Organizing Unstructured Image Collections using Natural Language

Mingxuan Liu[1] Zhun Zhong[4†] Jun Li[6] Gianni Franchi[3] Subhankar Roy[5] Elisa Ricci[1,2]
[1] University of Trento [2] Fondazione Bruno Kessler [3] ENSTA Paris, Institut Polytechnique de Paris
[4] Hefei University of Technology [5] University of Bergamo [6] Technical University of Munich

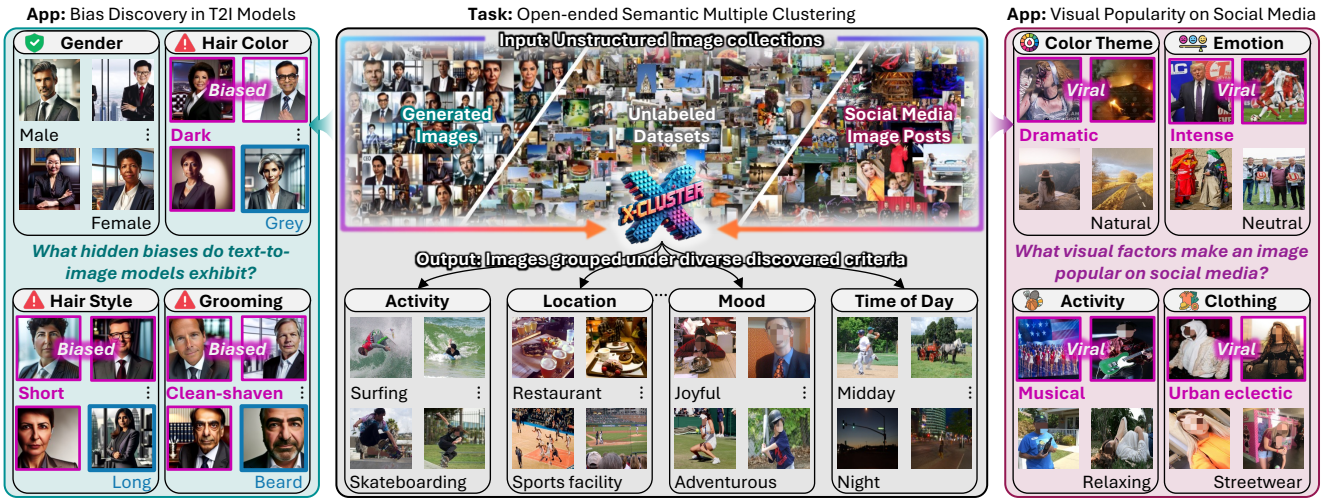

Figure 1. **Organizing unstructured image collections with $\mathcal{X}$-Cluster.** Mid: We propose $\mathcal{X}$-Cluster, which takes any unstructured image collection as input, automatically discovers multiple criteria (*e.g.*, `Activity` or `Location`) that meaningfully group the data, and outputs images organized into semantic clusters under each criterion, *without* any prior knowledge. We demonstrate $\mathcal{X}$-Cluster as a versatile tool for real-world analysis: *i)* Left - uncovers surprising *novel biases* (*e.g.*, "Hair color" or "Hair style") when applied to text-to-image (T2I) model outputs, and *ii)* Right - reveals *visual factors* (*e.g.*, "Dramatic" color or "Intensive" emotion) that drive social media posts virality.

Abstract

*In this work, we introduce and study the novel task of Open-ended Semantic Multiple Clustering (OpenSMC). Given a large, unstructured image collection, the goal is to automatically discover several, diverse semantic clustering criteria (*e.g.*, `Activity` or `Location`) from the images, and subsequently organize them according to the discovered criteria, without requiring any human input. Our framework,* ***$\mathcal{X}$-Cluster****: e****X****ploratory* ***Cluster****ing, treats text as a reasoning proxy: it concurrently scans the entire image collection, proposes candidate criteria in natural language, and groups images into meaningful clusters per criterion. This radically differs from previous works, which either assume predefined clustering criteria or fixed cluster counts. To evaluate $\mathcal{X}$-Cluster, we create two new benchmarks, COCO-4C and Food-4C, each annotated with four distinct grouping criteria and corresponding cluster labels. Experiments show that $\mathcal{X}$-Cluster can effectively reveal meaningful partitions on several datasets. Finally, we use $\mathcal{X}$-Cluster to achieve various real-world applications, including uncovering hidden biases in text-to-image (T2I) generative models and analyzing image virality on social media. Project page: https://oatmealliu.github.io/xcluster.html*

1. Introduction

When organizing a large collection of unlabelled images, a natural question arises: *how should we group them?* One could imagine many possible criteria, *i.e.*, based on `Activity`, `Location`, or even `Color`. Yet, it is often unclear which criterion, if any, best describes the dataset, or whether multiple valid grouping principles coexist. As a result, the seemingly simple task of *clustering images* becomes challenging, as it is influenced by both the visual appearance of the data and their underlying semantics. However, tackling this open-ended unsupervised task of *automatically uncovering diverse and interpretable substructures within large image collections* is pivotal for many applications, such as social media recommendation [13] and dataset auditing [4].

†Corresponding author.

Existing clustering approaches still heavily rely on an iterative, human-in-the-loop interpretation and refinement process. Typically, we begin by setting a few hyperparameters (*e.g.*, the number of grouping criteria or clusters) for Deep Clustering (DC) methods [11, 90] to get a single partition, or for Multiple Clustering (MC) methods [107] to produce several partitions showing different views of the data. We then inspect sample images from each cluster, hoping they correspond to meaningful categories (*e.g.*, "Surfing" or "Skateboarding") and, ideally, that all clusters follow a coherent criterion (*e.g.*, `Activity`). When such patterns fail to emerge, we tweak the hyperparameters and try again until the clusters finally make sense to us. This labor-intensive trial-and-error loop exists because *i)* the resulting clusters are not directly interpretable, being represented only as index assignments. *ii)* both DC and MC methods converge to solutions shaped by model inductive biases and hyperparameter settings, rather than the data's intrinsic semantics.

To enhance controllability and interpretability, recent studies have introduced Text-Conditioned Multiple Clustering (TCMC) [105, 106]. TCMC approaches employ Multimodal Large Language Models (MLLMs) [49, 70] to generate semantic clusters based on user-defined criteria and assign images accordingly, producing human-understandable cluster labels. However, these approaches assume that users already know meaningful ways to organize the dataset. As datasets grow in size and complexity, defining such criteria becomes increasingly unrealistic. Moreover, by relying on static preset criteria, this paradigm may overlook previously unknown grouping dimensions that organically emerge from ever-evolving data.

In this paper, we introduce the task of *Open-ended Semantic Multiple Clustering* (OpenSMC), the *first* task that aims to automatically generate *open-ended* and interpretable groupings of large, unstructured image collections *without* any human priors. Specifically, the goal of OpenSMC is to *discover* clustering criteria directly from the data and uncover their corresponding semantic clusters to organize images accordingly. This task is particularly challenging because *i)* it requires *concurrent* reasoning over all images to identify valid clustering criteria, and *ii)* it assumes *no* access to user knowledge about either the clustering criteria or the number of clusters. Tab. 1 summarizes the key differences between OpenSMC and other paradigms.

To address OpenSMC, we make the following contributions. First, we introduce **$\mathcal{X}$-Cluster**: e**X**ploratory **Cluster**ing, a novel training-free two-stage framework powered by MLLMs [43, 44, 49] and LLMs [61]. $\mathcal{X}$-Cluster consists of two consecutive modules: the Criteria Proposer and the Semantic Grouper. The Criteria Proposer employs a LLM to holistically reason over the entire image collection through textual representations to discover potential clustering criteria. For each discovered criterion, the Semantic Grouper then organizes images into distinct semantic substructures based on their criterion-related visual content. As shown in Fig. 1(mid), our $\mathcal{X}$-Cluster automatically discovers clustering criteria (*e.g.*, `Activity`, `Location`) and uncovers their corresponding semantic clusters (*e.g.*, "Surfing", "Skateboarding" under `Activity`), all expressed in human-interpretable natural language.

Table 1. **Comparison of different clustering paradigms.** Unlike DC, MC, and TCMC settings, the proposed OpenSMC task does not assume any prior knowledge and offers interpretable results.

		DC	MC	TCMC	OpenSMC
Prior	Knowledge # Criteria	✗	✓	✗	✗
	Text Criteria	✗	✗	✓	✗
	Knowledge # Clusters	✓	✓	✓	✗
Output	Multiple Clustering	✗	✓	✓	✓
	Interpretable	✗	✗	✓	✓
	Open-ended	✗	✗	✗	✓

As our second contribution, we introduce two realistic and large-scale benchmarks, COCO-4c and Food-4c, each annotated with ground-truth data for up to *four* clustering criteria. Using them, we comprehensively evaluate the effectiveness of our method in both criteria discovery and semantic grouping. As our third contribution, we demonstrate the versatility of $\mathcal{X}$-Cluster by applying it across diverse applications. When applied to occupation portrait images generated by text-to-image (T2I) generative models (Fig. 1(left)), it uncovers novel occupational biases, such as DALL·E3 [3] associating CEOs with "dark" and "short" hair, beyond well-known biases (*e.g.*, `Gender`). When applied to social media image posts, $\mathcal{X}$-Cluster finds that images featuring "dramatic" colors, "intense" emotions, or 'urban eclectic" clothing styles tend to attract greater popularity online. These findings show that $\mathcal{X}$-Cluster is a practical tool for understanding large-scale unstructured visual data, enabling the discovery of novel, unexpected patterns.

2. Related Work

Image Clustering. Deep clustering learns visual features and produces a *single* partition of an unlabeled dataset via self-supervision [11, 12, 81]. Multiple clustering extends this idea, seeking *multiple* non-redundant partitions with data augmentations, diversity losses, or subspace methods [34, 62, 78, 104, 107]. Despite steady progress, both paradigms share key limitations: *i)* their results are shaped by model inductive biases and training algorithms, limiting generalization beyond object-centric data and often misaligning with user intent or data semantics; and *ii)* clusters are produced as numeric indices rather than human-readable names. In contrast, $\mathcal{X}$-Cluster derives both meaningful criteria and cluster names directly from unlabeled data.

Text Criterion conditioned Multiple Clustering. TCMC lets users steer clustering by specifying the grouping crite-

ria. Learning-based approaches such as MMaP [105] and MSub [106] first use GPT-4 [1] to generate reference words (*e.g.*, fruits colors like "Red" or "Green") conditioned on the user-provided criterion (*e.g.*, *Color*-based fruits clustering). They then optimize learnable image embeddings by aligning with these criterion conditioned reference words. Training-free methods instead translate images into text. IC|TC [42] first captions each image with LLaVA [50] conditioned on the user's criterion, then uses GPT-4 to refine the captions and assign cluster names for a user-specified number of clusters. SSD-LLM [58] strengthens IC|TC by augmenting the prompt with the dataset's primary object labels. Similar ideas have also been explored in applications such as visual trend discovery [17], bias analysis [21, 27], and robot failure diagnosis [28]. While $\mathcal{X}$-Cluster likewise uses text as its reasoning medium, it fundamentally differs from TCMC in two key aspects: *i)* it *automatically* discovers the grouping criteria rather than relying on a user-supplied one; *ii)* it infers both the number and the names of clusters, requiring *no* user-specified parameters.

Topic Discovery. OpenSMC is also related to *Topic Discovery* [5, 22, 97] in NLP, which identifies latent themes [94, 110, 111] or events [66, 87] from *text corpora*. Our work similarly aims to uncover common themes from large, unstructured data but operates on *images*, which is more challenging since *i)* visual semantics are implicit, unlike text where meaning is explicit, and *ii)* no current vision model can reliably reason over large image sets.

3. Open-ended Semantic Multiple Clustering

Task Definition. Given a collection of unlabeled images $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}$, the goal of *Open-ended Semantic Multiple Clustering* is to build a system, $\mathcal{H}$, that automatically *i)* discovers a set of L grouping criteria $\mathcal{R} = \{R_l\}_{l=1}^{L}$ described in natural language, and *ii)* finds interpretable substructures $\mathcal{O}_l$ for each criterion by uncovering semantically meaningful clusters and assigns images to them. Formally, we define an OpenSMC system as:

$$\mathcal{H} : \mathcal{D} \mapsto \left\{ \mathcal{O}_l = \left\{ \mathcal{C}_k^l = \left(s_k^l, \mathcal{D}_k^l\right) \right\}_{k=1}^{K_l} \middle| \, R_l \right\}_{l=1}^{L},$$

where each cluster $\mathcal{C}_k^l$ is characterized by a semantic name s_k^l and a subset of images $\mathcal{D}_k^l \subset \mathcal{D}$ that share the same semantics. A criterion R_l refers to a *theme* for grouping images, such that all the clusters under R_l should align with the theme. As shown in Fig. 1(top), if R_l = `Activity`, each cluster under this criterion should collect images $\mathcal{D}_k^l$ that depict an activity, such as s_k^l = "Surfing". If R_l = `Location`, the same dataset should be organized into clusters like "Restaurant", "Sports facility", and so on.

An OpenSMC system should find $\mathcal{R}$ and $\mathcal{O}_l$ automatically, both expressed in natural language. In contrast to TCMC setting, where criteria $\mathcal{R}$ and the corresponding cluster counts K are *preset* by human operators.

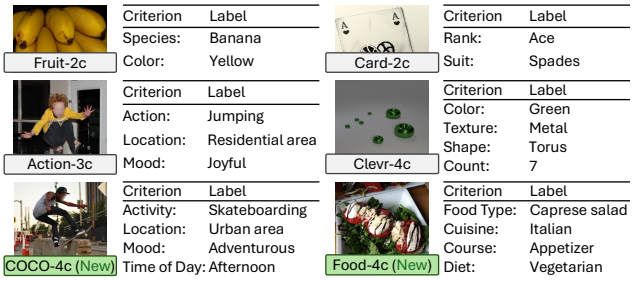

Figure 2. **OpenSMC benchmarks.** We introduce two **new** challenging benchmarks: **COCO-4c** and **Food-4c**. We show all annotated criteria and the corresponding labels for the example images.

Benchmark. Evaluating OpenSMC methods requires benchmarks that can be partitioned under multiple criteria. Currently, only a few benchmarks [107] support the evaluation of OpenSMC methods: Fruit-2c [64], Card-2c [36], Action-3c [42], and Clevr-4c [92]. As shown in Fig. 2, these benchmarks are limited by their object-centric nature with simple backgrounds (*e.g.* Fruit-2c), an insufficient number of criteria (*e.g.* up to three in Action-3c), and a lack of photorealism due to synthetic generation (*e.g.* Clevr-4c).

Given that the data encountered in real-world applications is more complex, we annotate and propose two *new* benchmarks for OpenSMC: Food-4c and COCO-4c. Food-4c is sourced from Food-101 [8], which includes 101 `Food type` (original annotations), along with new annotations for 15 `Cuisine` types, 5 `Courses` types, and 4 `Diet` preferences, totaling *four* clustering criteria. Additionally, we introduced COCO-4c using images from COCO-val [47], where we annotated *four* criteria with varying number of clusters: 64 `Activity`, 19 `Location`, 20 `Mood`, and 6 `Time of day`. Examples of these newly constructed benchmarks are shown in Fig. 2. Further details, such as cluster names and the annotation pipeline, are provided in Supp. D.

4. Method

The goal of an OpenSMC system is to first discover meaningful grouping criteria (or themes) from an unstructured image collection by finding commonalities among the images, and then group them into semantic clusters as per the discovered criteria. This is particularly a challenging task because it requires reasoning over the visual content of all images *simultaneously*. To address OpenSMC, we diverge from representation learning-based MC approaches [105, 106], as no existing model can yet encode large image sets and reason over them reliably. Instead, we convert the visual content of all images into text and use *text descriptions as a proxy* to discover the grouping criteria and the semantic substructures.

System Overview: As illustrated in Fig. 3, our proposed

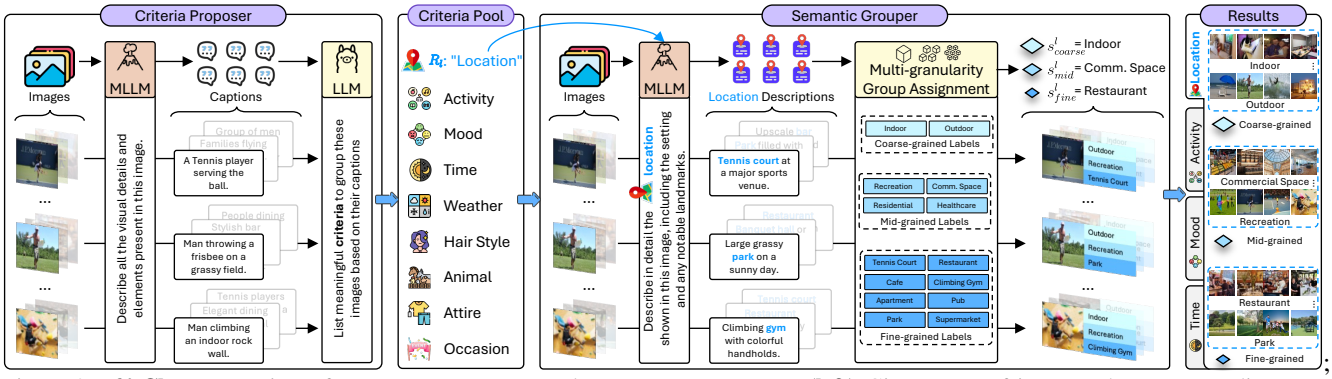

Figure 3. $\mathcal{X}$**-Cluster** consists of a *Criteria Proposer* and a *Semantic Grouper*. **(left)** Given a set of images, the Proposer discovers and outputs a pool of grouping criteria in natural language. **(right)** The Grouper subsequently extracts criterion-specific descriptions from images relevant to each criterion, discovers the underlying semantic clusters, and groups each image at three semantic granularity levels. **Results** shows an example, as how an unstructured image collection can be grouped into clusters of different semantic granularity corresponding to criterion "Location". See Supp. F for implementation and prompt details.

$\mathcal{X}$-Cluster is a two-stage framework that is composed of two modules: *Criteria Proposer* and *Semantic Grouper*. The Criteria Proposer processes the *entire* image set $\mathcal{D}$ to discover diverse common themes among the images and proposes grouping criteria $\mathcal{R}$ in natural language (*e.g.* `Location`). Once the criteria are proposed, the Semantic Grouper uncovers the substructure $\mathcal{O}_l$ of $\mathcal{D}$ by discovering distinct semantic clusters and assigning images to their respective clusters (*e.g.* "Tennis Court"), adhering to each criterion $R_l \in \mathcal{R}$. As the OpenSMC task operates without user priors to guide semantic granularity, we also design our Semantic Grouper to automatically discovers $\mathcal{O}_l$ across multiple granularity levels, from coarse (*e.g.*, "Outdoor") to mid (*e.g.*, "Recreation") to fine (*e.g.*, "Tennis Court"), and organizes the images accordingly. In this work, we explore *three* design choices for both the Proposer and the Grouper. Due to space constraints, only the main variant of $\mathcal{X}$-Cluster is illustrated in Fig. 3, while the illustrations of the alternative variants and additional implementation details, including the exact prompts, are provided in Supp. F. Next, we describe each variant in detail.

4.1. Criteria Proposer

As shown in Fig. 3(left), the Proposer takes as input a set of input images and generates distinct grouping criteria (or Criteria Pool) in natural language. Its core design principle is the ability to *concurrently reason* across a large set of images. Next, we explore three systematic approaches.

Caption-based Proposer (main): To enable reasoning over a large image set for criterion discovery, we first leverage a MLLM [49] to generate a comprehensive caption e_n for each image, converting its visual content into text representations $e_n = \text{MLLM}(\mathbf{x}_n)$. The resulting caption set $\{e_n\}_{n=1}^N$ serves as a rich and holistic semantic proxy for the image collection $\mathcal{D}$. Using these compact textual proxies, we then prompt a LLM [61, 70] to jointly analyze the aggregated visual content and propose multiple valid grouping criteria, denoted as $\widetilde{\mathcal{R}} = \text{LLM}(\{e_n\}_{n=1}^N)$. As an example, the LLM could use different cues such as "Tennis", "grassy field", "rock wall" in the captions (see Fig. 3) and its reasoning capability to discover the criterion "Location", since they are usually associated to a physical location, albeit locations may not explicitly appear in any caption.

Tag-based Proposer (alternative): Instead of using captions as textual proxies for reasoning, we further explore an approach that relies on image tags. Specifically, using the WordNet [63] vocabulary as the candidate tag set, we employ an open-vocabulary tagger (*e.g.*, CLIP [79]) to assign ten tags $t_{i,n}$ to each image as $\{t_{i,n}\}_{i=1}^{10} = \text{Tagger}(\mathbf{x}_n, \text{WordNet})$. These tags act as concise semantic descriptors that summarize the key elements present in each image. We then aggregate all assigned tags and prompt a LLM to analyze them jointly and propose grouping criteria as $\widetilde{\mathcal{R}} = \text{LLM}(\{\{t_{i,n}\}_{i=1}^{10}\}_{n=1}^N)$.

Image-based Proposer (alternative): Lastly, we explore an approach that reasons directly over images rather than their textual proxies. Since no existing model can reliably encode large image sets at once, we adopt a simple workaround: divide $\mathcal{D}$ into 64-image batches, stitch each batch into an 8×8 grid as a *single* composite image, and feed the resulting image grids to a MLLM. The model is prompted to propose grouping criteria from each grid, and we aggregate and deduplicate these subset proposals to obtain the final criteria set $\widetilde{\mathcal{R}}$.

Criteria Refinement: The accumulated criteria in $\widetilde{\mathcal{R}}$ may contain redundant or noisy entries, such as semantically overlapping concepts (*e.g.*, "Outdoor" *vs.* "Open space") or irrelevant ones (*e.g.*, "High resolution"). To clean them, we input all initially proposed criteria into a LLM, prompting it to consolidate similar ones and discard noise. This yields a refined criteria set $\mathcal{R} = \text{LLM}(\widetilde{\mathcal{R}})$, which is then stored in a pool for the subsequent substructure discovery stage.

4.2. Semantic Grouper

Each discovered criterion $R_l \in \mathcal{R}$ serves as a thematic indicator for a distinct semantic substructure $\mathcal{O}_l$ within the image set $\mathcal{D}$. To uncover these substructures, as shown in Fig. 3(right), the Grouper takes $\mathcal{D}$ and each criterion R_l as inputs, discovers cluster names $\{s_k^l\}_{k=1}^{K_l}$, and groups images $\mathcal{D}_k^l$ to their corresponding clusters. As a result, the interpretable substructure $\mathcal{O}_l = \{\mathcal{C}_k^l = (s_k^l, \mathcal{D}_k^l)\}_{k=1}^{K_l}$ emerges for each R_l. The core design of the Grouper focuses on *aligning* semantic substructure discovery with the given partitioning criterion. Like the Proposer, we explore three distinct approaches for the Grouper.

Furthermore, as clusters under a given criterion can be formed at varying semantic granularities based on user preferences, we have designed our Grouper to clusters $\mathcal{D}$ at three levels: coarse, middle, and fine-grained. This allows $\mathcal{X}$-Cluster to provide insights at different granularities. For example, under the `Cuisine` criterion, $\mathcal{X}$-Cluster can organize images at a coarse continental level (*e.g.*, "European" or "Asian"), a middle regional level (*e.g.*, "Mediterranean" or "Southeast Asian"), or a fine national level (*e.g.*, "Italian" or "Thai"). See Supp. F.2 for design details.

Caption-based Grouper (main): Given a target criterion R_l, we prompt the MLLM to generate criterion-specific captions that focus exclusively on the visual content relevant to R_l for each image, as $e_n^l = \text{MLLM}(\mathbf{x}_n, R_l)$. Next, we design a *Multi-granularity Group Assignment (MGA)* module that uses the LLM to group images into clusters across multiple semantic granularity levels through a three-step process: *i) Initial Naming*: The LLM assigns a provisional class name to each caption as $s_n^l = \text{LLM}(e_n^l, R_l)$, producing an initial set of names $\mathcal{S}_{\text{init}}^l$; *ii) Multi-granularity Cluster Refinement*: The LLM refines $\mathcal{S}_{\text{init}}^l$ into three structured granularity levels: $(\mathcal{S}_{\text{coarse}}^l, \mathcal{S}_{\text{mid}}^l, \mathcal{S}_{\text{fine}}^l) = \text{LLM}(\mathcal{S}_{\text{init}}^l, R_l)$, which serve as candidate cluster names; *iii) Final Assignment*: LLM assigns each image $\mathbf{x}_n$ to a cluster by linking its criterion-specific caption to the structured class names at different granularity levels as $(s_{\text{coarse}}^l, s_{\text{mid}}^l, s_{\text{fine}}^l) = \text{LLM}(e_n^l, \mathcal{S}_{\text{coarse}}^l, \mathcal{S}_{\text{mid}}^l, \mathcal{S}_{\text{fine}}^l)$. By aggregating these cluster assignments across $\mathcal{D}$ at different levels, we derive multi-granularity semantic substructures. As we will show in § 5.3, the Caption-based Grouper outperforms other alternatives, making it our main method.

Tag-based Grouper (alternative): Given a target criterion R_l, we prompt the LLM to generate a set of common categories (*e.g.*, "Commercial Space") related to the criterion $\mathcal{S}_{\text{mid}}^l = \text{LLM}(R_l)$ as the mid-grained tags. Following Liu et al. [52], we further query the LLM to infer potential super- and sub-categories (*e.g.*, "Indoor" and "Restaurant") for each mid-grained tag, thereby obtaining the corresponding coarse- and fine-grained tag sets, $\mathcal{S}_{\text{coarse}}^l$ and $\mathcal{S}_{\text{fine}}^l$. Finally, we employ an open-vocabulary image tagger [79] to assign the most relevant tag at each granularity level to each image, $(s_{\text{coarse}}^l, s_{\text{mid}}^l, s_{\text{fine}}^l) = \text{Tagger}(\mathbf{x}_n, \mathcal{S}_{\text{coarse}}^l, \mathcal{S}_{\text{mid}}^l, \mathcal{S}_{\text{fine}}^l)$, yielding multi-granularity substructures after aggregation.

Image-based Grouper (alternative): Given a target criterion R_l, we first prompt a LLM to generate a question q_l tailored to R_l, For *e.g.*, for the criterion `Mood` the generated question is: "What mood is conveyed by this image? Answer with an abstract, common, and specific category name, respectively". We then use q_l to guide a visual question answering (VQA) model [44] in directly inferring semantic cluster names and assignments for each image at different granularity levels as $(s_{\text{coarse}}^l, s_{\text{mid}}^l, s_{\text{fine}}^l) = \text{VQA}(\mathbf{x}_n, q_l)$.

5. Experiments

5.1. Experimental Protocol

Implementation Details: We run with our proposed $\mathcal{X}$-Cluster framework using: *i)* CLIP ViT-L/14 [79] as the Tagger, *ii)* LLaVA-NeXT-7B [49] as the MLLM, *iii)* Llama-3.1-8B [61] as the LLM, and *iv)* BLIP-2 Flan-T5$_{\text{XXL}}$ [44] as the VQA model. For the Image-based Proposer we use LLaVA-NeXT-Interleave-7B [43] as the MLLM due to its strong multi-image reasoning capability. Additionally, we explore a variant of the Image-based Grouper using LLaVA-NeXT-7B as the VQA model. We provide further details of $\mathcal{X}$-Cluster, including the exact prompt designs, in Supp. F.

Evaluation Metric for Criteria Discovery: We use True Positive Rate (TPR) [15] to evaluate the criteria discovery performance of different proposers. Specifically, we compute TPR as $\text{TPR} = \frac{|\mathcal{R} \cap \mathcal{Y}|}{|\mathcal{Y}|}$, measuring to what extent the predicted set covers the ground-truth criteria $\mathcal{Y}$. It is important to note that the number of grouping criteria is subjective and can be as extensive as one's preferences allow (open-ended), making False Positives hard to define. Thus, we use TPR as the primary metric. A higher TPR means better coverage of predicted criteria compared to the ground truth.

Evaluation Metrics for Substructure Uncovering: To assess each criterion-specific substructure uncovered by the Grouper, we evaluate its alignment with the ground-truth substructure along two dimensions: *i) Semantic Consistency:* For each image, we compute the semantic similarity between its assigned cluster name and the ground-truth label under the current criterion using Sentence-BERT. The average similarity across the dataset, reported as Semantic Accuracy (SAcc) [53], measures how well the predicted substructure aligns semantically with the ground truth. *ii) Structural Consistency:* We use clustering accuracy (CAcc) [30, 91] to measure the degree of structural match between the predicted and ground-truth substructures (clusters) using Hungarian matching algorithm [41].

Since the granularity of ground-truth annotations is unknown during OpenSMC evaluation, we select the predicted substructure with the highest CAcc for assessment. Unlike

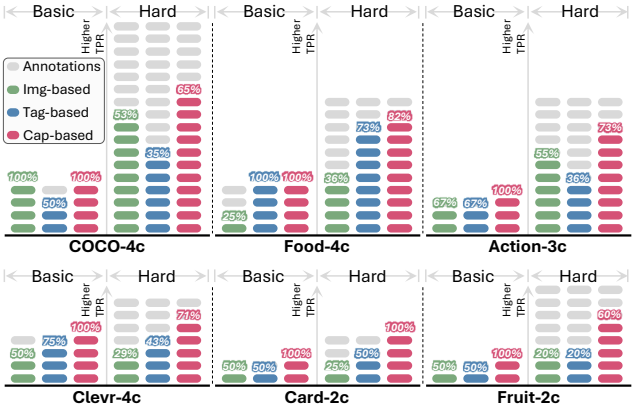

Figure 4. **Comprehensiveness Comparison of Criteria Proposers.** TPR performance of each proposer is evaluated against Basic and Hard ground-truth criteria, and visualized using a Progress Bar Chart. Each block represents one ground-truth criterion, with **Colored** blocks indicating successfully discovered criteria and **Gray** blocks representing undiscovered criteria.

TCMC methods [42, 106] that rely on ground-truth cluster counts for perfect matching, our strategy provides a fair and practical evaluation for open-ended OpenSMC systems.

5.2. Study of the Criteria Proposer

We also evaluate the performance of our design for the Proposer module. To properly assess his effectiveness, we realize that for complex datasets like COCO-4c, four ground-truth criteria may not cover all valid grouping options. Therefore, we expanded the ground-truth criteria for each of the six benchmarks in Sec. 3 using human annotators, resulting in {10, 4, 11, 7, 17, 11} distinct criteria for {Fruit-2c, Card-2c, Action-3c, Clevr-4c, COCO-4c, Food-4c}. We refer to the original per-image annotated criteria set (see Fig. 2) as **Basic** ground truth and the expanded set as **Hard** during evaluation. See Supp. D.2 for annotations.

Which Criteria Proposer Performs the Best? In Fig. 4, we compare different approaches for the Proposers in terms of the comprehensiveness of the discovered criteria using TPR. From Fig. 4, we observe that our caption-based Proposer discovers the most comprehensive criteria, making it the *closest* to the human-annotated set among all methods. It consistently outperforms other variants in both the Basic and Hard sets across all six benchmarks. Its superior performance is particularly evident under the Hard criteria set, where it surpasses the second-best Tag-based Proposer by +32.2% TPR. Intuitively, the Caption-based Proposer works better because captions capture more diverse and nuanced aspects of the image set, which further guides the LLM to comprehensively discover different grouping criteria. Contrarily, the Tag-based Proposer is less effective in complex benchmarks (*e.g.* COCO-4c and Action-3c) since tags provide less contextual and descriptive information. Similarly, the Image-based Proposer is subpar in terms of performance

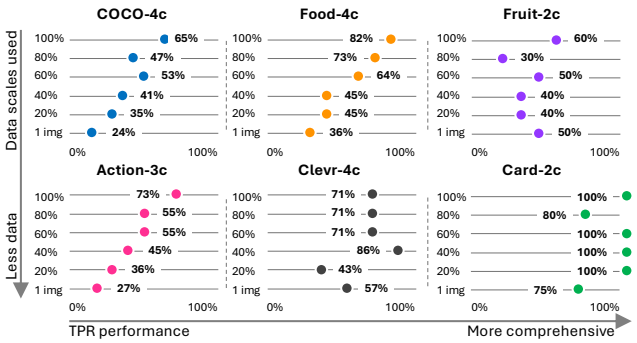

Figure 5. **Impact of Image Quantity on Criteria Discovery.** We evaluate the TPR performance of the Caption-based Proposer at different image scales against the Hard ground-truth criteria set.

since it is limited to reasoning over a small subset of images and loses visual details when combining images into a grid.

Impact of Image Quantity on Criteria Discovery: Fig. 5 shows the TPR performance of the Caption-based Proposer across different image scales. Interestingly, in *object-centric* benchmarks like Card-2c and Clevr-4c, satisfactory performance is achieved with just a *few* images. In fact, even a *single* image often suffices for reasonable criteria discovery, as object-centric datasets tend to have uniform structures, *i.e.*, seeing one playing card is enough to suggest criteria like `Suit`. However, this does not hold for more complex datasets like COCO-4c, Food-4c, and Action-3c, which feature diverse and realistic scenarios. Here, reducing the number of images leads to a clear drop in TPR performance, as capturing intricate and varied thematic criteria requires a larger image set. Since $\mathcal{X}$-Cluster operates *without* prior knowledge of the dataset, we use the *entire* dataset by default to ensure comprehensive discovery.

5.3. Study of the Semantic Grouper

Which Semantic Grouper Performs the Best? In Fig. 6, we evaluate different design choices for the Grouper using CAcc and SAcc for each criterion, determining the best performer based on Harmonic Mean (HM). To contextualize performance, we establish an oracle using CLIP ViT-L/14 in a zero-shot classification setup, where grouping criteria, cluster names, and the number of clusters are all *known*. We also use KMeans with ground-truth cluster numbers and visual features from CLIP-L/14, DINOv1-B/16 [45], and DINOv2-G/14 [71] as CAcc baselines.

From Fig. 6, we observe that the proposed Caption-based Grouper performs best, ranking first in 10 out of 15 tested criteria based on the HM across four benchmarks. It achieves an average CAcc of 59.9%, closely matching the oracle performance of 58.1%, highlighting the effectiveness of our text-driven approach. For SAcc, the Caption-based Grouper achieves an average of 60.5%, surpassing its counterparts, but falling short of the oracle 74.2% which benefits from exact ground-truth class names. This gap is expected

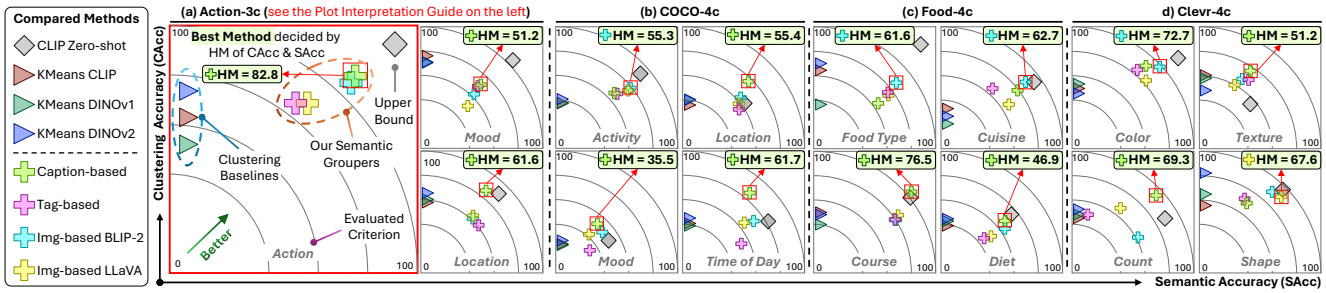

Figure 6. **Comparison of Semantic Groupers.** We report CAcc, SAcc, and their Harmonic Mean (HM) for different Semantic Groupers (✙) on the Basic criteria across four benchmarks. CLIP zero-shot classification (◆) serves as an oracle, while KMeans (▶) with strong visual features is used as a CAcc baseline. The best performer for each criterion, determined by HM, is highlighted in green. Overall, our *Caption-based Grouper* performs best, ranking first in *10 of 15* evaluated criteria. See Supp. H.2 for clustering visualizations.

Table 2. **Comparison with TCMC methods.** For each benchmark, we report the average CAcc (%) and SAcc (%) *across all criteria*. We provide CLIP L/14 zero-shot performance as the pseudo upper-bound reference (UB). **Note**: †-marked methods used the ground-truth criteria and the number of clusters (K_l) as prior input. MMaP and MSub do not build semantic clusters. See expanded results in Supp. H.3.

	COCO-4c		Food-4c		Clevr-4c		Action-3c		Card-2c		Fruit-2c		Avg	
	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc
UB	40.1	60.6	64.1	80.2	56.7	72.5	79.8	82.3	41.4	66.9	69.4	88.3	50.2	64.4
MMaP † [105]	33.9	-	43.8	-	62.8	-	60.6	-	36.9	-	51.0	-	48.2	-
MSub † [106]	36.0	-	47.3	-	**72.2**	-	64.3	-	39.6	-	54.4	-	52.3	-
IC|TC † [42]	48.9	**53.2**	**50.5**	61.7	58.3	36.8	76.4	56.3	**74.8**	81.2	63.3	55.1	**62.0**	57.4
SSD-LLM † [58]	41.6	52.1	47.5	55.5	54.8	37.6	**78.1**	52.9	67.3	76.3	62.0	46.8	58.6	53.6
$\mathcal{X}$-Cluster (Ours)	**51.2**	48.4	48.1	**64.9**	64.9	**54.3**	68.3	**60.6**	73.3	**84.3**	**65.1**	**61.1**	61.8	**62.3**

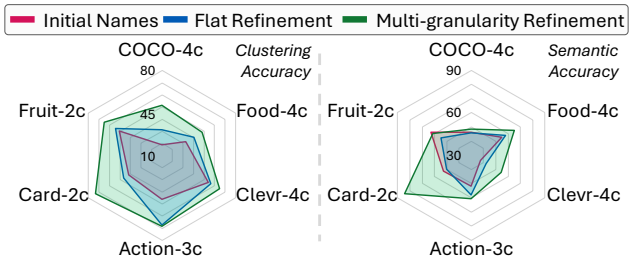

Figure 7. **Ablation study** of multi-granularity refinement.

due to the open nature of the semantic space, *i.e.*, terms like "Joyful", "Happy", and "Cheerful" often describe the same `Mood` but lack full semantic equivalence. The BLIP-2 Image-based Grouper ranks second. Its criterion-specific questions improve labeling accuracy, but per-image predictions can introduce noise in clustering.

Necessity of Multi-Granularity Cluster Refinement: To evaluate the effectiveness of multi-granularity cluster refinement design, we conduct controlled experiments using our Caption-based Grouper with three cluster naming strategies: *i) Initial Names*, where the initially assigned names are used as the final output; *ii) Flat Refinement*, where the LLM refines initial names into a single-level list with uniform granularity; and *iii) Multi-Granularity Refinement*, our proposed approach. As shown in Fig. 7, both refinement methods significantly improve clustering accuracy compared to using noisy initial names, highlighting the importance of granularity-consistent cluster names for revealing substructures. Moreover, our multi-granularity refinement outperforms flat refinement by enabling clustering at different levels of detail, providing greater flexibility in aligning with user-preferred grouping granularity.

5.4. Comparison with TCMC Methods

We first perform some experiments to compare our approach with state of the art TMTC methods: IC|TC [42], SSD-LLM [58], MMaP [105], and MSub [106]. Results are shown in Tab. 2. Unlike our fully automated $\mathcal{X}$-Cluster method, which discovers criteria through the Proposer and requires *no* pre-set cluster counts, *all TCMC methods used ground-truth text criteria and the number of clusters (K_l) as prior input*. The primary goal of this experiment is to evaluate dataset grouping performance. Our approach outperforms MMaP, MSub, and SSD-LLM, while achieving results comparable to IC|TC across six benchmarks. This demonstrates that our framework generates high-quality clusters for OpenSMC *without* requiring users to define criteria or cluster counts. Implementation details of the compared methods are provided in Supp. G.

Further Analysis of $\mathcal{X}$-Cluster is provided in the supplementary material: *i)* Supp. I presents qualitative results; *ii)* Supp. J examines failure cases; *iii)* Supp. L explores how $\mathcal{X}$-Cluster handles invalid (hallucinated) criteria; *iv)* Supp. M investigates model biases; *v)* Supp. N analyzes computational costs; *vi)* Supp. O studies system sensitivity to different MLLMs and LLMs; *vii)* Supp. K further investigates the impact of multi-granularity clustering; and *viii)* Supp. P explores improvements for handling fine-grained criteria.

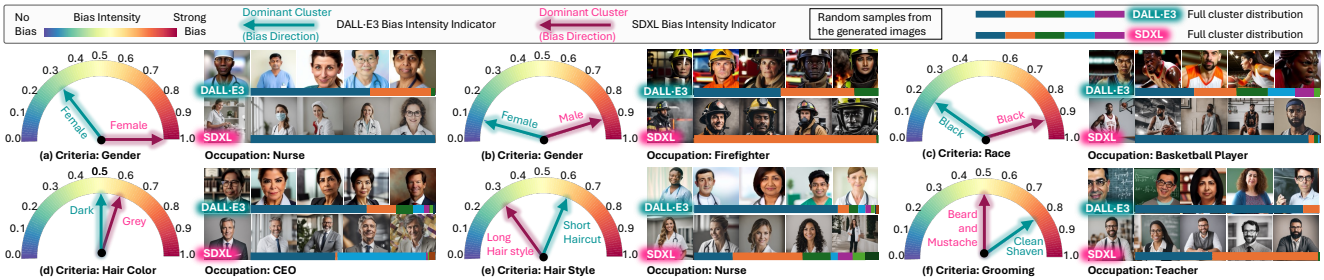

Figure 8. **Bias Discovery in T2I-Generated Images.** Bias intensity, dominant clusters, and example images are shown for few occupations.

6. Applications

We apply $\mathcal{X}$-Cluster to three applications, demonstrating its ability to generate novel, human-interpretable criteria for real-world analysis. Below, we present results for the first two applications, while additional results for the third application on confirming and mitigating gender bias in 162k CelebA [56] images are provided in Supp. Q.3.

6.1. Discovering Biases in T2I Diffusion Models

Do T2I models exhibit biases beyond the widely studied ones, such as gender and racial stereotypes? [65, 67] To investigate this, we selected *nine* occupations (e.g., Nurse, CEO) from prior studies [4, 6] and generated 100 images per occupation using the prompt "A portrait photo of a <OCCUPATION>" with DALL-E3 [3] and SDXL [76], resulting in 1.8k images. Applying $\mathcal{X}$-Cluster, we automatically identified 10 grouping criteria (bias dimensions) and their distributions for each occupation. To quantify bias, we measured the normalized entropy of each distribution [19] as bias intensity and identified the dominant cluster (the largest group) as the potential bias direction. We conducted a user study with 54 participants to validate our findings. $\mathcal{X}$-Cluster's predicted bias intensity closely matched human ratings with an Absolute Mean Error of 0.1396 (0–1 scale) and aligned with human-identified bias directions 72.3% of the time. User study details are provided in Supp. Q.1.

Findings: As shown in Fig. 8, our method identifies both well-known and novel biases in occupational images without relying on predefined categories. For instance, Fig. 8(a–c) reveals strong gender and racial imbalances in SDXL-generated images for roles like Nurse, Firefighter, and Basketball Player, exceeding official statistics [89]. In contrast, DALL·E3 exhibits improved bias mitigation, likely due to its built-in "guardrails" [69]. More notably, Fig. 8(d–f) highlights previously unrecognized bias dimensions. For example, SDXL strongly associates CEOs with "Grey" hair, while DALL·E3 favors "Dark" hair. Additionally, DALL·E3 shows stronger biases in `Hair style` and `Grooming` for occupations like Nurse (Fig. 8(e)) and Teacher (Fig. 8(f)). These findings suggest that while industrial T2I models with guardrails may address well-known biases, they may still overlook emerging or less-discussed ones, underscoring the need for broader bias analysis. For additional findings and experimental details, see Supp. Q.1.

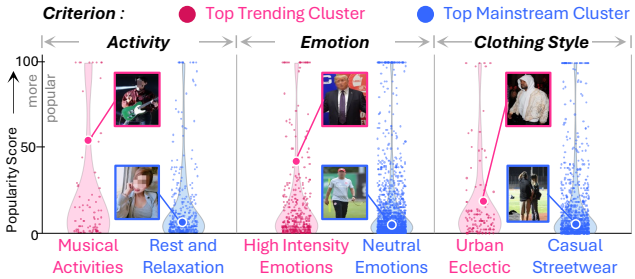

Figure 9. **Social media image popularity analysis**. We show the popularity score distributions for Top Trending (have *highest* average popularity score) and Top Mainstream (contain *most* images) clusters, discovered by $\mathcal{X}$-Cluster across three criteria.

6.2. Analyzing Social Media Image Popularity

What makes a photo popular? To explore this, we apply $\mathcal{X}$-Cluster to 4.1k Flickr photos from the SPID dataset [72], where popularity is measured by image view count. $\mathcal{X}$-Cluster discovered 10 grouping criteria and organizes photos into semantic clusters under each. Using the grouping results, Fig. 9 compares the sample popularity distributions of the Top Trending and the Top Mainstream clusters across three criteria.

Findings: As shown in Fig. 9, combining $\mathcal{X}$-Cluster's grouping with popularity scores provides a direct interpretation of the visual elements that drive trends versus those that define widely uploaded images. Interestingly, we find that trending elements often contrast with mainstream ones, such as "Musical activities" *vs.* "Rest and relaxation" or "High-intensity expressions" *vs.* "Neutral emotion". These results suggest that attention-grabbing visuals stand out due to novelty or intensity, especially in today's short attention span era [24, 59], underscoring $\mathcal{X}$-Cluster as a powerful tool for deep dataset analysis and understanding social behavior. For full findings and additional analysis, see Supp. Q.2.

7. Conclusion

We introduce the OpenSMC task and propose $\mathcal{X}$-Cluster, a system that discovers interpretable grouping criteria and substructures in image collections, effectively extracting valuable insights across six datasets and three applications.

Acknowledgments

This work was supported by the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398). We greatly acknowledge CINECA and the ISCRA initiative for providing high-performance computing resources. We also extend our gratitude to Feng Xue for his valuable suggestions on plot creation. M.L. warmly thanks Margherita Potrich for her unwavering support.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 3, 2, 9, 18, 27

[2] Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. Table-to-text generation and pre-training with tabt5. In *Findings of EMNLP*, 2022. 27

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. 2, 8, 23

[4] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023. 1, 8, 2, 23

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 2003. 3, 2

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016. 8, 2, 23

[7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Koulako Bala Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. 14

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 3, 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[10] Leonardo Bruni, Chiara Francalanci, and Paolo Giacomazzi. The role of multimedia content in determining the virality of social media information. *Information*, 2012. 26

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 2, 27

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2

[13] Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *KDD*, 2024. 1, 24

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[15] Gabriela Csurka, Tyler L Hayes, Diane Larlus, and Riccardo Volpi. What could go wrong? discovering and describing failure modes in computer vision. In *ECCV Workshop*, 2024. 5

[16] DeepLearning.AI. ChatGPT Prompt Engineering for Developers - DeepLearning.AI, 2024. 27

[17] Boyang Deng, Songyou Peng, Kyle Genova, Gordon Wetzstein, Noah Snavely, Leonidas Guibas, and Thomas Funkhouser. Visual chronicles: Using multimodal llms to analyze massive collections of images. In *arXiv:2504.08727*, 2025. 3

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 26

[19] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *CVPR*, 2024. 8, 23

[20] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *CVPR*, 2024. 2

[21] Lisa Dunlap, Joseph E Gonzalez, Trevor Darrell, Fabian Caba Heilbron, Josef Sivic, and Bryan Russell. Discovering divergent representations between text-to-image models. In *ICCV*, 2025. 3

[22] Anton Eklund and Mona Forsman. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *EMNLP: Industry Track*, 2022. 3, 2, 27

[23] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 2002. 27

[24] Zahra Farid. Why are social media images important?, 2024. 8

[25] Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. Visual fact checker: Enabling high-fidelity detailed caption generation. In *CVPR*, 2024. 14, 27

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 26

[27] Leander Girrbach, Stephan Alaniz, Genevieve Smith, and Zeynep Akata. A large scale analysis of gender biases in text-to-image generative models. *arXiv:2503.23398*, 2025. 3

[28] Aryaman Gupta, Yusuf Umut Ciftci, and Somil Bansal. From perception logs to failure modes: Language-driven semantic clustering of failures for robot safety. *arXiv:2506.06570*, 2025. 3

[29] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *ICCV*, 2023. 15

[30] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 2021. 5, 3

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 26

[32] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *CVPR*, 2017. 20

[33] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv:2401.03105*, 2024. 11

[34] Juhua Hu and Jian Pei. Subspace multi-clustering: a review. *Knowledge and information systems*, 2018. 2

[35] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv:2310.06825*, 2023. 2

[36] Kaggle. Cards Image Dataset-Classification, 2022. 3

[37] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 2

[38] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop*, 2011. 21

[39] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *CVPR*, 2024. 26

[40] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 27

[41] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 5, 3

[42] Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *ICLR*, 2024. 3, 6, 7, 2, 9, 15

[43] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *ICLR*, 2025. 2, 5, 4

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 5, 7

[45] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023. 6

[46] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *ICML*, 2024. 27

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 2

[48] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021. 26

[49] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 4, 5, 6, 7, 9, 15, 18, 27

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 3, 2, 9, 27

[51] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv:2402.00253*, 2024. 11

[52] Mingxuan Liu, Tyler L Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *CVPR*, 2024. 5, 7

[53] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *ICLR*, 2024. 5, 2, 11, 21, 23, 27

[54] Mingxuan Liu, Tyler L Hayes, Massimiliano Mancini, Elisa Ricci, Riccardo Volpi, and Gabriela Csurka. Test-time vocabulary adaptation for language-driven object detection. In *ICIP*, 2025. 7

[55] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *arXiv:2306.07272*, 2023. 2

[56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 8, 2, 26

[57] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *NeurIPS*, 2024. 27

[58] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *ECCV*, 2024. 3, 7, 9

[59] Kevin McSpadden. You now have a shorter attention span than a goldfish. *Time Magazine*, 14, 2015. 8

[60] Meta. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024. 2, 18, 27

[61] Meta. Introducing Llama 3.1: Our most capable models to date, 2024. 2, 4, 5, 6, 7, 15, 18, 27

[62] Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *CVPR*, 2023. 2

[63] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 4, 6, 7

[64] Horea Muresan and Mihai Oltean. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 2018. 3, 14

[65] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AAAI/ACM on AI, Ethics, and Society*, 2023. 8

[66] Nishanth Nakshatri, Siyi Liu, Sihao Chen, Dan Roth, Dan Goldwasser, and Daniel Hopkins. Using llm for improving key event discovery: Temporal-guided news stream clustering with event summaries. In *Findings of EMNLP*, 2023. 3

[67] Leonardo Nicoletti and Bass. Humans are biased. Generative AI is even worse, 2023. 8

[68] OpenAI. ChatGPT: A Large-Scale GPT-3.5-Based Model, 2022. 27

[69] OpenAI. DALL·E 2 pre-training mitigations, 2022. 8, 24

[70] OpenAI. Hello GPT-4o, 2024. 2, 4, 18

[71] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 6

[72] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Prediction of social image popularity dynamics. In *ICIAP*, 2019. 8, 24

[73] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2

[74] Jerry K Palmer and Jonathan S Gore. A theory of contrast effects in performance appraisal and social cognitive judgments. *Psychological Studies*, 2014. 26

[75] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024. 2

[76] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 8, 23

[77] Joseph R Priester, Utpal M Dholakia, and Monique A Fleming. When and why the background contrast effect emerges: thought engenders meaning by influencing the perception of applicability. *Journal of Consumer Research*, 2004. 26

[78] ZiJie Qi and Ian Davidson. A principled and flexible framework for finding alternative clusterings. In *SIGKDD*, 2009. 2

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 5, 2, 6, 23

[80] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 27

[81] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, S Yu Philip, and Lifang He. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[82] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, 2016. 26

[83] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. 22, 26

[84] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *CVPR*, 2023. 2

[85] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 7

[86] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv:2405.18870*, 2024. 19

[87] Sindhu Tipirneni, Ravinarayana Adkathimar, Nurendra Choudhary, Gaurush Hiranandani, Rana Ali Amjad, Vassilis N Ioannidis, Changhe Yuan, and Chandan K Reddy. Context-aware clustering using large language models. *arXiv:2405.00988*, 2024. 3

[88] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 2, 27

[89] U.S. Bureau of Labor Statistics. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity — bls.gov, 2021. [Accessed 26-Oct-2022]. 8

[90] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020. 2, 27

[91] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 5, 3

[92] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *NeurIPS*, 2024. 3

[93] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, 2014. 20

[94] Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. *TACL*, 2024. 3

[95] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011. 21

[96] Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. Protchatgpt: Towards understanding proteins with large language models. *arXiv:2402.09649*, 2024. 27

[97] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP conference short papers*, 2009. 3, 2

[98] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, 2024. 2

[99] Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In *EMNLP*, 2024. 14

[100] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 2, 27

[101] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 2023. 2

[102] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv:2408.08872*, 2024. 2, 18

[103] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *ICML*, 2023. 26

[104] Jiawei Yao, Enbei Liu, Maham Rashid, and Juhua Hu. Augdmc: Data augmentation guided deep multiple clustering. *Procedia Computer Science*, 2023. 2

[105] Jiawei Yao, Qi Qian, and Juhua Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *CVPR*, 2024. 2, 3, 7, 9

[106] Jiawei Yao, Qi Qian, and Juhua Hu. Customized multiple clustering via multi-modal subspace proxy learning. In *NeurIPS*, 2025. 2, 3, 6, 7, 9

[107] Guoxian Yu, Liangrui Ren, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Multiple clusterings: Recent advances and perspectives. *Computer Science Review*, 2024. 2, 3, 27

[108] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022. 26

[109] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014. 20

[110] Yuwei Zhang, Zihan Wang, and Jingbo Shang. Clusterllm: Large language models as a guide for text clustering. In *EMNLP*, 2023. 3

[111] Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. *NeurIPS*, 2024. 3

[112] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. ChatGPT asks, BLIP-2 answers: Automatic questioning towards enriched visual descriptions. *TMLR*, 2024. 2

[113] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. Mindstorms in natural language-based societies of mind. *arXiv:2305.17066*, 2023. 2