Organizing Unstructured Image Collections using Natural Language

Aude De Fornel, Hugo De Gieter, Romain Donné, Camille Ishac, Yael Juarez-Martinez

Institut Polytechnique de Paris, Télécom Paris

aude.defornel@telecom-paris.fr, hugo.degieter@telecom-paris.fr,
romain.donne@telecom-paris.fr, camille.ishac@telecom-paris.fr,
yjuarez-24@telecom-paris.fr

1 Introduction

In the age of visual information overload, digital platforms are increasingly challenged by the need to manage and structure large volumes of user-generated images. As the scale and diversity of visual data continue to grow, effective organization of such collections has become a critical concern, with broad implications for content retrieval, semantic search, trend analysis, and more.

Traditional image classification approaches typically rely on predefined categories or manual intervention to define sorting criteria. However, these methods quickly reach their limits when facing emerging themes, heterogeneous datasets, or semantic structures that do not align with existing taxonomies. In such cases, the rigidity of predefined labels limits both scalability and adaptability.

In this context, the work of Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci introduces an innovative approach: **X-Cluster**. Their method enables the discovery of semantic grouping criteria within large image collections without requiring human supervision. By leveraging the reasoning capabilities of multimodal large language models (MLLMs), X-Cluster automatically identifies thematic dimensions and clusters images accordingly, using natural language representations.

This report aims to reproduce the X-Cluster framework, assess its experimental robustness, and critically analyze the methodological design choices behind it. The work is conducted in the context of the ML Reproducibility Challenge, an initiative that promotes transparency, reproducibility, and critical evaluation in the machine learning community.

2 Scientific Background

2.1 From Classical to Multiple Semantic Clustering

Clustering is a foundational task in unsupervised learning, traditionally used to group data points based on a predefined similarity metric. Classical clustering methods, such as K-Means or DBSCAN, produce a single partition of the data space, assuming that there exists one "correct" way to segment the dataset. While effective in controlled settings, this assumption breaks down when applied to complex, real-world data, especially in the visual domain where a single image may contain multiple semantic dimensions.

To address this, Multiple Clustering (MC) approaches aim to uncover several meaningful partitions, each reflecting a distinct facet of the data. For example, a collection of food images might be clustered by cuisine, dietary category, or time of consumption. However, identifying these multiple,

interpretable groupings is far from trivial. Most existing methods require manual specification of clustering criteria, domain knowledge, or auxiliary supervision, which limits scalability and adaptability to new domains.

Moreover, evaluating multiple clustering results is inherently challenging, as there is rarely a unique ground truth to compare against. Ensuring that discovered partitions are both diverse and semantically coherent remains an open problem.

2.2 The Role of Multimodal Large Language Models (MLLMs)

Recent advances in Multimodal Large Language Models (MLLMs)—models capable of jointly reasoning over visual and textual inputs—offer new possibilities for semantic understanding in vision tasks. By associating images with rich natural language descriptions, MLLMs provide a bridge between low-level visual features and high-level conceptual knowledge.

This capability is central to the X-Cluster framework, which leverages MLLMs to:

- Propose clustering criteria directly expressed in natural language (e.g., "type of activity", "mood", "location");
- Automatically assign images to semantically meaningful clusters according to these criteria;
- Support open-ended exploration of large image datasets without relying on predefined labels.

Crucially, this is achieved in a fully unsupervised manner, making X-Cluster particularly suitable for organizing unstructured image collections at scale. The method departs from prior work by eliminating the need for predefined categories or manual tagging, shifting the clustering paradigm toward language-driven, interpretable organization of visual content.

3 Presentation of the X-Cluster Approach

The core task addressed by X-Cluster is **Open-ended Semantic Multiple Clustering** (OpenSMC). Unlike traditional clustering, which produces a single partition of data according to a predefined criterion, OpenSMC aims to uncover multiple, diverse, and semantically meaningful clusterings from unstructured image collections. Each clustering corresponds to a latent organizational criterion such as mood, setting, or object type, and is expressed in natural language without requiring human supervision or fixed labels.

Our implementation focuses specifically on the caption-based variant of X-Cluster, which consistently yields the most interpretable and coherent results. The system architecture is organized into two main components:

1. **Criteria Proposer:** Uses a multimodal large language model (MLLM) to analyze a subset of image captions and suggest possible clustering axes in the form of natural language criteria.
2. **Semantic Grouper:** Re-captions the images with a focus on each proposed criterion and then assigns them to semantically coherent clusters at three levels of granularity: coarse, middle, and fine.

Our pipeline is handled by dedicated Python modules. We use pre-trained vision-language models hosted on a local Ollama server, which we query using the `ollama.chat` function to generate captions, propose clustering criteria, and refine semantic groupings based on image data. We also use a second script based on Sentence-BERT for semantic evaluations. The outputs (e.g., CSVs and plots) are saved locally for each experiment, but not necessarily in a structured folder tree.

To evaluate the quality of clustering, we use the Food-4c benchmark derived from the Food-101 dataset. It allows comparison along multiple axes such as cuisine type, nutritional profile, and dish category. For each clustering criterion, we compute both CAcc (Clustering Accuracy, label-agnostic) and SAcc (Semantic Accuracy, embedding-based) to assess performance in terms of visual

consistency and semantic coherence. These metrics, along with cluster visualizations, provide an interpretable evaluation of the discovered organizational structures.

4 Implementation Environment

4.1 Pipeline Structure and Clustering Procedure

The `script_clustering.py` code implements the X-Cluster pipeline, designed to automatically discover semantic grouping criteria within an unlabelled image collection and to cluster these images according to each criterion at different levels of semantic granularity. The process is structured in three main steps.

First, the system extracts potential clustering criteria directly from the visual data. Each image is converted into a base64-encoded format and described using a multimodal language model (MLLM), specifically `LLaVA NeXT Video 7B`, which generates a global textual caption for the image content. These captions are then segmented into subsets to respect the context window limitations of the language model.

Next, a large language model (`LLaMA 3.1 8B`) is prompted to infer ten plausible clustering criteria from the caption sets. These candidates are refined through an additional step to remove redundancy and improve clarity. Once the semantic criteria are finalized, the pipeline enters the clustering stage. For each discovered criterion, the system generates criterion-specific descriptions of every image. These are used to assign initial cluster labels, which are then structured into a three-level hierarchy—coarse, intermediate, and fine—generated by the LLM.

Each image is thus assigned to a cluster at every granularity level for each criterion. The results are exported to a structured CSV file containing, for each image, its general caption and its cluster labels across all semantic axes.

4.2 Dependencies and Models Used

The pipeline relies on:

- **Ollama framework** for local interaction with both LLM and MLLM models.
- **Pillow** for image preprocessing.
- Standard Python libraries: `base64`, `json`, `csv`, `os`.
- Two pretrained models: `ManishThota/llava_next_video` for visual captioning, and `llama3.1:8b` for reasoning and clustering.

4.3 Dataset Download and Processing

We use the Food-101 dataset, a public culinary image collection with 101 categories of dishes, as testbed. Since it contains only raw images, we follow the original unsupervised setup. The main script loads data from `./data/food-101/images`, selects a random subset of 100 images, executes the captioning and clustering phases, and outputs a CSV file named `xcluster_food101_results.csv`.

4.4 Challenges and Solutions

GPU limitations made the pipeline resource-intensive. We mitigated this by:

- Reducing image count to 100,
- Caching intermediate results,
- Adjusting subset size to 400.

criterion	cluster_names
color_scheme	Vibrant And Colorful; Warm And Earthy Tones; Vibrant With Monochromatic Elements; Monochromatic Color Scheme; Earthy And Natural; Unknown; Vibrant Color Scheme
composition_style	Intricate; Unknown; Visually Appealing; Simple
contextual_setting	Formal Indoor Restaurant; Home Kitchen Setting; Casual Dining Establishment; Informal Restaurant Setting; Fine Dining Restaurant; Indoor Dining Restaurant; Restaurant-Like Home Cooking; Upscale Dining Restaurant; Fast-Food Restaurant Setting
cooking_method	Baked Or Roasted; Deep Frying With Oil; Grilled; Braised Or Simmered Method; Fried Or Pan-Seared; Combination Of Grilling And Steaming; Mixed Cooking Method; Steamed; Barbecue; Stir-Fry Cooking Method; Dip-Based; Unknown; Baked Sushi
food_category	Entree Or Main Course; Savory Course Or Appetizer; Multiple Course Meal Combination; Desserts; Baked Or Fried Foods; Seafood Main Course Dishes; Mixed Snacks Main Course; Unknown; Japanese Cuisine Sushi Main Courses
food_presentation_aesthetics	Formal And Elegant Presentations; Casual Dining Setting; Unknown; Artistic Arrangement; Elegant Food Presentation; Rustic And Casual Setting
lighting_and_ambiance	Warm And Cozy: Ambiance; Bright: With Warm Tones; Bright: Even Lighting; Dimly Lit: Cozy Atmosphere; With Warm Ambiance; Bright: Airy
meal_occasion	Dessert; Dinner; Unknown; Breakfast-Dessert Hybrid Meal; Appetizer Platter Or Snack Selection; Breakfast Or Light Lunch; Afternoon Tea Meal; Brunch; Dinner Or Casual Dining Occasion

Figure 1: Our discovered criteria

Table 5. **Full class names for Food-4c across the four basic criteria.**

Criterion	Food-4c
Food Type	"apple pie", "baby back ribs", "baklava", "beef carpaccio", "beef tartare", "beet salad", "beignets", "bibimbap", "bread pudding", "breakfast burrito", "bruschetta", "caesar salad", "cannoli", "caprese salad", "carrot cake", "ceviche", "cheesecake", "cheese plate", "chicken curry", "chicken quesadilla", "chicken wings", "chocolate cake", "chocolate mousse", "churros", "clam chowder", "club sandwich", "crab cakes", "creme brulee", "croque madame", "cup cakes", "deviled eggs", "donuts", "dumplings", "edamame", "eggs benedict", "escargots", "falafel", "filet mignon", "fish and chips", "foie gras", "french fries", "french onion soup", "french toast", "fried calamari", "fried rice", "frozen yogurt", "garlic bread", "gnocchi", "greek salad", "grilled cheese sandwich", "grilled salmon", "guacamole", "gyoza", "hamburger", "hot and sour soup", "hot dog", "huevos rancheros", "hummus", "ice cream", "lasagna", "lobster bisque", "lobster roll sandwich", "macaroni and cheese", "macarons", "miso soup", "mussels", "nachos", "omelette", "onion rings", "oysters", "pad thai", "paella", "pancakes", "panna cotta", "peking duck", "pho", "pizza", "pork chop", "poutine", "prime rib", "pulled pork sandwich", "ramen", "ravioli", "red velvet cake", "risotto", "samosa", "sashimi", "scallops", "seaweed salad", "shrimp and grits", "spaghetti bolognese", "spaghetti carbonara", "spring rolls", "steak", "strawberry shortcake", "sushi", "tacos", "takoyaki", "tiramisu", "tuna tartare", "waffles"
Cuisine	"japanese", "indian", "american", "greek", "spanish", "mexican", "italian", "vietnamese", "canadian", "korean", "chinese", "middle eastern", "french", "thai", "general"
Course	"appetizer", "main course", "side dish", "dessert", "breakfast"
Diet	"omnivore", "vegan", "vegetarian", "gluten free"

Figure 2: OpenSMC criteria from the original paper

Due to the absence of access to GPT-4 Vision (used in the paper), we relied on lighter models, which may have impacted clustering quality. We manually reviewed cluster outputs using CSV and image plots to verify coherence.

5 Experiments

5.1 Results on Food-4c

Our experimentation is based on the Food-101 dataset, which we used to reproduce the Food-4c benchmark introduced in the original X-Cluster paper. Due to hardware constraints, in particular the inference time required by Ollama when using multimodal and large language models locally, we limited our processing to a random subset of 100 images.

Despite this reduced scale, the full X-Cluster pipeline was executed end-to-end, resulting in the automatic extraction of eight clustering criteria in natural language: *Color Scheme*, *Composition Style*, *Contextual Setting*, *Cooking Method*, *Food Category*, *Food Presentation Aesthetics*, *Lighting and Ambiance*, and *Meal Occasion*. For each of these criteria, the system generated semantic clusters at an intermediate level of granularity, such as "Vibrant and Colorful" under *Color Scheme* or "Simple" under *Composition Style*.

criterion	cacc	sacc	n_clusters
Color Scheme	0.11	0.5512594950944185	7
Composition Style	0.09	0.5115245262781779	4
Contextual Setting	0.13	0.549360984582454	9
Cooking Method	0.18	0.5694153807684779	13
Food Category	0.16	0.5690217928215862	9
Food Presentation Aesthetics	0.12	0.5307113295197488	6
Lighting And Ambiance	0.11	0.5220342483520508	6
Meal Occasion	0.14	0.5722486861422658	9

Figure 3: Metric results on our clustering criteria

5.2 Qualitative Analysis of Generated Clusters

To evaluate the quality and coherence of the generated groupings, we developed a dedicated suite of analysis and evaluation scripts. The `cluster_utils.py` module handles post-processing: it cleans and normalizes cluster names, merges near-duplicate labels, and reshapes data for analysis. The `food101_cluster_analysis.py` script runs the evaluation pipeline, applying cleanup, showing visual clusters, and computing two metrics from the original paper: **CAcc** (Clustering Accuracy, label-agnostic) and **SAcc** (Semantic Accuracy, based on caption similarity).

These are implemented in `metrics.py`, and visualization support is provided by `visualization.py`. These tools rely on standard libraries like `pandas`, `numpy`, `scikit-learn`, `scipy`, and `matplotlib`. For caption embeddings, we use `sentence-transformers` (model: `all-MiniLM-L6-v2`).

The clusters show semantic structure despite no supervision. While CAcc scores are low (0.09–0.18), this is expected as we do not aim to reproduce Food-101 labels. SAcc scores between 0.51–0.57 show reasonable coherence. Some inconsistencies (e.g., "Vibrant Color Scheme" vs. "Vibrant and Colorful") suggest improved post-processing is needed.

5.3 Robustness to Noise and Perturbations

Qualitative evaluation confirms X-Cluster's strength in discovering semantic grouping dimensions—both visual and abstract. However, weaknesses include sensitivity to class imbalance (some clusters have few images) and poor image quality (which affects captions). Improvements could include confidence filtering, stronger models (e.g., GPT-4 Vision), or better label normalization.

Despite constraints, our implementation replicates the structure and functionality of the original pipeline, demonstrating feasibility and value in unsupervised semantic clustering.

5.4 Comparison with Original Paper Results

Direct comparison is limited because:

- We used the same MLLM (LLaVA-NeXT-7B), but not the same LLM (we used LLaMA via Ollama instead of GPT-4 or Claude).
- Our evaluation used only 100 images, vs. thousands in the original benchmark.

Still, we reproduced the pipeline's modular logic and qualitative behavior, confirming its soundness at small scale.

6 Critical Analysis

6.1 Theoretical and Practical Contributions

X-Cluster advances the field at the intersection of vision, clustering, and NLP. It automates semantic discovery in image corpora and:

- Generates human-interpretable criteria in natural language;
- Operates without labels or supervision;
- Produces multiple orthogonal partitions.

6.2 Limitations

Main limitations include:

1. **Scalability:** Slow and memory-intensive. One run on 100 images took over an hour.
2. **LLM Bias:** Criteria can reflect biases of the language model.
3. **Instability:** Captions can vary across runs, leading to inconsistent clusters.
4. **Naming Noise:** Redundant or vague cluster labels require manual correction.

6.3 Suggestions for Improvement

To mitigate issues:

- Use semantic embeddings to merge similar clusters;
- Add a human validation step;
- Upgrade to stronger models (e.g., Claude 3 or GPT-4 Vision);
- Add confidence-based pruning.

6.4 Relevance of Generated Criteria

Some criteria were concrete (e.g., Cooking Method), others subjective (e.g., Ambiance). Despite noise, coherence was decent (SAcc $\approx$ 0.55–0.57), showing that MLLM-guided clustering is viable.

6.5 Connection to Course Concepts

Our work involved:

- OpenSMC (multiple clusterings of one dataset);
- Sentence-BERT embeddings to compute semantic distance;
- Vision-language synergy through MLLMs.

7 Conclusion

This project was part of the ML Reproducibility Challenge. We reproduced X-Cluster using caption-based input, validated its outputs qualitatively and quantitatively, and demonstrated its ability to semantically cluster image datasets using only natural language.

Though limited by hardware and model availability, our reproduction showed expressive, meaningful clustering. X-Cluster opens interesting avenues for organizing visual data at scale, combining concepts from vision, language, and unsupervised learning.

Appendix I — Links

- https://oatmealliu.github.io/opensmc.html
- https://giannifranchi.github.io/
- https://github.com/OatmealLiu/OpenSMC
- Our repo: https://github.com/audedef/NLP_image_clustering

Appendix II — Cluster Visualizations

Figure 4: Clusters for the criterion: Color Scheme

Figure 5: Clusters for the criterion: Composition Style

Figure 6: Clusters for the criterion: Contextual Setting

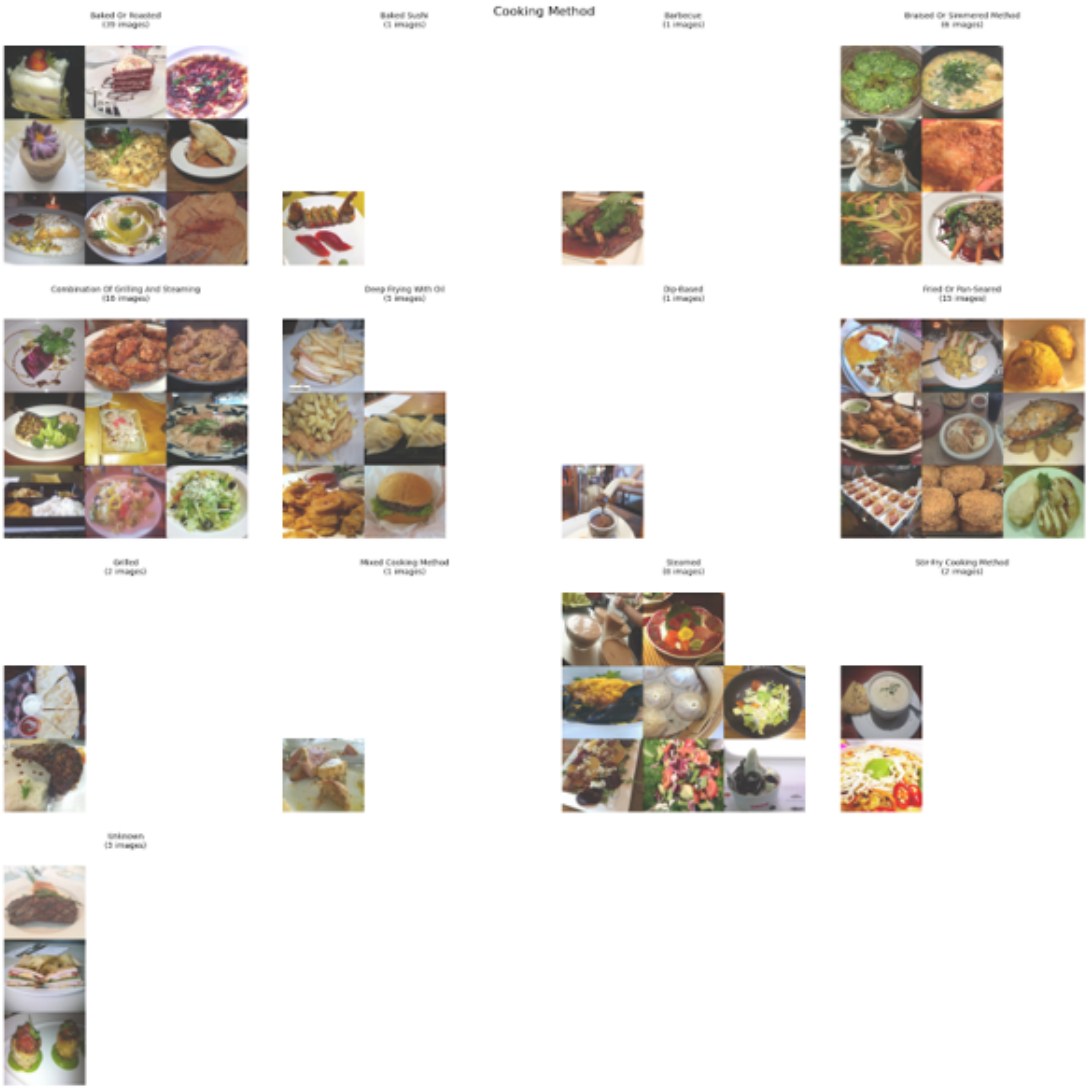

Figure 7: Clusters for the criterion: Cooking Method

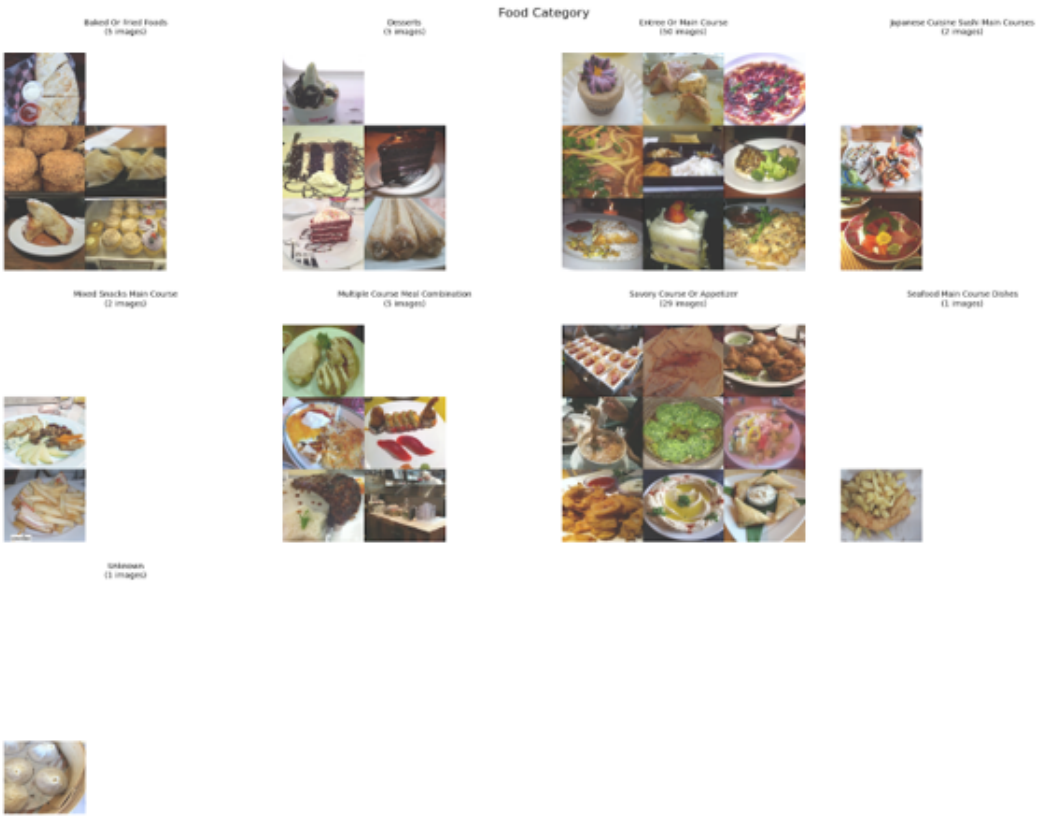

Figure 8: Clusters for the criterion: Food Category

Figure 9: Clusters for the criterion: Food Presentation Aesthetics

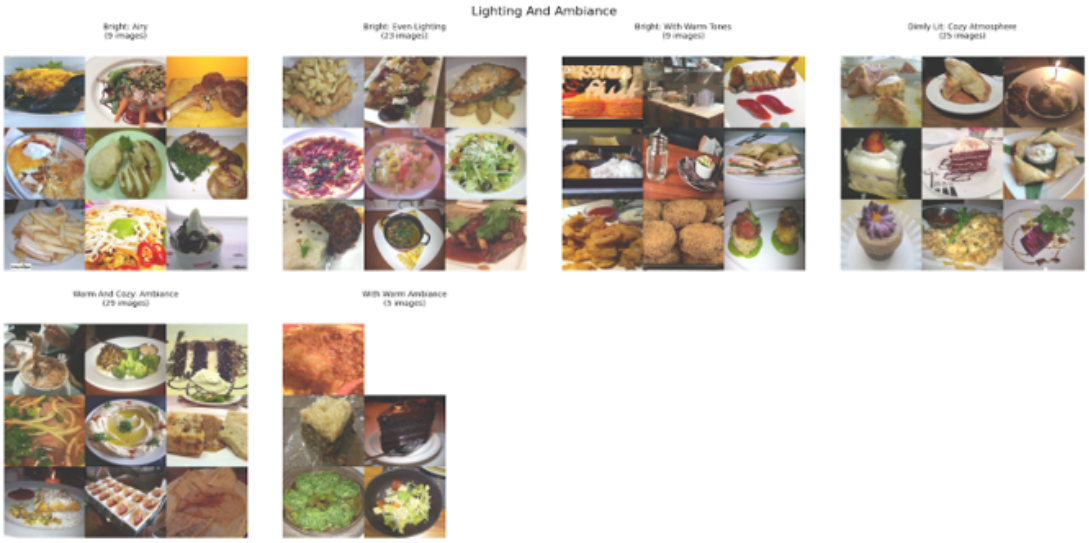

Figure 10: Clusters for the criterion: Lighting and Ambiance

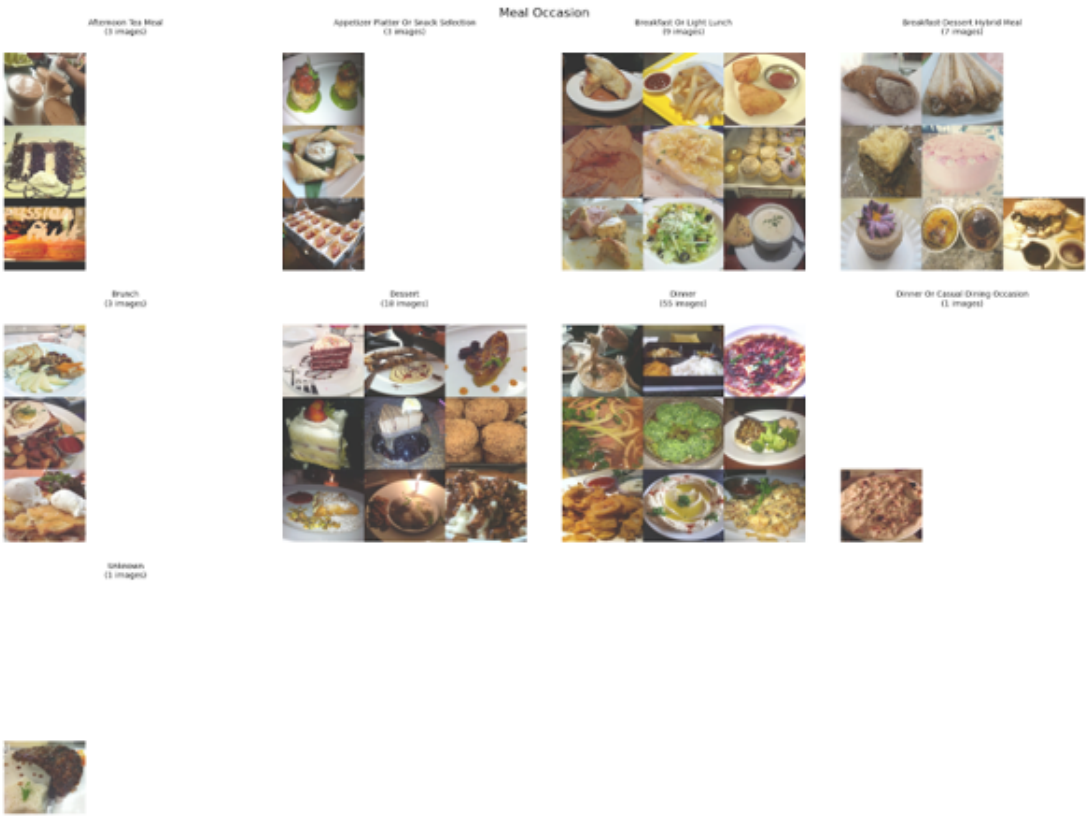

Figure 11: Clusters for the criterion: Meal Occasion