TU WIEN Informatics

SaCoDi: Semi-Automatic Concept Discovery Utilizing Multimodal Models and Interactive Refinement

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Visual Computing

eingereicht von

Dominik Wolf, BSc
Matrikelnummer 0925239

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Associate Prof. Dr.in techn. Manuela Waldner, MSc
Mitwirkung: Univ.Ass. Dipl.-Ing. Johannes Eschner, BSc

Wien, 10. August 2026

Dominik Wolf	Manuela Waldner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

TU WIEN Informatics

SaCoDi: Semi-Automatic Concept Discovery Utilizing Multimodal Models and Interactive Refinement

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Visual Computing

by

Dominik Wolf, BSc
Registration Number 0925239

to the Faculty of Informatics

at the TU Wien

Advisor: Associate Prof. Dr.in techn. Manuela Waldner, MSc
Assistance: Univ.Ass. Dipl.-Ing. Johannes Eschner, BSc

Vienna, August 10, 2026

Dominik Wolf	Manuela Waldner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Declaration of Authorship

Dominik Wolf, BSc

I hereby declare that I have written this thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work – including tables, maps and figures – which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.
I further declare that I have used generative AI tools only as an aid, and that my own intellectual and creative efforts predominate in this work. In the appendix "Overview of Generative AI Tools Used" I have listed all generative AI tools that were used in the creation of this work, and indicated where in the work they were used. If whole passages of text were used without substantial changes, I have indicated the input (prompts) I formulated and the IT application used with its product name and version number/date.

Vienna, August 10, 2026

Dominik Wolf

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Kurzfassung

Umfangreiche Bildsammlungen enthalten häufig wiederkehrende visuelle Konzepte. Ohne vorgegebene Klassen oder Taxonomien lassen sie sich jedoch nur schwer explorieren. Automatisch erzeugte Gruppen bilden nicht notwendigerweise jene Konzepte oder Granularitäten ab, die für eine konkrete Analyse relevant sind. Diese Arbeit stellt daher *Semi-Automatic Concept Discovery* (SaCoDi) vor, einen Human–AI-Workflow, der Bildsammlungen clustert und beschriftet und Nutzern erlaubt, Cluster zu teilen oder zusammenzuführen.

SaCoDi wurde als Prototyp in Python umgesetzt. Seine Kommandozeilenschnittstelle wurde für kontrollierte Experimente verwendet, während eine grafische Oberfläche die Inspektion und Verfeinerung von Clustern unterstützt. Verglichen werden eine deterministische WordNet-Beschriftung, die direkte Beschriftung durch ein multimodales Sprachmodell und die hybride Abstraktion von WordNet-Kandidaten durch ein Sprachmodell. Experimente auf STL-10, PACS und Food-101 untersuchen Labelqualität, den Einfluss von Inhalt und visueller Domäne, Verfeinerungsoperationen und Laufzeit.

Der direkte multimodale Ansatz erzielt in jeder Datensatzkonfiguration die höchste gewichtete Abdeckung, Exklusivität und semantische Kompatibilität und liegt bei der Klassenkompatibilität gleichauf oder vorn. Der WordNet-Ansatz ist reproduzierbar und überprüfbar, kann jedoch zu allgemeine oder falsche Begriffe auswählen. Der hybride Ansatz kann geeignetere Abstraktionen ableiten, erreicht aber nicht durchgehend die Spezifität des direkten Ansatzes. Alle initialen Ausgaben sind strukturell gültig. Lokales Reclustering erzielt in vier ausgewählten Split-Fällen die höchste Klassenreinheit, während hybride und visuelle Merges Abdeckung und Exklusivität unterschiedlich gewichten. Die semantische Beschriftung dominiert die Laufzeit; lokale Änderungen sind günstiger als eine vollständige Neuberechnung. Die Evaluation verwendet semantische Bewertungen des Autors, keine Benutzerstudie und kuratierte Klassifikationsdatensätze. Die Ergebnisse stützen daher die technische und methodische Machbarkeit von SaCoDi, nicht dessen Nutzen für reale Benutzergruppen.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Abstract

Large image collections often contain recurring visual concepts. Without predefined classes or taxonomies, however, they are difficult to explore. Automatically generated groups do not necessarily reflect the concepts or granularity relevant to a particular analysis. This thesis therefore introduces *Semi-Automatic Concept Discovery* (SaCoDi), a human–AI workflow that clusters and labels image collections while allowing users to split or merge clusters.

SaCoDi is implemented as a Python prototype. Its command-line interface was used for controlled experiments, while a graphical interface supports cluster inspection and refinement. Three labeling strategies are compared: deterministic WordNet labeling, direct labeling by a multimodal language model, and hybrid language-model abstraction of WordNet candidates. Experiments on STL-10, PACS, and Food-101 evaluate label quality, the effects of content and visual domain, refinement operations, and runtime.

The direct multimodal strategy achieves the highest weighted coverage, exclusivity, and semantic compatibility in every dataset configuration and the highest or tied-highest class compatibility. The WordNet strategy is reproducible and inspectable but can select overly broad or incorrect terms. The hybrid strategy can derive more suitable abstractions but does not consistently match the direct strategy's specificity. All initial outputs are structurally valid. Local reclustering yields the highest class purity in four selected split cases, while hybrid and visual merges balance coverage and exclusivity differently. Semantic labeling dominates runtime, and local edits are cheaper than complete recomputation. The evaluation uses single-author semantic judgments, no user study, and curated classification benchmarks. The findings therefore support SaCoDi's technical and methodological feasibility, not its effectiveness for end users.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Contents

Kurzfassung iv

Abstract v

Contents vi

1 Introduction 1
- 1.1 Problem Statement . . . 5
- 1.2 Aim and Scope . . . 5
- 1.3 Requirements . . . 6
- 1.4 Research Questions . . . 6
- 1.5 Methodological Overview . . . 7
- 1.6 Contributions . . . 8

2 Related Work 9
- 2.1 Concept Discovery in Image Collections . . . 10
- 2.2 Human-Centered AI for Concept Discovery . . . 10
- 2.3 Interactive Approaches to Image Collection Exploration . . . 12
- 2.4 Cluster Labeling . . . 12
- 2.5 Summary and Research Gap . . . 13

3 Method 15
- 3.1 Problem Setting . . . 16
- 3.2 Research Design and System Overview . . . 17
- 3.3 Graphical User Interface and Human Control . . . 18
- 3.4 Automatic Initialization . . . 20
- 3.5 Automatic Cluster Labeling . . . 22
- 3.6 Controlled Backend Comparison . . . 27
- 3.7 User-Directed Cluster Refinement . . . 27
- 3.8 Artifacts, Timing, and Reproducibility . . . 30

4 Experiments 32
- 4.1 Evaluation Design . . . 33
- 4.2 Preliminary Clustering Screening . . . 37

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.3 Initial Cluster Structures 37
4.4 Initial Label Quality Across Datasets 38
4.5 Split-Based Refinement 42
4.6 Merge-Based Abstraction 43
4.7 Computational Performance 44

5 Discussion **46**
5.1 RQ1: Cluster Label Quality 47
5.2 RQ2a: Split-Based Refinement 48
5.3 RQ2b: Merge-Based Abstraction 49
5.4 RQ3: Computational Performance 50
5.5 Implications for SaCoDi and Human–AI Interaction 50
5.6 Limitations and Validity 51
5.7 Future Work 52

6 Conclusion **53**

Overview of Generative AI Tools Used **56**

Glossary **57**

Acronyms **59**

Bibliography **61**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

Large portions of available data are unstructured, including text, images, and video [GH15]. These are different data *modalities*, meaning forms in which information is represented: written language as text, visual content as still images, or time-varying visual content as video. Such data can contain rich semantic information, but they are difficult to organize and analyze because they often lack explicit structures for querying, filtering, or aggregation [LAZ+24]. This is particularly challenging in data exploration workflows, where users need to identify higher-level concepts, relations, or recurring patterns in heterogeneous collections [ERT+17].

This thesis focuses on large image collections as one concrete form of unstructured data. Image collections are common in domains such as cultural heritage [RES26] and social media analysis [CSSL23]. However, these collections are often weakly structured or entirely unlabeled. Users therefore cannot rely on predefined metadata or domain-specific taxonomies—structured organizations of concepts and their relations—to search, filter, or aggregate their contents. Exploring such collections requires methods that help identify recurring visual patterns and describe them in human-interpretable terms.

In this thesis, the identification and labeling of recurring, human-interpretable patterns in an unlabeled image collection is referred to as *concept discovery*. These patterns are termed *visual concepts* and may correspond, for example, to objects, scenes, activities, or domain-specific motifs. Unlike traditional image classification, which assigns images to predefined target classes [SdRRSB13], concept discovery aims to identify and describe concepts that recur across a collection without assuming that the relevant categories are known in advance.

Recent advances in vision–language models (VLMs) provide a useful technical foundation for this task. Models such as Contrastive Language–Image Pre-training (CLIP) [RKH+21] represent data items as numerical vectors in an *embedding space*, where proximity reflects learned semantic similarity. CLIP uses separate image and text encoders that produce vectors of the same dimensionality. During training, the encoders are aligned so that matching images and texts are placed close together and unrelated pairs farther apart. Images and textual descriptions can therefore be compared in the resulting shared embedding space despite originating from different modalities. Figure 1.1 illustrates this shared representation. By relating visual and textual representations, these models support tasks such as zero-shot recognition [ZHJL24]. Zero-shot recognition refers to predicting classes for which no labeled examples were provided during task-specific training [FXJ+18].

Clustering provides a complementary approach for concept discovery because it can reveal groups of similar images without requiring predefined classes. When applied to image embeddings, clustering can produce groups that may correspond to recurring visual concepts. Recent work shows that multimodal representations can be used to organize image collections according to semantic criteria [CQC+23]. However, clustering alone does not make these groups interpretable. A cluster may contain visually or semantically related images, but the concept shared by those images is not necessarily apparent to the user.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

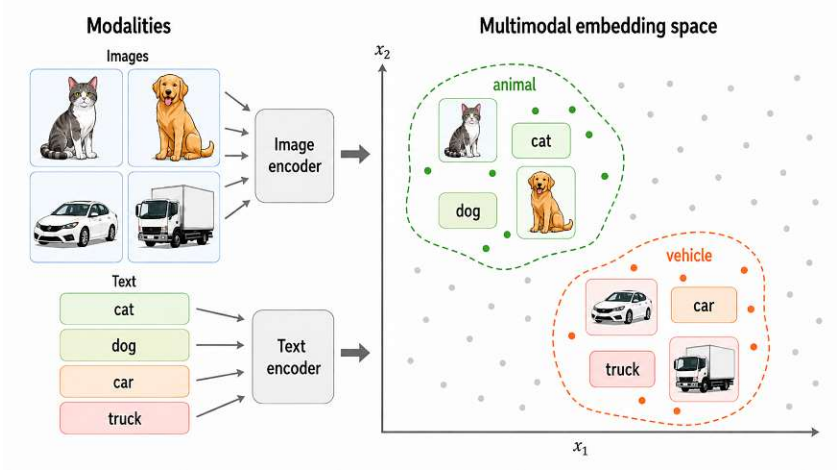


Figure 1.1: Schematic representation of a multimodal embedding space. Separate encoders map images and text into a shared space in which proximity indicates semantic similarity. Positions and distances are only illustrative.

This is a known challenge in cluster labeling. Clusters require concise and meaningful descriptions before they can effectively be interpreted, compared, and used in subsequent analysis [APK+16]. In this thesis, *cluster labeling* refers to assigning a concise, human-interpretable textual label to a cluster. The label summarizes the shared content of the cluster and allows it to be interpreted as a visual concept.

Neither zero-shot recognition nor clustering alone is sufficient for concept discovery at the collection level. Zero-shot recognition typically relates individual images to textual categories, but does not by itself identify which concepts recur across an entire collection. Clustering can reveal recurring groups, but does not describe them in human-interpretable terms. Concept discovery therefore requires both collection-level grouping and meaningful cluster labels.

A further challenge is that automatically generated clusters do not necessarily correspond to semantically coherent concepts. A cluster may contain several distinct concepts, while multiple clusters may represent the same or closely related concepts. Automatically generated clusters are therefore treated as preliminary representations of visual concepts that may require inspection and refinement.

This need for user involvement connects concept discovery to principles from visual analytics and human-in-the-loop machine learning. Visual analytics motivates the use of interactive visual representations that allow users to inspect automatically generated results [TC06]. Human-in-the-loop machine learning emphasizes that users should also be able to steer and iteratively refine these results rather than only accept fully automatic outputs [MRHPAR+23]. Together, these principles motivate a workflow in which generated clusters and labels remain open to user-directed refinement.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

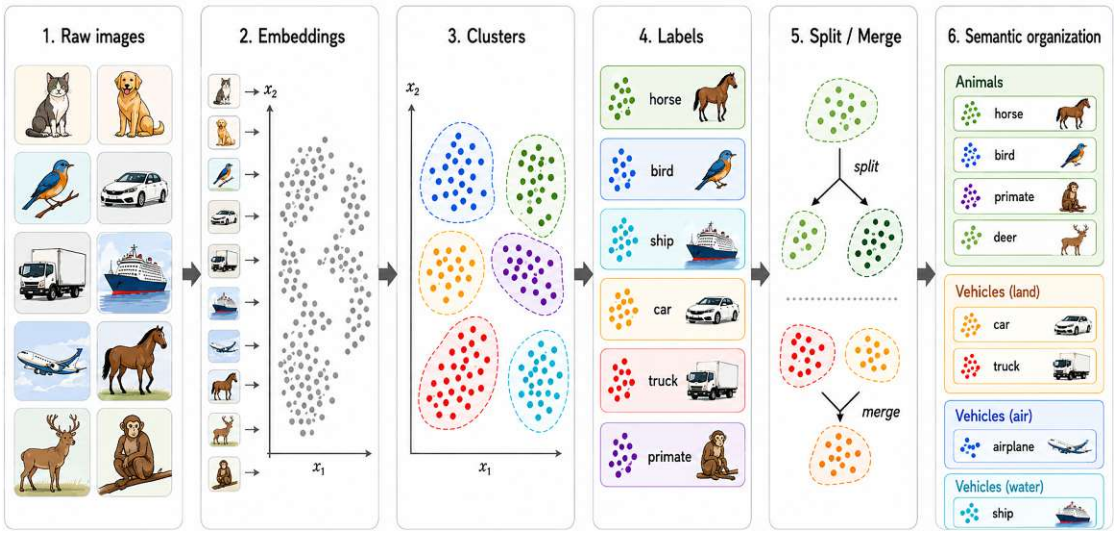


Figure 1.2: Overview of the proposed workflow. Raw image collections are embedded into a multimodal embedding space, clustered, and assigned cluster labels. The resulting clusters can then be refined through split and merge operations to produce a user-refined cluster structure.

This thesis addresses these challenges through *Semi-Automatic Concept Discovery* (SaCoDi), a workflow for discovering and refining visual concepts in unlabeled image collections. SaCoDi combines automatic clustering in a multimodal embedding space with cluster label generation and user-directed refinement. The system proposes an initial cluster structure and corresponding labels, while users can modify this structure through split and merge operations. Figure 1.2 provides an overview of the proposed workflow.

Within this workflow, the thesis investigates three strategies for cluster labeling. The first strategy is based on a lexical resource, namely WordNet [Mil95]. WordNet groups synonymous nouns, verbs, adjectives, and adverbs into *synsets*, sets that represent individual lexical concepts, and connects them through lexical and conceptual-semantic relations. One such relation is hypernymy: a *hypernym* is a more general concept linked to a more specific one; for example, *animal* is a hypernym of *horse*. In this thesis, WordNet is used as a general vocabulary for retrieving cluster labels by comparing cluster representations with WordNet concepts in the CLIP text embedding space. The relevant labels are therefore selected based on the contents of the generated clusters rather than specified as task-specific classes in advance. The second strategy uses a multimodal large language model (MLLM), a generative model that receives images together with textual instructions, to label a cluster from representative image samples. Unlike the WordNet-based strategy, this approach is not restricted to selecting labels from a fixed lexical vocabulary. The third strategy is hybrid: it first retrieves WordNet-based labels and provides them as text-only input to a large language model (LLM), which derives a summarizing or more general label.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

1.1 Problem Statement

This thesis addresses the following problem: given an unlabeled image collection, how can recurring visual concepts be identified, meaningfully labeled, and iteratively refined without assuming a predefined, task-specific label taxonomy?

This problem consists of three related challenges. First, images need to be grouped into coherent clusters that can represent recurring visual concepts without relying on predefined classes. Second, these clusters require meaningful labels that summarize their shared contents and make them interpretable to users. Third, the resulting cluster structure must be refinable when clusters contain multiple concepts or when several clusters represent semantically similar concepts.

1.2 Aim and Scope

The aim of this thesis is to develop and empirically evaluate SaCoDi as a semi-automatic workflow for concept discovery in unlabeled image collections. The thesis focuses on the generation and refinement of labels for image clusters represented in a multimodal embedding space.

A prototype is implemented to operationalize the SaCoDi workflow and enable an empirical comparison of the three labeling strategies. The strategies are evaluated with respect to label quality and computational performance during initial labeling and after split and merge operations.

The scope of the thesis is limited to image collections as the concrete data modality. This focus is motivated by the availability of mature VLMs and suitable image data. Although the underlying workflow may be conceptually applicable to other data modalities, such a transfer is not investigated.

The thesis does not aim to train a new VLM, MLLM, or clustering algorithm. Instead, it investigates how existing multimodal representations, lexical resources, language models, clustering methods, and refinement operations can be combined and evaluated for semi-automatic concept discovery.

The evaluation focuses on the quality of the generated cluster labels according to Requirements R1–R4, the effects of split and merge operations, and the computational performance addressed by R5 (Section 1.3). It does not evaluate the design of the visual interface, user experience, usability, or the analytical reasoning of users. The thesis is therefore positioned primarily as a contribution to data exploration and semi-automatic concept discovery in unlabeled image collections rather than as an evaluation of a visual analytics system.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.3 Requirements

The problem statement implies five requirements for labels and refinement operations in the SaCoDi workflow. These requirements make the intended qualities explicit and provide the basis for the evaluation criteria defined in Chapter 4.

R1 — Coverage.
A cluster label should describe all, or a large proportion, of the images contained in its target cluster. A label that describes only the majority class can therefore leave a substantial second concept undescribed.

R2 — Specificity and exclusivity.
A label should be as general as necessary to cover its target cluster, but specific enough to distinguish that cluster from other parts of the collection. The appropriate abstraction level consequently depends on the surrounding cluster structure.

R3 — Comprehensibility.
A label should be understandable to the anticipated users of an image-exploration workflow. Familiar terms are preferable to technically valid but uncommon lexical abstractions. Because this thesis does not include a user study, comprehensibility is assessed by the author using an explicit rubric and is not treated as evidence of actual user understanding.

R4 — Output robustness.
A labeling backend should return a syntactically valid and usable output with all required fields. This requirement separates semantic label quality from failures of structured generation or parsing.

R5 — Responsiveness.
Initial labeling and local refinement should complete within a time frame that permits iterative use. This thesis measures computational latency, but does not infer perceived responsiveness or usability from runtime alone.

1.4 Research Questions

To guide the development and evaluation of the SaCoDi workflow, this thesis addresses the following research questions:

- **RQ1 — Cluster Label Quality.**
  How do WordNet-based, MLLM-based, and hybrid cluster-labeling strategies compare with respect to coverage, exclusivity, class and semantic compatibility, specificity, author-assessed comprehensibility, and output robustness?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **RQ2a — Split-Based Refinement.**
  How do alternative split strategies affect the structural coherence and labels of clusters that contain multiple concepts?
- **RQ2b — Merge-Based Abstraction.**
  How well do alternative merge strategies derive a useful common abstraction from related source clusters?
- **RQ3 — Computational Performance.**
  What are the computational trade-offs of WordNet-based, MLLM-based, and hybrid cluster-labeling strategies during initial labeling and after split and merge operations?

1.5 Methodological Overview

To address these research questions, SaCoDi is first specified as a conceptual workflow and subsequently implemented as a prototype for empirical evaluation.

First, a literature review positions the work in relation to prior research on concept discovery, VLMs, image clustering, cluster labeling, lexical knowledge resources, and human-in-the-loop machine learning. Research in visual analytics is considered where it provides relevant principles for combining computational methods with user input. However, visual representation and interaction design are not treated as the main research objects of the thesis.

Second, the SaCoDi workflow is specified. This includes the representation of images in a multimodal embedding space, the generation of initial clusters, the three cluster labeling strategies, and the split and merge operations used for refinement.

Third, an experimental evaluation compares the three labeling strategies across different stages of the workflow and three datasets. STL-10 is a benchmark with ten broad object classes [CNL11]. PACS (Photo, Art Painting, Cartoon, and Sketch) contains seven object classes across four visual domains [LYSH17], while Food-101 contains 101 fine-grained food classes [BGVG14]. The evaluation considers initial cluster labeling and relabeling after split and merge operations. Label quality is assessed in relation to coverage, exclusivity, class and semantic compatibility, specificity, author-assessed comprehensibility, and output robustness. PACS additionally exposes the interaction between content and visual domain, while Food-101 tests a larger fine-grained vocabulary. Computational performance is assessed through runtime measurements for the different labeling strategies and workflow stages.

Finally, the thesis discusses the limitations of the proposed workflow and derives implications for cluster labeling, data exploration, and semi-automatic concept discovery in unlabeled image collections.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.6 Contributions

This thesis makes two main contributions.

First, it proposes *Semi-Automatic Concept Discovery* (SaCoDi), a workflow that combines multimodal image clustering, automatic cluster labeling, and user-directed split and merge operations to discover and refine visual concepts in unlabeled image collections. The workflow is operationalized through a prototype that implements the investigated labeling and refinement strategies.

Second, it provides empirical findings on the quality and computational performance of three cluster labeling strategies: deterministic WordNet labeling, direct MLLM labeling, and hybrid LLM labeling. These strategies are compared during initial cluster labeling and after split and merge operations, providing evidence about their respective strengths and trade-offs within the SaCoDi workflow.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Related Work

This chapter reviews related work on concept discovery, human-centered AI, interactive image collection exploration, and cluster labeling. It concludes by identifying the research gap addressed by this thesis.

2.1 Concept Discovery in Image Collections

Research on concept discovery in image collections includes methods for identifying recurring visual patterns, organizing images into semantic groups, and assigning human-interpretable descriptions to these groups. Existing approaches differ mainly in whether concepts are discovered automatically, guided by user input, or intended for tasks such as recognition and model interpretation.

Several methods discover visual concepts automatically. ConceptLearner learns concept detectors from weakly labeled image collections [ZJP15], while SegDiscover identifies recurring concepts through unsupervised segmentation and clustering [HCR22]. The ACE method extracts dataset-level visual concepts that help explain a model's predictions [GWZK19]. Zang et al. use VLMs and candidate concepts generated by language to discover visually recognizable concepts [ZYT+25]. These approaches identify reusable visual patterns, but primarily support recognition or model interpretation rather than the interactive organization of a particular image collection.

Other work focuses more directly on image clustering. Cai et al. combine visual and semantic representations to improve the semantic coherence of image clusters [CQC+23]. Their approach produces an automatic organization of the collection but does not provide mechanisms for users to interactively revise the resulting cluster structure.

Open-ended approaches instead aim to identify multiple possible organizations of a collection without predefined criteria. OpenSMC automatically proposes semantic criteria and generates labeled clusters for each of them [LZL+26]. Although this supports the discovery of previously unknown structures, the resulting clusters are generated automatically and provide limited support for direct revision after inspection.

Existing approaches provide limited support for workflows in which an AI system proposes an initial labeled organization and users directly refine its granularity through split and merge operations. This motivates the human-centered perspective discussed in the next section.

2.2 Human-Centered AI for Concept Discovery

Human-centered AI seeks to combine automated support with meaningful human control. Shneiderman argues that high levels of automation and human control are not mutually exclusive and that AI systems should augment human capabilities rather than replace human judgment [Shn20a, Shn20b]. Similarly, the human–AI interaction guidelines proposed by Amershi et al. emphasize that users should be able to understand system behavior, correct erroneous results, and influence future outputs [AWV+19]. Applied to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

concept discovery, these principles suggest complementary roles: AI methods can process large image collections and propose groups and descriptions, while users contribute contextual knowledge and assess whether the resulting concepts are meaningful and appropriately granular.

Existing systems implement this collaboration in different ways. ConceptExtract allows users to define visual concepts by labeling a small number of examples selected through active learning, an iterative process in which the model identifies informative examples for human annotation [ZXSR22]. The system reduces the required annotation effort, while users determine the semantic meaning of the concepts. However, its purpose is to interpret and diagnose neural-network predictions rather than to organize an entire image collection into a set of refinable concepts.

AnchorViz supports semantic data exploration and concept discovery during interactive machine learning by visualizing examples in relation to user-defined semantic anchors [SGR+19]. It demonstrates how domain knowledge and interactive visual evidence can expose classifier errors and help users refine a concept model. Its concepts are defined in support of classifier development, whereas SaCoDi starts from an unlabeled image collection and treats the generated cluster structure and labels themselves as the objects of refinement.

SNaCK combines human similarity judgments with machine-derived similarities to construct concept embeddings and demonstrates their use for qualitatively exploring an unstructured collection of pictographic characters [WKKB15]. Human input therefore shapes the representation itself, whereas SaCoDi asks users to refine an automatically proposed labeled cluster structure.

Interactive clustering provides more direct mechanisms for modifying automatically generated structures. Existing approaches incorporate user feedback through constraints, reassignment of individual items, or split and merge operations [BHR+20]. Interactive Document Clustering Revisited, for example, allows users to split clusters containing different topics and merge clusters that represent a common topic [SNMM18]. Such interactions incorporate human knowledge into an initially automatic organization, but the approach focuses on document clustering and does not investigate the generation and repeated evaluation of cluster labels.

Mixed-initiative visual analytics systems, in which the user and the computational system both actively contribute to the analysis, further demonstrate how automatically generated groups can remain open to human interpretation [PMCEA+22]. The tool Scholastic presents machine-generated clusters of textual material through an interactive interface and allows researchers to inspect and reorganize them during qualitative analysis [HMF+22]. The system treats clusters as provisional analytical results rather than final classifications. However, it does not address the discovery and labeling of visual concepts in image collections.

These approaches show that human input can support concept definition, cluster correction, and interpretation. However, limited work combines the automatic generation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of labeled image clusters with direct, iterative refinement of their semantic granularity. In particular, the role of different cluster-labeling strategies within a workflow based on inspection, splitting, merging, and relabeling remains insufficiently explored.

2.3 Interactive Approaches to Image Collection Exploration

Interactive systems for image collection exploration combine automated image analysis with visual representations that allow users to browse, inspect, and reorganize large collections. Early approaches primarily focused on making similarity structures navigable. The Semantic Image Browser arranges images according to automatically extracted semantic information and supports visual browsing and search within large image collections [YFH+06]. PEx-Image similarly uses projection-based visualizations to support user-driven exploration and classification of image collections [ENP+09]. AverageExplorer instead summarizes subsets of images through average images and allows users to interactively select and align images to investigate recurring visual structures [ZLE14].

Recent systems increasingly integrate textual representations generated by VLMs. Li et al. use automatically generated captions to expose semantic patterns and potential biases in large image datasets [LWA+24]. Their visual analytics system also allows users to inspect the relationship between images and captions and steer the caption-generation process. The interaction therefore affects the textual representation of the collection, but does not directly organize the images into refinable visual concepts.

Text-guided clustering gives users more direct control over the organization of a collection. IC|TC allows users to specify a clustering criterion in natural language and iteratively edit this criterion after inspecting the generated clusters [KPK+24]. ITGC similarly takes a user-provided criterion and automatically searches for interpretable visual concepts that support a corresponding clustering [ZA25]. These approaches align the generated organization with user intent, but interaction occurs primarily through the specification of a global textual criterion rather than through direct modification of individual clusters.

Existing systems therefore support image-collection exploration through visual navigation, semantic summaries, interactive captioning, and user-defined clustering criteria. However, they provide limited support for directly restructuring generated image groups and updating their labels in response to local split and merge operations.

2.4 Cluster Labeling

Cluster labeling assigns a concise description to a group of related data items so that users can interpret and compare the resulting clusters. The broader literature on interpretable clustering similarly emphasizes that cluster quality includes whether the resulting organization can be understood, not only its structural accuracy [HJD+24]. Most established labeling approaches have been developed for text collections, where

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

labels are extracted or generated from representative terms, phrases, or documents. Aker et al., for example, rank candidate labels for clusters of news comments using lexical and statistical features [APK+16]. Such approaches derive labels from the clustered content itself, but are difficult to transfer directly to image collections without first obtaining textual representations.

Knowledge-based methods instead select labels from structured lexical resources. Fukumoto et al. use hypernym relations from a machine-readable dictionary to generalize salient cluster terms [FS11]. Poostchi and Piccardi combine word embeddings with synonymy and hypernymy relations from WordNet [Mil95] to assign labels to word clusters [PP18]. These methods provide a controlled and reproducible label vocabulary, but their outputs are restricted by the coverage and granularity of the underlying lexical resource. For image clusters, VLM representations make it possible to apply a related strategy by comparing visual cluster representations with embedded lexical concepts.

Generative approaches formulate labeling as the production of a textual summary rather than the selection of an existing term. Earlier work generates topic labels from representative phrases or document summaries [MCNJ13, WW16]. More recently, LLMs have been evaluated for generating labels that summarize the shared content of text clusters and topics [PW25]. In the visual domain, MLLMs can process several images and generate human-readable descriptions for image groups, and recent image-collection methods use such models to produce interpretable cluster labels [LZL+26]. Generative methods are not limited to a fixed vocabulary, but their labels may vary across prompts and runs and can differ in specificity.

Lexical and generative approaches therefore offer complementary properties. Lexical resources constrain labels to a structured vocabulary, whereas language models can produce more flexible descriptions from representative cluster content. However, their relative label quality and computational cost have rarely been compared within the same image concept-discovery workflow. Moreover, limited work investigates hybrid strategies that retrieve structured candidate labels and subsequently use a language model to summarize or generalize them.

2.5 Summary and Research Gap

The reviewed work addresses several components of concept discovery in image collections. Automatic approaches can identify recurring visual patterns and organize images into semantically related groups. Human-centered and interactive systems allow users to inspect generated results, specify grouping criteria, or modify cluster structures. Cluster-labeling methods additionally provide lexical or generative descriptions that make automatically identified groups easier to interpret.

However, these capabilities are commonly investigated separately. Automatic concept-discovery methods provide limited support for revising the generated organization after inspection. Interactive image-exploration systems often focus on navigation, search,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

or the specification of a global grouping criterion rather than the direct refinement of individual clusters. Existing cluster-labeling research, in turn, primarily evaluates labels independently of an iterative workflow in which the underlying clusters are modified by users.

This leaves a gap for human-centered workflows in which an AI system proposes an initial labeled organization of an unlabeled image collection and users subsequently refine its granularity. In such a workflow, cluster labels must not only describe the initial organization but also remain meaningful after the cluster structure has changed.

This thesis addresses this gap through SaCoDi, which combines automatic image clustering and cluster labeling with user-directed refinement. It further compares deterministic WordNet, direct MLLM, and hybrid LLM labeling strategies with respect to label quality and computational performance during initial labeling and after split and merge operations.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Method

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

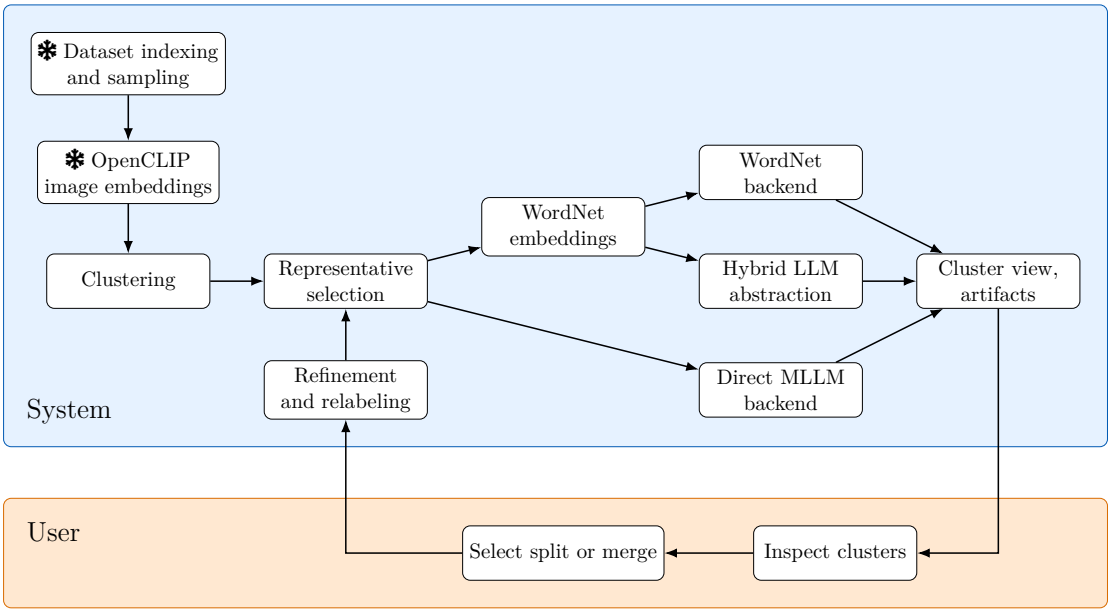


Figure 3.1: Operationalization of the SaCoDi workflow. The system generates an initial labeled cluster structure through shared preprocessing and one of three labeling backends. Users inspect the resulting clusters and can initiate split or merge operations. The system then recomputes the affected representatives and labels and presents the updated state.

This chapter describes the design of SaCoDi and its operationalization in a research prototype, which combines automatic image clustering and labeling with user-directed refinement. Through a graphical interface, users can inspect clusters and perform cluster refinement operations.

3.1 Problem Setting

SaCoDi addresses the problem of identifying, labeling, and refining recurring visual concepts in an unlabeled or weakly labeled image collection. Its input is a collection of images for which no task-specific concept taxonomy is assumed. Its initial output is a cluster structure together with one semantic label per cluster. Because this automatically generated structure may contain clusters that are too broad or separate clusters that represent related concepts, the initial result remains open to user-directed refinement. Figure 3.1 illustrates this progression and identifies the decisions made by the system and by the user.

The workflow assigns complementary responsibilities to the system and the user. The system processes the image collection, proposes an initial cluster structure, generates cluster labels, and recomputes affected results after a refinement operation. The user inspects the cluster contents and determines whether the current structure represents meaningful concepts at an appropriate granularity. When it does not, the user can

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

initiate a split of an overly broad cluster or a merge of related clusters. The resulting structure is therefore neither produced entirely manually nor treated as a final automatic classification.

Within this workflow, the prototype offers three labeling strategies under a shared pipeline:

1. a deterministic zero-shot CLIP–WordNet method,
2. a direct MLLM-based labeling method, and
3. a hybrid LLM method that uses the WordNet-derived candidate evidence from the first strategy but delegates the final abstraction step to an LLM.

This shared design isolates differences in semantic labeling behavior from differences in dataset preparation, embedding generation, clustering, or representative selection. It also allows the same labeling strategies to be studied after initial cluster generation and after user-directed changes.

3.2 Research Design and System Overview

The prototype is implemented in Python and follows a staged, artifact-oriented pipeline in which each processing step writes persistent outputs to a dedicated run directory. This design supports reproducibility, resumability, timing measurement, and controlled comparison between alternative labeling backends on the same data sample. It also preserves intermediate results for later inspection and analysis.

At a high level, the pipeline comprises the following stages, which are also depicted in Figure 3.1:

1. preprocessing through dataset indexing and sampling,
2. preprocessing through image embedding with OpenCLIP,
3. clustering in embedding space,
4. representative selection,
5. cluster labeling,
6. inspection of cluster contents by the user,
7. user-directed refinement through split and merge operations, and
8. automatic recomputation and presentation of the affected results.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The stages from dataset indexing to clustering are shared by all labeling methods. Semantic interpretation is then performed by one of three interchangeable backends: a deterministic WordNet backend, a direct MLLM backend, or a hybrid LLM backend that abstracts WordNet-derived candidates. The labeled cluster structure is presented for inspection. A user can then retain the current structure or initiate a split or merge operation, after which the system recomputes the affected representatives and labels and presents an updated state.

A central extension beyond initial cluster labeling is the support for derived-run refinement. Split and merge operations do not modify the original run in place. Instead, each edit creates a new run directory with updated cluster assignments, recomputed representatives, relabeled affected clusters, timing information, and preserved edit history. This design is important for auditability because it allows each refinement step to remain inspectable as a separate state. At the same time, it improves experimental comparability by ensuring that unaffected parts of the pipeline remain unchanged.

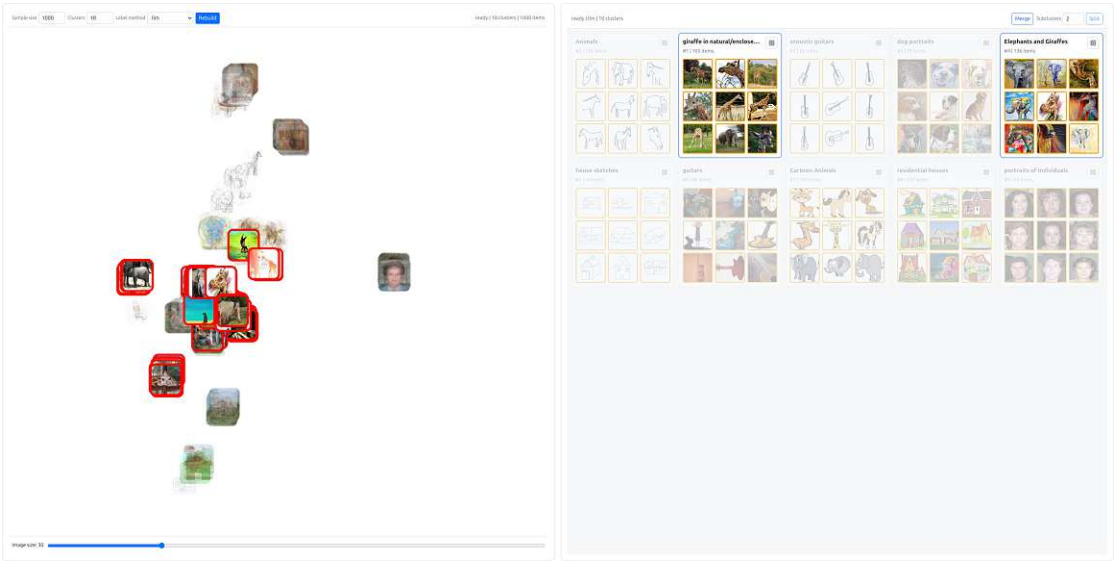
3.3 Graphical User Interface and Human Control

The prototype provides both a command-line interface (CLI) and a graphical user interface. The CLI exposes the SaCoDi pipeline and its configuration options for scripted execution and was used to run the controlled experiments and repeated runtime measurements reported in Chapter 4. For interactive use, SaCoDi was integrated into the JDE sandbox, a graphical data exploration framework of the Joint Human-Machine Data Exploration research project [Wal23]. The interface provides access to the SaCoDi workflow and presents the contents of the generated clusters so that users can inspect which images have been grouped together.

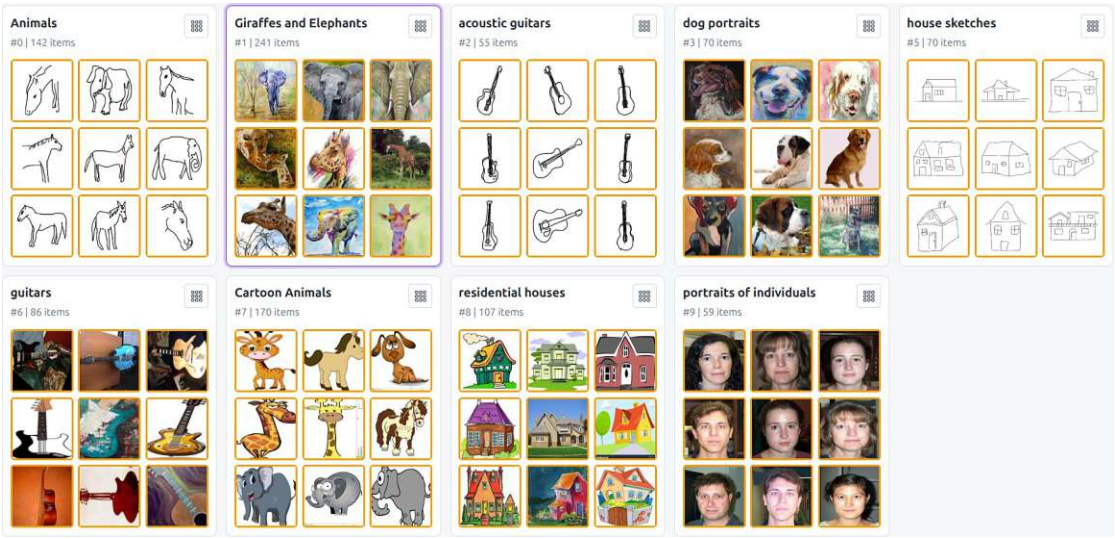
Figure 3.2 illustrates the interface with a PACS sample of 1,000 images, 10 clusters, and direct MLLM labeling. This configuration serves as an interface example and is separate from the experimental configurations reported in Chapter 4. In the main view, the complete sample is projected into two dimensions with Uniform Manifold Approximation and Projection (UMAP) [MHSG18] and displayed on the left. The generated clusters and their labels are shown as cards on the right. Two clusters are selected in Figure 3.2a; the images assigned to those clusters are simultaneously outlined in the UMAP view, linking the selected aggregate concepts to their positions in the collection.

Selection determines which refinement operation can be initiated. When one cluster is selected, the user specifies the desired number of subclusters and starts a split. When multiple clusters are selected, the merge command combines them. Figure 3.2b shows the cluster view after merging *giraffe in natural/enclosed habitat* and *Elephants and Giraffes*. The resulting cluster is labeled *Giraffes and Elephants*. To make the effect of an operation immediately visible, the interface briefly highlights each changed cluster: one result for a merge and as many result clusters as requested for a split. The interface then presents the updated persistent state produced by the corresponding refinement method.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Main interface before refinement. Two selected cluster cards on the right correspond to the images outlined in the UMAP projection on the left.



(b) Cluster view after the selected clusters have been merged. The new *Giraffes and Elephants* cluster is temporarily highlighted.

Figure 3.2: Graphical SaCoDi interface for inspecting and refining a PACS run. The controls support a user-directed transition from the selected source clusters in (a) to the merged cluster in (b).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.4 Automatic Initialization

The automatic initialization phase of SaCoDi transforms an image collection into an initial cluster structure and prepares the visual evidence required for labeling. It comprises dataset indexing and sampling, image embedding, clustering, and representative selection. These stages establish a common structural base before any of the three labeling strategies is applied.

3.4.1 Dataset Indexing and Sampling

The first preprocessing stage converts a dataset into a flat list of image records. Each record contains an image path together with available metadata such as class label, class id, or domain, which are used only for sampling and later evaluation but not for inferring the cluster labels.

Sampling is performed before embedding in order to control the number and distribution of images processed by the pipeline. The implementation uses class-balanced sampling, in which each available ground-truth class contributes the same number of images to the resulting subset. For datasets that additionally contain multiple domains, balancing can be performed across both classes and domains. In this case, the dataset is partitioned by class–domain combinations, and an equal number of images is sampled from each combination.

Sampling accepts a configurable random seed so that a selected image subset can be reproduced across runs.

3.4.2 Image Embedding

After indexing and sampling, OpenCLIP encodes each image as a dense vector in a semantic feature space [IWC+21, CBW+23]. The configured model checkpoint determines the required image preprocessing. The resulting embeddings serve as the shared visual representation for clustering, representative selection, zero-shot label scoring, and CLIP-based assignment during refinement. Together with indexing and sampling, this stage forms the shared preprocessing block of the pipeline.

OpenCLIP is multimodal because its image and text encoders map both modalities into the same vector space. Images and candidate label texts can therefore be compared through their vector similarity, but they retain different roles in the SaCoDi pipeline: KMeans receives only image embeddings, whereas cross-modal similarities support WordNet candidate retrieval and CLIP-based assignment during refinement.

The unnormalized image-embedding matrix is stored as a persistent artifact. The prototype also supports a per-image cache to avoid recomputing embeddings across runs that reuse the same images and model configuration.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Rationale. Using OpenCLIP embeddings provides a common semantic representation for all later stages. The same embedding space is used for global clustering, representative selection, and deterministic candidate-label retrieval. This makes the pipeline internally consistent and supports controlled comparison across backends.

3.4.3 Clustering in Embedding Space

Using x_i for the embedding of image i, the prototype partitions the image embeddings into clusters. The implementation supports several clustering backends, including KMeans [Llo82], MiniBatchKMeans [Scu10], HDBSCAN [CMZS15, MHA17], and agglomerative clustering [MC12].

The clustering backend can operate on the stored embeddings directly or on an L2-normalized view. Normalization was enabled in all evaluated runs:

$$\hat{x}_i = \frac{x_i}{\|x_i\|_2}. \tag{3.1}$$

This normalized view does not overwrite the stored embeddings. Cosine-based retrieval and refinement steps separately L2-normalize the image and text embeddings they require. For any two unit-normalized vectors, the dot product equals their cosine similarity.

The clustering stage produces a mapping

$$i \mapsto c_i \tag{3.2}$$

from each embedding index i to a cluster id c_i. The resulting assignments form the basis for representative selection and all later semantic labeling steps.

Methodological role. Clustering provides the initial structure on which semantic labeling and refinement operate, but it is not itself the main research contribution of the thesis. The availability of multiple clustering backends allows a suitable method to be selected without changing the downstream SaCoDi stages.

3.4.4 Representative Selection

After clustering, the system selects representative images for each cluster. These representatives serve two purposes. First, they provide a compact visual summary of a cluster. Second, in the controlled comparison, they define the evidence subset used by all three labeling backends. Representative images also support visual split and merge operations.

For a cluster C_k, the centroid is computed as

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i. \tag{3.3}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

For CLIP-based retrieval later in the pipeline, let $E_k \subseteq C_k$ denote the image-evidence subset used to score candidate labels. Depending on the configuration, this subset can contain the complete cluster or selected representatives; in the controlled comparison, it contains the same six representatives provided to the direct MLLM (Table 4.2). The prototype first L2-normalizes the embeddings in E_k, averages them, and then L2-normalizes the result. This retrieval centroid is denoted by $\hat{\mu}_k$.

This distinction is intentional: μ_k denotes the centroid of the complete cluster used for representative-selection logic, whereas $\hat{\mu}_k$ denotes the L2-normalized centroid of the configured evidence subset used for CLIP-style label retrieval.

SaCoDi selects a small set of representative images that combines typical examples near the cluster center with images reflecting visual variation within the cluster. This provides compact evidence of both the dominant cluster content and its internal diversity. The representative counts used in the controlled comparison are specified in Table 4.2.

In addition to representatives, the stage also computes cluster-level metadata such as cluster size and available class or domain distributions. These summaries are stored in debugging artifacts and reports, but were not supplied to the labeling backends in the controlled comparison. The resulting representatives provide compact visual evidence used both for automatic labeling and human inspection.

3.5 Automatic Cluster Labeling

After the initial cluster structure has been generated, the system assigns textual labels that make the proposed visual concepts interpretable. These labels summarize the shared cluster contents and help users assess whether a cluster represents a meaningful concept or requires refinement. The core methodological comparison concerns three alternative labeling backends. All three operate on the same clustered sample and representative summaries, but they differ in how they derive cluster-level semantics.

Figure 3.3 makes these differences explicit. The deterministic path retrieves candidate concepts and resolves them through the WordNet hierarchy without a language model. The direct path sends representative images to the MLLM. The hybrid path uses the same WordNet candidate retrieval as the deterministic path, but an LLM performs only the final text-based abstraction.

3.5.1 Deterministic WordNet Backend

The WordNet [Mil95] backend is the deterministic labeling path. It combines zero-shot retrieval using CLIP [RKH+21] image–text embeddings with semantic aggregation over the WordNet noun hierarchy.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

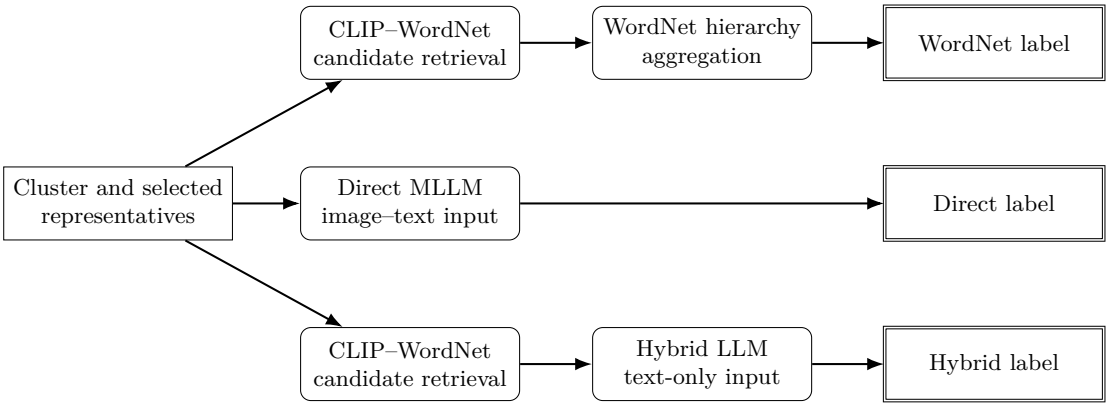


Figure 3.3: Information flow of the three cluster-labeling strategies. The direct MLLM receives representative images and textual instructions, whereas the hybrid LLM receives only WordNet candidate-label strings. The deterministic WordNet strategy does not use a generative language model.

Its internal pipeline consists of six steps:

1. constructing a visual WordNet vocabulary,
2. embedding label texts with the CLIP text encoder,
3. scoring candidate labels against cluster image embeddings,
4. retaining plausible candidate labels,
5. aggregating retained candidates upward through WordNet, and
6. selecting a final hypernym-like cluster label.

Vocabulary construction. The vocabulary is derived from selected WordNet noun subtrees. A *root* is the top synset of a subtree admitted to the candidate vocabulary; including or excluding a root therefore includes or removes all concepts reachable below that synset.

Label embeddings. For each candidate label l_j, the system creates several prompt templates such as *a photo of a {label}*, *an image of a {label}*, and *a picture of a {label}*. These prompts are embedded with the CLIP text encoder. The resulting prompt vectors are averaged and normalized:

$$t_j = \text{normalize}\left(\frac{1}{|P(l_j)|}\sum_{p\in P(l_j)} f_{\text{text}}(p)\right), \tag{3.4}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

where $P(l_j)$ denotes the prompt set of label l_j, and $f_{\text{text}}(p)$ is the L2-normalized vector produced from prompt p by the OpenCLIP text encoder. The resulting unit vector t_j represents candidate label l_j in the shared image–text embedding space.

Visual ranking of candidate labels. This step assigns each WordNet candidate label a numerical similarity value so that the labels can be ranked and weak candidates can subsequently be discarded. It does not evaluate the quality of a final cluster label. For each candidate l_j, the backend compares its text embedding t_j with the normalized evidence centroid $\hat{\mu}_k$ and the normalized embeddings $\hat{x}_i$ of the images in E_k. Because the compared vectors are unit normalized, the following dot products correspond to cosine similarities; a higher value indicates stronger visual–text alignment. Centroid similarity is defined as

$$s_j^{\text{centroid}} = \hat{\mu}_k^\top t_j, \tag{3.5}$$

while mean similarity is

$$s_j^{\text{mean}} = \frac{1}{|E_k|} \sum_{x_i \in E_k} \hat{x}_i^\top t_j. \tag{3.6}$$

The evaluated combined mean–centroid scoring mode uses both similarities. The parameter $\alpha \in [0, 1]$ controls their relative influence: larger values place more weight on the mean image-to-label similarity, while smaller values place more weight on centroid similarity.

$$s_j^{\text{combined}} = \alpha s_j^{\text{mean}} + (1 - \alpha) s_j^{\text{centroid}}. \tag{3.7}$$

The evaluated configuration uses $\alpha = 0.7$, giving the mean and centroid similarities weights of 0.7 and 0.3, respectively (Table 4.2). Because $\hat{\mu}_k$ is the normalized mean of the same evidence embeddings used by s_j^{mean}, centroid and mean similarities are proportional across candidates within a cluster. Consequently, α does not change their within-cluster ranking; it changes the score scale and can therefore affect the subsequent absolute threshold and retention-margin decisions.

Candidate retention. After ranking labels by the resulting similarity values, the backend retains only candidates that satisfy a minimum threshold and remain within a configurable margin of the best value. These retained candidates form the semantic evidence for the later hypernym-selection step. The ranking and retention parameters are stored in the run configuration so that the retrieval process remains reproducible.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

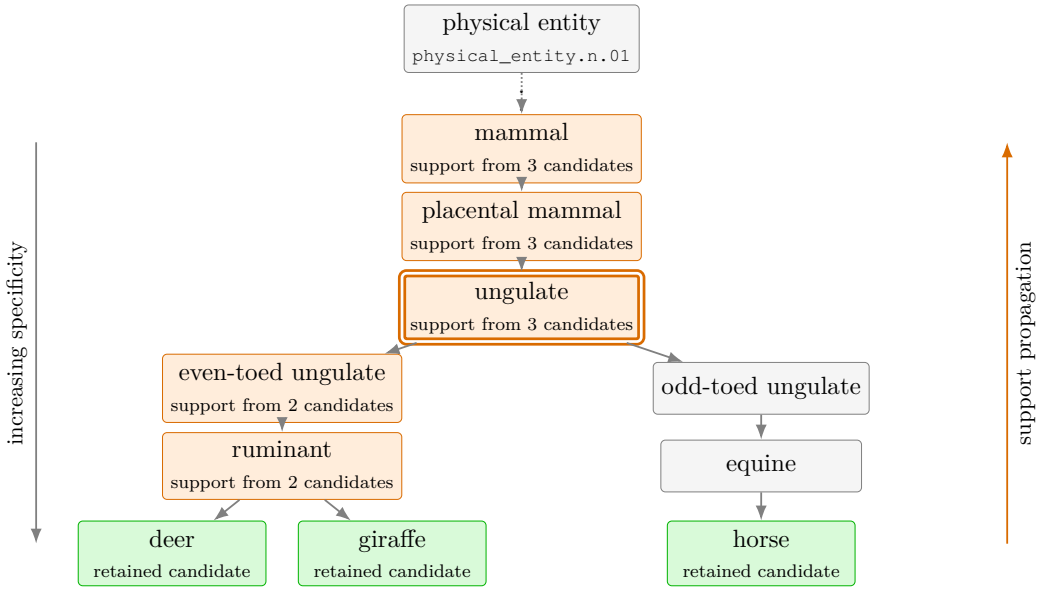


Figure 3.4: Example excerpt of the WordNet hierarchy used for hypernym aggregation. Edges point from general hypernyms to more specific concepts, called hyponyms, whereas retained-candidate support propagates upward. Green nodes are retained candidates, orange nodes accumulate support from multiple candidates, and the double-bordered node marks a possible shared abstraction. Intermediate nodes between *physical entity* and *mammal* are omitted for readability.

Hypernym aggregation. Each retained candidate contributes weighted support to its WordNet ancestors. Support decays with path distance, which makes nearby ancestors more influential than highly general ones. The final selected hypernym balances support, depth, supporter agreement, and penalties for excessive generality.

Figure 3.4 illustrates this aggregation on an excerpt of the WordNet noun hierarchy. In the example, *deer*, *giraffe*, and *horse* are retained candidates. Their support propagates toward more general ancestors: *ruminant* receives support from two candidates, while *ungulate* receives support from all three and represents a possible shared abstraction. The displayed counts denote contributing candidates rather than the distance-weighted score used by the implementation. WordNet is a directed acyclic graph in which a synset can have more than one hypernym path; the figure shows only the paths relevant to this example.

Methodological role. This backend serves as the deterministic baseline. Given fixed artifacts and parameters, it is reproducible and its lexical evidence remains inspectable, which makes it suitable for comparison against more flexible but less transparent generative approaches.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.5.2 Direct MLLM Backend

The direct MLLM backend labels a cluster directly from its representative evidence using a generative model with multimodal input capabilities. In contrast to the deterministic backend, it does not rely on a predefined lexical hierarchy. Instead, textual instructions prompt the model to infer a suitable cluster label from the provided visual evidence.

Each request combines compact textual instructions with selected representative images. The instructions ask the model to return a concise, concrete cluster label and a short description in JSON form. When the visual evidence is weak, the model is instructed to choose the broadest defensible label. In the evaluated configuration, six representative images were provided per cluster; generated captions and cluster metadata such as size, class or domain distributions, and distances to the centroid were not included (Table 4.2).

The requested JSON structure contains a label and a short description. The parser additionally accepts optional alternative-label and evidence fields and normalizes the response to the prototype's common label structure. Because generative outputs are not guaranteed to be well-formed, the implementation includes retry logic, JSON repair, and fallback extraction when parsing fails.

Methodological role. This backend represents the direct generative approach. It is the most flexible of the three labeling paths and can produce natural descriptions directly from visual evidence, but it is also the least deterministic. It is therefore particularly relevant for studying trade-offs between human-readable output, robustness, and methodological transparency.

3.5.3 Hybrid LLM Backend

The hybrid LLM backend combines deterministic candidate retrieval with LLM-based abstraction. It reuses the WordNet retrieval pipeline up to retained candidate generation, but replaces the final hierarchy-based selection step with a language-model decision.

Instead of receiving raw cluster images, the language model is given a compact textual summary of retained candidate concepts associated with a cluster. These are the WordNet labels that remain after similarity ranking and the threshold and margin filters described in Section 3.5.1. In the evaluated configuration, only the label strings of the five highest-ranked retained candidates per cluster are passed to the model; candidate scores and images are not included. The model is then asked to choose a broader hypernym-like label that best summarizes this candidate evidence.

This design isolates the final semantic abstraction step:

- candidate retrieval remains CLIP–WordNet based,
- final category abstraction is delegated to the LLM.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Methodological role. This backend occupies an intermediate position between the deterministic and direct generative paths. It allows the thesis to study whether generative reasoning is most useful for turning constrained candidate evidence into an interpretable abstraction rather than for deriving the complete cluster semantics directly from images.

3.6 Controlled Backend Comparison

A central methodological requirement is that the alternative system-generated labels remain comparable. If different labeling methods operate on different subsets of evidence, differences in output may reflect differences in input rather than differences in the labeling logic itself. To address this, the comparison restricts the CLIP–WordNet candidate retrieval shared by the deterministic and hybrid backends to the same representative subset that is shown to the direct MLLM backend.

This mode does not make the methods identical, because the MLLM reasons over image inputs while the deterministic backend operates in embedding space. Nevertheless, it aligns the evidence budget more closely and therefore provides a more controlled basis for comparison.

3.7 User-Directed Cluster Refinement

After inspecting the automatically generated cluster structure, users can refine its semantic granularity through split and merge operations. The user determines which part of the structure requires modification, while the system executes the selected refinement method, recomputes affected representatives and labels, and returns an updated state. The operations are implemented as derived-run transformations: the original run is never modified in place. Instead, an edit produces a new run directory with updated assignments, recomputed representatives, relabeled affected clusters, timing artifacts, and a record of edit provenance.

A *derived run* is therefore a persistent successor state linked to its unchanged source run. Figure 3.5 contrasts the two transformations. A split replaces one selected cluster with finer subclusters; a merge replaces several selected clusters with one broader cluster. In both cases, the user selects the target structure and the system performs the transformation and relabeling.

This design has two methodological advantages. First, it supports auditability because each refinement step remains inspectable and can be related to the user action that produced it. Second, it allows concept refinement to be studied as an evolving process rather than as a destructive modification of a single run.

3.7.1 Split Operations

Users initiate a split when inspection indicates that a cluster contains more than one meaningful concept or is otherwise too broad for the intended analysis. The system then

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

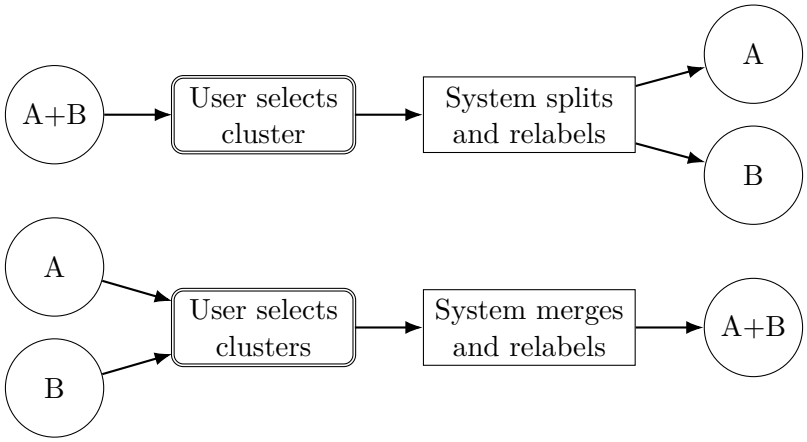


Figure 3.5: User-directed refinement operations. For example, A may represent *cat*, B may represent *dog*, and A+B may represent the broader concept *animal*. Double-bordered nodes represent human decisions; the following structural transformation and relabeling are performed by the system and persisted as a derived run.

divides the selected source cluster into finer subclusters using either *local reclustering* or *visual MLLM splitting*.

Split Strategy: Local Reclustering

The *local reclustering* method treats all images in the selected source cluster as a local dataset. It then reclusters only this subset of images into a specified number of subclusters. The resulting local assignments are mapped back into the global run structure. Representatives and labels are recomputed only for the affected clusters.

This method relies directly on the local embedding structure. It performs another clustering step and therefore depends on the quality of the embedding space and the chosen number of subclusters.

Split Strategy: Visual MLLM

The *visual MLLM* split method uses representative images from the selected source cluster. The MLLM is asked to propose a specified number of subgroup labels. These subgroup labels are then used as text targets, and all images in the source cluster are assigned to the proposed subgroups using CLIP similarity.

In this method, the MLLM proposes the semantic split dimensions but does not directly assign every image. The final assignment remains embedding-based. Consequently, the proposed split dimension depends on the representative evidence shown to the model and may differ from the dominant structure of the full cluster.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.7.2 Merge Operations

Users initiate a merge when two or more clusters should be represented by a shared broader concept. The system combines the selected clusters into one derived cluster and relabels the result. Merge operations therefore support semantic abstraction by deriving a broader concept from several more specific source clusters. The prototype supports three merge-labeling strategies: *deterministic WordNet*, *visual MLLM*, and *hybrid LLM*.

At a high level, a merge operation:

1. unions the selected source clusters at the assignment level,
2. recomputes representatives for the merged cluster,
3. derives a new semantic label for the merged cluster,
4. relabels only the affected derived cluster, and
5. preserves all unaffected clusters unchanged.

Merge Strategy: Deterministic WordNet

The *deterministic WordNet* merge method takes the union of all images from the selected source clusters. It recomputes WordNet candidate evidence for the merged image set and then applies the normal WordNet labeling strategy as described in Section 3.5.1.

This method is deterministic and image-based. It stays closest to the original WordNet labeling pipeline and does not depend on existing cluster labels. However, when visually diverse clusters are merged, the common lexical ancestor may become too broad to be useful.

Merge Strategy: Visual MLLM

The *visual MLLM* merge method builds a balanced representative sample across the selected source clusters. These representative images are sent to the MLLM, which is asked to generate one merged label for the combined visual evidence.

This method is useful when the visual commonality between clusters is clearer than the existing labels. Because it uses image evidence, it can sometimes recover a better abstraction than a purely label-based method. However, it can also be biased by the representative sample and may produce descriptive labels that enumerate visible categories rather than deriving a true common abstraction.

Merge Strategy: Hybrid LLM

The *hybrid LLM* merge method does not use images directly. Instead, it uses the source labels of the selected clusters as textual evidence and asks the LLM to produce a broader common hypernym-like label.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Operation	Method	Evidence type	Mechanism and constraint
Split	*Local reclustering*	Image embeddings	Derives local groups without predefined subgroup labels; requires a target cluster count
Split	*Visual MLLM*	Representative images and CLIP similarity	Generates subgroup labels before assignment; depends on representative evidence
Merge	*Deterministic WordNet*	Merged image set and WordNet evidence	Recomputes lexical evidence over the union; constrained by the hierarchy
Merge	*Visual MLLM*	Representative images from selected clusters	Generates one label from balanced visual evidence
Merge	*Hybrid LLM*	Stored source labels	Abstracts stored source labels; inherits their errors

Table 3.1: Implemented methods for executing user-directed split and merge operations.

This method avoids recomputing image-based candidate evidence and does not require multimodal image input. Its main constraint is its dependence on the existing source labels: generic, overly specific, or incorrect source labels can be inherited by the merged abstraction.

3.7.3 Comparison of Refinement Methods

Table 3.1 summarizes the system methods available for executing user-directed split and merge operations and the type of evidence each method uses.

User initiation. In the graphical interface, split and merge operations are explicitly initiated by the user after inspecting the cluster contents; the system does not autonomously decide which concepts must be refined.

3.8 Artifacts, Timing, and Reproducibility

Each run persists an extensive artifact set, including configuration files, dataset index files, embeddings, cluster assignments, representative summaries, label outputs, debugging information, timing information, and HTML reports. This artifact-oriented design is important for thesis work because it preserves intermediate decisions and makes later inspection, debugging, and comparison possible.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

These artifacts include:

- `dataset_index.jsonl`, a JSON Lines file containing one indexed image record and its metadata per line;
- `clusters.json`, containing cluster assignments and cluster-level information;
- `representatives.json`, containing representative-image selections and cluster summaries;
- `cluster_labels.json`, containing labels, descriptions, alternatives, and evidence from the labeling backends;
- `run_timing.json`, containing timing measurements for full runs and derived-run edit operations; and
- an HTML report for qualitative inspection.

The prototype additionally records stage timings in a dedicated timing artifact. For the labeling stage, work shared by multiple backends is measured separately from backend-specific work. This is especially important for comparing the deterministic WordNet and hybrid LLM backends, which share the same candidate-retrieval stage but differ in the final decision logic.

The prototype further includes multiple cache layers, including image-embedding caches, text-embedding caches, and WordNet-related caches. These mechanisms reduce redundant computation across runs and support incremental reuse of previously generated artifacts.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Experiments

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Dataset	Images	Classes	Cluster settings	Sampling
STL-10	10,000	10	$K = 10$; $K = 5$ for splits	Combined train and test split; 1,000 images per class
PACS	2,240	7	$K = 7$ and $K = 28$	All four domains; 80 images per class–domain combination
Food-101	10,100	101	$K = 101$	Test split; 100 images per class

Table 4.1: Datasets and balanced samples used in the cross-dataset evaluation.

This chapter describes the experimental design and reports the evaluation of SaCoDi. The cross-dataset experiments compare label quality and refinement behavior across three image datasets: STL-10, PACS, and Food-101. Separate STL-10 studies examine the choice of clustering method and compare runtime across three hardware configurations. All experiments were executed through the command-line interface (CLI); the graphical interface was not evaluated as a user interface.

Both generative backends were instantiated with the Qwen3-VL 4B checkpoint, a generative vision–language model that supports textual and visual input [BCC+25, Qwe25]. The same model was used in two roles: with image–text input for the direct MLLM backend and with text-only input for the hybrid LLM backend. These terms therefore distinguish the model's role in the pipeline rather than different models. LM Studio was used to load the model and serve requests through a local OpenAI-compatible API [LM 26, Ope26].

4.1 Evaluation Design

4.1.1 Datasets and Sampling

The three datasets expose different challenges. STL-10 provides natural images from ten broad object classes and supports a controlled baseline [CNL11]. PACS, named after its four visual domains—photo, art painting, cartoon, and sketch—contains seven object classes and therefore reveals whether the embedding space organizes images by content, visual domain, or both [LYSH17]. Food-101 contains images of 101 fine-grained dish categories and tests a larger, domain-specific vocabulary [BGVG14]. The differences in class granularity and visual style are illustrated in Figure 4.1. Ground-truth class and domain metadata are used to construct the balanced samples and, after inference, for evaluation and case selection; they are not supplied to clustering or labeling. Table 4.1 summarizes the resulting samples and cluster settings.

The PACS settings represent two hypotheses about useful granularity. With $K = 7$, the number of clusters equals the number of content classes. With $K = 28$, it equals the number of class–domain combinations. The comparison is not intended to recover a known taxonomy automatically; it tests which distinctions dominate the shared embedding space.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

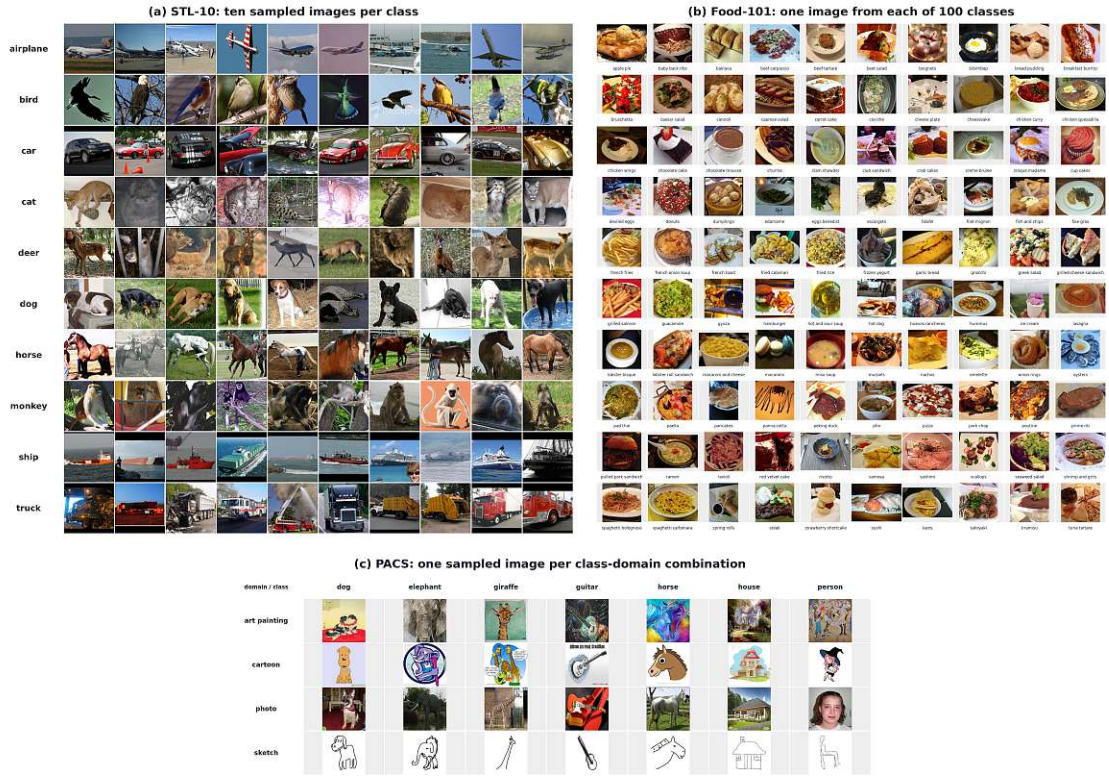


Figure 4.1: Labeled examples from the balanced samples used in the cross-dataset evaluation. (a) STL-10 shows ten sampled images per class, arranged by class rows. (b) Food-101 shows one sampled image for each of 100 distinct classes; only *waffles* is omitted from the 101-class dataset. (c) PACS shows one sampled image for every class–domain combination, with classes in columns and domains in rows.

4.1.2 Common Configuration and Provenance

Table 4.2 records the main configuration recovered from the run artifacts. The same sampled records, embeddings, assignments, and representatives are reused by all three labeling backends when labeling the initial cluster structure. For each cluster, representative selection stores three central images and six images selected to cover visual variation. The controlled comparison uses the three central images and the first three diversity-oriented images. These six images are passed to the direct MLLM, and their embeddings are scored during the CLIP–WordNet candidate retrieval shared by the deterministic and hybrid backends. The three labeling strategies therefore derive their outputs from the same visual evidence.

The result artifacts contain 196 runs. They include four initial cluster structures, repeated initial-label timings, the selected refinements, and repeated STL-10 edit timings. The semantic evaluation comprises 438 reviewed initial cluster labels and 62 reviewed refinement labels. Each row retains the generated surface label, cluster composition,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Component	Configuration
Embedding	OpenCLIP `ViT-B-32`, checkpoint `laion2b_s34b_b79k`; checkpoint-provided transforms without custom preprocessing; unnormalized image embeddings persisted; batch size 64
Clustering	KMeans; L2 normalization enabled; random seed 42; K according to Table 4.1
Representative evidence	Nine images selected per cluster: three central and six diversity-oriented; the direct MLLM and CLIP–WordNet retrieval both use the same three central and three diversity-oriented images
WordNet vocabulary	Root: `physical_entity.n.01`
Candidate retrieval	Combined mean–centroid similarity with mean weight $\alpha = 0.7$; 20 highest-scoring labels considered; candidates retained if their similarity is at least 0.15 and within 0.08 of the best score; candidate labels limited to four tokens
Direct MLLM	Image–text input; temperature 0; six images per cluster; maximum image size 128 pixels
Hybrid LLM abstraction	Text-only input; label strings of at most the five highest-ranked WordNet candidates remaining after retrieval filtering; temperature 0

Table 4.2: Common configuration of the cross-dataset evaluation. All values are taken from the persisted `run_config.json` artifacts.

covered ground-truth classes, and structured author judgments.

4.1.3 Experimental Factors and Measures

Table 4.3 maps the independent variables and dependent measures to the research questions. Initial-label comparisons hold the cluster structure and evidence budget fixed within each dataset configuration. Split and merge cases are selected post hoc with ground truth, but the operations themselves receive no ground-truth labels.

Structural measures. For a clustering with index sets I_k and ground-truth class y_i, weighted purity is

$$\text{purity}_{\text{weighted}} = \frac{1}{N} \sum_{k=1}^{K} \max_{y} |\{i \in I_k : y_i = y\}|. \tag{4.1}$$

Adjusted mutual information (AMI), homogeneity, and completeness complement purity. Homogeneity is high when each cluster contains only members of a single ground-truth class, whereas completeness is high when all members of a class are assigned to the same cluster. Both measures range from 0 to 1, with higher values indicating stronger agreement. For PACS, these measures are computed separately against class, domain, and their Cartesian product. Purity is used to describe local split effects; it is not treated as a merge objective because merging deliberately combines concepts.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Experiment	Independent variables	Controlled variables	Dependent variables	RQ
Initial labeling	Backend; dataset; K	Sample, embeddings, assignments, representatives, evidence budget	Coverage, exclusivity, class and semantic compatibility, specificity, comprehensibility, valid output	RQ1
Split refinement	Split method; selected source cluster	Source run, target count, labeling configuration	Purity, AMI, resulting-label quality, runtime	RQ2a RQ3
Merge abstraction	Merge method; selected source clusters	Source run and stored evidence	Coverage, exclusivity, specificity, comprehensibility, valid output, runtime	RQ2b RQ3
Hardware comparison	GPU	STL-10 sample, $K = 10$, fixed generative-model configuration, and cached pipeline	Labeling and end-to-end latency	RQ3

Table 4.3: Experimental factors, dependent measures, and research-question mapping.

Coverage and exclusivity. Let $A(l, y) \in \{0, 1\}$ be an author-coded relation indicating whether label l covers class y. For cluster C_k, label coverage is

$$\text{coverage}(l, C_k) = \frac{\sum_{i \in I_k} A(l, y_i)}{|I_k|}. \tag{4.2}$$

Exclusivity measures how strongly the covered classes are concentrated in the target cluster:

$$\text{exclusivity}(l, C_k) = \frac{\sum_{i \in I_k} A(l, y_i)}{\sum_{i=1}^{N} A(l, y_i)}. \tag{4.3}$$

Exclusivity considers all images in the evaluated sample whose ground-truth class is covered by label l. It reports the proportion of these images that occur in cluster C_k; the denominator was nonzero for every reviewed label. Dataset-level coverage and exclusivity are reported as unweighted means over clusters, giving each proposed concept equal importance. Coverage is additionally reported as a cluster-size-weighted mean, giving each image equal importance.

Author-coded criteria. The author reviewed all outputs with the fixed rubric in Table 4.4. Class compatibility accepts the majority class, a transparent synonym, or a phrase that unambiguously identifies that class. Semantic compatibility additionally accepts a suitable hypernym. Class compatibility is a per-cluster labeling measure: it does not measure whether all dataset classes are recovered uniquely. Specificity and comprehensibility use *good*, *partial*, and *poor*; the aggregate tables report the proportion

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Criterion	Positive decision	Partial or negative decision
Class compatibility	Clearly denotes the cluster's majority class	Hypernym, different class, or unrelated phrase
Semantic compatibility	Appropriate class term, synonym, description, or hypernym	Does not describe the dominant cluster semantics
Comprehensibility	Familiar, concise, and clear without specialist lexical knowledge	Ambiguous, unusually technical, or unnecessarily verbose
Specificity	Narrowest useful label that covers the coded concepts	Omits a target concept or includes substantially broader concepts
Valid output	Required structured fields and usable primary label are present	Missing or malformed required output

Table 4.4: Rubric used for the structured author assessment. It makes the decisions auditable but does not provide inter-rater reliability or evidence of user understanding.

rated *good*. Valid output requires a usable primary label and the expected structured fields.

Runtime protocol. Quality and runtime are evaluated separately. Repeated measurements reuse caches for images, text embeddings, and WordNet data. Unless stated otherwise, runtime measurements were obtained on an RTX PRO 6000. The main downstream comparison excludes indexing and image embedding, because refinement starts from an existing run. Candidate retrieval is shared by the deterministic and hybrid paths, so their totals overlap.

4.2 Preliminary Clustering Screening

KMeans, MiniBatchKMeans, HDBSCAN, and agglomerative clustering were screened on STL-10 as a preliminary method-selection study. Table 4.5 summarizes this method-selection study.

KMeans offered a fixed, user-controllable cluster count, the shortest or near-shortest runtime, and the strongest purity in the screened configurations. It was therefore fixed for the subsequent backend comparison. This screening selects an experimental baseline; it is not a general benchmark of clustering algorithms.

4.3 Initial Cluster Structures

Table 4.6 characterizes the structures presented to the labeling backends. STL-10 is strongly class-aligned, although one dog/horse cluster remains mixed and the airplane class is divided across two clusters. Food-101 is substantially harder: its weighted purity is 0.674 despite setting K to the number of classes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Method	Images	Clusters	Time	Purity	Main observation
Agglomerative	5,000	10	2.480 s	0.292	Imbalanced clusters and singletons
HDBSCAN	5,000	5	13.640 s	0.422	Large mixed animal cluster and noise
KMeans	5,000	10	**0.238 s**	**0.977**	Strong class alignment
MiniBatchKMeans	5,000	10	0.255 s	0.884	Mixed animal cluster
KMeans	10,000	10	**0.591 s**	**0.882**	Mixed dog/horse cluster
MiniBatchKMeans	10,000	10	0.668 s	0.869	Several mixed clusters

Table 4.5: Preliminary clustering-method screening on STL-10. HDBSCAN's cluster count excludes noise.

Dataset	Target metadata	K	Purity	AMI	Homogeneity	Completeness
STL-10	Class	10	**0.882**	0.896	0.883	0.909
PACS	Class	7	0.564	0.557	0.555	0.563
PACS	Domain	7	0.614	0.426	0.510	0.369
PACS	Class–domain	7	0.250	0.681	0.540	**0.938**
PACS	Class	28	0.873	0.649	0.876	0.522
PACS	Domain	28	**0.979**	0.558	**0.942**	0.400
PACS	Class–domain	28	0.852	**0.907**	0.904	0.921
Food-101	Class	101	0.674	0.721	0.752	0.758

Table 4.6: Structure of the evaluated initial clusterings. Bold values highlight the strongest PACS signals, not values directly comparable across different target metadata.

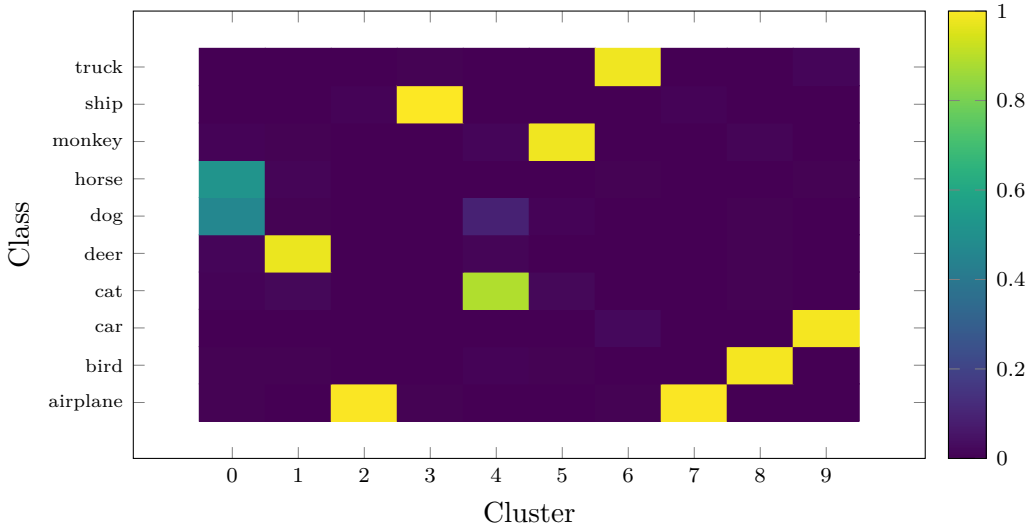
Figure 4.2 shows two complementary cluster-composition matrices. The STL-10 matrix exposes the mixed horse/dog cluster and the duplicated airplane structure. The PACS matrix shows that most $K = 28$ clusters are dominated by one visual domain. Figure 4.3 then isolates the PACS AMI values to make the interaction between content and domain visible.

The PACS result answers the content-versus-domain question more precisely than either purity or a two-domain comparison alone. At $K = 7$, neither factor cleanly explains the structure, although the combined target has the highest AMI. At $K = 28$, domain purity reaches 0.979 and the class–domain AMI reaches 0.907. The clusters are therefore not purely content based: visual domain strongly partitions the space, while content differentiates clusters within that domain-sensitive organization.

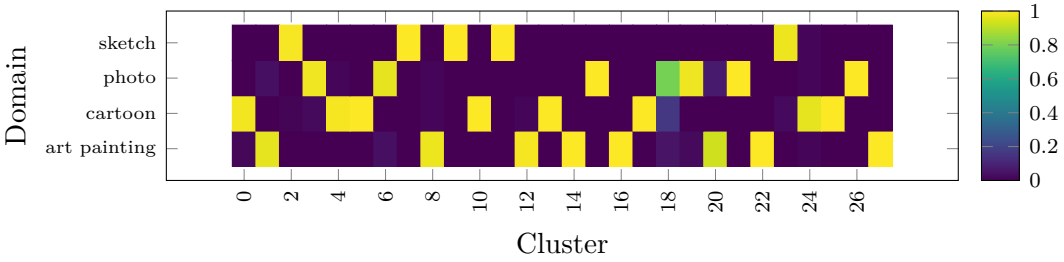
4.4 Initial Label Quality Across Datasets

Table 4.7 aggregates all 438 reviewed labels for the initial cluster structures. Each configuration contributes one row per backend. Values are macro means except for weighted coverage, which weights cluster-level coverage by cluster size.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) STL-10 class fractions.



(b) PACS $K = 28$ domain fractions.

Figure 4.2: Cluster-composition heatmaps from the evaluated runs. Cell color represents the fraction of a cluster assigned to the respective ground-truth category.

Figure 4.4 visualizes the two measures most directly connected to semantic usefulness across all configurations: weighted coverage and semantic compatibility.

The direct MLLM backend has the highest weighted coverage, exclusivity, and semantic-compatibility score in every configuration. The gap is smallest on STL-10, whose broad object classes are well represented in WordNet, and largest on Food-101, where labels must distinguish fine-grained dishes. Food-101 also exposes the difference between class and semantic compatibility: a useful label such as *salad* may cover several visually related food classes without reproducing the majority class name.

The hybrid LLM backend improves semantic compatibility over deterministic WordNet on PACS $K = 28$ and Food-101, but its constrained candidate evidence does not guarantee class-level specificity. Its comprehensibility is also dataset dependent: all ten STL labels are rated good, compared with 39.3% on PACS $K = 28$. WordNet remains deterministic and returns valid structured output in all cases, but its low Food-101 exclusivity and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

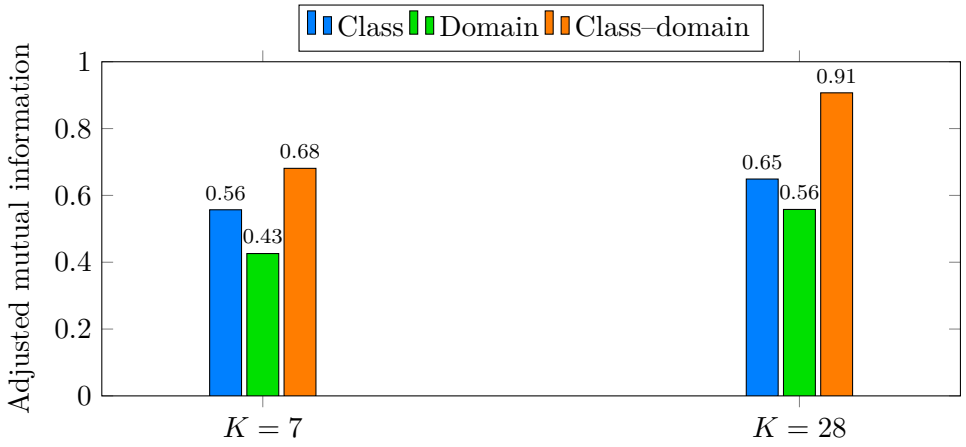


Figure 4.3: PACS cluster agreement with content class, visual domain, and their combined labels. The $K = 28$ result is most strongly explained by class–domain combinations.

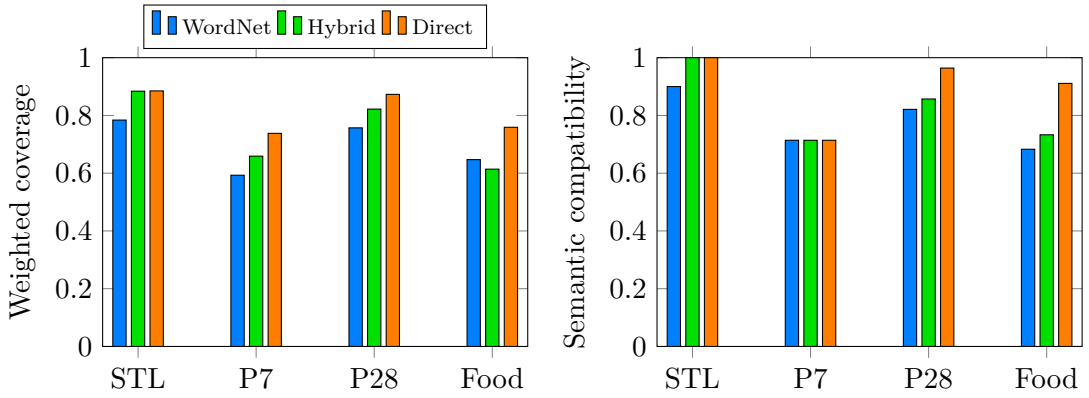
Data	Backend	Cov.	W. cov.	Excl.	Class	Sem.	Compr.	Spec.	Valid
STL $K = 10$	WordNet	0.831	0.784	0.784	**0.800**	0.900	**1.000**	0.800	**1.000**
	Hybrid	0.929	0.884	0.809	0.700	**1.000**	**1.000**	**0.900**	**1.000**
	Direct MLLM	**0.931**	**0.885**	**0.817**	**0.800**	**1.000**	0.900	0.800	**1.000**
PACS $K = 7$	WordNet	0.632	0.593	0.486	**0.429**	**0.714**	**1.000**	0.286	**1.000**
	Hybrid	0.705	0.659	0.438	0.143	**0.714**	0.429	0.429	**1.000**
	Direct MLLM	**0.776**	**0.738**	**0.522**	**0.429**	**0.714**	**1.000**	**0.571**	**1.000**
PACS $K = 28$	WordNet	0.784	0.757	0.170	0.643	0.821	0.893	0.536	**1.000**
	Hybrid	0.831	0.822	0.159	0.107	0.857	0.393	0.250	**1.000**
	Direct MLLM	**0.898**	**0.873**	**0.198**	**0.679**	**0.964**	**0.929**	**0.893**	**1.000**
Food $K = 101$	WordNet	0.641	0.647	0.111	0.030	0.683	0.743	0.198	**1.000**
	Hybrid	0.625	0.614	0.232	0.158	0.733	**0.941**	0.505	**1.000**
	Direct MLLM	**0.762**	**0.759**	**0.451**	**0.416**	**0.911**	0.901	**0.584**	**1.000**

Table 4.7: Initial label quality. *Compr.* and *Spec.* are proportions rated good. Best values within each dataset configuration are bold.

specificity show that a formally valid lexical parent can be too broad for exploration.

Table 4.8 lists the STL-10 surface labels because this controlled case makes individual backend choices and the effects of imperfect cluster structure directly inspectable. The truck-dominated cluster is the clearest deterministic failure: WordNet selects *horse*, although trucks account for 97.7% of the cluster. The stored trace shows that truck-related evidence was present: *truck* is the third-ranked visual candidate. However, the retained terms *wheeler* and *workhorse* map to equine WordNet senses. Because the final hypernym score also rewards depth in the hierarchy, their combined support makes the specific ancestor *horse* outrank broader vehicle-related alternatives. The error therefore results from ambiguous candidate strings combined with the specificity preference during

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



(a) Coverage weighted by cluster size. (b) Author-coded semantic compatibility.

Figure 4.4: Cross-dataset comparison of the three labeling backends for the initial cluster structures. P7 and P28 denote the two PACS configurations.

Cluster	Size	Majority	Purity	WordNet	Hybrid	Direct MLLM
0	1,932	horse	0.512	horse	Horse	horses
1	995	deer	0.972	deer	Deer	deer
2	471	airplane	0.996	airplane	Airplane	airplanes
3	998	ship	0.997	ship	Ship	Cargo ships
4	1,083	cat	0.882	cat	Cat	cats and a lion
5	985	monkey	0.980	primate	Primate	monkeys on structures
6	1,011	truck	0.977	horse	Truck	commercial vehicles
7	530	airplane	0.994	airplane	Aircraft	seaplanes
8	1,004	bird	0.985	bird	Bird	birds on branches
9	991	car	0.988	car	Vehicle	cars

Table 4.8: Raw labels for the final STL-10 $K = 10$ structure.

hierarchy aggregation, rather than from absent truck evidence.

Conversely, *primate* is semantically appropriate for a monkey-dominated cluster but is not a class-compatible match. This is a mismatch with the stricter class-compatibility criterion rather than a semantically incorrect label. The mixed horse/dog cluster limits coverage for all majority-oriented labels. Moreover, *airplane* appears as the majority class of two clusters while *dog* appears in none; high per-cluster compatibility therefore must not be interpreted as complete recovery of all ten classes.

All 438 initial-label outputs are structurally valid. Robustness differences are therefore not observed in this controlled run. This result concerns the generated artifacts, not the stability of repeated free-form generation or perceived quality in the graphical interface.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Case	Source-cluster composition	Target clusters	Source purity	Local reclustering	Visual MLLM
STL horse/dog	horse 49.7% dog 47.7% other 2.6%	2	0.497	**0.969**	0.507
STL ship/truck	ship 50.4% truck 48.9% other 0.7%	2	0.504	**0.993**	0.991
PACS sketch cluster	horse 19.1% dog 18.9% elephant 18.9% giraffe 18.9% person 18.9% other 5.2%	5	0.191	**0.730**	0.574
Food soups	lobster bisque 43.2% clam chowder 42.7% other 14.1%	2	0.432	**0.779**	0.648

Table 4.9: Weighted class purity before and after the evaluated split operations. Horizontal bars show the class composition of each source cluster; infrequent classes are grouped as *other*.

4.5 Split-Based Refinement

Four mixed source clusters were selected post hoc as controlled stress cases: two near-balanced STL-10 clusters at $K = 5$, the lowest-purity PACS $K = 7$ cluster, and a two-class Food-101 soup cluster. Both split methods are given the same requested number of child clusters for each case. Local reclustering partitions the source-cluster embeddings directly. In the visual MLLM method, the model instead inspects representative images and generates the requested number of subgroup labels. CLIP then assigns every image from the source cluster to the generated label with the highest image–text similarity. Table 4.9 reports the structural results.

Figure 4.5 presents the same comparison graphically.

Local reclustering produces the highest purity in all four cases. The effect is strongest when the selected number of subclusters corresponds to structure already separated in the embedding space. This does not imply that local reclustering always finds the user-intended distinction; both the case and target count are supplied externally.

The visual MLLM method demonstrates the complementary failure mode. For STL horse/dog it proposes *Horses with White Markings* and *Solid-Colored Horses*, splitting by visual subtype rather than the two dominant classes. For ship/truck it proposes *Cargo Ships* and *Truck Trailers*, and its purity is almost identical to local reclustering. On PACS it recovers only part of the five content classes from a sketch-dominated source cluster. On Food-101 it proposes tomato-based and creamy vegetable soups, neither of which covers the annotated lobster-bisque and clam-chowder classes. A visually plausible

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

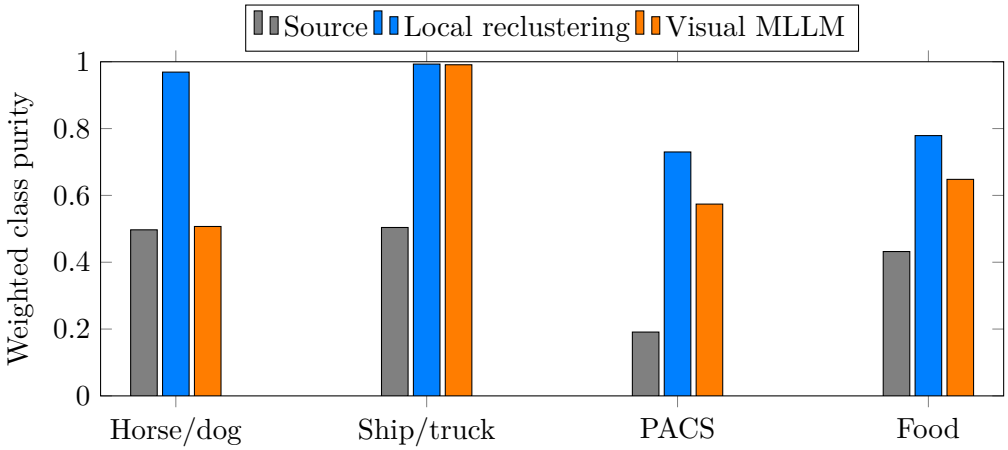


Figure 4.5: Structural effect of the two split methods in the four selected cases.

split dimension is therefore not necessarily the relevant semantic dimension.

Relabeling after local reclustering also remains backend dependent. In the STL horse/dog case, the direct MLLM correctly names the separated concepts, whereas deterministic WordNet labels one child *person*. In the PACS case, the backends produce labels at different levels of granularity. The structural operation and subsequent semantic labeling must therefore be evaluated separately.

4.6 Merge-Based Abstraction

Six merge cases were selected post hoc as controlled abstraction cases. STL-10 provides three pairwise cases and one four-class transport merge. PACS merges four guitar clusters selected across all domains, testing whether a shared content concept survives domain-specific partitioning. Food-101 merges the visually similar pho and ramen clusters.

For the hybrid LLM merge, the prompt consumes the direct-MLLM labels from the source run. This source-label backend is held fixed; alternatives are not compared. Table 4.10 exposes the raw labels, and Table 4.11 aggregates the reviewed criteria.

The hybrid LLM merge has the highest mean coverage, while the visual MLLM merge has the highest exclusivity. Their differences are visible in the Food-101 case: *Asian noodle soups* covers both pho and ramen, whereas *Vietnamese noodle soups* is exclusive but covers only the pho portion. The deterministic WordNet merge produces a correct but overly broad *dish* label there and fails on two STL vehicle pairs by selecting *person*. Its deterministic procedure is therefore not sufficient protection against incorrect abstraction when the retrieved evidence is unsuitable.

The PACS merge succeeds for all three methods. Four domain-separated clusters are mapped back to a common guitar concept with coverage 1.0 and exclusivity 0.981. This complements the initial PACS finding: domain can dominate the initial partition while

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Case	Source concepts	Deterministic WordNet	Hybrid LLM	Visual MLLM
STL cat/bird	cat, bird	living thing	wild animals in natural habitats	cats and birds
STL car/truck	car, truck	person	vehicles	commercial vehicles
STL ship/truck	ship, truck	person	Transport vehicles	trucks and cargo vessels
STL transport	car, truck, ship, airplane	vessel	Vehicles	aircraft and vehicles
PACS guitar	guitar in four domains	stringed instrument	guitars	guitars and related instruments
Food soups	pho, ramen	dish	Asian noodle soups	Vietnamese noodle soups

Table 4.10: Raw labels produced by the three merge methods for the six selected cases.

Method	Coverage	Exclusivity	Good specificity	Good comprehensibility	Valid
Deterministic WordNet	0.542	0.391	0.167	**1.000**	**1.000**
Hybrid LLM	**0.980**	0.731	**0.500**	**1.000**	**1.000**
Visual MLLM	0.901	**0.926**	**0.500**	0.833	**1.000**

Table 4.11: Macro averages over the six reviewed merge cases.

a user-directed merge restores a content-level concept. All 18 merge outputs pass the structured-output validity criterion, but the six manually selected cases do not estimate performance over arbitrary cluster combinations.

4.7 Computational Performance

4.7.1 Initial Labeling and Local Edits

On the RTX PRO 6000, labeling all three backends for the final STL-10 configuration takes 9.390 ± 0.165 seconds over ten repetitions. The direct MLLM component takes 3.230 ± 0.040 seconds. The deterministic total takes 4.656 ± 0.088 seconds and the hybrid total 3.877 ± 0.146 seconds, but both include the same 2.385 ± 0.078 second candidate-retrieval stage and must not be added.

Food-101 shows how backend cost grows with 101 clusters. Measured separately over three repetitions, WordNet requires 6.008 ± 0.018 seconds, the hybrid backend 19.855 ± 1.218 seconds, and the direct MLLM 39.075 ± 0.626 seconds. This difference is consistent with the number of cluster-level semantic decisions being more influential at this stage than

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Dataset and operation	Local reclustering/ Deterministic WordNet	Hybrid LLM	Visual MLLM
STL split, horse/dog	6.266 ± 0.238 s	–	2.089 ± 0.071 s
STL split, ship/truck	6.337 ± 0.103 s	–	2.093 ± 0.119 s
PACS split	7.155 s	–	2.047 s
Food-101 split	6.307 s	–	2.180 s
STL merges, observed range	5.111–5.195 s	0.726–0.778 s	0.825–0.933 s
PACS merge	4.685 s	0.313 s	0.547 s
Food-101 merge	5.095 s	0.820 s	1.081 s

Table 4.12: Runtime of the final refinement cases on the RTX PRO 6000. STL values use ten repetitions; PACS and Food-101 values are single quality-run timings and are not used for variance claims.

the 10,100 image embeddings, which are already available. Table 4.12 complements these initial-label measurements with the local edit costs.

The visual MLLM split is faster because it proposes labels and assigns images once, whereas local reclustering is followed by representative recomputation and all three labeling backends. For merges, the hybrid LLM method is fastest; deterministic WordNet recomputes lexical evidence over the merged image set. These measurements support local artifact reuse, but they measure backend latency rather than perceived interface responsiveness.

4.7.2 Hardware Comparison

To isolate the effect of serving hardware, the cached STL-10 pipeline was executed on three systems while holding the dataset sample, $K = 10$ cluster structure, and generative-model configuration constant. Figure 4.6 normalizes the mean labeling-stage latency to the RTX PRO 6000. The RTX 4090 requires 1.17 times and the RTX 3080 Ti 1.67 times the labeling latency of this reference system. The comparison therefore quantifies the relative hardware effect without treating the absolute measurements as portable performance estimates.

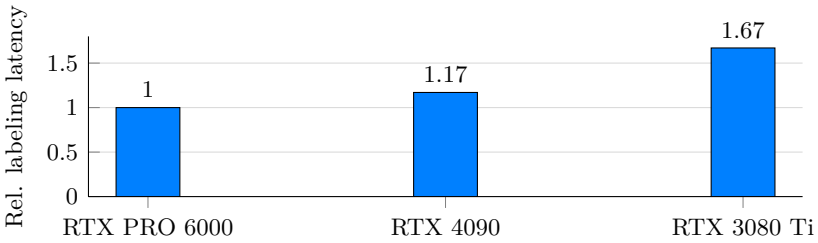


Figure 4.6: Labeling-stage latency across the three hardware configurations, normalized to the RTX PRO 6000. Lower values are better.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CHAPTER 5

Discussion

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

This chapter interprets the experimental results in relation to the research questions and the requirements defined in Section 1.3. The experiments evaluate generated structures, labels, refinements, and backend latency through the CLI. The graphical interface operationalizes human inspection and control, but its usability and effect on analytical reasoning remain outside the evaluated scope.

5.1 RQ1: Cluster Label Quality

RQ1 asks how deterministic WordNet, direct MLLM, and hybrid LLM labeling compare with respect to coverage, exclusivity, class and semantic compatibility, specificity, author-assessed comprehensibility, and output robustness. Across STL-10, both PACS configurations, and Food-101, the direct MLLM obtains the highest weighted coverage, exclusivity, and semantic-compatibility score and the highest or tied-highest class compatibility. The result is consistent across broad object classes, domain-shifted depictions, and fine-grained food categories, but the absolute values vary substantially with the dataset and cluster structure.

5.1.1 Label Quality Depends on Cluster Structure

STL-10 provides the easiest labeling setting. Its broad object classes are strongly separated by the embedding, and both generative paths reach full semantic compatibility. Nevertheless, the large mixed horse/dog cluster cannot be described completely by a single majority-oriented label. Moreover, two clusters are airplane dominated while no cluster is dog dominated. This outcome is possible because KMeans minimizes variation within the embedding-space clusters but does not enforce a one-to-one correspondence between those clusters and the ten ground-truth classes. The labels *airplanes* and *seaplanes* suggest that visual subtype structure divided the airplane class while dogs and horses remained grouped. Per-cluster class compatibility therefore describes the usefulness of labels for existing clusters, not unique recovery of the dataset taxonomy.

PACS exposes a different dependency. At $K = 28$, its structure is strongly domain sensitive, with domain purity 0.979 and class–domain AMI 0.907. A label can correctly describe the visible object while remaining incomplete as a description of why the cluster is separate from another cluster with the same object in a different domain. This explains the low exclusivity values despite relatively high coverage. At $K = 7$, mixed class and domain effects reduce both structural coherence and label coverage.

Food-101 is the most difficult initial-labeling setting. Its weighted structural purity is 0.674, and visually related dishes often share clusters. The gap between class compatibility and semantic compatibility is consequently large. A term such as *salad* or *soup* can describe several covered food classes without reproducing the majority class name. The direct MLLM handles these fine-grained concepts best among the evaluated backends, but its weighted coverage of 0.759 remains well below complete cluster description.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.1.2 Complementary Backend Behavior

WordNet offers repeatable decisions and inspectable lexical evidence, but the hierarchy and retrieval stage do not guarantee an appropriate result. The label *horse* for a truck-dominated STL-10 cluster is a categorical error, not merely a granularity mismatch. As the trace discussed in Section 4.4 shows, string-based visual retrieval can admit terms whose selected WordNet senses belong to a different semantic branch; the hierarchy score can then amplify this ambiguity by preferring a deep, specific ancestor. In Food-101, WordNet more commonly selects correct but broad ancestors, which produces semantic compatibility without useful specificity or exclusivity.

The hybrid LLM path improves semantic compatibility over WordNet on PACS $K = 28$ and Food-101. It can turn constrained candidates into more natural abstractions, but its result remains bounded by candidate quality. This is visible in its low class compatibility on the two PACS configurations. The direct MLLM is less constrained and produces the strongest aggregate scores, although several surface forms are descriptive phrases rather than normalized category names.

All 438 initial-label outputs pass the structured validity criterion. The experiment therefore shows no robustness difference under the evaluated serving configuration. It does not establish that generative outputs are deterministic or that fallback parsing will remain reliable in other prompts and models.

Answer to RQ1. Within the evaluated configurations, the direct MLLM provides the strongest overall coverage, exclusivity, class compatibility, and semantic compatibility. WordNet is the most inspectable but can be too broad or categorically wrong, while the hybrid LLM path often turns constrained candidate evidence into more suitable abstractions without consistently reaching class-level specificity. All reviewed initial-label outputs are structurally valid, so no output-robustness difference is observed in this configuration. None of the backends compensates for an unsuitable cluster structure, and comprehensibility remains a single-author assessment rather than user evidence.

5.2 RQ2a: Split-Based Refinement

RQ2a concerns how split methods affect mixed clusters. Local reclustering produces the highest class purity in each of the four selected cases: two near-balanced STL-10 clusters, one sketch-dominated PACS cluster, and one mixed Food-101 soup cluster. The result extends the earlier STL-only observation to two different sources of heterogeneity, but remains conditional on post hoc case selection and a supplied target count.

The PACS case is particularly informative for human–AI interaction. The selected source cluster is almost entirely composed of sketches, yet contains five substantial content classes. A user who wants content concepts may therefore reject the domain-induced cluster and request a five-way split. Local reclustering raises purity from 0.191 to 0.730.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The operation does not discover that analytical intention; it executes a user-specified correction to the automatic organization.

The visual MLLM can identify a useful split axis, as in the STL ship/truck case, where its purity reaches 0.991. It can also select an unintended but visually coherent distinction. For horse/dog it proposes two horse appearance groups, and for Food-101 it proposes soup appearances that do not correspond to the dominant annotated dishes. Representative evidence is thus sufficient to suggest possibilities, but not to determine which distinction is relevant to the analyst.

Structural and semantic success are separate. Local reclustering can isolate classes while a subsequent label backend still names a child incorrectly, as demonstrated by WordNet's *person* label in the separated STL animal case. Conversely, a broader semantic label may be appropriate even when it is not class compatible. An interactive workflow must expose both the child images and the generated labels for validation.

Answer to RQ2a. Local reclustering most strongly improves class-based coherence in all four selected cases. Visual MLLM splitting succeeds when the representatives expose the desired distinction, but can choose a different semantic axis. The evidence supports local reclustering for these cases, not a universal ranking of split algorithms, and it leaves the target granularity as a human decision.

5.3 RQ2b: Merge-Based Abstraction

RQ2b asks whether alternative merge methods derive useful common abstractions. Across six cases, the hybrid LLM merge has the highest mean coverage at 0.980, whereas the visual MLLM merge has the highest mean exclusivity at 0.926. Both outperform the deterministic WordNet merge on the aggregate specificity judgment. The comparison demonstrates a practical trade-off: broad coverage is not identical to a label that distinguishes the merged concept from the rest of the collection.

The Food-101 merge makes this distinction concrete. *Asian noodle soups* covers pho and ramen with high exclusivity. *Vietnamese noodle soups* is more exclusive but covers primarily pho, while WordNet's *dish* covers both sources but almost every other food class as well. A useful merge label must therefore be judged against both selected source concepts and surrounding clusters.

PACS shows how merging can counter a domain-driven initial organization. Four highly pure guitar clusters associated with different visual domains are merged successfully by all three methods. The result is not evidence that the automatic system knows which domains should be collapsed. It shows that, once a user selects related clusters, the available abstraction methods can restore a shared content concept.

The hybrid LLM merge consumes the direct-MLLM labels stored for the source clusters. Its performance consequently reflects both the merge model and those upstream labels.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Keeping the source-label backend fixed makes the reported cases reproducible but does not establish that it is the optimal input representation.

Answer to RQ2b. The hybrid LLM and visual MLLM merge strategies derive useful abstractions in most selected cases, with different coverage–exclusivity trade-offs. The deterministic WordNet merge succeeds for the cross-domain guitar case but is frequently too broad or incorrect in the STL and Food-101 cases. Generated merge labels remain proposals that require inspection, particularly when the intended common concept depends on analytical context.

5.4 RQ3: Computational Performance

RQ3 concerns the computational trade-offs of initial labeling and local refinement. Semantic labeling dominates downstream runtime. On the RTX PRO 6000, labeling all three backends in the final STL-10 runs requires approximately 9.4 seconds, while KMeans remains below one second in the preliminary clustering screening. On the same system, Food-101 further shows that the number of cluster-level decisions matters: direct MLLM labeling of 101 clusters requires approximately 39.1 seconds, compared with 6.0 seconds for WordNet.

Derived-run operations reuse existing embeddings and assignments. Visual MLLM splits complete in roughly two seconds in the evaluated cases, whereas local reclustering followed by all three labeling paths takes approximately six to seven seconds. Hybrid LLM merges remain below one second in most cases; deterministic WordNet merges are slower because candidate evidence is recomputed over the merged image set. This supports the artifact-reuse design as an implementation strategy.

Runtime depends strongly on hardware. Relative to the RTX PRO 6000, labeling-stage latency increases by a factor of 1.17 on the RTX 4090 and 1.67 on the RTX 3080 Ti. These measurements characterize backend computation rather than perceived responsiveness; progressive visual analytics instead emphasizes maintaining a short feedback loop between the analyst and the system [SPG14, FP16].

Answer to RQ3. The main computational trade-off is between deterministic lexical processing and per-cluster MLLM inference. Runtime grows primarily with the number of semantic decisions and the serving hardware, while clustering is comparatively inexpensive. Local derived runs are cheaper than full recomputation because they update only affected concepts.

5.5 Implications for SaCoDi and Human–AI Interaction

The results support SaCoDi as a mixed-initiative workflow rather than a fully automatic taxonomy-construction method. The system supplies scale by embedding, clustering,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

labeling, and computing alternatives. The user supplies contextual judgment by deciding whether a group is meaningful, choosing the intended granularity, initiating a split or merge, and accepting or rejecting the result.

The cross-dataset findings make this division of responsibility concrete. PACS demonstrates that an embedding may organize data by visual domain even when content categories are the analytical target. Food-101 demonstrates that one broad label may cover several visually related classes without being sufficiently exclusive. These are not errors that can always be resolved from the data alone: the appropriate concept system depends on what the user wants to investigate.

Backend disagreement can be presented as evidence rather than only as a ranking problem. Agreement among constrained and generative paths may increase confidence in a label, while disagreement such as *horse*, *Truck*, and *commercial vehicles* can signal a cluster that deserves inspection. The interface should preserve this uncertainty and provenance instead of presenting one generated label as ground truth.

Split and merge operations provide complementary control over granularity. Persistent derived runs preserve the relationship between user action and system response and make alternative structures comparable. The CLI evaluation characterizes the computational cost of these operations under the tested configuration. A user study is still necessary to determine whether the graphical workflow actually improves understanding, trust, or analytical efficiency.

5.6 Limitations and Validity

5.6.1 Dataset Scope

STL-10, PACS, and Food-101 cover different class granularities and forms of visual variation, but all three are curated image-classification benchmarks with known classes and balanced evaluation samples. Real exploratory collections may contain multiple objects per image, continuous or overlapping concepts, long-tailed frequencies, duplicates, and domain terminology absent from WordNet or model training. The findings therefore support cross-dataset technical feasibility, not unrestricted generalization to arbitrary image collections.

Ground-truth classes also privilege class-level concepts. An open-ended label may be analytically useful without matching the dataset taxonomy, while a class-compatible label can still omit a minority concept. Reporting coverage, exclusivity, semantic compatibility, and PACS domain structure reduces this bias but does not eliminate it.

5.6.2 Author Judgment and Case Selection

All semantic relations, specificity decisions, and comprehensibility ratings are coded by the author. The structured rubric makes the assessment criteria explicit, but provides neither independent agreement nor evidence about the intended user group. The split

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and merge cases are selected post hoc using ground truth. This creates controlled stress cases rather than an unbiased estimate across all possible edits.

5.6.3 Configuration and Runtime

Results depend on the OpenCLIP checkpoint, representative selection, prompts, retrieval thresholds, WordNet vocabulary, generative-model checkpoint, and serving setup. The run artifacts persist the model and checkpoint identifiers but not the exact installed OpenCLIP library version, which limits environment-level reproduction. Temperature zero reduces but does not eliminate generative variation. The semantic evaluation reviews one output per configuration rather than a distribution over repeated generations.

The preliminary clustering screening and hardware comparison are kept separate from the semantic quality results. Runtime repetitions use warm caches, and absolute latency depends on hardware. Relative patterns are therefore more portable than the reported values.

5.7 Future Work

The highest-priority extension is a user study of the graphical workflow with representative analysts. Testing an implemented visualization system with members of its target user group provides evidence about whether its data and operation abstractions support their work [Mun14]. Human studies are also needed because automatic metrics do not capture all user-experience requirements of foundation-model-assisted visual analytics [HJSM25]. For SaCoDi, such a study should test whether users recognize unsuitable clusters, interpret alternative backend labels, select useful refinements, and understand the provenance of generated abstractions.

Further technical evaluation should use naturally imbalanced and multi-label collections, repeat clustering across random seeds, and sample refinement cases independently of ground truth. Multiple annotators would permit agreement analysis for specificity and comprehensibility.

Automatic refinement suggestions are another extension. This would constitute visual-analytics guidance: the system proposes possible next steps while the user retains control over the direction of the analysis [CGM+17, CGM19]. In SaCoDi, internal embedding variance, backend disagreement, low coverage, and similarity between cluster labels could identify candidate splits or merges. Because the appropriate granularity depends on the analytical task, suggestions should be accompanied by supporting evidence and require user confirmation.

Finally, structured MLLM output should use stricter schema enforcement, validation, and retry policies. Candidate abstractions should be checked against both visual and textual evidence before presentation, especially when a plausible primary label can coexist with unsupported auxiliary evidence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Conclusion

This thesis introduced *Semi-Automatic Concept Discovery* (SaCoDi), a human–AI workflow for organizing unlabeled image collections into labeled visual concepts and refining their granularity through user-directed split and merge operations. SaCoDi does not assume a task-specific taxonomy. Instead, it combines multimodal embeddings, clustering, three complementary labeling strategies, and persistent refinement states.

The workflow was operationalized in a Python prototype with two access paths. A command-line interface supports scripted experiments and repeated measurements, while a graphical interface enables cluster inspection and user-directed refinement. The prototype is an instrument for implementing and evaluating SaCoDi rather than a separate research contribution.

Regarding RQ1, the direct MLLM achieves the strongest aggregate coverage, exclusivity, class compatibility, and semantic compatibility across STL-10, PACS, and Food-101. WordNet provides repeatable and inspectable lexical evidence but can select overly broad or incorrect concepts. The hybrid LLM strategy can turn constrained candidate evidence into more suitable abstractions without consistently reaching the direct method's specificity. All reviewed initial-label outputs are structurally valid, so no output-robustness difference is observed in the evaluated configuration. All three strategies remain dependent on the cluster structure, and comprehensibility is assessed by the author rather than target users.

Regarding RQ2a, local reclustering produces the highest class purity in the four selected split cases. The visual MLLM succeeds when representative evidence exposes the intended distinction, but it can instead propose another visually coherent split dimension. Regarding RQ2b, the hybrid LLM merge provides the highest mean coverage and the visual MLLM merge the highest mean exclusivity across six selected cases. The successful PACS guitar merge demonstrates how user-directed abstraction can reunite a content concept separated by visual domain.

Regarding RQ3, semantic labeling dominates downstream runtime, while clustering is comparatively inexpensive. Runtime grows mainly with the number of clusters requiring semantic decisions and depends on the serving hardware. Derived-run refinement remains cheaper than complete recomputation because it reuses source artifacts and updates only affected concepts.

The thesis makes two contributions. First, it specifies SaCoDi as a semi-automatic concept-discovery workflow that assigns complementary responsibilities to the system and the user. Second, it provides empirical findings about the semantic quality and computational cost of deterministic WordNet, direct MLLM, and hybrid LLM strategies during initial labeling and refinement across three datasets.

The evidence remains bounded by balanced classification benchmarks, post hoc case selection, single-author semantic coding, and the absence of a user study. The results therefore establish technical and methodological feasibility across several controlled data conditions, not general effectiveness for arbitrary collections or analysts. Future work

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

should prioritize user-centered evaluation, independently sampled refinements, multiple semantic annotators, and less curated image collections.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

AI-based tools were used during the preparation of this thesis.

I used OpenAI ChatGPT[1] (GPT-5) for brainstorming, structuring ideas, and refining text passages. Since English is not my first language, ChatGPT was additionally used to improve wording, grammar, and readability. ChatGPT was also used to generate the teaser image and the schematic representation of the multimodal embedding space included in Chapter 1. The conceptual decisions, experimental design, interpretation of results, and final responsibility for the thesis remain my own.

I also used OpenAI Codex[2] (GPT-5-Codex) as a programming assistant during the prototyping phase of the implemented system. Its use was limited to supporting code development, debugging, and exploring implementation alternatives. Generated code was reviewed, adapted, and tested by me before being integrated into the prototype.

Google NotebookLM[3] was used to search, organize, and summarize selected sources provided to the tool. It supported the identification of relevant information within the literature and the comparison of content across sources.

No AI tool was used as an autonomous source of scientific claims. All content was critically reviewed and edited by me.

[1] OpenAI, *ChatGPT*. Available at: `https://chatgpt.com/`. Accessed: August 9, 2026.
[2] OpenAI, *Codex*. Available at: `https://openai.com/codex/`. Accessed: August 9, 2026.
[3] Google, *NotebookLM*. Available at: `https://notebooklm.google.com/`. Accessed: August 9, 2026.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Glossary

class compatibility Author-coded criterion indicating whether a generated label clearly denotes the majority ground-truth class of its target cluster, including transparent synonyms.

cluster labeling Assignment of a concise, human-interpretable textual description to a cluster so that its proposed concept can be inspected and compared.

comprehensibility Author-coded assessment of whether a label is familiar, concise, and understandable without specialist lexical knowledge.

concept discovery Identification and labeling of recurring, human-interpretable patterns in an unlabeled collection without assuming that the relevant categories are known in advance.

coverage Proportion of the images in a target cluster whose author-coded ground-truth classes are described by the generated label.

exclusivity Degree to which the dataset instances of the classes covered by a label are concentrated in its target cluster rather than distributed across other clusters.

hypernym A more general concept related to a more specific concept; for example, *animal* is a hypernym of *horse*.

lexical resource A structured collection of words or concepts and their relations. SaCoDi uses WordNet as a source of candidate labels and hypernym relations.

merge A user-directed refinement operation that replaces several selected source clusters with one combined cluster and a newly generated label.

modality A form in which information is represented, such as written text, still images, or video.

multimodal embedding space A shared vector space in which inputs from different modalities are represented as comparable numerical vectors. In SaCoDi, OpenCLIP uses separate image and text encoders whose outputs are aligned so that semantically related images and texts are close together.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

OpenCLIP An open-source implementation and collection of CLIP-style vision–language models used by SaCoDi to compute image and text embeddings.

Qwen3-VL The generative vision–language model used in the 4B configuration for both generative backends. It receives images and text in the direct MLLM backend and text only in the hybrid LLM backend.

representative image An image selected from a cluster to provide compact visual evidence for automatic labeling, inspection, or refinement.

semantic compatibility Author-coded criterion indicating whether a label appropriately describes the dominant semantics of a cluster, including suitable descriptions and hypernyms.

specificity Author-coded assessment of whether a label is as narrow as possible while still covering the relevant concepts in its target cluster.

split A user-directed refinement operation that replaces one selected source cluster with a requested number of finer subclusters and corresponding labels.

synset A WordNet set of synonymous words that represents one lexicalized concept and participates in relations such as hypernymy.

visual concept A recurring, human-interpretable visual pattern, such as an object, scene, activity, or domain-specific motif.

WordNet A lexical database that groups words into synsets and connects them through lexical and conceptual relations; SaCoDi uses its noun hierarchy as a candidate-label vocabulary.

zero-shot recognition Prediction of classes for which no labeled examples were supplied during task-specific training.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

AI Artificial intelligence; computational methods that perform tasks commonly associated with intelligent behavior. In SaCoDi, AI components propose cluster structures and labels, while users retain control over refinement decisions.

AMI Adjusted mutual information; a chance-corrected measure of agreement between two partitions, used to compare cluster assignments with class, domain, or combined class–domain labels.

API Application programming interface; a defined software interface through which components exchange requests and responses.

CLI Command-line interface; the scriptable interface used to execute the controlled experiments and repeated runtime measurements.

CLIP Contrastive Language–Image Pre-training; a family of vision–language models that maps images and texts into a shared embedding space.

GPU Graphics processing unit; the accelerator on which the embedding and model-inference workloads are executed.

JSON JavaScript Object Notation; the structured serialization format used for model responses and persisted artifacts.

LLM Large language model; a language model used with textual input. In this thesis, LLM specifically denotes the text-only abstraction role in the hybrid backend.

MLLM Multimodal large language model; a generative large language model that accepts more than one modality. In this thesis, the MLLM receives representative images and textual instructions to generate labels directly from visual evidence.

RQ Research question; one of the questions that structures the experimental design and interpretation of the thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

SaCoDi Semi-Automatic Concept Discovery; the proposed human–AI workflow for clustering, labeling, inspecting, and refining visual concepts in unlabeled image collections.

UMAP Uniform Manifold Approximation and Projection; the dimensionality-reduction method used to display image embeddings in the graphical interface.

VLM Vision–language model; the broad category of models that jointly process or represent visual and textual information. An OpenCLIP model is used as a VLM in this thesis; a VLM need not be a generative language model and is therefore not synonymous with MLLM.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[APK+16] Ahmet Aker, Monica Lestari Paramita, Emina Kurtić, Adam Funk, Emma Barker, Mark Hepple, and Robert Gaizauskas. Automatic label generation for news comment clusters. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 61–69, 2016.

[AWV+19] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[BCC+25] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[BGVG14] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461. Springer, 2014.

[BHR+20] Juhee Bae, Tove Helldin, Maria Riveiro, Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, and Göran Falkman. Interactive clustering: A comprehensive review. *ACM Comput. Surv.*, 53(1), February 2020.

[CBW+23] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[CGM+17] Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jörg Schulz, Marc Streit, and Christian Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2017.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[CGM19] Davide Ceneda, Theresia Gschwandtner, and Silvia Miksch. A review of guidance approaches in visual data analysis: A multifocal perspective. *Computer Graphics Forum*, 38(3):861–879, 2019.

[CMZS15] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51, 2015.

[CNL11] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 215–223, 2011.

[CQC+23] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering, 2023.

[CSSL23] Yan Chen, Kate Sherren, Michael Smit, and Kyung Young Lee. Using social media images as data in social science research. *New Media & Society*, 25(4):849–871, 2023.

[ENP+09] Danilo M. Eler, Marcel Y. Nakazaki, Fernando V. Paulovich, Davi P. Santos, Gabriel F. Andery, Maria Cristina F. Oliveira, João Batista Neto, and Rosane Minghim. Visual analysis of image collections. *The Visual Computer*, 25(10):923–937, October 2009.

[ERT+17] A. Endert, W. Ribarsky, C. Turkay, B.L. William Wong, I. Nabney, I. Díaz Blanco, and F. Rossi. The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum*, 36(8):458–486, March 2017.

[FP16] Jean-Daniel Fekete and Romain Primet. Progressive analytics: A computation paradigm for exploratory data analysis, 2016.

[FS11] Fumiyo Fukumoto and Yoshimi Suzuki. Cluster labelling based on concepts in a machine-readable dictionary. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1371–1375, 2011.

[FXJ+18] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[GH15] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[GWZK19] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019.

[HCR22] Haiyang Huang, Zhi Chen, and Cynthia Rudin. Segdiscover: Visual concept discovery via unsupervised semantic segmentation, 2022.

[HJD+24] Lianyu Hu, Mudi Jiang, Junjie Dong, Xinying Liu, and Zengyou He. Interpretable clustering: A survey. *arXiv preprint arXiv:2409.00743*, 2024.

[HJSM25] Maeve Hutchinson, Radu Jianu, Aidan Slingsby, and Pranava Madhyastha. Foundation model assisted visual analytics: Opportunities and challenges. *Computers & Graphics*, 130:104246, 2025.

[HMF+22] Matt-Heun Hong, Lauren A. Marsh, Jessica L. Feuston, Janet Ruppert, Jed R. Brubaker, and Danielle Albers Szafir. Scholastic: Graphical human-ai collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.

[IWC+21] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Zenodo, 2021. Version 0.1.

[KPK+24] Sehyun Kwon, Jaden Park, Minkyu Kim, Jaewoong Cho, Ernest K. Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 1393–1432, 2024.

[LAZ+24] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model, 2024.

[Llo82] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LM 26] LM Studio. LM Studio as a Local LLM API Server. `https://lmstudio.ai/docs/developer/core/server`, 2026. Accessed: 2026-08-09.

[LWA+24] Yiran Li, Junpeng Wang, Prince Aboagye, Chin-Chia Michael Yeh, Yan Zheng, Liang Wang, Wei Zhang, and Kwan-Liu Ma. Visual analytics for efficient image exploration and user-guided image captioning. *IEEE Transactions on Visualization and Computer Graphics*, 30(6):2875–2887, 2024.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[LYSH17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, 2017.

[LZL+26] Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci. Organizing unstructured image collections using natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 8907–8918, June 2026.

[MC12] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[MCNJ13] Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Shafiq Joty. Towards topic labeling with phrase entailment and aggregation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 179–189, 2013.

[MHA17] Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[MRHPAR+23] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, April 2023.

[Mun14] Tamara Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[Ope26] OpenAI. OpenAI API Reference: Chat. `https://developers.openai.com/api/reference/resources/chat`, 2026. Accessed: 2026-08-09.

[PMCEA+22] I. Pérez-Messina, D. Ceneda, M. El-Assady, S. Miksch, and F. Sperrle. A typology of guidance tasks in mixed-initiative visual analytics environments. *Computer Graphics Forum*, 41(3):465–476, 2022.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[PP18] Hanieh Poostchi and Massimo Piccardi. Cluster labeling by word embeddings and wordnet's hypernymy. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 66–70, 2018.

[PW25] Andrew Piper and Sophie Wu. Evaluating large language models for narrative topic labeling. In *Proceedings of the 5th international conference on natural language processing for digital humanities*, pages 281–291, 2025.

[Qwe25] Qwen Team. Qwen3-VL: Official repository. `https://github.com/QwenLM/Qwen3-VL`, 2025. GitHub repository. Accessed: 2026-04-26.

[RES26] Karina Rodriguez Echavarria and Myrsini Samaroudi. Expanding perspectives to improve access to visual archives through multimodal image enrichment. *J. Comput. Cult. Herit.*, 19(1), January 2026.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[Scu10] D. Sculley. Web-scale K-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178, 2010.

[SdRRSB13] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[SGR+19] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. Anchorviz: Facilitating semantic data exploration and concept discovery for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 10(1):1–38, 2019.

[Shn20a] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[Shn20b] Ben Shneiderman. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020.

[SNMM18] Ehsan Sherkat, Seyednaser Nourashrafeddin, Evangelos E. Milios, and Rosane Minghim. Interactive document clustering revisited: A visual

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

analytics approach. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 281–292, New York, NY, USA, 2018. Association for Computing Machinery.

[SPG14] Charles D. Stolper, Adam Perer, and David Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, December 2014.

[TC06] J.J. Thomas and K.A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[Wal23] Manuela Waldner. Joint human-machine data exploration, 2023. FWF Project P36453.

[WKKB15] Michael J. Wilber, Iljung S. Kwak, David Kriegman, and Serge Belongie. Learning concept embeddings with combined human-machine expertise, 2015.

[WW16] Xiaojun Wan and Tianming Wang. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305, 2016.

[YFH+06] Jing Yang, Jianping Fan, Daniel Hubball, Yuli Gao, Hangzai Luo, William Ribarsky, and Matthew Ward. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, pages 191–198, 2006.

[ZA25] Bingchen Zhao and Oisin Mac Aodha. Interpretable text-guided image clustering via iterative search, 2025.

[ZHJL24] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024.

[ZJP15] Bolei Zhou, Vignesh Jagadeesh, and Robinson Piramuthu. Conceptlearner: Discovering visual concepts from weakly labeled image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[ZLE14] Jun-Yan Zhu, Yong Jae Lee, and Alexei A. Efros. Averageexplorer: interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics*, 33(4):1–11, july 2014.

[ZXSR22] Zhenge Zhao, Panpan Xu, Carlos Scheidegger, and Liu Ren. Human-in-the-loop extraction of interpretable concepts in deep learning models.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

IEEE Transactions on Visualization and Computer Graphics, 28(1):780–790, 2022.

[ZYT+25] Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts, 2025.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.