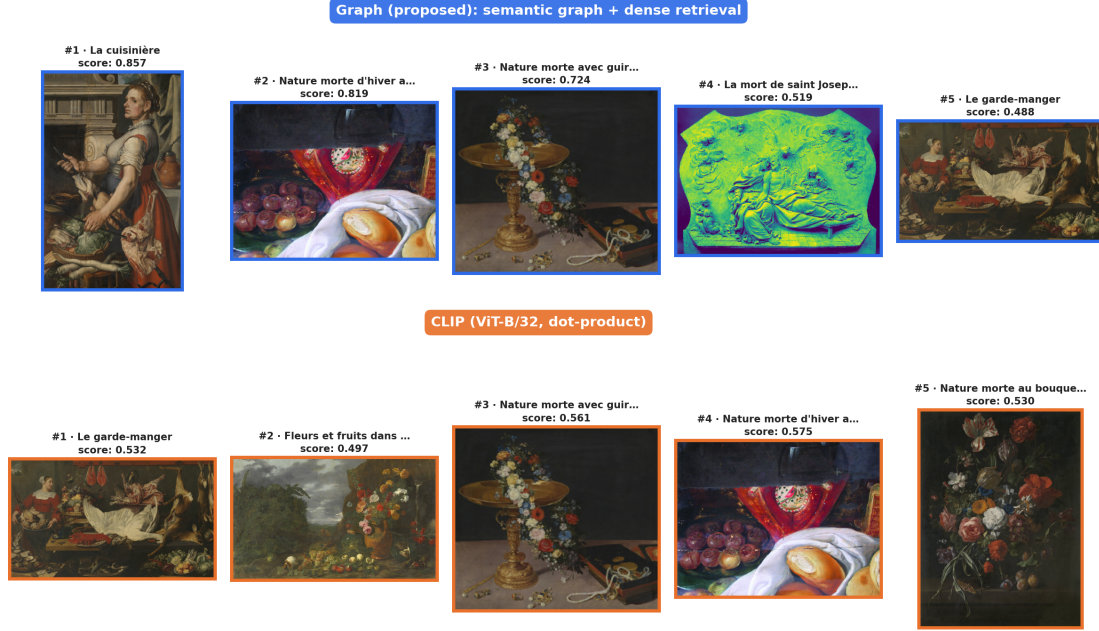


Figure 6.8: Qualitative comparison for Query 4 (une nature morte avec des fruits): top-five artworks from the proposed method (top row) and CLIP (bottom row).

Taken together, the qualitative results show that the two systems agree on at least one and often several artworks per query despite operating on fundamentally different representations, while their disagreements are interpretable: CLIP responds to direct visual resemblance, whereas the proposed method reasons over curated iconographic labels with hierarchy. This, combined with the inference times reported above, supports using the proposed graph-based method as an interpretable alternative to a purely visual baseline rather than as a direct competitor on raw speed.

Chapter 7

Limitations and Future Perspectives

Despite the promising preliminary results described in Chapter 6, several limitations remain, and some future work can be identified. First discussing the limitations of the proposed framework, follows a discussion about perspectives for future improvements.

7.1 Limitations

7.1.1 Loss of syntactic structure through tokenisation

A first limitation stems from the way input stories and artwork descriptions are processed. As described in Chapter 4, all text is tokenised into individual words after lowercasing, punctuation removal, and stopword filtering. While this representation is what makes the mapping onto graph nodes tractable, it discards word order and syntactic structure, and with them part of the meaning of the original sentence. Relations such as negation, the distinction between subject and object, or the association between a modifier and the word it qualifies are lost. Compositional cues of this kind can be important for correctly interpreting a personal narrative, and their absence may lead the system to retrieve an artwork that shares vocabulary with the query without genuinely matching its meaning. A representation that preserves more of the sentence structure, for instance by encoding multi-word expressions or short spans rather than isolated tokens, could reduce this loss.

7.1.2 Nature of the *iconographicInterpretation* field

A further limitation of the proposed framework concerns the nature of the *iconographicInterpretation* field itself, which serves as the primary textual anchor for

both the semantic graph and the emotion assignment. In practice, this field does not always describe the visual content of the artwork in a direct, literal way. In many cases it extends beyond what is depicted on the canvas to include broader historical context, biographical information about the artist, or commentary on the circumstances surrounding the work's creation. As a result, tokens extracted from this field may reflect the historical or authorial framing of the piece rather than its actual visual or emotional content. This introduces a form of noise into the graph: a caption label or a dominant emotion derived from contextual narrative rather than depicted subject matter can mislead retrieval, since the system implicitly assumes a tight coupling between text and image that does not always hold. Ideally, one would use a set of textual descriptions restricted to the artwork itself.

7.1.3 Variable emotional performance

As shown in Section 6.3, the emotional alignment achieved by Phase Two obtained a mean opinion score of only 3.1/5 (median 3), with ratings spread across the whole scale. This insufficient and inconsistent performance can be partly explained by the fact that each artwork is assigned a single dominant emotion during preprocessing, whereas in reality an artwork may simultaneously evoke several, sometimes conflicting, emotions. Reducing this rich emotional content to one label inevitably discards nuance that could be relevant. This limitation is particularly important in an art-therapy setting, where the emotional appropriateness of a suggestion is arguably as important as its semantic relevance.

A further source of imprecision is the projection onto nine emotion categories. Emotions freely entered by users (such as *nostalgie*, *fatigue*, *calme* or *bonheur*) do not belong to the predefined scheme and are mapped onto the closest of the nine categories in the embedding space. This projection necessarily loses information: two distinct feelings may collapse onto the same category, and a nuanced emotional state may be forced into a label that only approximates it. Part of the variable emotional performance observed in Section 6.3 is therefore attributable not only to the "single dominant emotion" assumption on the artwork side, but also to this coarse discretisation on the query side. Emotional evaluation is intrinsically difficult because of emotional subjectivity, emotional uncertainty, the cultural variability of emotions, and the fact that a single artwork admits multiple possible interpretations.

7.1.4 Small testing samples

Both testing protocols presented in Chapter 6 relied on a limited number of participants: ten for the controlled retrieval test, and ten for the baseline comparison and emotional alignment evaluation. While these samples were sufficient to obtain

an initial indication of the algorithm’s behaviour, they remain too small to draw statistically robust conclusions. No significance testing was carried out, and results such as the 80% recovery success rate or the 3.1/5 mean emotional score should be interpreted as preliminary rather than conclusive. Similarly, the CLIP benchmark in Section 6.4 used only four test queries over a subset of 1,000 artworks, which does not guarantee that the same behaviour would hold across a wider variety of queries, artistic styles, or museum collections.

7.1.5 Heavy dependence on heuristic thresholds

Our framework relies on a number of thresholds and hyperparameters determined heuristically rather than through systematic optimisation. These include the similarity threshold of 0.45 used to map query tokens onto graph nodes, the fallback weight multiplier of $\times 0.6$, the maximum propagation cutoff of 4.0 used in Dijkstra’s algorithm and the cosine-similarity threshold of 0.5 used to build visual-similarity edges for undescribed artworks. These values were chosen from empirical observation during development; they were not validated through a formal validation, or sensitivity analysis. As a result, the robustness of the system to these choices remains uncertain, and its performance could improve, or degrade unexpectedly, if the system were applied to a different dataset or a different linguistic and cultural context without fine-tuning these parameters.

A systematic ablation study would be especially valuable here: measuring retrieval quality with the hierarchy removed, without co-occurrence edges, without IDF weighting, without degree regularisation, and without the visual-similarity links would quantify what each component actually contributes, and would help distinguish essential mechanisms from those that could be simplified away.

7.1.6 Risk of triggering distressing memories

A limitation specific to the therapeutic context is the risk of triggering unwanted memories or emotions. Reminiscence is not always unrisks: a retrieved artwork could evoke a traumatic episode, and in a vulnerable population this could cause real distress. The system currently has no mechanism for vulnerability detection or for identifying content likely to be associated with trauma, and it applies the same retrieval logic regardless of the emotional fragility of the user.

This is a serious concern that reinforces the need for a human professional to mediate every interaction.

7.1.7 Ethical considerations

Because the intended application touches on vulnerable users and personal memories, several ethical issues deserve explicit attention before any deployment. Lets separate them in terms of the system’s ethical strengths and weaknesses.

Strengths A deliberate strength of our design is that it keeps a **human in the loop**: the system is intended to support, not replace, the judgement of caregivers and therapists, who remain responsible for deciding whether and how a retrieved artwork is shown to a patient. Combined with the **transparent reasoning** discussed in Chapter 5, this makes the system auditable and its suggestions contestable, which is a meaningful ethical advantage over opaque automated tools.

Weaknesses The main ethical weaknesses concern the emotional dimension. The emotion modelling is oversimplified (a single dominant label per artwork, projected onto nine categories), and there is a genuine potential for harmful emotional triggering, discussed separately in Section 7.1.6.

Privacy The stories submitted by relatives are personal and potentially sensitive data. Any deployment beyond research would need to address, at minimum, **data retention** (how long memories are stored), **deletion rights** (allowing users to remove their data at any time), informed **consent**, and the **secure storage** of memories. These requirements are not merely legal: they concern the dignity and autonomy of vulnerable users.

Prototype status Finally, it should be noted that the artefact developed in this thesis is a **research prototype and not a finished medical device**. It has not been evaluated with patients, it has not undergone any clinical validation, and it must not be understood as a diagnostic or therapeutic instrument. A design with humans in the loop should be considered and informed consent should be prerequisites for any use beyond research.

7.2 Future Perspectives

Lets organise future work along two levels: higher-level, conceptual and application directions on the one hand, and lower-level technical improvements on the other.

7.2.1 Higher-level and application directions

Integration with previous work

A natural first direction would be to merge the present system with the work of Victor Rijks [27]. His contribution is complementary to our graph-based, interpretable retrieval. Combining the two could yield a system that is both richer in data and transparent in its reasoning.

Clinical validation with patients and caregivers

All the evaluations conducted in this thesis involved general participants describing artworks or personal memories, rather than actual patients, their relatives, or professional therapists. An important next step would be a clinical study conducted directly within an art-therapy setting, involving real patients and caregivers, to assess whether the tool genuinely supports the therapeutic goals outlined in Chapter 2: triggering autobiographical memories, inducing emotional engagement, and facilitating meaningful interaction between patients and relatives. Such a study would also allow the evaluation of outcomes beyond retrieval accuracy or emotional-alignment scores, such as mood improvement, level of engagement, or the quality of the conversations sparked by the retrieved artworks.

Emotional safety studies

Given the risk of triggering distressing memories (Section 7.1.6), a dedicated line of work would concern emotional-safety studies: mechanisms for trauma-aware filtering, vulnerability detection, and defined exclusion of sensitive themes.

Bias auditing and formal explainability evaluation

Logically following from Chapter 5, future work could turn the qualitative discussion of bias and explainability into a systematic evaluation. This includes bias auditing of retrieval pathways (quantifying, for instance, gendered or cultural associations encoded in the co-occurrence edges) and a formal explainability evaluation measuring whether the traced paths are genuinely useful to caregivers.

Serious games

As mentioned in Chapter 2, serious games combine entertainment with therapeutic or cognitive objectives, and have shown promise for stimulating memory, attention, and social interaction in elderly and dementia populations. A promising extension would be to embed the artwork-retrieval engine within a serious game, turning the reminiscence process into a more playful and engaging experience.

Mobile application

In a real art-therapy or family setting, a relative or caregiver is more likely to have a smartphone or tablet at hand than a desktop computer. Packaging the system as a mobile application, with a simple interface adapted to elderly users and non-technical caregivers, would make it easier to submit a memory and view the retrieved artwork. Because the current inference already runs well under the five-second target, most of the heavy computation could be performed on a server and exposed through a lightweight mobile client.

A formal theory of taste

A more abstract but plausible direction would be to move beyond semantic and emotional alignment toward an explicit modelling of individual taste. Where the current framework retrieves artworks that match the content and mood of a narrative, a further step would be to recommend artworks that resonate with a user’s aesthetic preferences, learned from their prior reactions and choices. This raises questions that are simultaneously mathematical and philosophical: taste may be formalised as a metric structure over the space of artworks, yet doing so inevitably confronts long-standing debates on whether aesthetic judgement is subjective, intersubjective, or grounded in shared perceptual invariants. Encoding taste in the graph, for instance through user specific preference nodes or personalised edge weights, would allow recommendations to be traced and audited in the same interpretable manner as the rest of the pipeline. Given the conceptual and empirical difficulty of even defining a ground truth for taste, this line of work would first take the form of a feasibility study, aimed at clarifying which aspects of aesthetic preference can be meaningfully quantified and which resist formalisation, before any predictive system is built.

7.2.2 Technical improvements

Richer emotion detection

The oversimplified emotion model could be replaced by a dedicated emotion-detection system. In particular, the approach developed by Sébastien Lugan on the emotional response to art¹ is a promising candidate for producing a more faithful emotional characterisation of artworks than the single dominant emotion assignment used here.

¹<https://siamese.sd2.net/blog/2026/0216/when-art-touches-the-mind.html>

Implementation of multi-threading

Finally, the benchmark presented in Section 6.4 highlighted two aspects of performance that could be improved through parallelisation. First, the indexing phase exhibited very high relative variability (approximately 118%), due in part to the sequential embedding of every unique token and the sequential construction of the NetworkX graph; distributing these computations across multiple CPU cores or GPU threads could accelerate indexing. Second, the inference phase runs Dijkstra’s algorithm independently from each query token to all image nodes; since these shortest-path computations are independent, they are natural candidates for parallel execution using Python’s `multiprocessing` module.

Chapter 8

Conclusion

To conclude, this master’s thesis set out to answer the following question: can graph-based retrieval provide a more interpretable alternative to CLIP-like retrieval for memory-triggering applications, in a context where being able to explain a recommendation matters as much as the recommendation itself?

Our results suggest that it can. By representing images, captions, hierarchical museum labels, and emotions as nodes in a single heterogeneous graph, and by using Dijkstra’s algorithm to score artworks against a query narrative, the reasoning of the retrieval mechanism obtained can be traced from beginning to end.

The two-phase design (semantic retrieval followed by emotional retrieval) made the algorithm more tractable, and it loosely mirrors how a human curator or therapist might approach the same task.

The evaluations support this design while remaining modest in scope. The controlled retrieval test reached an 80% success rate, with the two failures both involving portraits of specific historical figures, for which participants’ descriptions were understandably generic. Against the lexical baseline, participants preferred our system’s semantic retrieval in 7 of 10 comparisons, which is encouraging but, again, based on a small sample. Emotional alignment is the weakest link: a mean opinion score of 3.1/5 shows that Phase Two works, but inconsistently, most likely because reducing an artwork’s emotional content to a single dominant label discards real nuance.

Compared with CLIP, our system is not faster and uses considerably more memory, but it stays well under the five-second target set in Chapter 4. The main goal here was not to directly compete with CLIP on a speed level, but rather to offer an interpretable alternative and, from this standpoint, our algorithm has completed its mission.

Taken together, these findings suggest that, for a specific use case where transparency is a requirement, trading a modest amount of performance for a clear decision path is a reasonable trade-off. It should be noted that the system explains

its own reasoning process, not necessarily human memory formation, as discussed in Chapter 5.

The limitations laid out in Chapter 7 are non-negligible, and they define what would need to happen before this tool could be used in an actual clinical setting.

More tuning, more testing, and more clinical validation would be needed before this framework could genuinely help a patient and their family. The evaluations reported here, however, point in a promising direction.

Beyond the empirical results, this work positions itself at the intersection of several research strands. Relative to retrieval systems similar to CLIP, it trades a degree of raw performance for a transparent decision path. Relative to the literature on explainable AI [44], it offers a concrete case where the explanation is structural. Relative to cultural-heritage AI and responsible AI [45], it treats bias auditing and human oversight as important design concerns.

Use of AI

Claude AI was used to correct typos, incoherent formulation and translate some sentences. It was also used to incorporate clean comments into the code.

Code repository

The full code for this master’s thesis is available on: https://github.com/Safiyand149/Master_Thesis_Safiya_Nouidei.

Bibliography

- [1] Francesco Angelucci et al. “Integrating AI in fighting advancing Alzheimer: diagnosis, prevention, treatment, monitoring, mechanisms, and clinical trials”. en. In: *Current Opinion in Structural Biology* 87 (Aug. 2024), p. 102857. ISSN: 0959440X. DOI: [10.1016/j.sbi.2024.102857](https://doi.org/10.1016/j.sbi.2024.102857). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0959440X24000848>.
- [2] Phoebe Institute on Aging. *The Visual Arts and Alzheimer’s Disease*. Online video. 2015. URL: https://youtu.be/_OkBjTNiR1Y?si=MladGOUchMaz0OVf.
- [3] Sheung-Tak Cheng. “Dementia Caregiver Burden: a Research Update and Critical Analysis”. In: *Current Psychiatry Reports* 19.9 (2017), p. 64. ISSN: 1523-3812. DOI: [10.1007/s11920-017-0818-2](https://doi.org/10.1007/s11920-017-0818-2). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5550537/>.
- [4] John D. Omura et al. “Modifiable Risk Factors for Alzheimer Disease and Related Dementias Among Adults Aged 45 Years — United States, 2019”. In: *MMWR. Morbidity and Mortality Weekly Report* 71.20 (May 2022), pp. 680–685. ISSN: 0149-2195, 1545-861X. DOI: [10.15585/mmwr.mm7120a2](https://doi.org/10.15585/mmwr.mm7120a2). URL: http://www.cdc.gov/mmwr/volumes/71/wr/mm7120a2.htm?s_cid=mm7120a2_w.
- [5] Abigail Dove et al. “The impact of diabetes on cognitive impairment and its progression to dementia”. eng. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 17.11 (Nov. 2021), pp. 1769–1778. ISSN: 1552-5279. DOI: [10.1002/alz.12482](https://doi.org/10.1002/alz.12482).
- [6] Teresa Pardo-Moreno et al. “Therapeutic Approach to Alzheimer’s Disease: Current Treatments and New Perspectives”. eng. In: *Pharmaceutics* 14.6 (May 2022), p. 1117. ISSN: 1999-4923. DOI: [10.3390/pharmaceutics14061117](https://doi.org/10.3390/pharmaceutics14061117).
- [7] Kirsten Hötting and Brigitte Röder. “Beneficial effects of physical exercise on neuroplasticity and cognition”. en. In: *Neuroscience & Biobehavioral Reviews* 37.9 (Nov. 2013), pp. 2243–2257. ISSN: 01497634. DOI: [10.1016/j.neubiorev.2013.04.005](https://doi.org/10.1016/j.neubiorev.2013.04.005). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0149763413001012>.

- [8] Bob Woods et al. “Cognitive stimulation to improve cognitive functioning in people with dementia”. In: *The Cochrane Database of Systematic Reviews* 2023.1 (Jan. 2023), p. CD005562. ISSN: 1469-493X. DOI: [10.1002/14651858.CD005562.pub3](https://doi.org/10.1002/14651858.CD005562.pub3). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9891430/>.
- [9] Shan Wang et al. “The efficacy of reminiscence therapy on the cognition of older patients with cognitive impairment or dementia: a meta-analysis based on regulatory factors”. en. In: *Aging Clinical and Experimental Research* 38.1 (Jan. 2026), p. 45. ISSN: 1720-8319. DOI: [10.1007/s40520-025-03300-4](https://doi.org/10.1007/s40520-025-03300-4). URL: <https://doi.org/10.1007/s40520-025-03300-4>.
- [10] Dahlia W. Zaidel. “Art and brain: insights from neuropsychology, biology and evolution”. eng. In: *Journal of Anatomy* 216.2 (Feb. 2010), pp. 177–183. ISSN: 1469-7580. DOI: [10.1111/j.1469-7580.2009.01099.x](https://doi.org/10.1111/j.1469-7580.2009.01099.x).
- [11] Manuela Campisi et al. “Mitigating cellular aging and enhancing cognitive functionality: visual arts-mediated Cognitive Activation Therapy in neurocognitive disorders”. eng. In: *Frontiers in Aging Neuroscience* 16 (2024), p. 1354025. ISSN: 1663-4365. DOI: [10.3389/fnagi.2024.1354025](https://doi.org/10.3389/fnagi.2024.1354025).
- [12] Tiffany Tong, Jonathan H. Chan, and Mark Chignell. “Serious Games for Dementia”. en. In: *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. Perth, Australia: ACM Press, 2017, pp. 1111–1115. ISBN: 9781450349147. DOI: [10.1145/3041021.3054930](https://doi.org/10.1145/3041021.3054930). URL: <http://dl.acm.org/citation.cfm?doid=3041021.3054930>.
- [13] Diana Castilla et al. “Process of design and usability evaluation of a telepsychology web and virtual reality system for the elderly: Butler”. en. In: *International Journal of Human-Computer Studies* 71.3 (Mar. 2013), pp. 350–362. ISSN: 10715819. DOI: [10.1016/j.ijhcs.2012.10.017](https://doi.org/10.1016/j.ijhcs.2012.10.017). URL: <https://linkinghub.elsevier.com/retrieve/pii/S1071581912001772>.
- [14] Bruno Bouchard et al. “Developing serious games specifically adapted to people suffering from Alzheimer”. In: ed. by Minhua Ma et al. Vol. 7528. Berlin: Springer, 2012, pp. 243–254. DOI: [10.1007/978-3-642-33687-4_21](https://doi.org/10.1007/978-3-642-33687-4_21). URL: http://dx.doi.org/doi:10.1007/978-3-642-33687-4_21.
- [15] Dawei Huang et al. “From Large Language Models to Large Multimodal Models: A Literature Review”. en. In: *Applied Sciences* 14.12 (June 2024), p. 5068. ISSN: 2076-3417. DOI: [10.3390/app14125068](https://doi.org/10.3390/app14125068). URL: <https://www.mdpi.com/2076-3417/14/12/5068>.

- [16] Karan Singhal et al. “Large language models encode clinical knowledge”. In: *Nature* 620.7972 (2023), pp. 172–180. ISSN: 0028-0836. DOI: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10396962/>.
- [17] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. arXiv:1506.02640. May 2016. DOI: [10.48550/arXiv.1506.02640](https://doi.org/10.48550/arXiv.1506.02640). URL: <http://arxiv.org/abs/1506.02640>.
- [18] Rahima Khanam and Muhammad Hussain. *YOLOv11: An Overview of the Key Architectural Enhancements*. arXiv:2410.17725. Oct. 2024. DOI: [10.48550/arXiv.2410.17725](https://doi.org/10.48550/arXiv.2410.17725). URL: <http://arxiv.org/abs/2410.17725>.
- [19] Gary Marchionini. “Exploratory search: from finding to understanding”. en. In: *Communications of the ACM* 49.4 (Apr. 2006), pp. 41–46. ISSN: 0001-0782, 1557-7317. DOI: [10.1145/1121949.1121979](https://doi.org/10.1145/1121949.1121979). URL: <https://dl.acm.org/doi/10.1145/1121949.1121979>.
- [20] Emilie Palagi et al. “A Survey of Definitions and Models of Exploratory Search”. en. In: *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*. Limassol Cyprus: ACM, Mar. 2017, pp. 3–8. ISBN: 9781450349031. DOI: [10.1145/3038462.3038465](https://doi.org/10.1145/3038462.3038465). URL: <https://dl.acm.org/doi/10.1145/3038462.3038465>.
- [21] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020. Feb. 2021. DOI: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020). URL: <http://arxiv.org/abs/2103.00020>.
- [22] Alberto Baldrati et al. “Exploiting CLIP-based Multi-modal Approach for Artwork Classification and Retrieval”. In: (2023). DOI: [10.48550/ARXIV.2309.12110](https://doi.org/10.48550/ARXIV.2309.12110). URL: <https://arxiv.org/abs/2309.12110>.
- [23] Yi Li et al. “A closer look at the explainability of Contrastive language-image pre-training”. In: *Pattern Recognition* 162 (2025), p. 111409. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2025.111409>. URL: <https://www.sciencedirect.com/science/article/pii/S003132032500069X>.
- [24] Beichen Zhang et al. *Long-CLIP: Unlocking the Long-Text Capability of CLIP*. en. Mar. 2024. URL: <https://arxiv.org/abs/2403.15378v3>.
- [25] Matthijs Douze et al. *The Faiss library*. arXiv:2401.08281. Oct. 2025. DOI: [10.48550/arXiv.2401.08281](https://doi.org/10.48550/arXiv.2401.08281). URL: <http://arxiv.org/abs/2401.08281>.
- [26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. *Billion-scale similarity search with GPUs*. arXiv:1702.08734. Feb. 2017. DOI: [10.48550/arXiv.1702.08734](https://doi.org/10.48550/arXiv.1702.08734). URL: <http://arxiv.org/abs/1702.08734>.

- [27] Victor Rijks. “Preserving cultural heritage with AI: Addressing data challenges for museums and memory support”. eng. In: (2025). URL: <https://hdl.handle.net/2078.2/43800>.
- [28] Augusto Barbosa Villar Silva and Marc Bejjani. “ArtEvoke: AI-powered platform to trigger memories in Alzheimer’s patients using art imagery”. eng. In: (2025). URL: <https://hdl.handle.net/2078.2/43752>.
- [29] Zonghan Wu et al. “A Comprehensive Survey on Graph Neural Networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), pp. 4–24. ISSN: 2162-2388. DOI: [10.1109/TNNLS.2020.2978386](https://doi.org/10.1109/TNNLS.2020.2978386). URL: <https://ieeexplore.ieee.org/document/9046288/>.
- [30] Yanbei Jiang, Krista A. Ehinger, and Jey Han Lau. *KALE: An Artwork Image Captioning System Augmented with Heterogeneous Graph*. en. Sept. 2024. URL: <https://arxiv.org/abs/2409.10921v1>.
- [31] Muhammad Ahsan Khan. “A Comprehensive Study of Dijkstra’s Algorithm”. In: *SSRN Electronic Journal* (2023). URL: <https://api.semanticscholar.org/CorpusID:263005984>.
- [32] mihaly. *Packet-Switching Networks*. en. Oct. 2014. URL: <https://www.slideserve.com/mihaly/packet-switching-networks>.
- [33] Chen Gao et al. “A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions”. en. In: *ACM Transactions on Recommender Systems* 1.1 (Mar. 2023), pp. 1–51. ISSN: 2770-6699, 2770-6699. DOI: [10.1145/3568022](https://doi.org/10.1145/3568022). URL: <https://dl.acm.org/doi/10.1145/3568022>.
- [34] *Thésaurus iconographique, système descriptif des représentations de François Garnier / Ministère de la Culture*. fr-FR. June 2020. URL: <https://www.culture.gouv.fr/thematiques/musees/pour-les-professionnels/conserver-et-gerer-les-collections/informatiser-les-collections-d-un-musee-de-france/vocabulaires-scientifiques-du-service-des-musees-de-france/thesaurus-iconographique-systeme-descriptif-des-representations-de-francois-garnier>.
- [35] Mathieu Ciancone et al. *MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis*. arXiv:2405.20468. June 2024. DOI: [10.48550/arXiv.2405.20468](https://doi.org/10.48550/arXiv.2405.20468). URL: <http://arxiv.org/abs/2405.20468>.
- [36] Jing Zhang et al. “Web image annotation based on Tri-relational Graph and semantic context analysis”. en. In: *Engineering Applications of Artificial Intelligence* 81 (May 2019), pp. 313–322. ISSN: 09521976. DOI: [10.1016/j.engappai.2019.02.018](https://doi.org/10.1016/j.engappai.2019.02.018). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0952197619300429>.

- [37] *Find Shortest Paths from Source to all Vertices using Dijkstra's Algorithm*. en. Nov. 2012. URL: <https://www.geeksforgeeks.org/dsa/dijkstras-shortest-path-algorithm-greedy-algo-7/>.
- [38] *ArtEmis: Affective Language for Visual Art*. URL: <https://artemisdataset.org>.
- [39] Leilani H. Gilpin et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: (2018). DOI: [10.48550/ARXIV.1806.00069](https://arxiv.org/abs/1806.00069). URL: <https://arxiv.org/abs/1806.00069>.
- [40] Cynthia Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: (2018). DOI: [10.48550/ARXIV.1811.10154](https://arxiv.org/abs/1811.10154). URL: <https://arxiv.org/abs/1811.10154>.
- [41] Alon Jacovi and Yoav Goldberg. *Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* 2020. DOI: [10.48550/ARXIV.2004.03685](https://arxiv.org/abs/2004.03685). URL: <https://arxiv.org/abs/2004.03685>.
- [42] “Bias types and patterns: a typology applied to Europeana use cases”. In: (). DOI: [10.5281/ZENODO.14514378](https://zenodo.org/doi/10.5281/zenodo.14514378). URL: <https://zenodo.org/doi/10.5281/zenodo.14514378>.
- [43] *What Is MOS Score? Mean Opinion Score Explained*. en-US. Mar. 2026. URL: <https://scienceinsights.org/what-is-mos-score-mean-opinion-score-explained/>.
- [44] Bowen Long et al. *Explainable AI the Latest Advancements and New Trends*. 2025. DOI: [10.48550/ARXIV.2505.07005](https://arxiv.org/abs/2505.07005). URL: <https://arxiv.org/abs/2505.07005>.
- [45] Marina Paolanti, Emanuele Frontoni, and Roberto Pierdicca. “Towards trustworthy AI in cultural heritage”. en. In: *npj Heritage Science* 14.1 (Mar. 2026), p. 131. ISSN: 3059-3220. DOI: [10.1038/s40494-026-02403-z](https://www.nature.com/articles/s40494-026-02403-z). URL: <https://www.nature.com/articles/s40494-026-02403-z>.

Appendix A

Detailed test results by participant

Test 1: Recovering the Target Image

Participant 1

- **Artwork:** Le déjeuner - Isidore Verheyden
- **Input (FR):** Une petite fille boit un café ou un thé avec des biscuits. Un bouquet de fleurs décore la table. Elle est seule assise sur sa chaise.
- **Translation (EN):** A little girl is drinking coffee or tea with biscuits. A bouquet of flowers decorates the table. She is sitting alone on her chair.
- **Result:** success (first image)

Participant 2

- **Artwork:** Louis Faider à trois ans - Louis Gallait
- **Input (FR):** Une fille avec une canne, à côté d'un chien noir et gris. Elle est habillée avec une veste à boutons.
- **Translation (EN):** A girl with a cane, next to a black and grey dog. She is dressed in a jacket with buttons.
- **Result:** failure

Participant 3

- **Artwork:** Vase de fleurs - Jacob van Walscapelle
- **Input (FR):** Je vois un bouquet de fleurs avec des géraniums, un mélange de couleurs.

- **Translation (EN):** I see a bouquet of flowers with geraniums, a mix of colors.
- **Result:** success (second image)

Participant 4

- **Artwork:** Vase de fleurs - Jacob van Walscapelle
- **Input (FR):** Il y a un bouquet de fleurs formé de fleurs jaunes, blanches et roses et d'une pivoine. L'ambiance est sombre.
- **Translation (EN):** There is a bouquet of flowers made up of yellow, white, and pink flowers, and a peony. The mood is dark.
- **Result:** success (first image)

Participant 5

- **Artwork:** Le déjeuner - Isidore Verheyden
- **Input (FR):** Une fille en robe orange boit du thé avec des fleurs dans un vase à côté d'elle. Elle mange son goûter.
- **Translation (EN):** A girl in an orange dress is drinking tea with flowers in a vase next to her. She is eating her snack.
- **Result:** success (first image)

Participant 6

- **Artwork:** Louis Faider à trois ans - Louis Gallait
- **Input (FR):** Un petit garçon de noir vêtu portant un chapeau pose avec son chien. Il tient dans sa main droite une canne en bois.
- **Translation (EN):** A little boy dressed in black, wearing a hat, poses with his dog. He holds a wooden cane in his right hand.
- **Result:** success (first image)

Participant 7

- **Artwork:** Un port en Orient - Bonaventura Peeters I
- **Input (FR):** Je vois un port avec beaucoup de bateaux. Il y a plusieurs drapeaux.
- **Translation (EN):** I see a port with many boats. There are several flags.
- **Result:** success (first image)

Participant 8

- **Artwork:** La veillée. Jeanne la Folle - Gustaf Wappers
- **Input (FR):** Il y a une dame assise par terre, en colère. L'image est en noir et blanc. Il y a aussi une bougie et des rideaux.
- **Translation (EN):** There is a lady sitting on the ground, angry. The image is in black and white. There is also a candle and curtains.
- **Result:** failure

Participant 9

- **Artwork:** Un port en Orient - Bonaventura Peeters I
- **Input (FR):** Il y a un château fort, des bateaux avec des drapeaux de différents pays et des nuages. Les bateaux accostent.
- **Translation (EN):** There is a fortified castle, boats with flags from different countries, and clouds. The boats are docking.
- **Result:** success (first image)

Participant 10

- **Artwork:** La veillée. Jeanne la Folle - Gustaf Wappers
- **Input (FR):** Une femme assise avec un regard noir, un crucifix et une bougie sur la table ; l'image est en nuances de gris.
- **Translation (EN):** A seated woman with a dark look, a crucifix, and a candle on the table; the image is in shades of grey.
- **Result:** success (first image)

Test 2: Comparison with Baseline

Participant 1

- **Input (FR):** Des enfants jouent dans un jardin fleuri ensoleillé. L'un fait de la balançoire, les autres jouent au ballon.
- **Translation (EN):** Children are playing in a sunny, flowery garden. One is on the swing, the others are playing ball.
- **Result phase 1:** Failure
- **Emotion:** Joie
- **Result phase 2:** 4/5

Participant 2

- **Input (FR):** J'aime jouer au ping-pong avec mon cousin. Il a 11 ans, les cheveux bouclés et aime jouer aux échecs.
- **Translation (EN):** I like playing ping-pong with my cousin. He is 11 years old, has curly hair, and likes playing chess.
- **Result phase 1:** Success
- **Emotion:** Joie
- **Result phase 2:** 3/5

Participant 3

- **Input (FR):** Je jouais aux Legos dans ma chambre puis je les rangeais dans les tiroirs de mon lit.
- **Translation (EN):** I used to play with Legos in my room, then I would put them away in my bed's drawers.
- **Result phase 1:** No preference
- **Emotion:** Amusement
- **Result phase 2:** 3/5

Participant 4

- **Input (FR):** Dans la cuisine, les gens s'activaient. Une bonne odeur de viande rôtie se répandait.
- **Translation (EN):** In the kitchen, people were bustling about. A good smell of roasted meat was spreading.
- **Result phase 1:** Success
- **Emotion:** Nostalgie
- **Result phase 2:** 5/5

Participant 5

- **Input (FR):** Un ogre aimait les pommes. Un jour, il tomba sur un pommier avec des pommes luisantes. Il décida d'en cueillir et de faire une tarte.
- **Translation (EN):** An ogre loved apples. One day, he came across an apple tree with shiny apples. He decided to pick some and make a pie.
- **Result phase 1:** Failure
- **Emotion:** Fatigue
- **Result phase 2:** 2/5

Participant 6

- **Input (FR):** C'est la fête du village, en été. La foule danse au son de la musique de l'orchestre.
- **Translation (EN):** It's the village festival, in summer. The crowd dances to the sound of the orchestra's music.
- **Result phase 1:** Success
- **Emotion:** Bonheur
- **Result phase 2:** 4/5

Participant 7

- **Input (FR):** Lorsqu'on était enfants, mon frère et moi jouions dans les parcs. Il y avait beaucoup de gens qui promenaient leurs chiens et mon frère s'est fait mordre.
- **Translation (EN):** When we were children, my brother and I used to play in the parks. There were a lot of people walking their dogs, and my brother got bitten.
- **Result phase 1:** Success
- **Emotion:** Peur
- **Result phase 2:** 4/5

Participant 8

- **Input (FR):** Mon papy rêvait d'être astronome. La nuit, il observait les étoiles et les planètes.
- **Translation (EN):** My grandpa dreamed of being an astronomer. At night, he would observe the stars and the planets.
- **Result phase 1:** Success
- **Emotion:** Calme
- **Result phase 2:** 1/5

Participant 9

- **Input (FR):** J'allais cueillir des fleurs avec ma tante dans les champs, près de mon village.
- **Translation (EN):** I used to go pick flowers with my aunt in the fields, near my village.
- **Result phase 1:** Failure
- **Emotion:** Joie
- **Result phase 2:** 2/5

Participant 10

- **Input (FR):** Ma cousine a toujours aimé les chevaux et les sports équestres.
- **Translation (EN):** My cousin has always loved horses and equestrian sports.
- **Result phase 1:** Success
- **Emotion:** Nostalgie
- **Result phase 2:** 3/5

Appendix B

Pseudo code of the main algorithm

Algorithm 2: Graph-based artwork retrieval from a narrative	
Input: Story query q , target emotion e , artwork dataset	
Output: One retrieved artwork	
1	Build the graph (once);
2	Preprocess every artwork description into filtered tokens (caption labels);
3	Extract hierarchical labels from the museum fields, with a depth per label;
4	Assign one dominant emotion to each artwork (closest emotion embedding);
5	Count label co-occurrences to weight label-label edges;
6	Build a graph with image, caption, label, and emotion nodes;
7	Link description-less artworks to visually similar ones (ResNet50);
8	Convert every edge weight into a distance ($\text{dist} = 1/\text{weight}$);
9	Phase 1: Semantic retrieval;
10	Tokenise q , add its detected emotion, map each token to graph nodes;
11	foreach <i>mapped query label</i> l do
12	Run Dijkstra from l ; each reachable image gains $\frac{\text{IDF}(l) \cdot \text{sim}(l) \cdot \text{boost}(l)}{1 + \text{dist}}$;
13	end
14	Apply a coverage bonus and normalise by $\sqrt{\text{deg}(\text{image})}$;
15	Also score artworks by dense similarity (full-text embeddings);
16	Fuse graph and dense scores; keep the top 5 artworks with an image;
17	Phase 2: Emotional alignment;
18	foreach <i>candidate in the top 5</i> do
19	Score = similarity(candidate emotion, e);
20	Penalise heavily if its valence is opposite to e ;
21	end
22	return the candidate with the highest score;

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl