Organizing Unstructured Image Collections using Natural Language

Supplementary Material

Table of Contents

Overview

This supplementary material provides additional details supporting the implementations, experiments, findings, and discussions in the main paper.

In Supp. A, we provide a reproducibility statement to ensure the transparency and replicability of our work. Additionally, in Supp. B, we present an ethics statement to discuss and address the potential ethical implications and concerns associated with the proposed task and methodology, highlighting possible societal impacts and mitigation strategies.

Supp. C expands on Related Work, covering relevant tasks and methods. Supp. D describes the benchmarks used in our study, including the construction of the newly proposed COCO-4c and Food-4c datasets, as well as the process for creating hard ground-truth criteria for proposer evaluation. Supp. E details the evaluation metrics used in this work.

Supp. F presents the prompts and implementation details for $\mathcal{X}$-Cluster, covering both Criteria Proposers and Semantic Groupers. Supp. G provides implementation specifics of the compared methods. Additional quantitative results, including evaluations of the Criteria Proposer, Semantic Grouper, and comparisons with other clustering methods, are in Supp. H.

For qualitative analysis, Supp. I presents further visualizations of predicted clusters, while Supp. J examines failure cases. Supp. K investigates the impact of multi-granularity clustering. Supp. L explores the effect of invalid (hallucinated) criteria on system performance, and Supp. M studies the influence of foundation model hallucinations and biases.

Supp. N analyzes the computational cost and runtime of $\mathcal{X}$-Cluster. Supp. O and Supp. P extend our analysis with studies on system sensitivity and fine-grained image collections. Supp. Q provides additional findings, implementation details, and user study results for the three explored applications. Supp. R discusses how LLMs enhance image clustering. Finally, Supp. S outlines potential future research directions for our proposed OpenSMC task.

The Table of Contents on the next page outlines the main topics in this supplementary material, with hyperlinks for direct navigation to each section.

A. Reproducibility Statement

Upon publication, we will *open-source* all essential resources for reproducing this work. Specifically, we will provide the full code implementation of $\mathcal{X}$-Cluster, along with the exact prompts used in each module. Additionally, we will release our two newly proposed benchmarks, COCO-4c and Food-4c, including annotations for each grouping criterion. Lastly, we will provide the code for our evaluation protocol, experiments, and application studies.

B. Ethics Statement

We do not anticipate any immediate negative societal impacts from our work. However, we encourage future researchers building on

this work to remain vigilant, as we have, about the potential for $\mathcal{X}$-Cluster, which integrates LLMs and MLLMs–particularly their human-like reasoning abilities– to be used both for good and for harm.

The motivation behind our studies on biases in existing datasets and text-to-image (T2I) generative model outputs is to *reveal and address* these biases that objectively exist in the datasets and models. We emphasize that our aim is *to study and mitigate these issues*, and in doing so, we *do not create* any new biases or disturbing content. Specifically, in Sec. 6, we use well-established benchmarks, such as CelebA [56], for our study of dataset bias, and for bias discovery in T2I generative models, we select occupation-related subjects known to be associated with biases from prior studies [4, 6]. However, we acknowledge that our methodology and findings could potentially be misused by malicious actors to promote harmful narratives or discrimination against certain groups. We strongly oppose any such misuse or misrepresentation of our work. Our research is conducted with the aim of advancing technology while prioritizing public welfare and well-being.

For the creation of our two new benchmarks, COCO-4c and Food-4c, we sourced images exclusively from the COCO-val-2017 [47] and Food-101 [8] datasets, strictly adhering to their licensing agreements. Additionally, we utilized voluntary human annotators for proposing valid grouping criteria and creating annotations along these criteria, rather than employing annotators from crowdsourcing platforms. This decision was made to ensure sustainability, fair compensation, and high-quality work, as well as to safeguard the psychological well-being of participants. Similarly, for our user study on T2I model bias evaluation, we recruited voluntary participants via questionnaires to collect human evaluation results. The user study was conducted entirely anonymously, with participants providing informed consent. Our project, including data annotation and the user study involving human subjects, was approved by the Ethical Review Board of our university.

Lastly, we emphasize that our proposed framework, $\mathcal{X}$-Cluster, relies on open-source LLMs and MLLMs, allowing full deployment on local machines. We refrain from using APIs from industrial LLMs or MLLMs, both to ensure reproducibility and to protect data privacy.

C. Additional Related Work

Topic Discovery. The setting of open-ended semantic multiple clustering (OpenSMC) is also related to the field of Topic Discovery [5, 22, 97] in natural language processing, which aims to identify textual themes from large *text corpora* (*e.g.*, documents). Our work shares motivational similarities with topic discovery because both tasks seek to find common, thematic concepts from large volumes of data. In contrast, our work focuses on discovering thematic criteria from large *visual content*. However, indeed, the core challenges of OpenSMC and topic discovery are highly similar: they both require systems that can concurrently reason over large volumes of data. Nevertheless, OpenSMC is an even more challenging task than topic discovery for two reasons: *i)* semantics are not explicitly expressed in images, whereas they are in text; *ii)* there is currently no vision model that can encode large sets of images and reliably reason over them. Thus, in this work, we translate images to text and use text as a proxy to elicit the large-scale reasoning capability of large language models [61].

Multimodal Large Language Model. Recent advancements in multimodal large language models (MLLMs) have been driven by the availability of large-scale vision-language aligned training data. The typical paradigm [50] involves using a pre-trained large language model (LLM) [14, 35, 60, 61] alongside a pre-trained vision encoder [79]. A projector is learned to align visual inputs with the LLM in the embedding space, which enhances visual understanding by utilizing the reasoning capabilities of LLMs. Several models have achieved significant success in zero-shot image captioning and visual question answering (VQA), including BLIP-2 [44], BLIP-3 [102], Kosmos-2 [75], and the LLaVA series [43, 49, 50]. In our proposed $\mathcal{X}$-Cluster framework, we employ MLLM primarily as a *zero-shot* image parser, converting visual information into text and using this text as a proxy to elicit LLMs for reasoning over large image collections and discovering grouping criteria. Additionally, we leverage the multi-image reasoning capability of LLaVA-NeXT-Interleave [43] to establish a baseline image-based proposer for the OpenSMC task, while utilizing BLIP-2 with customized prompts in a VQA style [84, 112] as the image-based grouper to form semantic clusters linked to specific visual content within the images.

Large Language Model. In the era of large language models (LLMs) advancement [73], modern LLMs, such as the Llama series [60, 61, 88], Vicuna [14], Mistral-7B [35], and the GPT series [9], have demonstrated remarkable zero-shot capabilities in tasks involving text analysis, completion, generation, and summarization. With advanced prompting techniques like Chain-of-Thought (CoT) [100], the reasoning abilities of LLMs can be further enhanced. In the proposed $\mathcal{X}$-Cluster framework, we design CoT prompts (see Supp. F) to harness the text generation and summarization capabilities of Llama-3.1 as a reasoning engine. This aids $\mathcal{X}$-Cluster in several key areas: discovering grouping criteria from large sets of image captions, automatically prompting VQA models, generating criterion-specific tags, uncovering cluster semantics, and grouping images based on their captions. Unlike prior works [113] that focus on set difference captioning [20], fine-grained concept discovery [53], or video understanding [98], we leverage LLMs to tackle the challenging open-ended semantic multiple clustering task. While IC|TC [42] also uses the LLM (GPT-4 [1]) for grouping visual data, our proposed $\mathcal{X}$-Cluster differs in two key aspects: *i)* $\mathcal{X}$-Cluster does *not* require user-defined grouping criteria or the number of clusters, and *ii)* $\mathcal{X}$-Cluster provides *multi-granularity* outputs to meet various user preferences.

Text-Driven Image Retrieval. Given a query text (*e.g.*, "sofa" or "person wearing a blue T-shirt"), text-driven image retrieval methods [37, 55, 101] aim to find images from an image collection that are relevant to the query. In other words, in the scenario we are considering, given the image collection and a list of text queries, one can organize images according to the text using text-driven image retrieval techniques. In this context, the query can be considered as a sort of "cluster name". However, this differs significantly from the proposed task of open-ended semantic multiple clustering (OpenSMC), because OpenSMC requires both discovering the textual criteria and the corresponding textual clusters. Thus, without knowing text queries as prior information, text-driven image retrieval methods are not able to accomplish OpenSMC.

D. Benchmark Details

D.1. Construction of COCO-4c and Food-4c

To create high-quality benchmarks for COCO-4c and Food-4c, we designed a four-step annotation pipeline:

(1) Criteria Identification: We first split COCO-val-2017 [47] and Food-101 [8] images into batches of 100. Each batch was stitched into a 10×10 grid to form a single image. These grid images were then distributed to 5 human annotators, who were tasked with identifying grouping criteria. For each dataset, we selected the 4 most frequently occurring criteria, as shown in Tab. 1, to proceed with per-image annotation.

(2) Label Candidate Generation: To facilitate the annotation process, we used GPT-4V [1] to generate an initial list of candidate labels for each criterion. Specifically, for each criterion of COCO-4c and Food-4c, GPT-4V was prompted to assign a label that reflected the criterion for each image. This resulted in a list of criterion-specific label candidates for each dataset.

(3) Image Annotation: Next, 10 human annotators were tasked with assigning a label from the criterion-specific candidates to each image in COCO-4c and Food-4c for each criterion. The entire annotation process took 25 days to complete.

(4) Label Merging: Image annotation is inherently subjective, with annotators potentially assigning different labels for the same criterion. For example, one annotator might label the `Mood` criterion as "Happy", while another might label it as "Joyful" or "Delightful". To resolve such discrepancies, we used majority voting to determine the final label for each image. Specifically, the most frequently assigned label among the 10 annotators was chosen as the final label for each criterion.

Following these steps, we constructed COCO-4c and Food-4c. *Note that we used the official COCO-val-2017 [47] and Food-101 [8] images for our benchmarks and did not collect any new images. We adhered strictly to the licenses of the datasets during their creation.* The exact number of classes is presented in Tab. 1. Additionally, the annotated class names for each criterion of COCO-4c are provided in Tab. 2, and for Food-4c in Tab. 3.

D.2. Details on Hard Grouping Criteria Annotation

In Tab. 4, we present the additional annotated **Hard** grouping criteria ground truth alongside the **Basic** criteria for each benchmark.

While we have established more rigorous and challenging benchmarks such as COCO-4c and Food-4c, which feature up to *four* distinct grouping criteria, these annotated criteria sets do not encompass all potential grouping criteria within the image collections. This is particularly true for more complex and realistic datasets like COCO-4c, Food-4c, and Action-3c. As a result, the performance differences between different criteria proposers on these basic criteria, as shown in Fig. 4, tend to be close to each other, limiting our understanding of each proposer's ability to generate comprehensive grouping criteria.

To address this limitation, we employed human annotators to further identify and propose grouping criteria across the six benchmarks, resulting in a more extensive ground-truth set for each benchmark. This provides a better basis for evaluating the comprehensiveness of the different proposers. We refer to this set of larger annotation criteria as the **Hard** criteria, in contrast to the **Basic** criteria, which involve per-image annotations. Note that for the **Hard** criteria, per-image label annotation is not provided due to the high cost of annotation. The procedure for obtaining the **Hard** grouping criteria is as follows:

(1) Criteria Discovery: We divided each dataset into batches of 100 images, displaying each batch in a 10×10 grid. Five human annotators were assigned to each batch and instructed to identify as many valid grouping criteria as possible. The proposed criteria from each annotator were then combined to form a comprehensive set of grouping criteria for the dataset.

(2) Criteria Merging: After collecting the annotated criteria from all five annotators, we aggregated the criteria and manually cleaned the set by merging semantically similar criteria (*e.g.*, `Location` and `Place`) and discarding binary grouping criteria, as the inclusion of binary criteria can result in an unmanageable number of grouping criteria for complex datasets.

By following this process, we developed a more comprehensive grouping criteria set as the **Hard** ground-truth for each benchmark, as shown in Tab. 4. This resulted in sets containing 8 criteria for Fruit-2c, 4 criteria for card, 11 criteria for Action-3c, 7 criteria for Clevr-4c, 17 criteria for COCO-4c, and 11 criteria for Food-4c. These expanded ground-truth sets enable us to more effectively evaluate the capabilities of various criteria discovery methods, providing a clearer understanding of different criteria proposers.

E. Further Details of Evaluation Protocol

Further Discussion on Clustering Accuracy (CAcc). Clustering Accuracy (CAcc) [30] is evaluated by applying the Hungarian algorithm [41] to determine the optimal assignment between the predicted cluster indices and ground-truth labels. As extensively discussed in the GCD [91] literature, if the number of predicted clusters (groups) exceeds the total number of ground-truth classes (groups), the extra clusters (not matched by the Hungarian algorithm) are assigned to a null set, and all instances in those clusters are considered incorrect during evaluation. On the other hand, if the number of predicted clusters is lower than the number of ground-truth classes, the extra classes are assigned to a null set, and all instances with those ground-truth labels are similarly considered incorrect. Thus, CAcc is maximized only when the number of predicted clusters matches the number of ground-truth clusters.

In the Open-ended Semantic Multiple Clustering (OpenSMC) task newly proposed in this work, we do not assume access to the ground-truth number of clusters as prior input. Consequently, our proposed method $\mathcal{X}$-Cluster does not rely on the ground-truth number of clusters to achieve an "optimal" CAcc with respect to the testing dataset. All clusters are automatically predicted by the $\mathcal{X}$-Cluster system. In stark contrast, in the comparison with criterion-conditioned clustering methods shown in Tab. 2, both IC|TC [42] and MMaP [105] use the ground-truth number of clusters as prior input.

F. Further Implementation and Prompt Details

In this section, we provide detailed descriptions of the exact prompts used in our framework, along with additional implemen-

Table 1. **Summary of number of classes for the basic criteria annotation across the six benchmarks.**

Dataset	Number of Images	Basic Criterion	Number of Classes
COCO-4c	5,000	Activity	64
		Location	19
		Mood	20
		Time of Day	6
Food-4c	25,250	Food Type	101
		Cuisine	15
		Course	5
		Diet	4
Action-3c	1,000	Action	40
		Location	10
		Mood	4
Clevr-4c	10,000	Color	10
		Texture	10
		Shape	10
		Count	10
Card-2c	8,029	Rank	14
		Suit	5
Fruit-2c	103	Species	34
		Color	15

Table 2. **Full class names for COCO-4c across the four basic criteria.**

Criterion	COCO-4c
Activity	"repairing a toilet", "playing volleyball", "playing guitar", "haircutting", "cutting a cigar", "kayaking", "applauding", "tying a tie", "playing basketball", "washing dishes", "gardening", "texting messages", "repairing a car", "peeing", "cleaning the floor", "writing on a book", "feeding a horse", "singing", "baking", "hiking", "smoking", "riding an elephant", "pouring liquid", "waving hands", "swimming", "meditating", "fixing a bike", "cutting vegetables", "walking a dog", "reading a book", "celebrating", "queuing", "cutting a cake", "brushing teeth", "playing soccer", "jumping", "snowboarding", "playing", "touching animals", "pushing a cart", "watching tv", "rowing a boat", "taking photos", "running", "flying a kite", "riding a horse", "playing video games", "holding up an umbrella", "throwing a frisbee", "lying down", "riding a bike", "drinking", "cooking", "phoning", "chatting", "skiing", "driving", "surfing", "skateboarding", "playing baseball", "playing tennis", "using a computer", "posing", "eating"
Location	"amusement or theme park", "healthcare facility", "virtual or digital space", "educational institution", "industrial area", "historical landmark", "public event or gathering", "store or market", "underground or enclosed space", "transportation hub", "zoo", "water body", "office or workplace", "park or recreational area", "restaurant or dining area", "sports facility", "natural environment", "urban area or city street", "residential area"
Mood	"anxious", "sombre", "contemplative", "suspenseful", "serene", "nostalgic", "inspired", "whimsical", "romantic", "mysterious", "melancholic", "chaotic", "humorous", "vibrant", "peaceful", "energetic", "focused", "joyful", "relaxed", "adventurous"
Time of Day	"evening", "afternoon", "night", "morning", "indoor lighting", "midday"

tation details for the proposed Criteria Proposer in Supp. F.1 and the Semantic Grouper in Supp. F.2.

Further, we present a system overview illustrating all three variants of our Criteria Proposer and Semantic Grouper, namely the caption based (main), tag based, and image based versions, in Fig. 1.

F.1. Further Details of Criteria Proposer

Image-based Proposer: In Tab. 5, we present the exact prompt used in the image-based proposer for querying the MLLM LLaVA-NeXT-Interleave-7B [43]. Given a target image set, we first randomly shuffle the images and divide them into disjoint subsets, each containing 64 images. Each subset is then stitched into an 8×8 image grid, treated as a single image, and fed into the MLLM. For each subset, the MLLM is prompted to propose 5 distinct grouping criteria for organizing the images within that subset, using the prompt shown in Tab. 5. After iterating through all subsets, we take the union of the criteria proposed for each subset as the discovered criteria for the target image set. Finally, we deduplicate the discovered criteria and accumulate them into the criteria pool.

Tag-based Proposer: In Tab. 6, we present the exact prompt

Table 3. **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"

Table 4. **Annotated criteria for the six benchmarks.** The basic criteria are annotated on per-image level for each benchmark, while the hard criteria (those not in the basic criteria) are further exhaustively annotated by human annotators for further evaluating the performance of the rule proposer in OpenSMC task.

COCO-4c	
Basic criteria **Total: 4**	**Hard criteria** **Total: 17**
Activity	Activity
Location	Location
Mood	Mood
Time of Day	Time of Day
	Interaction
	Number of People Present
	Group Dynamics
	Clothing Style
	Occasion or Event Type
	Photo Style
	Type of Animal Present
	Weather
	Type of Primary Object
	Continent
	Age or Age Composition
	Race or Race Composition
	Gender or Gender Composition

Food-4c	
Basic criteria **Total: 4**	**Hard criteria** **Total: 11**
Food Type	Food Type
Cuisine	Cuisine
Course	Course
Diet	Diet
	Tableware Type
	Presentation Style
	Color Palette
	Setting/Theme
	Primary Taste
	Primary Ingredient
	Cooking Method

Action-3c	
Basic criteria **Total: 3**	**Hard criteria** **Total: 11**
Action	Action
Mood	Mood
Location	Location
	Clothing Style
	Number of People Present
	Age or Age Composition
	Race or Race Composition
	Occasion or Event Type
	Group Dynamics
	Lighting Condition
	Gender or Gender Composition

Clevr-4c	
Basic criteria **Total: 4**	**Hard criteria** **Total: 7**
Color	Color
Texture	Texture
Shape	Shape
Count	Count
	Spatial Positioning
	Count of Surface
	Complexity of Geometry

Card-2c	
Basic criteria **Total: 2**	**Hard criteria** **Total: 4**
Rank	Rank
Suit	Suit
	Color
	Illustration Style

Fruit-2c	
Basic criteria **Total: 2**	**Hard criteria** **Total: 8**
Species	Species
Color	Color
	Size
	Seasonality
	Primary Taste
	Texture
	Ripeness
	Fruit Quantity and Arrangement

Table 5. **Prompts for the MLLM in the image-based proposer for criteria proposing.**

Prompt purpose	Prompt
System Prompt	You are a helpful AI assistant
Input Explanation	This image contains 64 individual images arranged in 8 columns and 8 rows.
Goal Explanation	I am a machine learning researcher trying to identify all the possible clustering criteria or rules that could be used to group these images so I can better understand my data.
Task Instruction	Your job is to carefully analyze the entire set of the 64 images, and identify five distinct clustering criteria or rules that could be used to cluster or group these images. Please consider different characteristics.
Output Instruction	Please write a list of the 5 identified clustering criteria or rules (separated by bullet points "*").
Task Reinforcement	Again, I want to identify all the possible clustering criteria or rules that could be used to group these images. List the 5 distinct clustering criteria or rules that you identified from the 64 images. Answer with a list (separated by bullet points "*"). Your response:

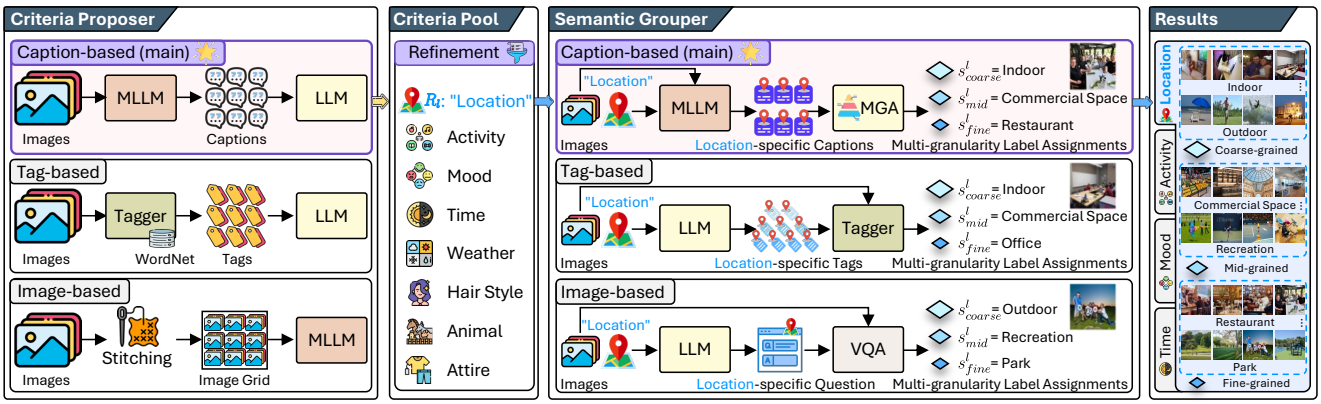


Figure 1. **All three variants of the proposed $\mathcal{X}$-Cluster framework.** We explore different design choices for both the Criteria Proposer (**left**) and the Semantic Grouper (**right**), and designate the best-performing Caption-based system (marked with ⭐) as the main $\mathcal{X}$-Cluster configuration in our experiments.

Table 6. **Prompts for the LLM used in the tag-based proposer for criteria proposing.** We embed the exact image captions by replacing the placeholders `"{TAGS}"` in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	The following are the tagging results of a set of images in the format of "Image ID: tag 1, tag 2, ..., tag 10". These assigned tags reflect the visible semantic content of each image:
Tag Embedding	Image 1: `"{TAGS}"` Image 2: `"{TAGS}"` ... Image N: `"{TAGS}"`
Goal Explanation	I am a machine learning researcher trying to figure out the potential clustering or grouping criteria that exist in these images. So I can better understand my data and group them into different clusters based on different criteria.
Task Instruction	Please analyze these images by using their assigned tags. Come up with an array of distinct clustering criteria that exist in this set of images.
Output Instruction	Please write a list of clustering criteria (separated by bullet points "*").
Task Reinforcement	Again, I want to figure out what are the potential clustering or grouping criteria that I can use to group these images into different clusters. List an array of clustering or grouping criteria that often exist in this set of images based on the tagging results. Answer with a list (separated by bullet points "*"). Your response:

used in the tag-based proposer for querying the LLM Llama-3.1-8B [61]. For a given target image set, we first utilize an open-vocabulary tagger, CLIP ViT-L/14 [79], to assign 10 related natural language tags to each image. These tags are selected from the WordNet [63] vocabulary, which contains 118k English synsets, and represent the semantic content of the images. We employ the standard prompt "`A photo of {concept}`" provided by CLIP for image tagging. Next, we embed the assigned tags into the prompt shown in Tab. 6 to carry the semantics of the entire image set and query the LLM to propose grouping criteria. The criteria proposed by the LLM are then added to the criteria pool. Note that in this case, we embed the tags for the entire dataset into a single prompt for criteria proposal, without reaching the LLM context length limits (*e.g.*, 128k for Llama-3.1-8B) for the datasets used in our experiments. However, for larger datasets, it may be necessary to split the dataset into subsets, prompt the LLM for each subset, and use the union of the proposed criteria as the final output.

Caption-based Proposer: We present the prompt used in the caption-based proposer for the MLLM LLaVA-NeXT-7B [49] in Tab. 7, and the prompt for the LLM Llama-3.1-8B [61] in Tab. 8. Specifically, we first use the MLLM with a general prompt to generate detailed descriptions for each image in the target dataset, effectively translating the visual information into natural language. The generated captions are then *randomly shuffled* and split into disjoint subsets, each containing 400 captions. Next, we embed the captions from each subset into the prompt shown in Tab. 8 and use it to query the LLM to propose grouping criteria for the images represented by the captions. After iterating through all subsets, we take the union of the proposed criteria across subsets as the discovered criteria for the target image set. Finally, we deduplicate these criteria and add them to the criteria pool. Due to the context window limitations of LLMs, embedding all captions into a single prompt is infeasible. To address this, we limit each subset to 400 captions, which results in approximately 115k tokens per subset. This strategy allows us to remain within the context length limits of modern LLMs (*e.g.*, 128k tokens for both Llama-3.1 and

Table 7. **Prompts for the MLLM in the caption-based proposer for generating detailed descriptions of the images.**

Prompt purpose	Prompt
System Prompt	You are a helpful AI assistant
Task Instruction	Describe the following image in detail.

Table 8. **Prompts for the LLM used in the caption-based proposer for criteria proposing.** We embed the exact image captions by replacing the placeholders `"{CAPTION}"` in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	The following are the result of captioning a set of images:
Caption Embedding	Image 1: `"{CAPTION}"` Image 2: `"{CAPTION}"` ... Image N: `"{CAPTION}"`
Goal Explanation	I am a machine learning researcher trying to figure out the potential clustering or grouping criteria that exist in these images. So I can better understand my data and group them into different clusters based on different criteria.
Task Instruction	Come up with ten distinct clustering criteria that exist in this set of images.
Output Instruction	Please write a list of clustering criteria (separated by bullet points "*").
Task Reinforcement	Again I want to figure out what are the potential clustering/grouping criteria that I can use to group these images into different clusters. List ten clustering or grouping criteria that often exist in this set of images based on the captioning results. Answer with a list (separated by bullet points "*"). Your response:

GPT-4o) while maximizing the number of samples per query to effectively propose clustering criteria.

Criteria Pool Refinement: In Tab. 9, we present the exact prompt used for criteria pool refinement when querying the LLM Llama-3.1-8B [61]. Since the accumulated criteria pool $\widetilde{\mathcal{R}}$ may contain highly similar or noisy clustering criteria, we embed the criteria from the pool into the prompt shown in Tab. 9 and ask the LLM to merge similar criteria and rephrase their names to enhance clarity. This process yields a refined set of grouping criteria, which is then passed to the next stage for image grouping.

F.2. Further Details of Semantic Grouper

Image-based Grouper: In Tab. 10, we present the prompt used to query the LLM Llama-3.1-8B [61] for automatically generating criterion-specific VQA questions for the image-based grouper. The objective at this stage is to condition the VQA model BLIP-2 Flan-T5$_{\text{XXL}}$ [44] to label each image across three different semantic granularity levels based on a specific criterion. To guide the VQA model effectively, a criterion-specific question is required.

Rather than manually creating these questions, we embed the target criterion into the prompt shown in Tab. 10 and query the LLM to automatically generate high-quality, criterion-specific questions. These questions are then used to direct the VQA model, enabling it to accurately label each image according to the visual content relevant to the target criterion.

Tag-based Grouper: We present the prompts used in the tag-based grouper for querying the LLM Llama-3.1-8B. The prompt for generating criterion-specific tags is shown in Tab. 11, while the prompts for generating coarse-grained and fine-grained tags are shown in Tab. 12 and Tab. 13, respectively.

In the tag-based grouper, we begin by embedding the target criterion into the prompt from Tab. 11 to generate criterion-specific tags at a middle granularity. To enhance the diversity and coverage of the tags, we query the LLM 10 times and take the union of the generated tags after deduplication as candidates [54]. Following the SHiNe framework [52], for each middle-grained tag, we further embed it into the prompts from Tab. 12 and Tab. 13 to generate 3 super-categories (coarse-grained) and 10 sub-categories (fine-grained) for each tag.

After generating coarse and fine-grained categories for all middle-grained tags, we take the union of the super-categories as the coarse-grained tag candidates and the union of the sub-categories as the fine-grained tag candidates. Lastly, we use the open-vocabulary tagger CLIP ViT-L/14 to assign the most relevant tags to each image based on cosine similarity, using candidates from each granularity level. After tagging all the images, we group those sharing the same tag into clusters, yielding the clustering result. Note that we do not utilize lexical databases such as WordNet [63] or ConceptNet [85] for tag generation, as they do not support free-form input and may not capture certain discovered criteria.

Caption-based Grouper: We first present the MLLM prompt used for LLaVA-NeXT-7B [49] to generate criterion-specific captions in Tab. 14. Following this, we present the LLM Llama-3.1-8B prompts used in the caption-based grouper for the *Initial Naming* step in Tab. 15, the *Multi-granularity Cluster Refinement* step in Tab. 16, and the *Final Assignment* step in Tab. 17.

Specifically, we begin by generating criterion-specific captions for each image using LLaVA-NeXT-7B with the prompt shown in Tab. 14. For each image, we then embed its criterion-specific caption and the relevant criterion into the LLM prompt shown in Tab. 15, querying the LLM to assign an initial name based on the target criterion. Once the initial names for all images in the dataset

Table 9. **Prompts for the LLM used in Proposed Criteria Refinement step** We embed the exact initially discovered criteria by replacing the placeholders "{CRITERION}" in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	I am a machine learning researcher working with a set of images. I aim to cluster this set of images based on the various clustering criteria present within them. Below is a preliminary list of clustering criteria that I've discovered to group these images:
Criteria Embedding:	* Criterion 1: "{CRITERION}" * Criterion 2: "{CRITERION}" ... * Criterion L: "{CRITERION}"
Goal Explanation	My goal is to refine this list by merging similar criteria and rephrasing them using more precise and informative terms. This will help create a set of distinct, optimized clustering criteria.
Task Instruction	Your task is to first review and understand the initial list of clustering criteria provided. Then, assist me in refining this list by: * Merging similar criteria. * Expressing each criterion more clearly and informatively.
Output Instruction	Please respond with the cleaned and optimized list of clustering criteria, formatted as bullet points (using "*"). Your response:

Table 10. **Prompts for the LLM used in the image-based grouper for automatic criterion-specific VQA question generation.** We embed the exact discovered criterion by replacing the placeholder "{CRITERION}" in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Goal Explanation	Hello! I am a machine learning researcher focusing on image categorization based on the aspect of "{CRITERION}" depicted in images.
Task Instruction	Therefore, I need your assistance in designing a prompt for the Visual Question Answering (VQA) model to help it identify the "{CRITERION}" category in a given image at three different granularity. Please help me design and generate this prompt using the following template: "Question: [Generated VQA Prompt Question] Answer (reply with an abstract, a common, and a specific category name, respectively):". The generated prompt should be simple and straightforward.
Output Instruction	Please respond with only the generated prompt using the following format "* Answer *". Your response:

Table 11. **Prompts for the LLM used in the tag-based grouper for generating middle-grained criterion-specific tags.** We embed the exact discovered criterion by replacing the placeholder "{CRITERION}" in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Goal Explanation	Hello! I am a machine learning researcher focusing on image categorization of a certain aspect. I'm interested in generating a list of tags specifically for categorizing the types of "{CRITERION}" depicted in images.
Task Instruction	Please provide a list of potential "{CRITERION}" category names. Please generate diverse category names. Do not include too general or specific category names such as "Sports".
Output Instruction	Please respond with the list of category names. Each category should be formatted as follows: "* Category Name". Your response:

Table 12. **Prompts for the LLM used in the tag-based grouper for generating coarse-grained criterion-specific tags.** We embed the exact discovered criterion and middle-grained category by replacing the placeholder "{CRITERION}" and "{MIDDLE-GRAINED CATEGORY NAME}" in the prompt, respectively.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Task Instruction	Generate a list of three more abstract or general "{CRITERION}" super-categories that the following "{CRITERION}" category belongs to and output the list separated by "&" (without numbers): "{MIDDLE-GRAINED CATEGORY NAME}"
Output Instruction	Your response:

Table 13. **Prompts for the LLM used in the tag-based grouper for generating fine-grained criterion-specific tags.** We embed the exact discovered criterion and middle-grained category by replacing the placeholder "{CRITERION}" and "{MIDDLE-GRAINED CATEGORY NAME}" in the prompt, respectively.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Task Instruction	Generate a list of ten more detailed or specific "{CRITERION}" sub-categories of the following "{CRITERION}" category and output the list separated by "&" (without numbers): "{MIDDLE-GRAINED CATEGORY NAME}"
Output Instruction	Your response:

Table 14. **Prompts for the MLLM used in the caption-based grouper for generating criterion-specific captions.** We embed the exact discovered criterion by replacing the placeholder "{CRITERION}" in the prompt.

Prompt purpose	Prompt
System Prompt	You are a helpful AI assistant.
Task Instruction	Analyze the image focusing specifically on the "{CRITERION}". Provide a detailed description of the "{CRITERION}" depicted in the image. Highlight key elements and interactions relevant to the "{CRITERION}" that enhance the understanding of the scene.
Output Instruction	Your response:

Table 15. **Prompts for the LLM used in the caption-based grouper at the *Initial Naming* step for initially assigning a criterion-based category name to the image based on its criterion-specific caption.** We embed the exact discovered criterion and the corresponding criterion-specific caption by replacing the placeholder "{CRITERION}" and "{CRITERION-SPECIFIC CAPTION}" in the prompt, respectively.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	The following is the description about the "{CRITERION}" of an image:
Caption Embedding	"{CRITERION-SPECIFIC CAPTION}"
Goal Explanation	I am a machine learning researcher trying to assign a label to this image based on what is the "{CRITERION}" depicted in this image.
Task Instruction	Understand the provided description carefully and assign a label to this image based on what is the "{CRITERION}" depicted in this image.
Output Instruction	Please respond in the following format within five words: "*Answer*". Do not talk about the description and do not respond long sentences. The answer should be within five words.
Task Reinforcement	Again, your job is to understand the description and assign a label to this image based on what is the "{CRITERION}" shown in this image. Your response:

are obtained, we embed these names along with the target criterion into the prompt in Tab. 16 to query the LLM for cluster name refinement across three semantic granularity levels: coarse, middle, and fine.

Finally, for each image, we embed the target criterion, its criterion-specific caption, and cluster candidates from each granularity level into the prompt shown in Tab. 17, and use this to query the LLM for final cluster assignment at each granularity level.

G. Further Details of the Compared Methods

In this section, we provide the implementation details of the compared methods, IC|TC [42] and MMaP [105].

Implementation details of IC|TC [42]: In the original implementation of IC|TC, LLaVA-1.5 [50] was used as the MLLM, and GPT-4-2023-03-15-preview [1] as the LLM. However, since the GPT-4-2023-03-15-preview API has been deprecated, we re-implemented IC|TC using the state-of-the-art MLLM LLaVA-NeXT-7B [49] and the latest version of GPT-turbo-2024-04-09 as the LLM, while strictly adhering to the original IC|TC prompt design in our experiments to ensure a fair comparison.

Implementation Details of SSD-LLM [58]: Following similar setup of IC|TC, we reproduced and compared with SSD-LLM using GPT-turbo-2024-04-09 as the LLM and LLaVA-NeXT-7B as the MLLM. Since SSD-LLM requires a primary class name for each benchmark, we provided the ground-truth primary class in its prompt: "Food" for Food-4c, "Object" for Clevr-4c, "Person" for Action-3c, "Playing card" for Card-2c, and "Fruit" for Fruit-2c. For COCO-4c, which consists of everyday life scenes and lacks a consistent primary class, we used "Object" as a neutral placeholder in SSD-LLM's prompt.

Implementation details of MMaP and MSub [105, 106]: We closely followed the training configuration outlined in the original MMaP and MSub paper. Specifically, GPT-turbo-2024-04-09 was used as the LLM to generate reference words for each dataset. We then prompt-tuned CLIP-ViT/B32 using Adam with a momentum of 0.9, training the model for 1,000 epochs for each criterion across all datasets. Hyperparameters were optimized accord-

Table 16. **Prompts for the LLM used in the caption-based grouper at the *Multi-granularity Cluster Generation* step for refining the initially assigned names to a structured three granularity levels.** We embed the exact discovered criterion and the initially assigned name categories by replacing the placeholder "{CRITERION}" and "{MIDDLE-GRAINED CATEGORY NAME}" in the prompt, respectively.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	The following is an initial list of "{CRITERION}" categories. These categories might not be at the same semantic granularity level. For example, category 1 could be "cutting vegetables", while category 2 is simply "cutting". In this case, category 1 is more specific than category 2.
Category Embedding	* "{MIDDLE-GRAINED CATEGORY NAME}" * "{MIDDLE-GRAINED CATEGORY NAME}" ... * "{MIDDLE-GRAINED CATEGORY NAME}"
Task Instruction	These categories might not be at the same semantic granularity level. For example, category 1 could be "cutting vegetables", while category 2 is simply "cutting". In this case, category 1 is more specific than category 2. Your job is to generate a three-level class hierarchy (class taxonomy, where the first level contains more abstract or general coarse-grained classes, the third level contains more specific fine-grained classes, and the second level contains intermediate mid-grained classes) of "{CRITERION}" based on the provided list of "{CRITERION}" categories. Follow these steps to generate the hierarchy.

Sub-task Instruction	Follow these steps to generate the hierarchy:
	Step 1 - Understand the provided initial list of "{CRITERION}" categories. The following three-level class hierarchy generation steps are all based on the provided initial list.
	Step 2 - Generate a list of abstract or general "{CRITERION}" categories as the first level of the class hierarchy, covering all the concepts present in the initial list.
	Step 3 - Generate a list of middle-grained "{CRITERION}" categories as the second level of the class hierarchy, in which the middle-grained categories are the subcategories of the categories in the first level. The categories in the second-level are more specific than the first level but should still cover and reflect all the concepts present in the initial list.
	Step 4 - Generate a list of more specific fine-grained "{CRITERION}" categories as the third level of the class hierarchy, in which the categories should reflect more specific "{CRITERION}" concepts that you can infer from the initial list. The categories in the third-level are subcategories of the second-level.
	Step 5 - Output the generated three-level class hierarchy as a JSON object where the keys are the level numbers and the values are a flat list of generated categories at each level, structured like:
	{
	"level 1": ["categories"],
	"level 2": ["categories"],
	"level 3": ["categories"]
	}

Output Instruction	Please only output the JSON object in your response and simply use a flat list to store the generated categories at each level. Your response:

Table 17. **Prompts for the LLM used in the caption-based grouper at the *Final Assignment* step.** We embed the exact discovered criterion and the refined category names from each granularity level, by replacing the placeholder "{CRITERION}" and "{CANDIDATE CATEGORY NAME}" in the prompt, respectively.

Prompt purpose	Prompt
System Prompt	You are a helpful assistant.
Input Explanation	The following is a detailed description about the "{CRITERION}" of an image.
Caption Embedding	"{CRITERION-SPECIFIC CAPTION}"
Task Instruction	Based on the content and details provided in the description, classify the image into one of the specified "{CRITERION}" categories listed below:
Candidate Category Embedding	"{CRITERION}" categories: * "{CANDIDATE CATEGORY NAME}" * "{CANDIDATE CATEGORY NAME}" ... * "{CANDIDATE CATEGORY NAME}"
Output Instruction	Ensure that your classification adheres to the details mentioned in the image description. Respond with the classification result in the following format: "*category name*". Your response:

ing to the loss score of MMaP, with the learning rate searched in {0.1, 0.05, 0.01, 0.005, 0.001, 0.0005}, weight decay in {0.0005,

0.0001, 0.00005, 0.00001, 0}, α and β in {0.0, 0.1, 0.2, ..., 1.0}, and λ fixed at 1 for all experiments. After training, KMeans, with the ground-truth number of clusters, was applied for each criterion and dataset to perform clustering.

H. Further Quantitative Experimental Results

In this section, we present additional numerical experiment results to supplement the figures in the main paper. In § H.1, we provide supplementary results for the evaluation of the Criteria Proposer in our framework. In § H.2, we present additional results for the evaluation of the Semantic Grouper across various criteria on the six tested benchmarks. Furthermore, we include expanded results comparing our framework to prior criteria-conditioned clustering methods. Lastly, we present detailed results from the ablation study of the multi-granularity refinement component in § H.4.

H.1. Further Results for Criteria Proposer Study

We provide detailed numerical results corresponding to Fig. 4 in Tab. 18 and Fig. 5 in Tab. 19 for the six tested benchmarks.

Although captions generated by the MLLM may exhibit some information loss (*e.g.*, ignoring small objects or attributes) [33] and hallucinations (*e.g.*, introducing objects not present in the images) [51], these issues generally occur at the object or fine-grained attribute level. However, when reasoning about grouping criteria for OpenSMC task, the focus is on identifying general thematic elements shared across the image set. As a result, these minor inconsistencies in the captions do not hinder the LLM in our framework from effectively reasoning about grouping criteria, helping the Caption-based Proposer to achieve the best performance among all the studied design choices.

H.2. Further Results for Semantic Grouper Study

In this section, we present the expanded numerical results comparing different semantic groupers to supplement the summary in Fig. 6. Specifically, we provide detailed results for the evaluation of the six tested datasets as follows:

- COCO-4c (Fig. 6(a)) in Tab. 20
- Card-2c (Fig. 6(b)) in Tab. 21
- Action-3c (Fig. 6(c)) in Tab. 22
- Food-4c (Fig. 6(d)) in Tab. 23
- Fruit-2c (Fig. 6(e)) in Tab. 24
- Clevr-4c (Fig. 6(f)) in Tab. 25

In addition, we present the statistics of the predicted clusters at each granularity level in Tab. 26.

H.3. Further Comparative Results with TCMC Methods

We provide expanded results in Tab. 27 for each criterion and benchmark, detailing the comparison of criteria-conditioned clustering methods presented in Tab. 2 in the main paper.

H.4. Further Results for Multi-granularity Clustering Study.

We present expanded results in Tab. 28 for the ablation study on multi-granularity refinement, providing a detailed breakdown of the summary shown in Fig. 7 in the main paper.

I. Qualitative Analysis

In this section, we visualize the grouping results predicted by the best configuration of our proposed framework (Caption-based Proposer and Caption-based Grouper). Specifically, we present example clustering results across different criteria for COCO-4c in Fig. 2, Food-4c in Fig. 3, Action-3c in Fig. 4, Clevr-4c in Fig. 5, and Card-2c in Fig. 6. Additionally, we showcase example clustering results at different predicted granularity levels for COCO-4c in Fig. 7.

J. Failure Case Analysis

In Fig. 8, we present several failure cases from the best configuration of our proposed framework (Caption-based Proposer and Caption-based Grouper). As observed, our method frequently misassigns "Surfing" to the "Kayaking" cluster under the `Activity` criterion. Upon examining the intermediate criterion captions generated by the MLLM, we found that this error is largely due to the MLLM incorrectly describing a "Surfboard" as a "Kayak". This highlights the importance of the MLLM's ability to accurately describe images, as it is critical for the performance of our system. Potential improvements could include majority voting or model ensembling using different MLLM models.

Another issue arises in crowded scenes. When multiple people are present in an image, the model consistently assigns the `Mood` label "Communal" to the images. We speculate that this occurs because, in the presence of multiple people, the model struggles to accurately determine the mood of one key individual.

Finally, we observed that our method sometimes fails to distinguish subtle, fine-grained differences between images, leading to incorrect labels. For example, as shown in Fig. 8, "Edamame" or "Pho" are typical dishes from China, Vietnam, and Japan, but they may be presented differently depending on the cuisine. The "Edamame" shown in Fig. 8 is presented in a traditional Japanese style, yet our model incorrectly predicted it as Chinese cuisine. This oversight of fine-grained details could be improved by employing a more advanced prompting strategy [53].

K. Further Study on Multi-granularity Clustering

In this section, we provide a detailed study on how different levels of multi-granularity output from our $\mathcal{X}$-Cluster framework impact grouping results. Specifically, for the Action-3c dataset, we employed human annotators to label two additional granularity levels for the criteria `Action` and `Location`. For the `Action` criterion, we consider the original annotation as fine-grained (L3) and tasked annotators to name the action in the image using more abstract and general coarse-grained (L1) and middle-grained (L2) labels. For the `Location` criterion, we consider the original annotation as middle-grained (L2) and tasked annotators to provide both more abstract coarse-grained (L1) labels and more specific fine-grained (L3) labels. This process resulted in expanded ground-truth annotations at three distinct semantic granularity levels for both the `Action` and `Location` criteria of the Action-3c dataset.

Next, we quantitatively evaluated the multi-granularity grouping results at each predicted clustering granularity level against

Table 18. **Comparison of True Positive Rate (TPR) (%) for criteria proposers across the six OpenSMC benchmarks**. TPR performance is reported for both Basic and Hard ground-truth criteria. The best performance is highlighted in bold.

	COCO-4c		Food-4c		Action-3c		Clevr-4c		Card-2c		Fruit-2c		Average	
	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard
Image-based	**100.0**	52.9	25.0	36.4	66.7	54.6	50.0	28.6	50.0	25.0	50.0	20.0	56.9	36.2
Tag-based	50.0	35.3	**100.0**	72.7	66.7	36.4	75.0	42.9	50.0	50.0	50.0	20.0	65.3	42.9
Caption-based	**100.0**	**64.7**	**100.0**	**81.8**	**100.0**	**72.7**	**100.0**	**71.4**	**100.0**	**100.0**	**100.0**	**60.0**	**100.0**	**75.1**

Table 19. **Study of the impact of data scale on criteria discovery.** The Caption-based Proposer is used for criteria discovery, and TPR performance (%) is reported on the *Hard* ground-truth criteria sets across the six OpenSMC benchmarks for different data scales. The best performance is highlighted in bold.

Data scales	COCO-4c	Food-4c	Action-3c	Clevr-4c	Card-2c	Fruit-2c	Average
100%	**64.7**	**81.8**	**72.7**	71.4	**100.0**	**60.0**	**75.1**
80%	47.1	72.7	54.6	71.4	75.0	30.0	58.5
60%	52.9	63.6	54.6	71.4	**100.0**	50.0	65.4
40%	41.2	45.5	45.5	**85.7**	**100.0**	40.0	59.6
20%	35.3	45.5	36.4	42.9	**100.0**	40.0	50.0
1 img	23.5	36.4	27.3	57.1	75.0	50.0	44.9

Table 20. **Comparison of Semantic Groupers on COCO-4c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(a).

Methods	Activity			Location			Mood			Time of Day		
	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	62.6	73.5	67.6	34.3	51.5	41.1	22.4	43.3	29.5	40.6	74.1	52.4
KMeans CLIP	34.4	-	-	32.7	-	-	18.9	-	-	38.6	-	-
KMeans DINOv1	34.8	-	-	37.5	-	-	17.9	-	-	36.5	-	-
KMeans DINOv2	38.2	-	-	37.9	-	-	22.5	-	-	43.8	-	-
Img-based BLIP-2	48.7	64.1	**55.3**	39.6	48.0	43.4	30.2	37.5	33.4	40.7	60.3	48.6
Img-based LLaVA	46.5	61.8	53.1	34.0	46.3	39.2	28.0	24.7	26.3	39.4	51.7	44.7
Tag-based	43.2	51.5	47.0	28.6	46.6	35.5	13.0	25.6	17.2	19.3	48.8	27.7
Caption-based	44.1	48.9	46.4	55.2	55.6	**55.4**	38.1	32.6	**35.2**	67.6	56.7	**61.7**

Table 21. **Comparison of Semantic Groupers on Card-2c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(b).

Methods	Suit			Rank		
	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	47.9	69.5	56.7	35.0	64.2	45.3
KMeans CLIP	45.0	-	-	28.6	-	-
KMeans DINOv1	38.5	-	-	20.7	-	-
KMeans DINOv2	36.7	-	-	22.3	-	-
Img-based BLIP-2	66.7	77.7	**71.8**	47.5	54.4	50.7
Img-based LLaVA	36.8	65.8	47.2	24.6	49.8	32.9
Tag-based	39.2	32.9	35.8	22.3	39.1	28.4
Caption-based	54.5	73.6	62.6	92.1	95.1	**93.6**

each ground-truth annotation granularity level by measuring clustering accuracy (CAcc) and semantic accuracy (SAcc). The main caption-based $\mathcal{X}$-Cluster framework was used for this experiment. In Fig. 9, we report the Harmonic Mean of CAcc and SAcc for the `Action` and `Location` criteria of Action-3c, across each predicted clustering granularity level evaluated against each ground-truth annotation level. As clearly shown, the highest grouping performance consistently appears along the diagonal. This indicates that the best grouping performance is achieved when the predicted granularity *matches* the annotation granularity.

These experimental results not only highlight the importance of the multi-granularity output of our framework but also validate the effectiveness of our multi-granularity design in aligning with user-preferred granularities that is reflected by the annotations in

Table 22. **Comparison of Semantic Groupers on Action-3c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(c).

Methods	Action			Location			Mood		
	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	97.1	99.2	98.1	66.7	67.1	66.9	75.5	80.7	78.0
KMeans CLIP	62.3	-	-	58.3	-	-		-	-
KMeans DINOv1	49.3	-	-	61.4	-	-		-	-
KMeans DINOv2	75.7	-	-	67.6	-	-		-	-
Img-based BLIP-2	79.7	80.9	80.3	43.3	42.4	42.8	43.1	43.8	43.4
Img-based LLaVA	70.1	60.5	65.0	45.8	42.8	44.2	32.0	38.0	34.7
Tag-based	70.2	55.0	61.6	36.8	48.1	41.7	50.7	47.6	49.1
Caption-based	82.8	82.8	**82.8**	69.8	55.2	**61.6**	52.3	50.2	**51.2**

Table 23. **Comparison of Semantic Groupers on Food-4c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(d).

Methods	Food Type			Cuisine			Course			Diet		
	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	90.6	94.6	92.6	54.9	81.4	65.6	63.5	84.7	72.6	47.6	59.9	53.0
KMeans CLIP	66.1	-	-	29.8	-	-	49.5	-	-	36.9	-	-
KMeans DINOv1	33.6	-	-	15.3	-	-	38.1	-	-	41.4	-	-
KMeans DINOv2	72.7	-	-	22.5	-	-	47.6	-	-	43.4	-	-
Img-based BLIP-2	54.2	71.4	**61.6**	54.8	73.3	**62.7**	42.3	71.0	53.0	34.2	53.8	41.9
Img-based LLaVA	42.2	64.0	50.9	33.7	57.6	42.6	46.9	73.1	57.1	27.0	40.5	32.4
Tag-based	45.0	63.3	52.6	48.8	42.1	45.2	42.7	70.1	53.1	25.2	34.1	29.0
Caption-based	34.6	54.2	42.2	47.0	65.9	54.9	69.1	85.7	**76.5**	41.5	54.0	**46.9**

Table 24. **Comparison of Semantic Groupers on Fruit-2c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(e).

Methods	Species			Color		
	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	84.0	93.1	88.3	54.8	83.5	66.1
KMeans CLIP	67.1	-	-	39.6	-	-
KMeans DINOv1	53.8	-	-	36.0	-	-
KMeans DINOv2	71.2	-	-	36.7	-	-
Img-based BLIP-2	70.7	68.3	69.5	40.9	70.6	51.8
Img-based LLaVA	63.9	67.8	65.8	51.0	83.2	**63.2**
Tag-based	64.0	67.1	65.5	54.1	44.1	48.6
Caption-based	76.9	70.7	**73.7**	53.3	51.5	52.4

Table 25. **Comparison of Semantic Groupers on Clevr-4c.** We report Clustering Accuracy (CAcc), Semantic Accuracy (SAcc), and their Harmonic Mean (HM) in percentages (%). These results are plotted in Fig. 6(f).

Methods	Color			Texture			Count			Shape		
	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM	CAcc	SAcc	HM
CLIP Zero-shot	77.7	94.0	85.1	34.1	41.9	37.6	43.7	81.5	56.9	71.1	72.7	71.9
KMeans CLIP	48.8	-	-	61.4	-	-	44.2	-	-	56.1	-	-
KMeans DINOv1	53.0	-	-	58.4	-	-	47.5	-	-	67.0	-	-
KMeans DINOv2	44.1	-	-	46.9	-	-	52.5	-	-	87.0	-	-
Img-based BLIP-2	69.3	76.5	**72.7**	57.8	34.4	43.1	25.7	55.9	35.2	69.1	62.6	65.7
Img-based LLaVA	56.5	63.5	59.8	51.9	26.9	35.4	53.7	39.4	45.4	64.3	71.3	**67.6**
Tag-based	66.6	55.3	60.4	57.2	40.2	47.3	47.4	8.3	14.1	62.7	36.5	46.2
Caption-based	70.3	63.4	66.7	65.3	42.1	**51.2**	65.7	73.3	**69.3**	58.4	38.5	46.4

Table 26. **Summary of cluster counts across six benchmarks for the comparison of semantic groupers.** The results yield by the main Caption-based Grouper is reported. Specifically, we report: *i)* GT: the number of ground-truth clusters; *ii)* Pred-Init: predicted clusters from initial names; *iii)* Pred-Coarse: predicted coarse-grained clusters after multi-granularity refinement; *iv)* Pred-Middle: predicted middle-grained clusters after multi-granularity refinement; and *v)* Pred-Fine: predicted fine-grained clusters after multi-granularity refinement.

Dataset	Criteria	GT	Pred-Init	Pred-Corase	Pred-Middle	Pred-Fine
COCO-4c	Activity	64	203	12	23	52
	Location	19	145	7	14	28
	Mood	20	122	15	25	30
	Time of Day	6	96	2	8	31
Food-4c	Food Type	101	301	7	37	127
	Cuisine	15	141	9	18	53
	Course	5	97	4	12	78
	Diet	4	139	5	8	64
Action-3c	Action	40	71	8	15	51
	Location	10	82	5	10	67
	Mood	4	95	6	18	55
Clevr-4c	Color	10	25	6	12	17
	Texture	10	23	2	5	12
	Shape	10	22	5	11	14
	Count	10	11	2	4	11
Card-2c	Rank	14	147	4	7	16
	Suit	5	56	4	7	30
Fruit-2c	Species	34	54	8	25	38
	Color	15	66	5	15	39

these experiments.

L. Study on Handling Invalid Criteria

At the criteria refinement step, *invalid* grouping criteria (False Positives) may be proposed due to hallucinations from large language models (LLMs). While we did not observe hallucinated criteria being introduced during our experiments across six datasets and three application studies, it is important to further investigate the potential impact of such invalid criteria on the proposed $\mathcal{X}$-Cluster system.

To this end, we design and conduct a control experiment using the Fruit-2c dataset [64], where we *artificially* introduced two "hallucinated" invalid grouping criteria (False Positives), `Action` and `Clothing Style`, into the refined criteria pool. These invalid criteria were then used in the subsequent grouping process to evaluate their effect on our system. We apply the main Caption-based Grouper to group fruit images based on these "hallucinated" criteria.

The grouping results for the two invalid criteria are presented in Tab. 29. As observed, when processing invalid "hallucinated" criteria, nearly all images are assigned to a cluster named "Not visible" by our framework. This occurs because, in the absence of relevant visual content in the images, the MLLM-generated captions do not include descriptors corresponding to the invalid criteria. Consequently, the LLM creates a "Not visible" cluster and assigns the images to it. Since the system provides interpretable outputs, users can easily identify and disregard such invalid groupings. This control experiment highlights the robustness of our system against hallucination in practical scenarios.

M. Study on Model Hallucination and Bias

Model hallucination. LLM hallucination [99] typically occurs when LLMs are tasked with complex queries requiring world knowledge or factual information—for instance, answering a question like "Who was the 70th president of the United States?" might lead to a fabricated response. However, in our system, the use of LLMs is fully grounded in the visual descriptions (tags or captions) of the images. Consequently, the LLM output is strongly constrained to analyzing these visual descriptions, significantly reducing the likelihood of hallucination. That said, LLM hallucination can still have mild effects on clustering results. For example, as discussed in the failure case analysis in Sec. J, the LLM incorrectly grouped "Korean bibimbap" and "Vietnamese rice noodles" under "Chinese cuisine" (see Fig. 8). MLLMs also play a crucial role in our system, as they are responsible for translating images into text for subsequent processing steps. MLLM hallucination [99] typically involves incorrectly identifying the existence of objects, attributes, or spatial relationships within an image. However, since our proposed system operates at the *dataset level* rather than on a per-image basis, it is largely insensitive to such hallucinations, especially at the fine-grained visual detail level. Moreover, as our system is training-free, it can be further enhanced with LLM or MLLM hallucination mitigation techniques, such as the Visual Fact Checker [25], which we leave as a direction for future work.

Model Bias. Foundation models such as LLMs and MLLMs are known to inherit biases from their training data [7]. In our system, we addressed potential biases using Hard Positive Prompt-

Table 27. **Comparison with criteria-conditioned clustering methods on the six OpenSMC benchmarks.** We report Clustering Accuracy (CAcc) and Semantic Accuracy (SAcc)as percentages (%). Average (*Avg.*) CAcc and SAcc across different criteria on each dataset is also provided. For reference, we include the pseudo upper-bound (UB) performance of CLIP ViT-L/14 in zero-shot transfer, using ground-truth criteria and class names. Note that both IC|TC and MMaP utilize ground-truth criteria and the number of clusters for clustering. These expanded results correspond to Tab. 2.

Benchmark	Criterion	UB		IC|TC		SSD-LLM		MMaP		MSub		Ours	
		CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc	CAcc	SAcc
COCO-4c	Activity	62.6	73.5	51.3	53.2	44.0	52.1	33.8	-	35.9	-	44.1	48.9
	Location	34.3	51.5	58.5	54.0	51.2	52.9	35.3	-	37.4	-	55.2	55.6
	Mood	22.4	43.3	23.2	40.4	15.9	39.3	20.9	-	23.0	-	38.1	32.6
	Time of Day	40.6	74.1	62.8	65.2	55.5	64.1	45.7	-	47.8	-	67.6	56.7
	Avg.	40.1	60.6	48.9	**53.2**	41.6	52.1	33.9	-	36.0	-	**51.2**	48.4
Food-4c	Food Type	90.6	94.6	36.0	52.6	33.1	46.5	48.9	-	52.4	-	34.6	54.2
	Cuisine	54.9	81.4	46.8	42.4	43.9	36.3	31.7	-	35.2	-	47.0	65.9
	Course	63.5	84.7	70.5	89.5	67.6	83.4	48.6	-	52.1	-	69.1	85.7
	Diet	47.6	59.9	48.5	62.1	45.6	56.0	45.9	-	49.4	-	41.5	54.0
	Avg.	64.1	80.2	**50.5**	61.7	47.5	55.5	43.8	-	47.3	-	48.1	**64.9**
Clevr-4c	Color	77.7	94.0	51.2	43.2	47.8	44.0	75.3	-	84.7	-	70.3	63.4
	Texture	34.1	41.9	64.9	26.4	61.5	27.2	56.5	-	65.9	-	65.3	42.1
	Count	43.7	81.5	46.9	39.0	43.5	39.8	53.9	-	63.3	-	65.7	73.3
	Shape	71.1	72.7	70.0	38.7	66.6	39.5	65.5	-	74.9	-	58.4	38.5
	Avg.	56.7	72.5	58.3	36.8	54.8	37.6	62.8	-	72.2	-	**64.9**	**54.3**
Action-3c	Action	97.1	99.2	86.4	58.7	88.1	55.3	51.3	-	55.0	-	82.8	76.3
	Location	66.7	67.1	82.0	52.9	83.7	49.5	59.4	-	63.1	-	69.8	55.2
	Mood	75.5	80.7	60.8	57.4	62.5	54.0	71.0	-	74.7	-	52.3	50.2
	Avg.	79.8	82.3	76.4	56.3	**78.1**	52.9	60.6	-	64.3	-	68.3	**60.6**
Card-2c	Suit	47.9	69.5	54.9	65.6	47.5	60.7	41.3	-	44.0	-	54.5	73.6
	Rank	35.0	64.2	94.6	96.8	87.2	91.9	32.6	-	35.3	-	92.1	95.1
	Avg.	41.4	66.9	**74.8**	81.2	67.3	76.3	36.9	-	39.6	-	73.3	**84.3**
Fruit-2c	Species	84.0	93.1	69.3	66.9	68.1	58.6	58.8	-	62.2	-	76.9	70.7
	Color	54.8	83.5	57.2	43.3	56.0	35.0	43.3	-	46.7	-	53.3	51.5
	Avg.	69.4	88.3	63.3	55.1	62.0	46.8	51.0	-	54.4	-	**65.1**	**61.1**

ing techniques: *i) MLLM Bias Mitigation:* The MLLM is further prompted to generate criterion-specific captions that focus solely on describing the criterion-related content in each image. This approach constrains the MLLM from generating irrelevant content influenced by inherent biases; *ii) LLM Bias Mitigation:* Similarly, when prompting the LLM to assign image captions to clusters, we condition it to concentrate exclusively on the Criterion depicted in each image (see Tab. 15).

To validate the effectiveness of these bias mitigation techniques, we conducted a fair clustering experiment. Specifically, following Kwon et al. [42], we sampled images for four occupations (Craftsman, Laborer, Dancer, and Gardener) from the FACET [29] dataset, which contains images from 52 occupations. For each occupation, we selected 10 images of men and 10 images of women, totaling 80 images, ensuring a ground-truth gender proportion disparity of 0% for each occupation. Using our main $\mathcal{X}$-Cluster system, we grouped these images based on the criterion `Occupation` using three bias mitigation strategies: *i) No mitigation:* using general descriptions from the MLLM for LLM grouping; *ii) Our default hard positive prompting strategy:* using criterion-specific captions from the MLLM for LLM grouping; and *iii) Our default strategy with additional negative prompt:* adding a simple negative prompt, "Do not consider gender," to both the MLLM captioning and LLM grouping prompts.

In this experiment, non-biased result is defined as achieving equal gender proportions within each cluster. Tab. 30 presents the average gender ratios of the clustering results for each method across the four occupations. As observed, without bias mitigation, $\mathcal{X}$-Cluster exhibits noticeable gender bias in the studied occupations, with a gender disparity of 19.4%. However, our default bias mitigation techniques effectively reduce this disparity to 4.9%, achieving performance comparable to the addition of a manual negative prompt. This experiment demonstrates the effectiveness of our bias mitigation strategy and highlights the potential for further reducing model bias in our framework using more advanced techniques.

N. Computational Cost Analysis

The proposed $\mathcal{X}$-Cluster framework is training-free, requiring only inference processes. Specifically, our main framework (Caption-based) requires up to 31 GB of GPU memory to operate. All experiments reported in the paper were conducted on 4 Nvidia A100 40GB GPUs. In Tab. 31, we provide a detailed analysis of the computational efficiency of our main $\mathcal{X}$-Cluster framework (Caption-based Proposer and Caption-based Grouper) on the COCO-4c benchmark (5,000 images with four criteria) across various hardware configurations. For these experiments, we used LLaVA-NeXT-7B [49] as the MLLM and Llama-3.1-8B [61] as the LLM.

Table 28. **Ablation study of multi-granularity refinement on the six OpenSMC benchmarks.** We compare three ways of constructing cluster names: Initial Names (IN), Flat Refinement (FR), Multi-granularity Refinement (MR). We report Clustering Accuracy (CAcc) and Semantic Accuracy (SAcc)as percentages (%). Average (*Avg.*) CAcc and SAcc across different criteria on each dataset is also provided. These expanded results correspond to the plotting shown in Fig. 7.

Benchmark	Criterion	IN		FR		MR	
		CAcc	SAcc	CAcc	SAcc	CAcc	SAcc
COCO-4c	Activity	14.1	48.5	34.5	40.5	44.1	48.9
	Location	30.0	51.9	41.4	56.0	55.2	55.6
	Mood	6.6	34.7	21.9	32.1	38.1	32.6
	Time of Day	24.4	50.5	28.2	54.4	67.6	56.7
	Avg.	18.8	46.4	31.5	45.8	**51.2**	**48.4**
Food-4c	Food Type	33.9	52.4	35.5	54.3	34.6	54.2
	Cuisine	30.6	39.7	27.6	36.5	47.0	65.9
	Course	52.9	81.1	62.8	83.0	69.1	85.7
	Diet	14.0	46.6	36.8	58.2	41.5	54.0
	Avg.	32.9	55.0	40.7	58.0	**48.1**	**64.9**
Clevr-4c	Color	56.5	49.7	60.9	53.0	70.3	63.4
	Texture	56.5	26.0	60.9	33.0	65.3	42.1
	Count	56.5	39.6	56.5	40.8	65.7	73.3
	Shape	47.8	33.6	47.8	41.8	58.4	38.5
	Avg.	54.3	37.2	56.5	42.2	**64.9**	**54.3**
Action-3c	Action	72.2	63.6	90.5	63.0	82.8	76.3
	Location	46.0	50.4	65.9	59.3	69.8	55.2
	Mood	20.6	41.9	46.0	51.0	52.3	50.2
	Avg.	46.3	52.0	67.5	57.8	**68.3**	**60.6**
Card-2c	Suit	40.9	50.1	45.7	45.7	54.5	73.6
	Rank	43.0	55.1	47.7	54.6	92.1	95.1
	Avg.	42.0	52.6	46.7	50.2	**73.3**	**84.3**
Fruit-2c	Species	59.2	68.6	64.1	67.0	76.9	70.7
	Color	41.8	56.7	44.7	42.3	53.3	51.5
	Avg.	50.5	62.7	54.4	54.7	**65.1**	**61.1**

Table 29. **Study of the Influence of Invalid Grouping Criteria (False Positives) on the Fruit-2c Dataset.** We report the distribution of predicted groupings under the two "hallucinated" invalid grouping criteria. The main Caption-based Semantic Grouper is used for this experiment. †: For simplicity, all other minority clusters are grouped as "Others".

Predicted Clusters	Action (%)	Clothing Style (%)
Not visible	98.3	96.7
Others†	1.7	3.3

Table 30. **Average gender ratio and disparity** across the four studied occupations (Craftsman, Laborer, Dancer, and Gardener) from the FACET dataset. Images sampled from each occupation have an equal proportion of genders. Results from different bias mitigation strategies are reported.

Bias Mitigation Strategy	Male (%)	Female (%)	Gender Disparity (%)
Ground-truth	50.0	50.0	0.0
No mitigation	40.3	59.7	19.4
Ours (default)	47.6	52.5	4.9
Ours w. Negative prompt	48.4	51.6	3.2

As shown in Tab. 31, organizing 5,000 images based on all four discovered criteria can be completed by $\mathcal{X}$-Cluster in 29.1 hours on a single A100 GPU or 16.7 hours on a single H100 GPU. More importantly, most steps in our framework, such as per-image captioning and per-caption cluster assignment, are parallelizable across multiple GPUs, significantly accelerating the pro-

Figure 2. **Example predicted clusters of COCO-4c.**

cess. Therefore, when parallelizing the framework on 4 A100 or H100 GPUs, we achieve approximately a 4× speedup, reducing computational time to 7.6 hours on 4 A100 GPUs and 4.3 hours on 4 H100 GPUs.

O. System Sensitivity Analysis of Various MLLMs and LLMs

In Fig. 10, we perform a system-level sensitivity analysis using our default system configuration (caption-based proposer and caption-based grouper) to examine the impact of different MLLMs and LLMs on the system performance. Since all variants successfully

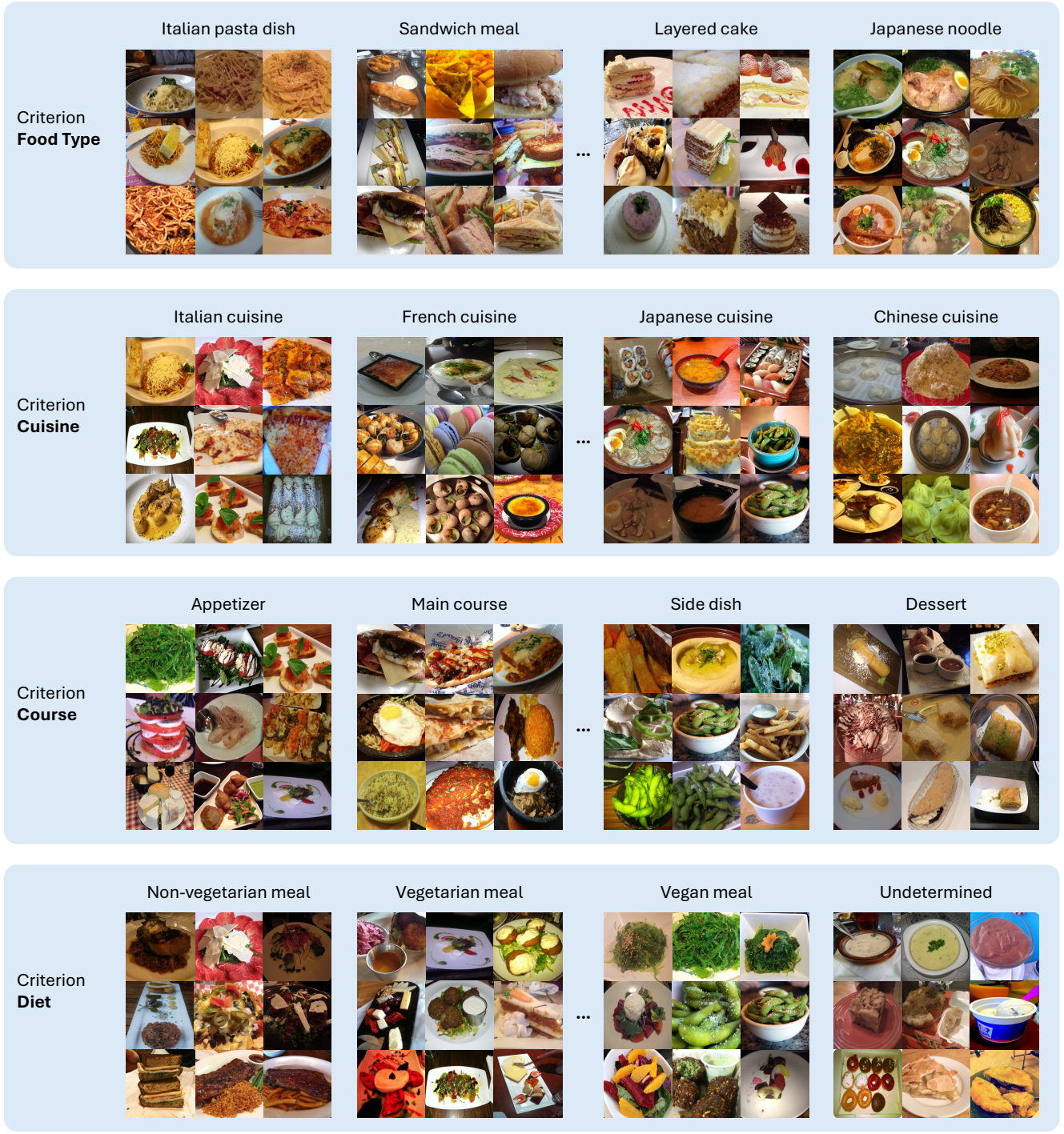


Figure 3. **Example predicted clusters of Food-4c.**

propose the basic criteria in each benchmark, we report the average clustering accuracy (CAcc) and semantic accuracy (SAcc) across various criteria for comparative analysis.

Specifically, in Fig. 10(a), we first fix the LLM in our system to Llama-3.1-8B [61] and assess the influence of various MLLMs: GPT-4V [1], BLIP-3-4B [102], and LLaVA-NeXT-7B [49]. Next, in Fig. 10(b), we set the MLLM to LLaVA-NeXT-7B and evaluate different LLMs: GPT-4-turbo [1], GPT-4o [70], Llama-3-8B [60], and Llama-3.1-8B.

Findings in Fig. 10(a) indicate a direct correlation between the size of the MLLM and the ability of our system to uncover sub-structures, highlighting the significant role of MLLMs in translating visual information into natural language. On the other hand, this scalability demonstrates that our system can enhance perfor-

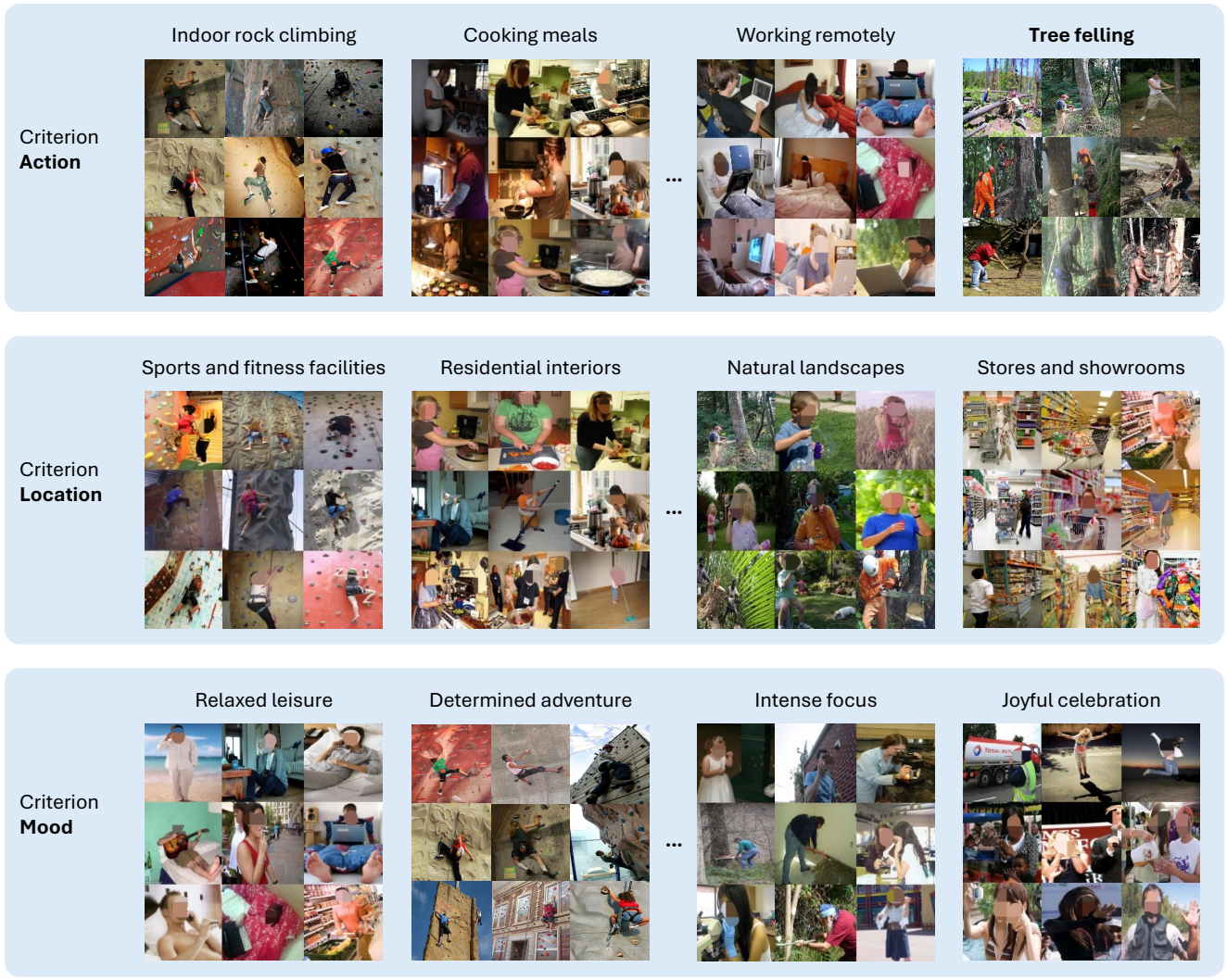


Figure 4. **Example predicted clusters of Action-3c.**

Table 31. **Computational cost analysis on the COCO-4c benchmark (5,000 images with four criteria).** We report the average and total time costs on various machines. The time costs were calculated for organizing all 5,000 images according to all the 4 criteria. Our main caption-based $\mathcal{X}$-Cluster framework is used in this experiment.

Method	Hardware	Average time cost (sec/img) ↓	Total time cost (hrs) ↓
$\mathcal{X}$-Cluster	1 Nvidia A100-40GB	20.9	29.1
	4 Nvidia A100-40GB	5.5	7.6
	1 Nvidia H100-80GB	12.0	16.7
	4 Nvidia H100-80GB	3.1	4.3

mance with more robust MLLMs, thanks to its training-free design, which ensures compatibility with any MLLM. Despite this, we use LLaVA-NeXT-7B as our default MLLM due to its *reproducibility*, being open-source and unaffected by API changes, and its capacity for local deployment, which *upholds privacy* by not exposing sensitive image data to external entities.

As for the LLMs, as depicted in Fig. 10(b), despite GPT-4-turbo showing marginally superior performance, the open-source Llama-3.1-8B achieves similar results across benchmarks, making it our default LLM. Notably, except for the Card-2c dataset, system performance remains largely consistent regardless of the power of the LLM. This consistency suggests that the reasoning task for OpenSMC, given the capabilities of modern LLMs to tackle complex problems [86], is relatively straightforward.

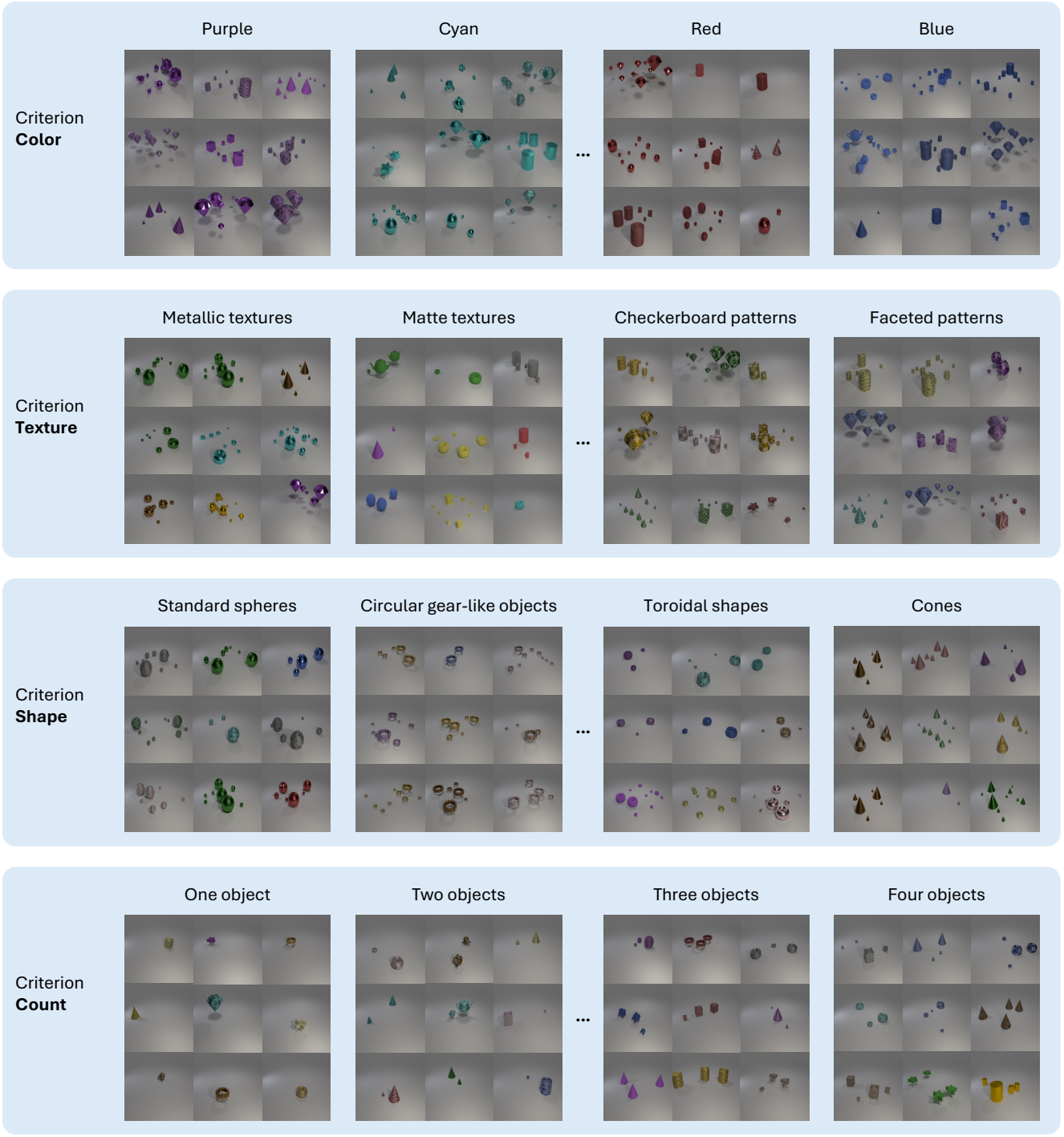


Figure 5. **Example predicted clusters of Clevr-4c.**

P. Study on Fine-grained Image Collections

Image collections may include fine-grained grouping criteria, such as `Bird species` in bird photography. Fine-grained criteria pose unique challenges for substructure discovery due to small inter-class differences and large intra-class variations [32, 93, 109]. This requires the model to detect subtle visual distinctions to accurately infer cluster names and guide the grouping process. The straightforward captioning process in our current framework may not fully capture these subtle visual nuances. However, the modular design of our framework allows for seamless integration of advanced cross-modal chain-of-thought (CoT) prompting strategies to address this issue.

We demonstrate this by enhancing our Caption-based Grouper

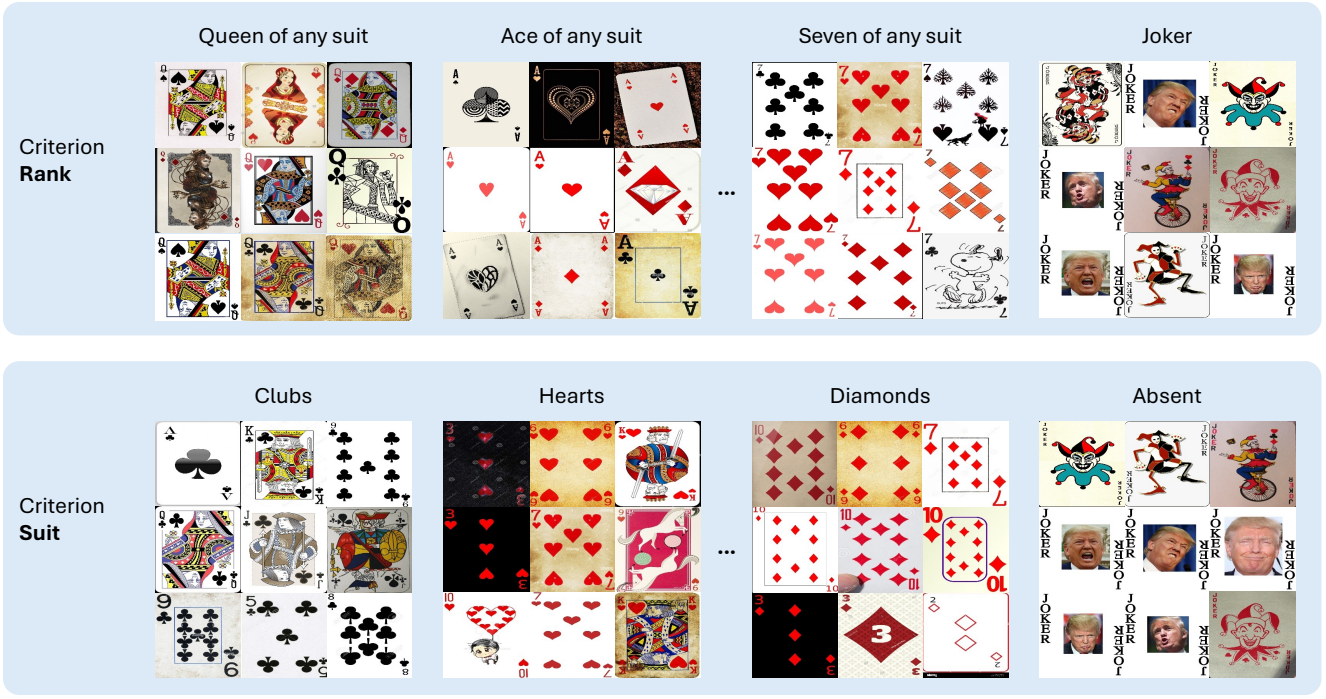


Figure 6. **Example predicted clusters of Card-2c.**

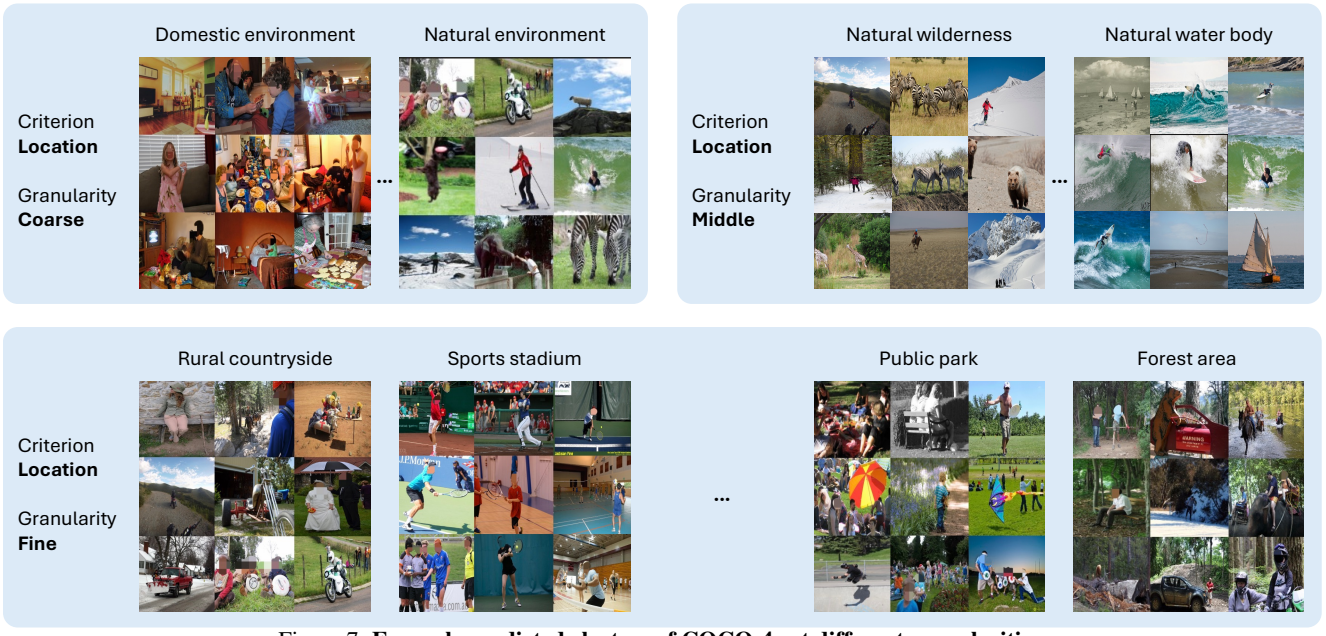
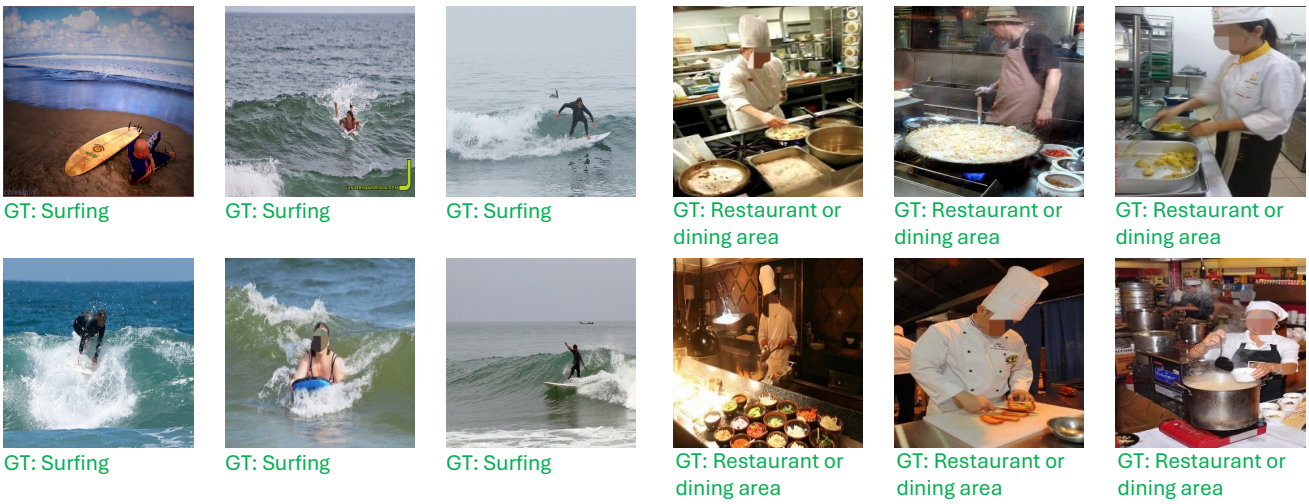


Figure 7. **Example predicted clusters of COCO-4c at different granularities.**

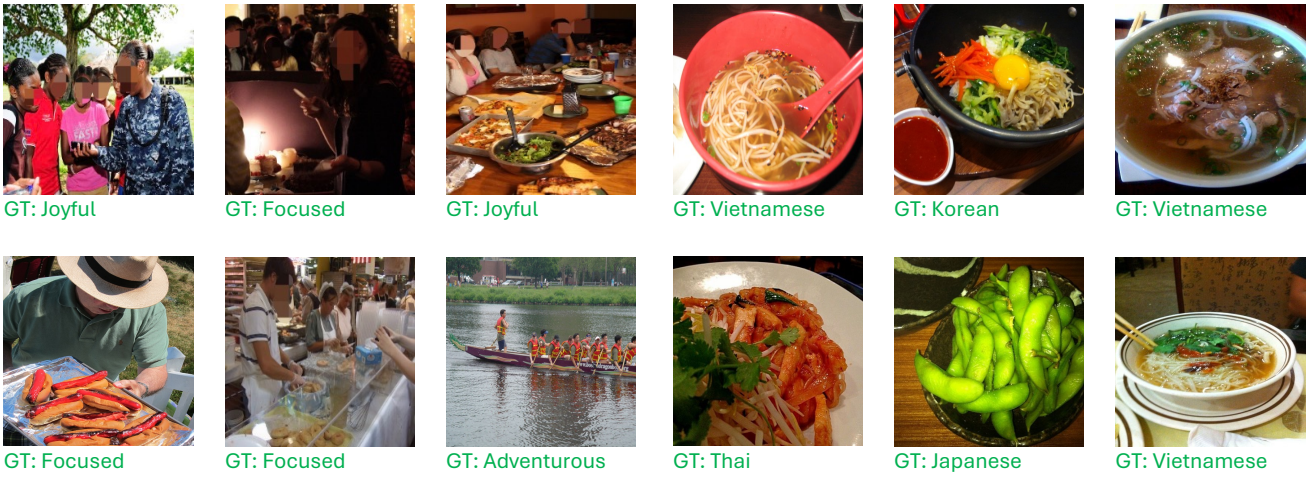
with FineR [53], a cross-modal CoT prompt method specifically designed for fine-grained visual recognition. When the proposer identifies fine-grained criteria, such as `Bird species`, the framework switches to a FineR-enhanced captioning strategy that provides more detailed attribute descriptions, such as "Wing color: Blue-grey," to enrich the captions and capture per-attribute visual characteristics to better support the subsequent substructure uncovering process.

We evaluate this on two image collections containing fine-grained criteria: CUB200 [95] and Stanford Cars196 [38]. Our framework successfully discovers the fine-grained criteria `Bird species` for CUB200 and `Car model` for Cars196. As shown in Tab. 32, when uncovering fine-grained substructures, integrating the FineR prompting strategy significantly improves performance by up to +15.0% CAcc and +12.2% SAcc, achieving results comparable to FineR itself. This demonstrates the flexibility of our



Benchmark:	COCO-4c
Criterion:	Activity
Predicted Cluster:	Kayaking

Benchmark:	Action-3c
Criterion:	Location
Predicted Cluster:	Professional workspaces



Benchmark:	Action-4c
Criterion:	Mood
Predicted Cluster:	Communal

Benchmark:	Food-4c
Criterion:	Cuisine
Predicted Cluster:	Chinese cuisine

Figure 8. **Failure case analysis.** We show wrongly predicted images with their ground-truth label for four clusters.

system, allowing future adaptations to specific application needs, such as fine-grained image collections.

Q. Further Details of the Application Study

In this section, we present additional implementation details, evaluation results, and findings for the application study discussed in Sec. 6 of the main paper. Specifically, § Q.3 offers further evaluation results and implementation details on using our predicted distribution to train a debiased model with GroupDRO [83]. § Q.1 outlines the implementation of the user study that assesses the alignment between predicted biases and human judgments, along with comprehensive findings for all studied occupations and identified criteria. Finally, § Q.2 provides additional insights from the analysis of social media image popularity.

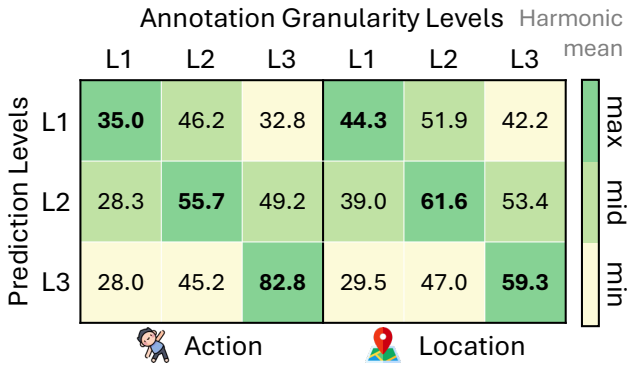


Figure 9. **Further study on the influence of multi-granularity clustering output.** We evaluate the CAcc and SAcc of the multi-granularity grouping results at each predicted clustering granularity level against each ground-truth annotation granularity level for the `Action` and `Location` criteria of the Action-3c dataset. The Harmonic Mean of CAcc and SAcc is reported for each granularity pair. L1, L2, and L3 represent the coarse-grained, middle-grained, and fine-grained levels, respectively, for both predictions and annotations.

Table 32. **Study of substructure discovery for fine-grained criteria.** We report clustering accuracy (CAcc) and semantic accuracy (SAcc) as percentages (%). The pseudo upper-bound (UB) performance is obtained using CLIP [79] ViT-L/14 in a zero-shot transfer setting with the ground-truth class names. †: We compare with FineR [53] without its post-class name refinement step to ensure a fair comparison.

	CUB200		Car196	
	CAcc	SAcc	CAcc	SAcc
UB	57.4	80.5	63.1	66.3
FineR†	44.8	64.5	**33.8**	**52.9**
Ours	30.1	56.7	21.3	35.9
Ours + FineR	**45.1**	**68.9**	31.1	47.3

Q.1. Further Details on Discovering Novel Bias in Text-to-Image Diffusion Models

Image Generation for the Subject Occupation: Following prior studies [4, 6], we selected nine occupations for our study: three stereotypically biased towards females (Nurse, Cleaning staff, Call center employee), three biased towards males (CEO, Firefighter, Basketball player), and three considered gender-neutral (Teacher, Computer user, Marketing coordinator). We used two state-of-the-art T2I diffusion model, DALL·E3 [3] and Stable Diffusion (SDXL) [76] to generate 100 images for each occupation for our study. This resulted in a total of 1,800 images. For each occupation, we provide some examples of images generated by DALL·E3 in Fig. 14, while provide some examples of images generated by SDXL in Fig. 15. We only used the simple prompt "A portrait photo of a <OCCUPATION>" for image generation for all occupations and did not include any potential biases in the prompt.

Bias Discovery and Quantification: We applied our method to 1,800 generated images and automatically identified 10 grouping criteria (bias dimensions) along with their predicted distributions for each occupation image set. For this study, we utilized the mid-granularity output of our system. To evaluate the biases, we first identified the dominant cluster for each criterion—the cluster containing the largest number of images—as the *bias direction*. We then calculated the normalized entropy of the distribution for each criterion of the occupation's images to determine the *bias intensity* score, following the method proposed by D'Incà et al. [19]:

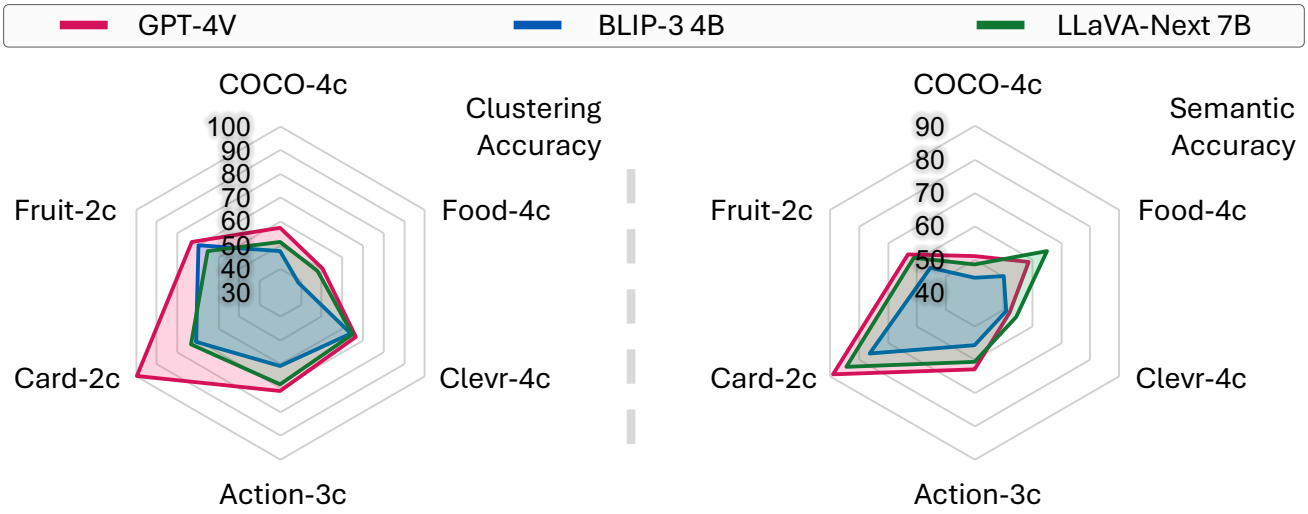
$$\mathcal{H}^l_{bias} = 1 + \frac{\sum_{c^l \in \mathcal{C}^l} \log(p(c^l|\mathcal{C}^l, \mathcal{D}_{\text{Occupation}}))}{\log(|\mathcal{C}^l|)} \quad (1)$$

where $\mathcal{D}_{\text{Occupation}}$ represents the generated images for each occupation, $\mathcal{C}^l$ denotes the clusters discovered under the l-th criterion, and $p(c^l|\mathcal{C}^l)$ is the probability of each cluster under the current distribution. The resulting score is bounded between $\mathcal{H}^l_{bias} \in [0, 1]$, where 0 indicates no bias towards a specific cluster (concept) under the evaluated criterion, and 1 indicates that the images are completely biased towards a particular cluster (concept) (*e.g.*, "Grey" hair color) within the current bias dimension (e.g., `Hair color`). We used the score defined in 1 to quantify the biases for each occupation across the 10 discovered grouping criteria. We report the bias intensity score for each occupation and each model across the 10 discovered grouping criteria in Fig. 11.

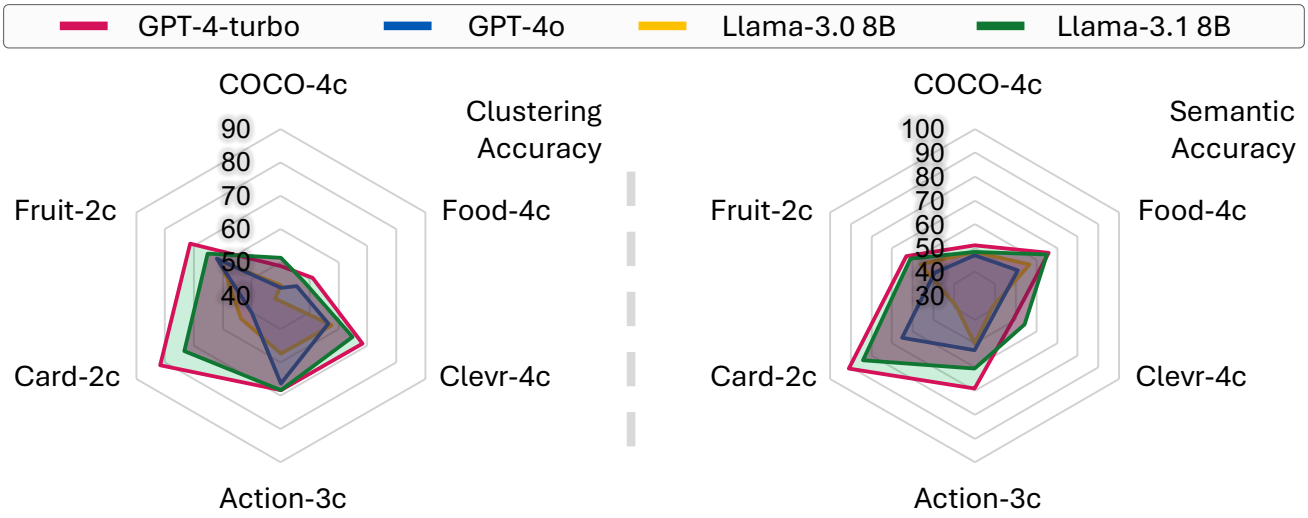
Human Evaluation Study Details: To assess the alignment between our method's predictions and human judgments on bias detection, we conducted a user study to gather human evaluation results for the generated images. As shown in the questionnaire example in Fig. 16, participants were presented with images generated by DALL-E3 and SDXL for each occupation and were asked to identify the bias direction (dominant class) for each of the 10 discovered criteria and rate the bias intensity on a scale from 0 to 10. We collected responses from 54 anonymous participants, resulting in 6 human evaluations for each occupation and each criterion.

The Absolute Mean Error (AME) between the bias intensity scores predicted by our system and those rated by humans (scaled to 0 to 1) was 0.1396. Additionally, our system's predicted bias directions aligned with human evaluations 72.3% of the time, with most discrepancies occurring in the criteria of "Age group," "Skin tone," and "Accessories worn." These findings indicate a strong correlation between our system's predictions and human judgments, validating the effectiveness of our approach. Detailed user study results are provided in § Q.1. We believe the discrepancies in certain criteria may be due to the influence of personal subjective cognition on respondents' answers. In Fig. 11, we present the human evaluation results, averaged across all participants for each model, occupation, and criterion, with the human ratings scaled from 0 to 1.

Complete Results and Additional Findings: In Fig. 11, we present the detailed bias detection results for each model, occupation, and criterion, alongside human evaluation scores for reference. A particularly interesting phenomenon emerges: *While DALL·E3 significantly outperforms SDXL on the* ***well-known*** *bias dimensions* (*e.g.*, `Gender`, `Race`, `Age`, and `Skin tone`), *both DALL·E3 and SDXL exhibit moderate to strong biases along the* ***novel*** *bias dimensions* (*e.g.*, `Hair color`, `Mood`, `Attire`, and `Accessories`).



(a) Sensitivity analysis of MLLMs
(Llama-3.1 8B is used as the LLM for all experiments)



(b) Sensitivity analysis of LLMs
(LLaVA-Next 7B is used as the MLLM for all experiments)

Figure 10. **Sensitivity analysis of different MLLMs and LLMs on the six OpenSMC benchmarks. Top (a):** We fix the LLM to Llama-3.1-8B and study the impact of different MLLMs. **Bottom (b):** We fix the MLLM to LLaVA-NeXT-7B and study the impact of different LLMs. The average clustering accuracy(%) across different criteria is reported on the **left**, while the average semantic accuracy(%) is reported on the **right**.

We speculate that DALL·E3's superior performance in mitigating well-known biases may be attributed to its "guardrails" [69], designed as part of its industrial deployment to avoid amplifying social biases via its easily accessible APIs. However, these guardrails do not prevent it from exhibiting biases along the novel dimensions discovered by our method, as these dimensions remain understudied. This observation highlights the importance of studying novel biases that could potentially exist in widely used T2I generative models to prevent further bias amplification.

Q.2. Further Details on Analyzing Social Media Image Popularity

With the rise of image-centric content on social media platforms like Instagram, Flickr, and TikTok, understanding what makes an image popular has become crucial for applications such as marketing, content curation, and recommendation systems. Traditional research often approaches image popularity as a regression problem [13, 72], utilizing metadata like hashtags, titles, or follower counts. However, the specific semantic visual elements that contribute to an image's popularity remain largely unexplored. In

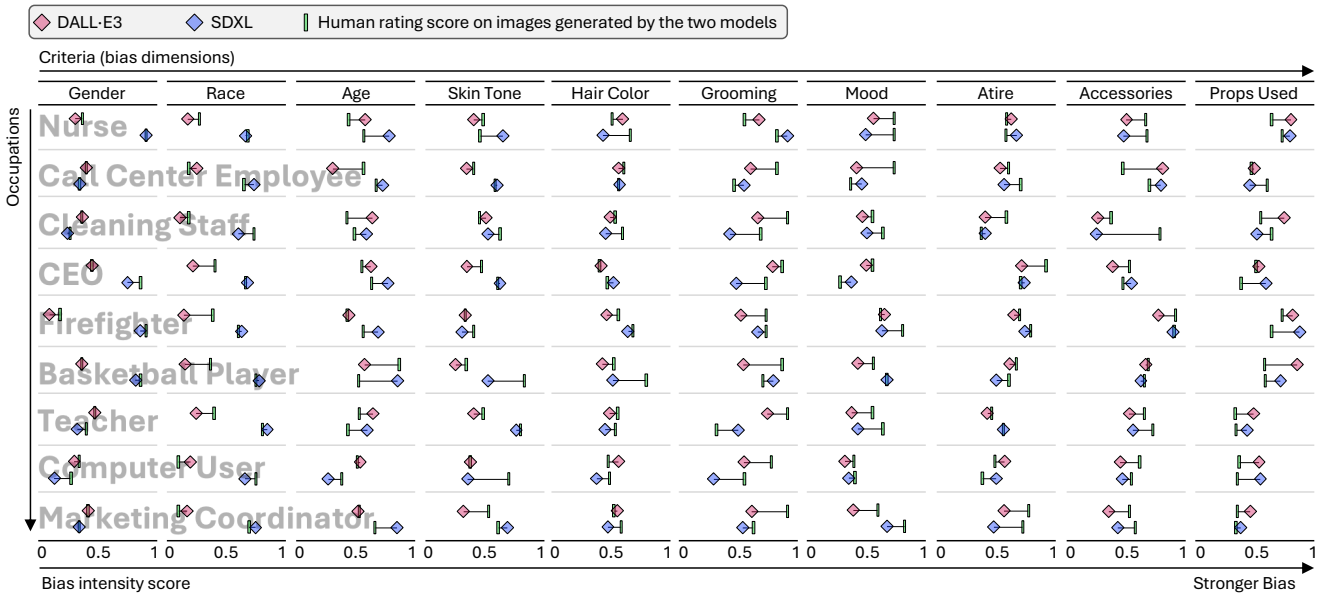


Figure 11. **Bias quantification results and human evaluation** for each occupation and criterion across the two studied T2I models, DALL·E3 and SDXL. The bias intensity score is reported.

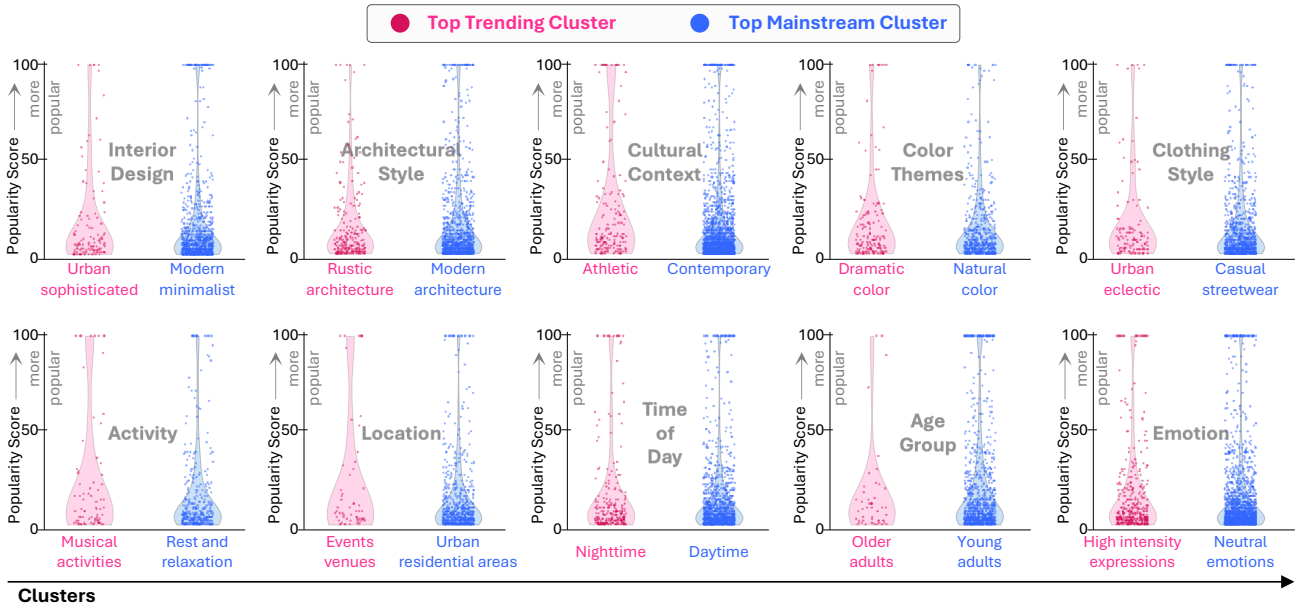


Figure 12. **Complete analysis of social media photo popularity on the SPID dataset.** We display the *Top Trending* and *Top Mainstream* clusters, along with the popularity distribution of data points within these clusters across all *ten* discovered criteria (in Grey).

this study, we applied our proposed method to automatically categorize social media images based on semantic visual elements across different dimensions (criteria). By analyzing these interpretable results alongside image popularity metrics (*e.g.*, number of views), we gained insights into the factors contributing to virality and identified common visual traits among popular images. These insights can provide valuable guidance for content creators and advertisers, enhancing productivity and informing strategic decision-making.

To expand on the discussion in Sec. 6 of the main paper, we present the complete findings across all ten discovered criteria in Fig. 12. Notably, we consistently observed a sharp semantic contrast between the visual elements in top trending images and those in the mainstream images across all ten criteria. For instance, there is a contrast between *Urban sophisticated* and *Modern minimalist* under `Interior Design`, *Rustic architecture* and *Modern architecture* under `Architecture Style`, and *Event venues* versus *Urban residential areas* under `Location`.

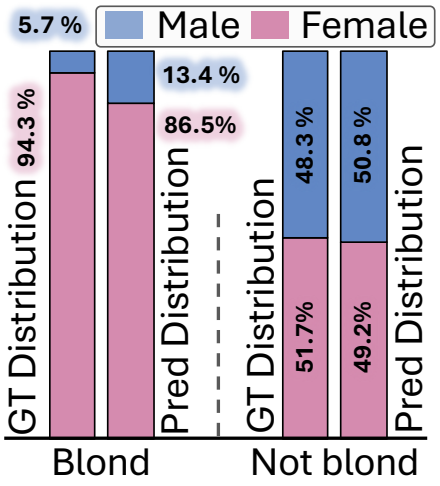


Figure 13. **Results of dataset bias discovery and mitigation.** Worst group and average accuracies(%) are reported.

This recurring observation reinforces the idea that viral (or trending) content tends to capture more attention, likely because it features novel, surprising, or striking visual elements. Humans are inherently attracted to stimuli that deviate from the norm [10, 74, 77]. On the other hand, widely uploaded yet "neutral" content is shared more often due to its familiarity and broad appeal, though it is less likely to provoke the strong emotional responses that fuel virality. We believe the insights generated by our method could offer valuable guidance to social media platform practitioners, helping them tailor their content more effectively to target audiences and gain a deeper understanding of social media image trends from various perspectives.

Q.3. Confirming and Mitigating Dataset Bias

Confirming and Mitigating Dataset Bias: Given an image collection that contains *spurious correlations* [26], we are curious whether we can proactively find this issue caused by data bias directly from the training images *without* relying on either the annotations [83] or *post hoc* misclassified images [39]. As a case study, we applied the proposed $\mathcal{X}$-Cluster framework to the 162k training images of the CelebA [56] dataset—a binary hair color classification dataset where the target label "Blond" is spuriously correlated with the demographic attribute "Female" in its training split.

Findings: As expected, our method identified the grouping criteria `Hair color` and `Gender`. Next, we analyzed the predicted gender distributions within the "Blond" and "Not Blond" (all other colors) clusters. As shown in Fig. 13, we observed that the gender distribution within the "Blond" cluster is highly skewed, with 86.5% of the images representing females, closely matching the ground-truth distribution (94.3%). Such an imbalance confirms the potential issue of spurious correlations between "Blond" and "Female". To further validate this observation, following B2T [39], we used our predicted distributions to train a debiased model with GroupDRO [83] and compared it with other unsupervised bias mitigation methods, including JTT [48], CNC [108], B2T, and GroupDRO trained with ground-truth labels. As shown in Tab. 33, our debiased model achieved robust performance, comparable to that of B2T, demonstrating the reliability of its discovered distributions.

Table 33. **Debiasing Results and Comparison on CelebA.** We use the groups discovered by $\mathcal{X}$-Cluster to train DRO and compare it with state-of-the-art debiasing methods. Additionally, we present DRO results using the ground-truth distribution (DRO+GT) for reference.

Method	Worst	Avg.
JTT	81.5	88.1
CNC	88.8	89.9
DRO+B2T	90.4	**93.2**
DRO+$\mathcal{X}$-Cluster	**90.9**	93.1
DRO+GT	89.7	93.6

Additional Evaluation: To further evaluate the prediction quality of our method for hair color and gender, we used the ground-truth labels from the CelebA dataset [56] to assess the classification accuracy of them. Our method achieved an impressive classification accuracy of 99.1% for gender and 87.4% for hair color on the 162,770 training images, demonstrating its effectiveness for uncovering gender and hair color substructures within the training set.

In addition, we quantified the *spurious correlation* between hair color and gender using the metric proposed by Yang et al. [103]. Specifically, given the correlated gender attribute distribution A and the target hair color distribution Y, we computed the normalized mutual information between A and Y to quantify the spurious correlation as:

$$I(A;Y) = \frac{2I(A;Y)}{H(A) + H(Y)} \tag{2}$$

where $H(A)$ and $H(Y)$ represent the normalized entropy of the gender and hair color distributions, respectively. A value of $H(A)$ or $H(Y)$ equal to 1 indicates a uniform distribution (*i.e.*, no class imbalance). We then used the ground-truth distribution from the dataset's labels and our predicted distribution to estimate the spurious correlation intensity using the score from 2. For gender and hair color, our method's predictions yielded a score of $I_{Pred} = 0.10$, which is nearly identical to the ground-truth score of $I_{GT} = 0.11$. This demonstrates that our method effectively identifies and confirms the bias directly from the training set.

Implementation Details of Training GroupDRO: To conduct debiased training using GroupDRO [83], we first used our predicted distribution to define four distinct training groups, rather than relying on the ground-truth distribution. We closely followed the training protocol outlined in B2T [39] and GroupDRO [83]. Specifically, we fine-tuned a ResNet-50 [31] model pre-trained on ImageNet [18], using the training split of the CelebA dataset [56]. The training was performed using the SGD optimizer [82] with a momentum of 0.9, a batch size of 64, and a learning rate of 1×10^{-5}. We applied a weight decay of 0.1 and set the group adjustment parameter to zero. The model was trained over 50 epochs. For evaluation, we reported both the average and worst-group test accuracies, selecting the model from the epoch that achieves the

highest worst-group accuracy on the validation set. The final evaluation and comparison results are provided in Tab. 33.

R. Why LLMs Improve Image Clustering?

The most compelling aspect of this work lies in our $\mathcal{X}$-Cluster framework's ability to transform large volumes of unstructured images into natural language and leverage the advanced text understanding and summarization capabilities of LLMs to tackle the challenging Open-ended Semantic Multiple Clustering (OpenSMC) task. This approach draws inspiration from the use of LLMs in the Topic Discovery task within the NLP domain [22]. Our core motivation is: "If LLMs can discover topics from documents and organize them, then by converting images into text, we can similarly use LLMs to organize unstructured images."

Traditional clustering methods [11, 23, 46, 90, 107] often depend on pre-defined criteria, pre-determined numbers of clusters, fixed feature representations (which require training), and are typically not interpretable. These limitations hinder their applicability to diverse datasets in open-world scenarios, as they demand significant human priors and retraining for each new dataset.

In contrast, LLMs [1, 60, 61, 68, 88] excel at understanding, summarizing, and reasoning over high-level semantics expressed in natural language across diverse domains (*e.g.*, everyday content, cultural knowledge, or medical content). Operating in a zero-shot [40], interpretable manner, LLMs are uniquely suited to the SMC task, which aims to discover meaningful and interpretable clustering criteria without requiring prior knowledge or training. By integrating LLMs with MLLMs [49] into the carefully designed $\mathcal{X}$-Cluster framework, we enable the discovery and refinement of clustering criteria directly from the dataset's content, followed by automatic grouping of the dataset. This design allows our framework to overcome the rigid assumptions of traditional clustering methods, making it automatic, generalizable, and training-free. Our approach provides a novel perspective, demonstrating how clustering tasks can evolve beyond traditional paradigms.

Challenges of employing LLMs to facilitate the SMC task. The main challenge of employing LLMs for the SMC task lies in accurately translating visual content from images into natural language that LLMs can effectively reason with. This is evident from the sensitivity analysis results in Supp. O: $\mathcal{X}$-Cluster's performance improves with larger or more powerful MLLMs (see Fig. 10 (a)), while it remains relatively insensitive to the specific choice of LLM (see Fig. 10 (b)). In other words, the quality of image captions generated by MLLMs is critical for the effective use of LLMs in the SMC task. Specifically, in the first stage of $\mathcal{X}$-Cluster (criteria proposal), captions need to be as comprehensive as possible to provide *rich* information for LLMs to discover grouping criteria. In the second stage (semantic grouping), criterion-specific captions should precisely capture relevant visual content to provide *accurate* information for assigning images to clusters.

To enhance caption quality, techniques such as MLLM model ensembling, prompt ensembling [50], or stronger models like GPT-4V [1] can improve comprehensiveness. For better precision, advanced prompting methods like CoT [100] or FineR [53] can capture nuanced details, while hallucination mitigation tools like Visual Fact Checker [25] can reduce noise caused by hallucinations. However, these techniques increase computational costs and framework complexity. In this work, we choose to keep $\mathcal{X}$-Cluster simple yet effective, and we outline these potential improvements for future practitioners.

S. Future Work

Closed-Loop Optimization. In this work, we designed our prompts following the Iterative Prompt Engineering methodology [16] introduced by Isa Fulford and Andrew Ng. In Supp. F, we provide the exact LLM and MLLM prompts used in our framework and break down each prompt to explain the objectives and purposes behind each design choice. These explanations cover elements such as system prompts, input formatting, task and subtask instructions, and output instructions. Our focus in this work is on creating a highly generalizable framework, $\mathcal{X}$-Cluster, and we do not perform any closed-loop, dataset-specific prompt optimizations. However, in future work or application scenarios where a labeled training/validation dataset is available, practitioners could build upon our design objectives. By leveraging our proposed evaluation metrics (see **??**) for each step, it would be possible to develop a Open-ended Semantic Multiple Clustering (OpenSMC) system with a closed-loop optimization pipeline to achieve improved performance.

$\mathcal{X}$-Cluster on Other Data Types. The core idea of our proposed framework, $\mathcal{X}$-Cluster, is to *use text as a proxy (or medium)* for reasoning over large volumes of unstructured data, generating human-interpretable insights at scale. As such, $\mathcal{X}$-Cluster can be directly applied to textual data (*e.g.*, documents). Moreover, since natural language is a highly versatile and widely-used medium of representation, $\mathcal{X}$-Cluster can be extended to other data types by converting these data into text (by replacing the captioning module with suitable tools) in future work, such as:

- **Audio Data:** Speech-to-Text models like Whisper [80] can convert audio data into text, enabling subsequent analysis with $\mathcal{X}$-Cluster.
- **Tabular Data:** Table-to-Text models, such as TabT5 [2], can translate tabular data into text, making it compatible with $\mathcal{X}$-Cluster. For tables containing figures, modern MLLMs like LLaVA-Next, which support both OCR and image-to-text capabilities, can handle these elements to create a unified textual representation for $\mathcal{X}$-Cluster.
- **Protein Structures:** Protein structure-to-text models, such as ProtChatGPT [96], can convert protein sequences into textual descriptions for analysis with $\mathcal{X}$-Cluster.
- **Point Cloud Data:** 3D captioning models, like Cap3D [57], can transform point cloud data or rendered 3D models into text, enabling their analysis using $\mathcal{X}$-Cluster.

We believe the versatile nature of $\mathcal{X}$-Cluster has the potential to open up a broad range of applications across diverse data modalities, fostering new directions in future research.

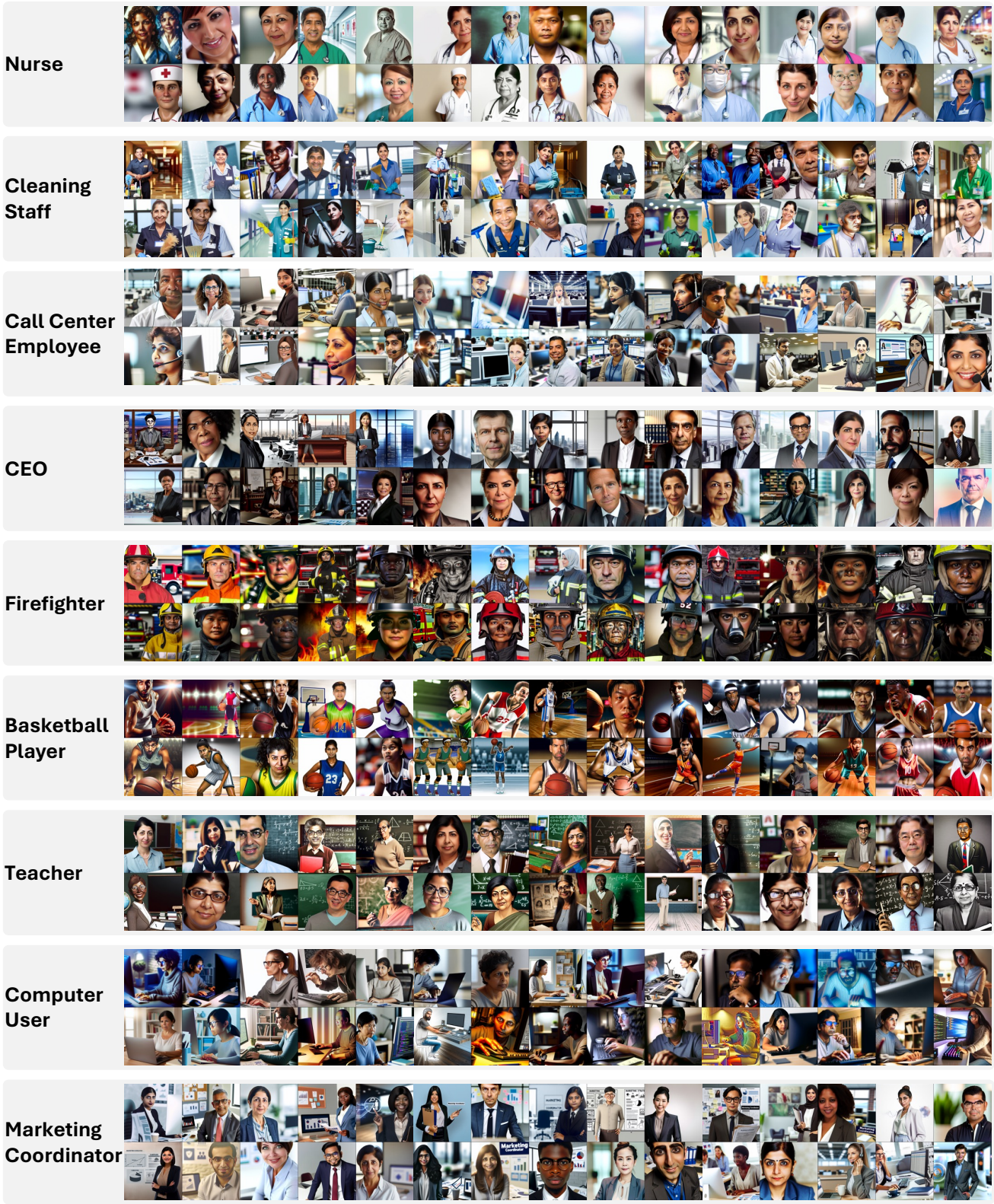


Figure 14. **Samples of DALL·E3 generated images.** For each occupation, the simple prompt "A portrait photo of a <OCCUPATION>", that does not reference any potential bias dimensions such as gender, race or hair color, is fed to DALL·E3 to generate 100 images. We present a random sample of 30 generated images.

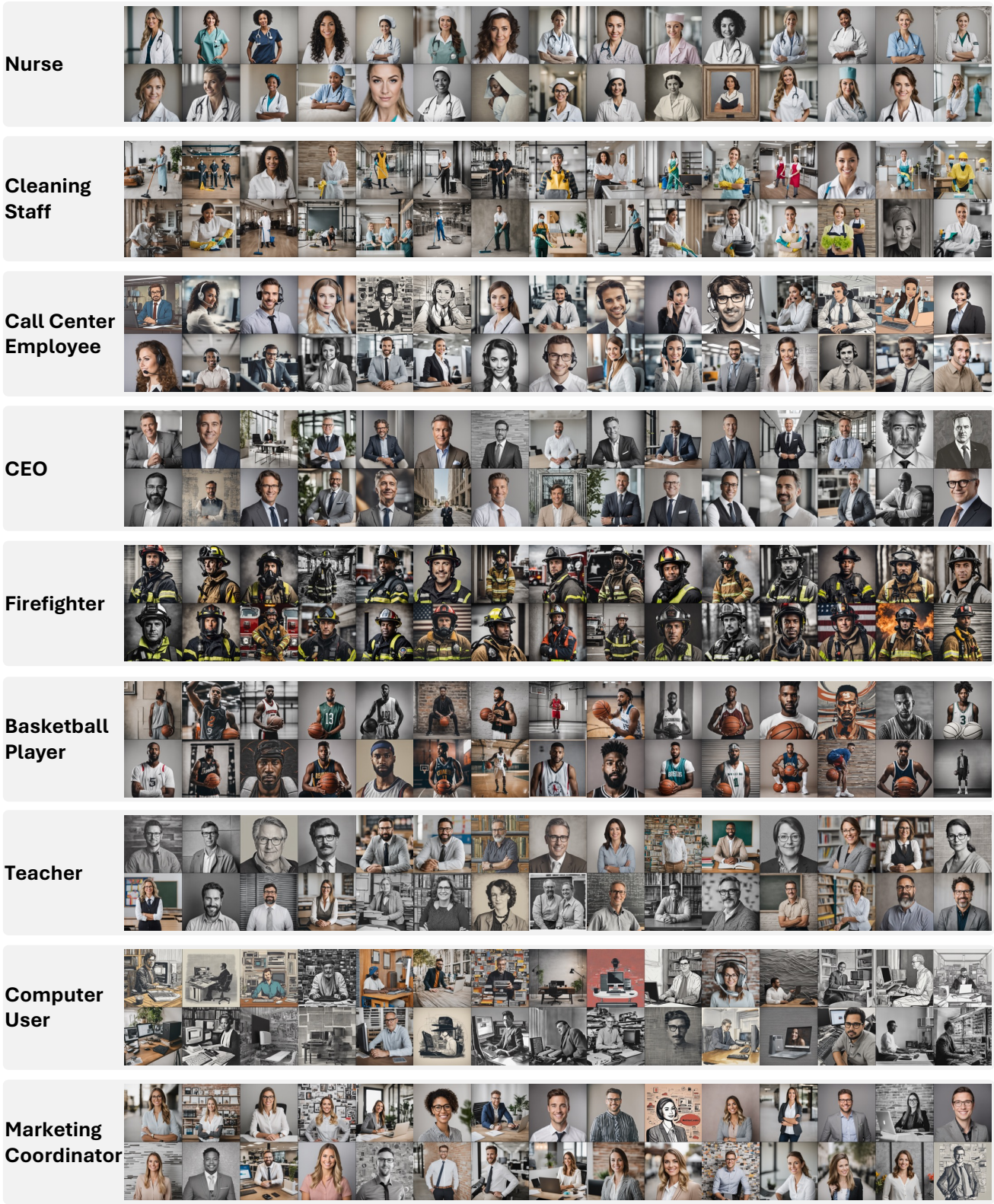
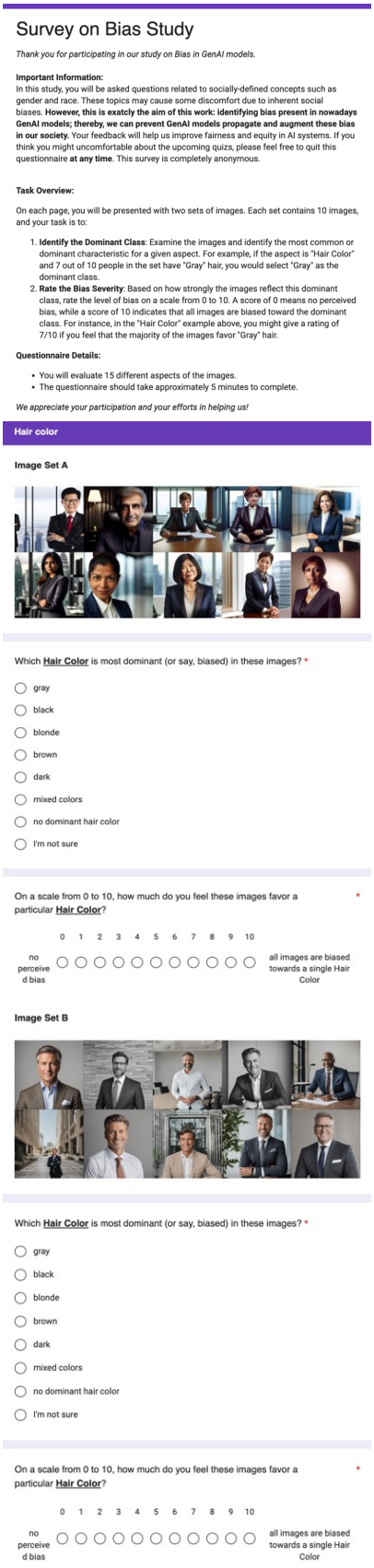


Figure 15. **Samples of SDXL generated images.** For each occupation, the simple prompt "A portrait photo of a <OCCUPATION>", that does not reference any potential bias dimensions such as gender, race or hair color, is fed to SDXL to generate 100 images. We present a random sample of 30 generated images.



Survey on Bias Study

Thank you for participating in our study on Bias in GenAI models.

Important Information:
In this study, you will be asked questions related to socially-defined concepts such as gender and race. These topics may cause some discomfort due to inherent social biases. **However, this is exatcly the aim of this work: identifying bias present in nowadays GenAI models; thereby, we can prevent GenAI models propagate and augment these bias in our society**. Your feedback will help us improve fairness and equity in AI systems. If you think you might uncomfortable about the upcoming quizs, please feel free to quit this questionnaire **at any time**. This survey is completely anonymous.

Task Overview:

On each page, you will be presented with two sets of images. Each set contains 10 images, and your task is to:

1. **Identify the Dominant Class**: Examine the images and identify the most common or dominant characteristic for a given aspect. For example, if the aspect is "Hair Color" and 7 out of 10 people in the set have "Gray" hair, you would select "Gray" as the dominant class.
2. **Rate the Bias Severity**: Based on how strongly the images reflect this dominant class, rate the level of bias on a scale from 0 to 10. A score of 0 means no perceived bias, while a score of 10 indicates that all images are biased toward the dominant class. For instance, in the "Hair Color" example above, you might give a rating of 7/10 if you feel that the majority of the images favor "Gray" hair.

Questionnaire Details:

- You will evaluate 15 different aspects of the images.
- The questionnaire should take approximately 5 minutes to complete.

We appreciate your participation and your efforts in helping us!

Hair color

Image Set A

Which Hair Color is most dominant (or say, biased) in these images? *

gray
black
blonde
brown
dark
mixed colors
no dominant hair color
I'm not sure

On a scale from 0 to 10, how much do you feel these images favor a particular Hair Color? *

0 1 2 3 4 5 6 7 8 9 10

no perceived bias — all images are biased towards a single Hair Color

Image Set B

Which Hair Color is most dominant (or say, biased) in these images? *

gray
black
blonde
brown
dark
mixed colors
no dominant hair color
I'm not sure

On a scale from 0 to 10, how much do you feel these images favor a particular Hair Color? *

0 1 2 3 4 5 6 7 8 9 10

no perceived bias — all images are biased towards a single Hair Color

Figure 16. **Example of the questionnaire for human evaluation study.**