

# Interactive Personalized Clustering: Navigating Beyond Rigid Criteria and Sample Ambiguity

Honglin Liu, Joey Tianyi Zhou, Peng Hu, Yunfan Li, Xi Peng

**Abstract**—Clustering methods have made significant strides in recent years, yet most of them follow an inherently static paradigm. Owing to the fixed optimization objective and the exclusive reliance on internal data, this static clustering paradigm faces two fundamental challenges: *i)* **Rigid Criteria**: prioritizing semantics under the dominant criterion while overlooking others, and *ii)* **Sample Ambiguity**: struggling to discriminate hard samples at cluster boundaries. Recognizing that these two challenges arise from the intrinsic data limitation, we propose incorporating external user interaction to guide the clustering process. To be specific, we present an Interactive Personalized Clustering paradigm (IPC) that consists of two stages, i.e., **Criterion Alignment** and **Grouping Refinement**, where each stage is dedicated to resolving one challenge. In the **Criterion Alignment** stage, IPC derives the representation aligned with the user-specified criterion via text basis transformation. Subsequently, in the **Grouping Refinement** stage, IPC identifies high-value sample pairs for friendly user feedback, utilizing three carefully designed losses to refine cluster assignments. Extensive experiments across 16 criteria demonstrate that our proposed IPC substantially outperforms state-of-the-art baselines, achieving a 23.6% average performance improvement. Furthermore, IPC exhibits robustness and generalizability across various LLMs and VLMs. The code is available at [XLearning-SCU/2026-TPAMI-IPC](https://github.com/XLearning-SCU/2026-TPAMI-IPC).

**Index Terms**—Deep Clustering, Personalized Clustering, Interactive Clustering.

## I. INTRODUCTION

**CLUSTERING** aims to group samples into different partitions sharing similar semantics. It plays a vital role in numerous real-world applications, including diverse domains such as bioinformatics [1], economics [2], medicine [3], physics [4], and social networks [5]. Traditional clustering methods [6], [7], which rely on shallow or hand-crafted features, often struggle to capture high-level semantics. Powered by the strong representation learning capability of deep neural networks, deep clustering methods [8], [9], [10], [11] have

This work was supported in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM610; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, 20826041K4049; in part by Sichuan Science and Technology Planning Project under Grant 2024NS-FTD0038; in part by China National Postdoctoral Program for Innovative Talents under Grant BX20250392.

Honglin Liu and Yunfan Li are with the College of Computer Science, Sichuan University, China. Peng Hu and Xi Peng are with the School of Artificial Intelligence, Sichuan University, China. Xi Peng is also with the Tianfu Jincheng Laboratory, China. E-mail: {tristanliuhl, penghu.ml, yunfanli.gm, pengx.gm}@gmail.com.

Joey Tianyi Zhou is with the Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore, and also with the Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore. E-mail: zhouty@cfar.a-star.edu.sg.

Corresponding Authors: Yunfan Li and Xi Peng.

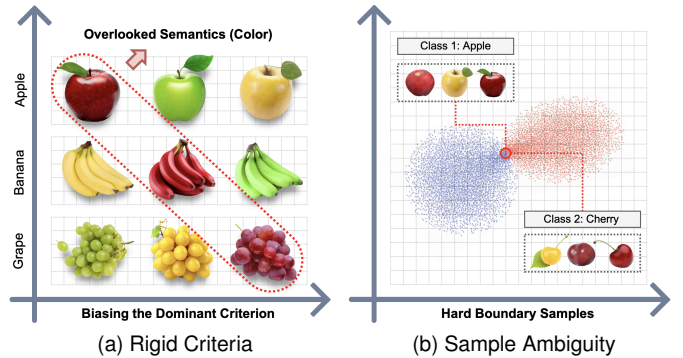


Fig. 1. Two key obstacles of the existing static clustering paradigm. (a) **Rigid Criteria**: In the absence of prior knowledge, the existing static clustering paradigm optimizes blindly for intra-cluster compactness and inter-cluster separability, which naturally drives it toward the most easily separable and discriminative “shape” or “category” semantics, while neglecting latent semantics under “color”, “scene”, and other criteria. (b) **Sample Ambiguity**: Constrained by the limited information of the data intrinsically, the existing static clustering paradigm suffers from a performance bottleneck, where easy samples are well separated, while hard boundary samples with subtle nuances remain challenging to distinguish.

gained unprecedented prominence. In recent years, inspired by the advancements in contrastive learning [12], [13], researchers have shifted towards employing data augmentations to construct self-supervised signals, thereby propelling the development of contrastive clustering frameworks [14], [15], [16], [17], [18].

Though achieving remarkable success, these clustering methods are inherently static, operating in an internally closed manner. We highlight that this static clustering paradigm encounters two key challenges: **Rigid Criteria** and **Sample Ambiguity**. As exemplified in Fig. 1, the static paradigm tends to bias toward dominant “category” semantics and fails to distinguish hard boundary samples, such as visually similar apples and cherries. Detailed elaboration on these two challenges is provided as follows: *i)* **Rigid Criteria**: On the one hand, the static clustering paradigm prioritizes the most salient semantics of images (typically “shape” or “category”), while neglecting other meaningful semantics (e.g., “color”, “count”, and “scene”). To be specific, owing to the absence of prior knowledge, the static clustering paradigm mindlessly pursues the objective of minimizing intra-cluster distances and maximizing inter-cluster separations. As expected, the model naturally optimizes toward the most easily achievable path, capturing the most salient “shape” or “category” semantics due to their significant discrepancies in both pixel and feature levels. This incurs the incapability of clustering under

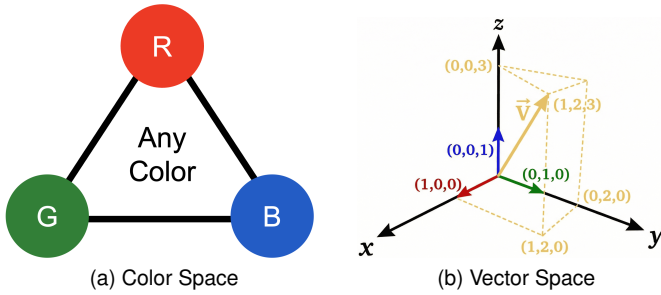


Fig. 2. Our observation of the basis concept across diverse criterion spaces. (a) Primary colors act as the basis for the color space, as any color can be obtained by mixing “red”, “green”, and “blue” in appropriate proportions. (b) Unit vectors span the Cartesian space, enabling any 3-D vector to be expressed as a linear combination of the  $x$ ,  $y$ , and  $z$  unit vectors. Extending this to a more generalized context, any criterion space can be spanned by its descriptive basis, allowing the semantics of a sample to be explicitly interpreted via the projection onto the basis. Motivated by this observation, we leverage the LLM to generate criterion-relevant textual descriptions serving as the basis, through which image features can be aligned to user-specified criteria.

other criteria. *ii) Sample Ambiguity:* On the other hand, the static clustering paradigm relies exclusively on uncovering semantics from the data itself. Nevertheless, according to the memorization effect [19] of neural networks, the model tends to first fit simple and distinct patterns, while remaining less sensitive to complex and ambiguous ones. For samples residing in overlapping semantic regions, the internal data alone fails to provide decisive discriminative cues. Without active intervention, the model’s capacity to resolve these hard boundary samples would reach an inherent plateau, leaving them improperly clustered. Inevitably, this static paradigm suffers from a performance bottleneck due to the limited information intrinsic to the data.

In fact, both the Rigid Criteria and Sample Ambiguity challenges stem from a common root: the static clustering paradigm is constrained to blindly discovering semantics from the data itself, which is akin to building a cart behind closed doors. Therefore, we propose to incorporate external user interaction to guide the clustering process, enabling clustering under different criteria flexibly and resolving the ambiguity of hard samples. Given that explicitly annotating class labels by the user is impractical, we resort to an affordable pairwise interaction, asking the user to determine whether a selected sample pair belongs to the same class.

To be specific, we present an interactive personalized clustering paradigm (IPC), which is comprised of two stages, i.e., Criterion Alignment and Grouping Refinement, with each stage targeting one challenge. In the Criterion Alignment stage, IPC aims to tackle the Rigid Criteria challenge through a training-free alignment process. The core intuition, as illustrated in Fig. 2, is that any semantic space can be spanned by its descriptive basis, much like primary colors compose the color space or unit vectors span the Cartesian space. Inspired by this, IPC first prompts an LLM to generate descriptive texts based on the user-specified criterion, which serve as a textual basis for the specific criterion space. By projecting image representations onto this basis, IPC obtains aligned representations that specifically emphasize the semantics required

by the user. Subsequently, IPC conducts the Grouping Refinement stage to resolve the Sample Ambiguity challenge. IPC first identifies high-value sample pairs by jointly evaluating their hardness, representativeness, and diversity. Specifically, representativeness is used to amplify the impact of user feedback, while diversity ensures that selections do not collapse into limited clusters, thereby minimizing interaction overhead. Upon receiving user feedback on these pairs, IPC refines the cluster assignments via dedicated positive and negative losses, complemented by a regularization term to prevent overfitting.

Overall, our main contributions can be summarized as follows:

- We propose interactive personalized clustering, an efficient and user-friendly paradigm that navigates beyond challenges of Rigid Criteria and Sample Ambiguity simultaneously.
- We present a frustratingly simple yet effective text basis transformation approach, which can convert image representations into criterion-specific ones that better capture the desired semantics.
- To keep the interaction cost within a human-acceptable level, we devise a valuable sample pair selection strategy that comprehensively considers their hardness, representativeness, and diversity.
- Extensive experiments on six datasets across 16 criteria demonstrate the robustness and effectiveness of our IPC, among which we contribute additional manual annotations for two widely-used benchmarks based on personalized criteria.

## II. RELATED WORK

In this section, we briefly introduce four relevant research areas, including deep clustering, multiple clustering, semi-supervised clustering, and interactive approaches.

### A. Deep Clustering

Benefiting from the powerful feature extraction ability, deep clustering methods [9], [10], [11], [20] have witnessed substantial advances, achieving impressive performance on large-scale and high-dimensional real-world data. In the past few years, advances in self-supervised learning [12], [13] have motivated a wave of contrastive clustering methods [14], [15], [16], [17], [21], which leverage data augmentation to construct strong self-supervision signals. Building on this framework, a pseudo-labeling mechanism [22], [23], [24], [25], [26] has been applied to utilize the high-confidence predictions to further boost the clustering performance.

Though making great progress, we point out that these deep clustering methods follow a static paradigm, resorting to blindly mining the internal semantics. As discussed in Section I, due to the optimizing path dependency and the information ceiling of internal data, this static clustering paradigm encounters two key obstacles: Rigid Criteria and Sample Ambiguity. In contrast, we propose incorporating external user interaction to capture the semantics under user-specified criteria and rectify hard samples at cluster boundaries. By efficient Criterion Alignment and Grouping Refinement, we overcome these two obstacles at a human-acceptable cost.

## B. Multiple Clustering

While most clustering methods yield only a single fixed partition, multiple clustering methods uncover several distinct and non-redundant groupings, enabling a richer understanding of hidden data structures from multiple perspectives. Traditional multiple clustering methods can be broadly categorized into two types. One begins with a given reference clustering and operates upon it to generate alternatives [27], [28], [29]. The other initiates with multiple candidate subspaces or objectives, discovering several distinct clusterings concurrently [30], [31], [32]. Nevertheless, these approaches, whether sequential or joint, generate outcomes that are uncontrollable and random to some extent. Inspired by the data augmentation in contrastive learning methods, researchers achieve clustering under specific criteria by tailoring the corresponding data augmentation [33], [34]. However, such methods rely heavily on one-by-one augmentation design for every criterion, making them hard to extend to complex real-world scenarios where numerous potential criteria may exist.

The advent of Large Language Models (LLMs) and Vision-Language Models (VLMs) opens up a promising direction for multiple clustering, as they are capable of interpreting free-form language instructions. The most straightforward application is to directly query LLMs and VLMs to assign class labels to samples under a given criterion [35], [36], [37]. Clearly, this naive approach could lead to a prohibitive computational cost [38]. Besides, due to the strong inductive bias inherent in the pre-trained LLMs and VLMs, direct optimization to achieve multiple clustering [39] is challenging in practice. In contrast to these methods, our proposed IPC is effective and scalable. By querying the LLM only once to establish the textual basis for a specific criterion, IPC achieves seamless criterion alignment through text basis transformation, which is training-free.

## C. Semi-supervised Clustering

Unlike standard unsupervised clustering, semi-supervised clustering integrates a limited amount of supervision signal to instruct the partitioning process, thereby significantly boosting the clustering performance. Based on the form of the supervision signal, these methods can be divided into label-wise and pair-wise types. The former exploits a small set of labeled samples to directly supervise cluster assignments [40], [41], [42], while the latter introduces supervision through relational constraints, i.e., must-link and cannot-link, to provide auxiliary information [43], [44], [45].

Despite the advantages of semi-supervised clustering, existing label-wise and pair-wise methods still face inherent limitations. To be specific, on the one hand, label-wise methods are impractical for real-world scenarios, where we may only access the clustering criterion rather than explicit class labels, or involve a large number of categories, making full annotation prohibitively expensive. On the other hand, due to the lack of effective selection strategies, most pair-wise methods require a large number of pairs to capture the desired semantics. This also leads to high annotation costs, especially for clustering under non-salient criteria. In contrast,

our proposed IPC aligns representations with the user-specified clustering criterion from the outset, and further identifies high-value sample pairs for user-friendly interaction, efficiently achieving the ideal clustering result.

## D. Interactive Approaches

Interactive approaches have been explored in a wide range of areas, where user command or feedback plays a central role to steer model behavior, improving model outputs or representations. For instance, in retrieval fields, recent works [46], [47] leverage multimodal interaction to refine retrieval results, enabling more flexible and accurate search compared to static similarity matching. Beyond retrieval, interactive vision editing methods [48], [49] illustrate how minimal user instruction can effectively guide complex visual transformations. Moreover, by introducing expert domain knowledge, interactive paradigms help ensure that medical model outputs align with clinical requirements [50], [51].

Taken together, these studies highlight the great potential of applying interaction to improve model performance or efficiency. In contrast, our proposed IPC explores leveraging interaction to play a dual role: it not only improves the performance of distinguishing hard samples, but also enables the capability of clustering under arbitrary criteria, which is unattainable for static clustering paradigms.

## III. METHOD

In this section, we illuminate the proposed Interactive Personalized Clustering (IPC) paradigm, with details of its two stages, i.e., Criterion Alignment and Grouping Refinement. As Fig. 3 depicts, in the Criterion Alignment stage, IPC begins by prompting the LLM with the user-specified criterion to obtain relevant descriptive texts, thereby establishing the text basis. After that, IPC aligns the image representation with the specific criterion through text basis transformation. Following this, IPC proceeds to Grouping Refinement, where it queries the user on the class consistency of carefully selected high-value sample pairs. These pairs are identified by integrating their hardness, representativeness, and diversity, where the latter two are introduced to reduce the interaction cost. Finally, based on the resulting user feedback, IPC utilizes the positive, negative, and regularization losses to refine the cluster assignments.

### A. Criterion Alignment

1) *Basis Construction*: In mathematics, a basis is defined as a set of vectors that span the entire space, typically requiring linear independence<sup>1</sup>. For instance, the set of canonical vectors  $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$  constitutes a basis of the three-dimensional Cartesian coordinate system, as any vector in this space can be represented as a linear combination of these components. Similarly, the primary colors, i.e., “red”, “green”, and “blue”, form a basis for the color space, as these three colors are capable of covering the entire gamut of hues. From a

<sup>1</sup>In this paper, we do not enforce strict linear independence, permitting redundancy in the constructed basis.

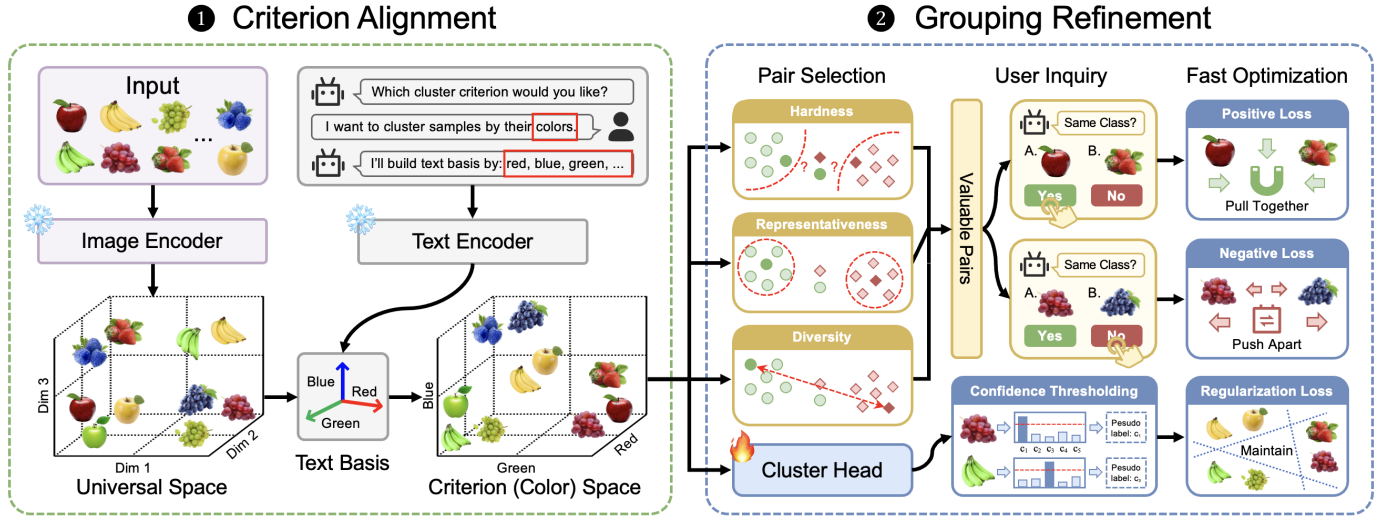


Fig. 3. The framework of our proposed IPC, which consists of Criterion Alignment and Grouping Refinement stages. In the Criterion Alignment stage, IPC first interacts with the user to confirm the desired criterion, leveraging the LLM to acquire related descriptive texts that serve as the basis. Further, IPC utilizes this text basis to project the original universal feature into the criterion space, obtaining the criterion-aligned feature. Then, in the Grouping Refinement stage, IPC mines the most valuable pairs by integrating their hardness, representativeness, and diversity. Then, IPC solicits the user to determine the class consistency for each highly valuable pair. Finally, based on the user feedback, IPC utilizes positive and negative losses to refine the grouping, together with a regularization term to maintain cluster boundaries.

broader perspective, a descriptive basis can be used to span the criterion space, with each sample’s semantics reflected by its projection onto this basis. Building on this insight, we leverage an LLM to generate criterion-specific textual descriptions, functioning as a basis, enabling image features to be aligned with user-specified criteria.

To be specific, IPC first confirms the clustering criterion  $C$  with the user. After that, IPC asks the LLM to produce the descriptive texts  $\mathbf{W} = \{w_1, w_2, \dots, w_t\}$  related to  $C$ , namely,

$$\mathbf{W} = \text{LLM}(P_1, C), \quad (1)$$

where  $P_1$  denotes the LLM prompt template constructed as follows:

Please generate common expressions to describe the  $C$ , as many as possible.

The full prompts can be found in the supplementary material.

Based on the above-mentioned prompt, the LLM generates semantically relevant texts  $\mathbf{W}$  aligned with the given criterion, converting the abstract criterion into specific descriptive content.

The descriptive texts  $\mathbf{W}$  are then fed into the VLM text encoder to compute their normalized representation  $\mathbf{T}$ , namely,

$$\mathbf{T} = \text{VLM}_{\text{text}}(P_2, C, \mathbf{W}), \quad (2)$$

where the VLM prompt  $P_2$  is constructed as follows:

Objects with the  $C$  of  $w_1$ .  
 Objects with the  $C$  of  $w_2$ .  
 ...  
 Objects with the  $C$  of  $w_t$ .

Notably, the word “Objects” could be substituted with more specific descriptions if we can access the prior of the dataset.

By introducing the text representation  $\mathbf{T}$  as the basis, the semantic space of the user-specified criterion becomes

explicitly structured. This stands in contrast to the original universal space, which lacks a semantically transparent basis, while each dimension of  $\mathbf{T}$  conveys a clear and interpretable real-world meaning.

2) *Representation Projection*: Once the text basis is acquired, IPC utilizes it to align the image representation with the user-specified criterion.

Specifically, IPC begins by encoding the images  $\mathbf{I}$  with the VLM image encoder, yielding their normalized feature representation  $\mathbf{X}$ , namely,

$$\mathbf{X} = \text{VLM}_{\text{image}}(\mathbf{I}). \quad (3)$$

After that, IPC aligns the image representation  $\mathbf{X}$  with the user-specified criterion  $C$  via text basis transformation, as given by the following equation:

$$\mathbf{E} = \mathbf{X}\mathbf{T}^\top, \quad (4)$$

where  $\mathbf{E}$  signifies the criterion-aligned representation. The rationality behind this manipulation lies in the coherence between visual and textual modalities within the VLM embedding space.

To present an intuitive understanding of Criterion Alignment, we exemplify it through alignment with the “color” criterion, as shown in Fig. 3.

As discussed in Section I, the original image representation  $\mathbf{X}$  is dominated by the most significant “shape” or “category” semantics, which could lead to a poor clustering result under the “color” criterion.

To capture the desired semantics, IPC first prompts the LLM to generate descriptions corresponding to the “color” criterion. For simplicity, we assume that the LLM outputs descriptive texts  $\mathbf{W}$  which contain only 3 words {“red”, “green”, “blue”}. From Eq. (2), IPC can establish the text basis as

$$\mathbf{T} = [t_1^\top, t_2^\top, t_3^\top]^\top, \quad (5)$$

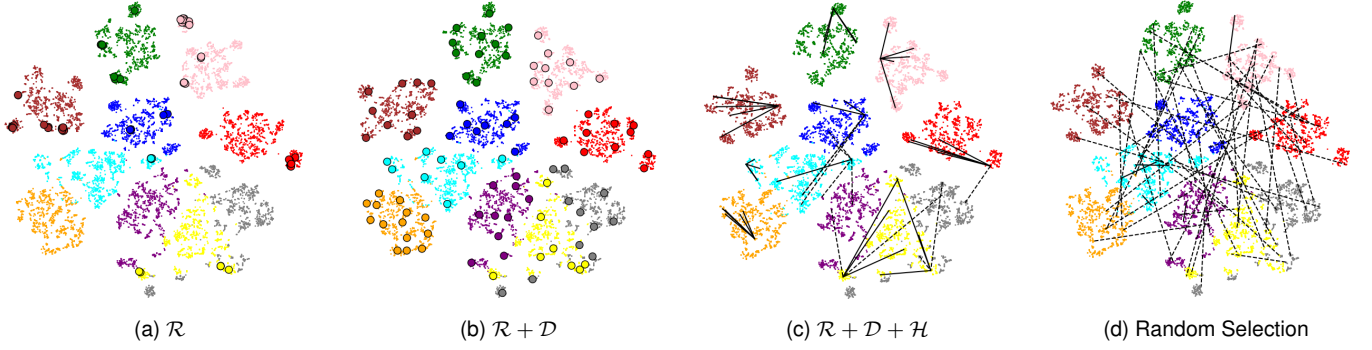


Fig. 4. Visualizations on the “color” criterion for the Clevr4-10k dataset, where solid and dashed lines denote pairs with identical and distinct labels, respectively. (a) Samples selected by representativeness solely, which tend to collapse into a few dense clusters. (b) Valuable samples selected by representativeness and diversity. (c) Valuable pairs selected by IPC, consisting of dispersed intra-class pairs and adjacent inter-class pairs. (d) Randomly selected pairs.

where  $\{t_1, t_2, t_3\}$  denote the VLM textual representations of “red”, “green”, and “blue”, respectively.

Then, IPC projects  $\mathbf{X}$  into the “color” criterion space spanned by  $\mathbf{T}$ . For the  $i$ -th image  $x_i$ , IPC can obtain the criterion-aligned representation  $e_i$  via Eq. (4), i.e.,

$$e_i = x_i \mathbf{T}^\top = [x_i \cdot t_1, x_i \cdot t_2, x_i \cdot t_3]. \quad (6)$$

As illustrated in Eq. (6), the three dimensions of  $e_i$  represent its similarity to the “red”, “green”, and “blue” representations, respectively. Given the consistency of the VLM visual and textual modalities, the  $i$ -th image characterized primarily by “red” will exhibit a larger value in the first dimension of  $e_i$  compared to the others. Consequently,  $e_i$  offers superior expressiveness over  $x_i$  regarding the “color” criterion. In effect, this completes the alignment of the original image representation with the “color” criterion.

## B. Grouping Refinement

1) *Value Mining*: During the initial Criterion Alignment stage, IPC obtains representations aligned with the user-specified criterion. The criterion-aligned representations are then passed to the Grouping Refinement stage, where a two-layer clustering head is employed to facilitate user interaction and further polish the cluster assignments.

As explicitly annotating class labels imposes an excessive burden on users, IPC adopts pair-wise user interaction to address challenges of rigid criteria and sample ambiguity. To ensure interaction costs remain practically feasible, IPC mines the highly valuable sample pairs, soliciting binary user feedback regarding their group membership. The value of a sample pair is assessed by three metrics: hardness, representativeness, and diversity. We provide a detailed interpretation of these three metrics as follows:

**Hardness.** As a general rule, sample pairs with significant differences tend to be assigned to distinct clusters, while those exhibiting high similarities are more likely to be grouped together. On the contrary, pairs near a certain similarity threshold are the most prone to confusion, which are also the most likely to be hard boundary sample pairs. Therefore, IPC prioritizes the selection of these hard pairs to improve

the clustering performance. Formally, the hardness metric is defined as follows:

$$\mathcal{H}(i, j) = -|e_i \cdot e_j - \tau_{sim}|, \quad (7)$$

where  $e_i$  is the criterion-aligned representation of the  $i$ -th sample, and  $\tau_{sim}$  is the  $1/K$ -th ( $K$  is the number of classes) quantile of the similarity distribution, rather than a fixed scalar. Intuitively, a larger hardness value indicates greater uncertainty in determining the cluster consistency for a sample pair. As Fig. 4c shows, these pairs primarily fall into two types: dispersed intra-class pairs and adjacent inter-class pairs, which are in line with the choice of  $\tau_{sim}$ .

Despite the benefits of rectifying hard pairs, the quadratic computational complexity  $\mathcal{O}(N^2)$  of the hardness metric remains a significant bottleneck, hindering its scalability to large-scale real-world datasets. To overcome this limitation, IPC circumvents exhaustive computation by first selecting valuable samples and then identifying their counterparts based on the hardness metric.

**Representativeness.** To maximize the utility of user feedback, IPC favors samples residing in dense regions. The rationale is that a single correction within such a region can propagate to a multitude of adjacent samples, thereby acting as a force multiplier for each user interaction. Mathematically, the representativeness for the  $i$ -th sample is defined by the distances of its neighbors as follows:

$$\mathcal{R}_i = -\log \sum_{l=1}^Q \|e_i - e_{i(l)}\|_2^2, \quad (8)$$

where  $e_{i(l)}$  denotes the  $l$ -th nearest neighbor of  $e_i$ , and  $Q$  is the number of nearest neighbors set empirically to 20. A larger representativeness value signifies a more compact local structure that implies a greater potential corrective impact.

**Diversity.** Empirical observations reveal, however, that an excessive focus on representativeness can cause the selection to converge within a narrow subset of clusters, leading to a severe distributional imbalance, as shown in Fig. 4a. To address this problem, IPC incorporates the diversity metric to encourage a more uniform distribution of selection among different clusters. Notably, this metric is implemented via a

greedy selection process, where the latest selection is conditioned on the deviation of all prior selections. To be specific, in each iteration, the diversity of the  $i$ -th sample is computed as follows:

$$\mathcal{D}_i = \min_{s \in \mathbf{S}} \log(1 - e_i \cdot e_s), \quad (9)$$

where  $\mathbf{S}$  signifies the indices of all prior selections.

By synthesizing the representativeness and diversity metrics, IPC derives the overall value of the  $i$ -th sample:

$$\mathcal{V}_i = \mathcal{R}_i + \mathcal{D}_i. \quad (10)$$

To bolster robustness against potential sampling bias, IPC further pairs each high-value sample with its  $B$  hardest counterparts based on their hardness.

The following theorem upholds the optimality of IPC, guaranteeing that samples with the highest values are consistently prioritized. The complete justification is deferred to the Supplementary Material.

**Theorem 1.** *IPC sequentially selects samples in descending order of value, i.e.,*

$$\mathcal{V}_{s_m}^m \geq \mathcal{V}_{s_{m+1}}^{m+1}, \quad \forall m \in [1, M/B - 1], \quad (11)$$

where  $s_m$  denotes the index of the  $m$ -th selected sample, and  $\mathcal{V}_i^m$  denotes the value of the  $i$ -th sample in the  $m$ -th selection.

In summary, IPC operates via a two-step process to mine the top  $M$  valuable pairs. It first selects  $M/B$  high-value anchor samples by maximizing Eq. (10). Subsequently, for each selected anchor, IPC identifies its  $B$  counterparts from the entire dataset with the  $B$  largest hardness scores via Eq. (7), thereby forming the final set of high-value sample pairs  $\mathbf{P}$ .

2) *Fast Optimization.* For each identified highly valuable pair, IPC solicits the user to determine whether the two samples in this pair belong to the same class (positive feedback  $f_+$ ), or different classes (negative feedback  $f_-$ ).

Based on positive and negative feedback, IPC employs the corresponding positive loss  $\mathcal{L}_p$  and negative loss  $\mathcal{L}_n$  to provide supervisory signals. Besides, to prevent overfitting, a regularization term  $\mathcal{L}_r$  is incorporated. The overall loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_n + \mathcal{L}_r, \quad (12)$$

where three loss terms are detailed below.

**Positive Loss.** For positive feedback, IPC pulls the two samples in the pair closer by the following positive loss:

$$\mathcal{L}_p = -\frac{1}{2|f_+|} \sum_{(i,j) \in f_+} \sum_{c=1}^K (p_i^c \log p_j^c + p_j^c \log p_i^c), \quad (13)$$

where  $|f_+|$  refers to the number of positive feedback pairs,  $K$  is the number of classes, and  $p_i^c$  denotes the probability of the  $i$ -th sample belonging to the  $c$ -th cluster.

**Negative Loss.** For negative feedback, IPC pushes the two samples in the pair apart by the following negative loss:

$$\mathcal{L}_n = -\frac{1}{2|f_-|} \sum_{(i,j) \in f_-} \sum_{c=1}^K (p_i^c \log(1 - p_j^c) + p_j^c \log(1 - p_i^c)), \quad (14)$$

---

### Algorithm 1 Interactive Personalized Clustering (IPC)

---

**Input:** LLM prompt  $P_1$ , VLM prompt  $P_2$ , image dataset  $\mathbf{I}$ , dataset size  $N$ , number of pairs to be selected  $M$ , number of counterparts per sample  $B$ , cluster head  $h$ .

**Output:** Clustering results.

```
// Criterion Alignment
1: The user specifies the clustering criterion  $C$ .
2: Ask the LLM to generate the descriptive texts  $\mathbf{W}$  related to  $C$  via Eq. (1).
3: Compute the text basis  $\mathbf{T}$  via Eq.(2).
4: Compute the original image representation  $\mathbf{X}$  via Eq. (3).

5: Transform  $\mathbf{X}$  into the criterion-aligned representation  $\mathbf{E}$  based on  $\mathbf{T}$  via Eq. (4).
// Grouping Refinement
6: Initialize the selected pair indices set  $\mathbf{P} = \{\}$ , the selected sample indices set  $\mathbf{S} = \{\}$ .
7: for  $i \in [1, N]$  do
8:   Compute the representativeness score  $\mathcal{R}_i$  by Eq. (8).
9:   Initialize the diversity score  $\mathcal{D}_i = 0$ , since no sample has been selected.
10:  Compute the value score  $\mathcal{V}_i$  by Eq. (10).
11: end for
12: for  $m \in [1, M/B]$  do
13:   Select the  $s_m$ -th sample that has the highest value outside  $\mathbf{S}$ , and update  $\mathbf{S} = \mathbf{S} \cup \{s_m\}$ .
14:   Update the diversity score by Eq. (9) and the value score by Eq. (10) for samples outside  $\mathbf{S}$ .
15:   Select  $B$  samples that have the  $B$  highest hardness scores with the  $s_m$ -th sample via Eq. (7), forming pairs with the  $s_m$ -th sample and adding them to  $\mathbf{P}$ .
16: end for
17: For each pair in  $\mathbf{P}$ , solicit the user to determine whether the two samples in the pair belong to the same cluster.
18: Utilize the user feedback to optimize the cluster head  $h$  via Eq. (12).
19: Obtain the clustering results by  $h$ .
```

---

where  $|f_-|$  refers to the number of negative feedback pairs.

**Regularization Loss.** Due to the extremely small number of selected pairs ( $M = 50/500$ ), compared to the entire dataset ( $N(N-1)/2$ ), there exists an overfitting risk if only the two aforementioned loss terms are used. Therefore, to safeguard the overall cluster boundaries, IPC utilizes the following regularization loss to retain confident predictions:

$$\mathcal{L}_r = -\frac{\sum_{i=1}^N \mathbb{1}[p_i^{\hat{c}} > \tau] \log p_i^{\hat{c}}}{\sum_{i=1}^N \mathbb{1}[p_i^{\hat{c}} > \tau]}, \quad \hat{c} = \operatorname{argmax}_i p_i, \quad (15)$$

where  $\mathbb{1}[\text{condition}]$  denotes the indicator function that equals 1 if condition is true and 0 otherwise,  $\tau = 0.9$  is the confidence threshold.

The complete procedure is outlined in Algorithm 1.

Last but not least, it should be pointed out that IPC is fast and scalable:

- The Criterion Alignment stage is inherently training-free. Formulated as a text basis transformation, it is computa-

tionally equivalent to a single matrix multiplication, thus ensuring high operational efficiency.

- The pair selection and user annotation process is a one-time procedure performed at initialization. Moreover, the number of sample pairs queried is kept minimal (50 or 500), which is a negligible fraction of  $O(N^2)$  possible combinations, making IPC user-friendly and scalable.
- The optimization is fast and lightweight. Only the parameters of the two-layer cluster head are updated, while the VLM backbone remains frozen.

#### IV. EXPERIMENTS

In this section, we first introduce the experimental benchmarks and evaluation metrics, followed by the implementation details of the proposed IPC. Then, we assess its performance compared with state-of-the-art clustering baselines. Furthermore, we conduct thorough ablation studies and parameter analysis to demonstrate its effectiveness and robustness.

##### A. Datasets and Evaluation Metrics

We evaluate IPC on 6 benchmarks that contain 16 clustering criteria, as listed in Table I. To be specific, Clevr4-10k [52] is a synthetic 3D modeling dataset consisting of 10,531 samples and categorized by four distinct data partition criteria, i.e., “texture”, “shape”, “color”, and “count”. Cards [33] comprises 8,029 poker card samples and is structured around two criteria: “number” and “suits”. Fruit360 [33] consists of 4,856 fruit images, organized along two criteria: “color” and “species”. CIFAR-10 [53] is a classic image dataset consisting of 60,000 low-resolution images in ten classes, where the samples can be organized by the naive “object” criterion or the “scene” criterion introduced in [54]. Stanford Cars [55] is a fine-grained dataset for vehicle recognition, containing 16,185 images annotated with the original “brand” criterion. Additionally, we provide manual annotations for “color” and “type” criteria to facilitate personalized clustering. CUB [56] is a specialized dataset for bird species recognition, containing 11,788 challenging bird images. In addition to the original “species” labels, we introduce manual annotations for “action” and “scene” criteria to facilitate clustering along diverse semantic dimensions.

Following standard clustering practice, we evaluate performance using three widely used metrics: Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI), with higher scores indicating better performance.

##### B. Implementation Details

In the Criterion Alignment stage, we adopt GPT-4o [57] as the default LLM to generate the descriptive texts and report the results for other LLMs in Section IV-D. Subsequently, in the Grouping Refinement stage, we add a two-layer cluster head  $h$  to the criterion-aligned representation and update the parameters of  $h$  only. Moreover, we train for 100 epochs using the Adam optimizer, with the initial 10 epochs serving as a warm-up that only uses the regularization loss. The number of pairs to be selected  $M$  is set to 500 by default, but is reduced to

TABLE I  
A SUMMARY OF THE USED DATASETS.

| Dataset       | Criteria (Number of Classes)                 | Size   |
|---------------|----------------------------------------------|--------|
| Clevr4-10k    | Texture(10), Shape(10), Color(10), Count(10) | 10,531 |
| Cards         | Number(13), Suits(4)                         | 8,029  |
| Fruit360      | Color(4), Species(4)                         | 4,856  |
| CIFAR-10      | Object(10), Scene(2)                         | 60,000 |
| Stanford Cars | Brand(196), Color(8), Type(8)                | 16,185 |
| CUB           | Species(200), Action(6), Scene(9)            | 11,788 |

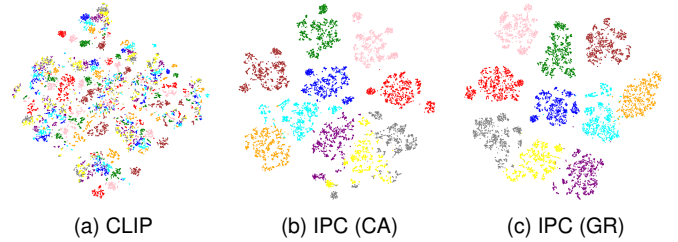


Fig. 5. Visualizations on the “color” criterion for the Clevr4-10k dataset, showing results from CLIP, IPC after Criterion Alignment (CA), and IPC after Grouping Refinement (GR). To ensure an objective visualization, the t-SNE plots are colored by the ground-truth labels rather than the commonly used predicted clusters.

50 for the Fruit360 and CIFAR-10 datasets due to their relative simplicity. Besides, the number of counterparts per sample  $B$  is set to 5 by default. To make a fair comparison, we adopt the ViT-B/32 [58] version of CLIP [59] as the backbone for our IPC and all baselines. We also report the results for other VLM backbones in Section IV-D. In addition, the parameters of the baselines remain the same as in their own settings. Thanks to the lightweight design of IPC, we find it sufficient to complete all experiments through a single NVIDIA RTX 3090 GPU.

##### C. Main Results

To evaluate the performance of the proposed IPC, we conduct a comparison against nine representative baselines: CLIP [59], TAC [60], TEMI [61], Multi-Map [39], CMClusters [34], Cop-Kmeans [43], DCGMM [44], Vol-MaxDCC [45], and PCL [54]. Among them, the former five are unsupervised, while the others are semi-supervised.

As shown in Table II, IPC achieves remarkable clustering results, with a 23.6% improvement in average performance over the previous state-of-the-art method Multi-Map. Moreover, previous methods exhibit pronounced instability across various criteria, whereas IPC achieves consistent and substantial gains on all of them, underscoring its superior robustness and generalizability.

The significant performance gap can be attributed to two main factors. Firstly, IPC devises an easy yet effective Criterion Alignment mechanism to obtain criterion-aligned representation, while prior methods directly manipulate the original representation. To show the power of this mechanism, we take the “color” criterion for the Clevr4-10k dataset as an example. As illustrated by the t-SNE [62] visualization in Fig. 5, the cluster distributions have been significantly improved by this mechanism, exhibiting clearer separation and more compact

TABLE II  
CLUSTERING PERFORMANCE ON MULTI-CRITERIA DATASETS.

| Dataset       | Criterion | Metric | CLIP  | TAC   | TEMI  | Multi-Map | CM-Clusts | Cop-Kmeans | DC-GMM | VolMax-DCC | PCL   | IPC (ours) |
|---------------|-----------|--------|-------|-------|-------|-----------|-----------|------------|--------|------------|-------|------------|
| Clevr4-10k    | Texture   | NMI    | 1.16  | 3.68  | 0.66  | 3.77      | 15.00     | 1.00       | 5.62   | 2.79       | 0.18  | 22.63      |
|               |           | ACC    | 13.13 | 17.19 | 12.63 | 17.25     | 21.91     | 12.90      | 17.18  | 15.51      | 11.34 | 35.52      |
|               |           | ARI    | 0.43  | 1.81  | 0.25  | 1.81      | 6.24      | 0.35       | 2.48   | 1.35       | 0.10  | 14.56      |
|               | Shape     | NMI    | 73.90 | 72.85 | 80.13 | 67.48     | 91.76     | 73.18      | 72.44  | 77.79      | 91.35 | 91.04      |
|               |           | ACC    | 72.75 | 81.56 | 79.76 | 66.01     | 88.63     | 68.55      | 76.55  | 78.83      | 95.18 | 95.47      |
|               |           | ARI    | 63.79 | 68.32 | 71.84 | 57.40     | 86.04     | 61.06      | 64.97  | 73.78      | 90.17 | 90.91      |
|               | Color     | NMI    | 0.86  | 8.37  | 8.47  | 56.83     | 23.81     | 0.36       | 36.28  | 12.93      | 0.18  | 96.26      |
|               |           | ACC    | 12.26 | 20.78 | 16.27 | 56.46     | 21.90     | 11.92      | 31.88  | 20.02      | 11.59 | 98.24      |
|               |           | ARI    | 0.28  | 4.53  | 3.26  | 45.73     | 9.55      | 0.07       | 20.45  | 9.50       | 0.10  | 96.23      |
|               | Count     | NMI    | 9.56  | 3.03  | 1.58  | 11.38     | 10.19     | 9.95       | 10.60  | 8.49       | 3.14  | 21.99      |
|               |           | ACC    | 18.98 | 15.43 | 14.12 | 20.13     | 18.79     | 19.49      | 19.60  | 18.20      | 15.53 | 31.17      |
|               |           | ARI    | 5.69  | 1.63  | 0.77  | 7.67      | 5.10      | 5.75       | 4.79   | 5.19       | 1.79  | 14.95      |
| Cards         | Number    | NMI    | 16.18 | 16.97 | 16.51 | 16.32     | 13.96     | 16.57      | 20.85  | 17.13      | 19.65 | 35.82      |
|               |           | ACC    | 19.21 | 22.71 | 20.61 | 20.61     | 17.47     | 19.13      | 22.79  | 22.33      | 19.78 | 43.39      |
|               |           | ARI    | 8.24  | 8.77  | 10.34 | 7.95      | 7.60      | 7.97       | 10.82  | 8.75       | 7.60  | 25.62      |
|               | Suits     | NMI    | 16.94 | 8.90  | 18.05 | 14.02     | 16.63     | 18.22      | 37.98  | 33.60      | 18.68 | 53.97      |
|               |           | ACC    | 44.06 | 38.55 | 49.05 | 46.65     | 40.38     | 45.32      | 56.11  | 59.60      | 42.62 | 82.41      |
|               |           | ARI    | 13.33 | 9.04  | 15.16 | 11.08     | 15.05     | 13.94      | 28.43  | 33.11      | 15.34 | 58.76      |
| Fruit360      | Color     | NMI    | 36.19 | 32.59 | 21.15 | 54.47     | 28.71     | 39.75      | 39.33  | 57.55      | 20.23 | 58.90      |
|               |           | ACC    | 48.39 | 53.40 | 47.06 | 65.70     | 57.76     | 56.73      | 54.37  | 57.56      | 49.54 | 75.97      |
|               |           | ARI    | 22.64 | 22.13 | 19.19 | 45.64     | 27.31     | 26.49      | 26.30  | 41.17      | 14.16 | 54.64      |
|               | Species   | NMI    | 49.35 | 40.39 | 27.06 | 50.37     | 47.99     | 51.72      | 57.88  | 61.97      | 50.93 | 82.47      |
|               |           | ACC    | 61.55 | 60.67 | 51.75 | 61.82     | 49.59     | 62.97      | 76.46  | 76.94      | 54.88 | 92.30      |
|               |           | ARI    | 32.39 | 29.43 | 21.20 | 34.20     | 24.28     | 38.93      | 50.04  | 48.78      | 38.38 | 78.84      |
| CIFAR-10      | Object    | NMI    | 70.96 | 83.30 | 75.41 | 70.36     | 64.01     | 70.87      | 74.63  | 70.95      | 76.40 | 79.70      |
|               |           | ACC    | 70.75 | 91.90 | 84.75 | 71.02     | 72.81     | 70.45      | 78.02  | 71.70      | 86.12 | 87.71      |
|               |           | ARI    | 58.25 | 83.10 | 70.86 | 59.00     | 53.96     | 56.98      | 66.48  | 60.89      | 73.67 | 74.61      |
|               | Scene     | NMI    | 1.08  | 4.55  | 0.17  | 1.76      | 10.90     | 1.07       | 2.15   | 15.75      | 5.19  | 26.30      |
|               |           | ACC    | 58.12 | 62.60 | 50.42 | 60.56     | 70.13     | 58.11      | 60.10  | 73.29      | 63.58 | 77.80      |
|               |           | ARI    | 2.41  | 6.34  | 0.41  | 3.95      | 16.11     | 2.40       | 3.94   | 21.67      | 7.35  | 30.05      |
| Stanford Cars | Brand     | NMI    | 66.09 | 61.69 | 33.56 | 65.73     | 29.40     | -          | 66.20  | 62.17      | 34.95 | 70.58      |
|               |           | ACC    | 36.91 | 31.08 | 69.94 | 36.56     | 6.77      | -          | 38.75  | 32.02      | 7.09  | 45.92      |
|               |           | ARI    | 25.96 | 20.66 | 2.67  | 25.55     | 1.56      | -          | 27.51  | 20.60      | 1.04  | 31.59      |
|               | Color     | NMI    | 3.35  | 2.86  | 2.27  | 27.96     | 31.22     | 3.08       | 3.27   | 4.62       | 9.19  | 64.91      |
|               |           | ACC    | 20.24 | 21.04 | 17.63 | 40.03     | 45.39     | 19.04      | 20.06  | 21.31      | 22.49 | 78.02      |
|               |           | ARI    | 1.55  | 1.67  | 0.88  | 15.90     | 31.64     | 1.15       | 1.49   | 2.10       | 5.21  | 60.90      |
|               | Type      | NMI    | 31.61 | 33.58 | 26.43 | 31.00     | 6.61      | 35.65      | 34.82  | 34.48      | 13.86 | 51.27      |
|               |           | ACC    | 38.02 | 46.64 | 35.90 | 37.75     | 22.08     | 41.24      | 42.95  | 44.29      | 22.77 | 64.19      |
|               |           | ARI    | 19.25 | 23.10 | 15.93 | 18.85     | 3.39      | 23.89      | 22.62  | 24.17      | 6.68  | 45.44      |
| CUB           | Species   | NMI    | 66.97 | 64.75 | 56.44 | 64.95     | 36.21     | -          | 65.96  | 58.29      | 68.31 | 68.25      |
|               |           | ACC    | 40.30 | 35.68 | 22.17 | 37.85     | 9.20      | -          | 40.61  | 28.45      | 18.21 | 45.50      |
|               |           | ARI    | 28.80 | 24.79 | 14.74 | 26.52     | 2.61      | -          | 28.55  | 16.42      | 11.86 | 29.20      |
|               | Action    | NMI    | 16.14 | 0.85  | 31.44 | 17.55     | 13.06     | 13.49      | 14.73  | 21.72      | 3.20  | 35.81      |
|               |           | ACC    | 29.41 | 23.56 | 2.99  | 39.18     | 34.38     | 27.04      | 38.73  | 28.48      | 22.90 | 67.96      |
|               |           | ARI    | 5.65  | 0.32  | 2.51  | 9.40      | 6.38      | 3.81       | 6.88   | 8.91       | 0.20  | 36.65      |
|               | Scene     | NMI    | 23.60 | 1.11  | 33.07 | 15.18     | 21.70     | 21.69      | 24.57  | 24.00      | 8.69  | 45.36      |
|               |           | ACC    | 32.85 | 20.67 | 4.22  | 40.14     | 33.62     | 29.97      | 36.20  | 27.88      | 22.99 | 67.90      |
|               |           | ARI    | 13.02 | 0.85  | 3.23  | 14.41     | 11.67     | 10.83      | 14.34  | 12.92      | 3.72  | 65.06      |
| Average       |           |        | 29.22 | 28.95 | 26.35 | 34.83     | 28.76     | 27.45      | 34.54  | 33.96      | 26.42 | 58.39      |

\* Cop-Kmeans fails to converge on datasets with a large number of classes.

structures. Secondly, unlike IPC, which actively selects high-value sample pairs, these semi-supervised clustering methods lack an effective strategy to utilize informative interactions, leading to inefficient use of the limited annotation budget. As a combined effect of these two factors, these semi-supervised clustering methods require substantial annotation costs, which are often impractical in practice. This also helps explain why these semi-supervised clustering methods can even underperform the unsupervised approaches.

#### D. Ablation Study and Parameter Analysis

To further demonstrate the robustness and effectiveness of the proposed IPC, we conduct a thorough ablation study and parameter analysis, from the overall two stages to the components in each stage. For clarity, the values reported in the following tables are the per-dataset average, calculated across all criteria specific to each respective dataset.

1) *Effectiveness of Criterion Alignment and Grouping Refinement*: We first validate the overall effectiveness of the Criterion Alignment stage (CA) and the Grouping Refinement

TABLE III  
EFFECTIVENESS OF CRITERION ALIGNMENT (CA) AND GROUPING REFINEMENT (GR).

| CA | GR | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|----|----|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|    |    | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| ✓  | ✓  | 21.37        | 29.28        | 17.55        | 16.56        | 31.64        | 10.79        | 42.77        | 54.97        | 27.52        | 28.05        |
|    |    | 50.93        | 55.81        | 43.45        | 32.68        | 47.90        | 25.38        | 57.16        | 68.16        | 45.07        | 47.39        |
|    |    | 34.44        | 42.85        | 29.55        | 28.18        | 42.05        | 22.31        | 45.61        | 60.92        | 37.75        | 38.18        |
| ✓  | ✓  | <b>57.98</b> | <b>65.10</b> | <b>54.16</b> | <b>44.90</b> | <b>62.90</b> | <b>42.19</b> | <b>70.69</b> | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |

TABLE IV  
CLUSTERING PERFORMANCE USING DIFFERENT LLMs.

| LLM         | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|             | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| GPT-4o      | 57.98        | 65.10        | 54.16        | 44.90        | <b>62.90</b> | <b>42.19</b> | 70.69        | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |
| DeepSeek-V3 | 58.05        | 63.96        | 53.49        | 41.69        | 60.85        | 37.73        | 68.99        | 82.69        | 62.47        | 58.88        |
| Gemini 2.5  | 57.53        | 64.55        | 54.51        | <b>44.95</b> | 62.13        | 41.69        | 67.52        | 82.64        | 64.19        | 59.97        |
| Claude 4    | <b>59.87</b> | <b>66.36</b> | <b>55.05</b> | 42.93        | 60.54        | 39.71        | <b>72.19</b> | 80.84        | 61.74        | 59.91        |
| Llama 3     | 54.21        | 61.67        | 50.52        | 42.58        | 61.50        | 39.38        | 68.06        | 82.93        | 60.73        | 57.95        |
| Qwen 3      | 54.14        | 61.34        | 51.31        | 43.38        | 62.92        | 38.83        | 69.32        | 82.92        | 63.19        | 58.59        |

TABLE V  
CLUSTERING PERFORMANCE USING DIFFERENT VLMs.

| VLM                   | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|-----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                       | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| CLIP                  | 21.37        | 29.28        | 17.55        | 16.56        | 31.64        | 10.79        | 42.77        | 54.97        | 27.52        | 28.05        |
| <b>CLIP + IPC</b>     | <b>57.98</b> | <b>65.10</b> | <b>54.16</b> | <b>44.90</b> | <b>62.90</b> | <b>42.19</b> | <b>70.69</b> | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |
| ALIGN                 | 22.96        | 31.22        | 21.14        | 9.18         | 24.62        | 4.41         | 46.12        | 59.53        | 34.34        | 28.17        |
| <b>ALIGN + IPC</b>    | <b>61.29</b> | <b>66.07</b> | <b>56.38</b> | <b>41.84</b> | <b>58.74</b> | <b>40.10</b> | <b>65.62</b> | <b>74.41</b> | <b>56.72</b> | <b>57.91</b> |
| MetaCLIP              | 22.56        | 30.04        | 19.26        | 16.53        | 29.39        | 11.02        | 41.17        | 54.33        | 27.07        | 27.93        |
| <b>MetaCLIP + IPC</b> | <b>56.30</b> | <b>63.80</b> | <b>53.84</b> | <b>52.84</b> | <b>69.17</b> | <b>49.16</b> | <b>65.33</b> | <b>78.76</b> | <b>60.87</b> | <b>61.12</b> |
| BLIP2                 | 23.75        | 31.72        | 21.36        | 27.79        | 36.15        | 17.66        | 50.49        | 59.93        | 36.29        | 33.90        |
| <b>BLIP2 + IPC</b>    | <b>61.46</b> | <b>64.99</b> | <b>55.45</b> | <b>66.77</b> | <b>81.22</b> | <b>64.98</b> | <b>71.41</b> | <b>76.04</b> | <b>61.67</b> | <b>67.11</b> |

TABLE VI  
THE AVERAGE PERFORMANCE FOR DIFFERENT VLM PROMPTS.

| Prompt | NMI          | ACC          | ARI          | Average      |
|--------|--------------|--------------|--------------|--------------|
| i      | <b>44.89</b> | <b>53.17</b> | <b>37.43</b> | <b>45.16</b> |
| ii     | 44.02        | 50.80        | 36.33        | 43.72        |
| iii    | 44.73        | 52.67        | 36.95        | 44.78        |
| iv     | 43.37        | 50.23        | 35.61        | 43.07        |
| v      | 42.69        | 49.57        | 34.72        | 42.33        |

stage (GR). As shown in Table III, by obtaining criterion-aligned representations, CA substantially enhances the ability to characterize criterion-specific semantics, with an average performance improvement of 69% relatively. Subsequently, by mining high-value sample pairs and efficient user interaction, GR further boosts the discriminability of hard boundary samples. By integrating these two stages, their collaborative effects deliver a remarkable improvement of 117% relatively.

2) *Impact of different LLMs*: In the Criterion Alignment stage, IPC adopts GPT-4o as the default LLM to generate descriptive texts related to the user-specified criterion. In practice, however, IPC is LLM-agnostic and compatible with a wide range of mainstream models. Table IV exhibits the stable

performance using six mainstream LLMs: GPT-4o, DeepSeek-V3 [63], Gemini 2.5 [64], Claude 4 [65], Llama 3 [66], and Qwen 3 [67]. The low standard deviation (1.11) underscores the robustness of IPC, confirming its efficacy is not dependent on the choice of a specific LLM.

3) *Impact of different VLMs*: As for the VLM, IPC adopts the widely-used ViT-B/32 version of CLIP as the default backbone. To explore the impact of different VLMs on the proposed IPC, we conduct experiments on more VLMs, including ALIGN [68], MetaCLIP [69], and BLIP2 [70]. One can observe from Table V that IPC achieves significant performance improvements on all criteria for all four VLMs. Moreover, IPC acquires better clustering results when using the more advanced BLIP2 as the backbone, demonstrating its scalability and generalizability.

4) *Impact of different VLM prompts*: Prompt i) used in Section III is not mandatory, and we provide the following alternative VLM prompts that can be used as replacements: ii) A photo with the  $C$  of  $w$ , iii) Itap with the  $C$  of  $w$ , iv) Art with the  $C$  of  $w$ , and v) A cartoon with the  $C$  of  $w$ .

We conduct experiments on the Clevr4-10k and Cards [33] datasets, reporting the average performance across all six clustering criteria in the Criterion Alignment stage. As evidenced

TABLE VII  
CLUSTERING PERFORMANCE WITH DIFFERENT PAIR SELECTION STRATEGIES.

| $\mathcal{R}$ | $\mathcal{D}$ | $\mathcal{H}$ | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|---------------|---------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|               |               |               | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| ✓             | ✓             | ✓             | 55.84        | 60.53        | 50.29        | 40.87        | 57.75        | 36.36        | 62.05        | 74.20        | 52.43        | 54.48        |
|               |               |               | 52.51        | 58.49        | 45.96        | 42.34        | 57.64        | 37.51        | 70.15        | 82.89        | 64.03        | 56.84        |
|               |               |               | 55.72        | 61.32        | 49.84        | 40.20        | 50.00        | 33.43        | 68.85        | 83.46        | 62.90        | 56.19        |
|               |               |               | 56.79        | 63.11        | 52.36        | 43.75        | 60.63        | 40.03        | 67.00        | 81.67        | 63.95        | 58.81        |
| ✓             | ✓             | ✓             | 55.74        | 61.38        | 49.93        | 44.44        | 61.09        | 40.10        | <b>71.10</b> | 81.29        | 64.76        | 58.87        |
| ✓             | ✓             | ✓             | 51.90        | 58.64        | 46.54        | 43.14        | 60.58        | 38.82        | 66.49        | 80.32        | 59.75        | 56.24        |
| ✓             | ✓             | ✓             | <b>58.54</b> | 64.21        | 54.15        | 41.25        | 55.94        | 35.41        | 62.49        | 74.04        | 56.89        | 55.88        |
| ✓             | ✓             | ✓             | 57.98        | <b>65.10</b> | <b>54.16</b> | <b>44.90</b> | <b>62.90</b> | <b>42.19</b> | 70.69        | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |

TABLE VIII  
CLUSTERING PERFORMANCE WITH DIFFERENT COMBINATIONS OF LOSS TERMS.

| $\mathcal{L}_p$ | $\mathcal{L}_n$ | $\mathcal{L}_r$ | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|-----------------|-----------------|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                 |                 |                 | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| ✓               | ✓               | ✓               | 50.93        | 55.81        | 43.45        | 32.68        | 47.90        | 25.38        | 57.16        | 68.16        | 45.07        | 47.39        |
|                 |                 |                 | <b>58.71</b> | 62.07        | 52.35        | 35.28        | 35.52        | 25.52        | 68.83        | 75.57        | 60.66        | 52.72        |
|                 |                 |                 | 51.48        | 59.80        | 49.03        | 29.33        | 51.14        | 26.43        | 47.36        | 63.54        | 38.35        | 46.27        |
|                 |                 |                 | 51.41        | 56.84        | 43.97        | 37.42        | 51.14        | 29.12        | 59.67        | 72.07        | 48.04        | 49.96        |
| ✓               | ✓               | ✓               | 57.35        | 63.08        | 52.93        | 42.37        | 59.95        | 38.17        | 68.47        | 79.95        | 62.06        | 58.26        |
| ✓               | ✓               | ✓               | 56.81        | 61.73        | 51.80        | 42.57        | 56.00        | 35.33        | 65.56        | 77.46        | 59.52        | 56.31        |
| ✓               | ✓               | ✓               | 52.40        | 59.52        | 48.77        | 36.25        | 56.16        | 32.61        | 56.38        | 68.82        | 48.33        | 51.03        |
| ✓               | ✓               | ✓               | 57.98        | <b>65.10</b> | <b>54.16</b> | <b>44.90</b> | <b>62.90</b> | <b>42.19</b> | <b>70.69</b> | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |

by Table VI, changing the prompt has a negligible impact on the results, where the standard deviation is merely 1.05.

5) *Effectiveness of the valuable pair mining strategy*: To evaluate our pair mining strategy, we conduct an ablation study by progressively removing its three key metrics, i.e., representativeness, diversity, and hardness. When hardness is omitted, random pair selection is applied to the samples pre-selected by representativeness and diversity.

As shown in Fig. 4 and Table VII, relying solely on representativeness causes selected samples to collapse into a few clusters, degrading performance. Incorporating diversity ensures a more balanced sample distribution across different clusters, while explicitly accounting for hardness prioritizes highly ambiguous pairs to achieve optimal results. Overall, our joint strategy outperforms random selection by 6.5%. Notably, this improvement is particularly meaningful given that only 500 sample pairs are selected for the first two datasets, and merely 50 pairs for Fruit360, which is negligible compared to the total number of possible sample pairs ( $O(N^2)$ ).

6) *Effectiveness of the loss terms*: We also conduct a complete ablation study on the three loss terms of IPC, i.e., positive loss  $\mathcal{L}_p$ , negative loss  $\mathcal{L}_n$ , and regularization loss  $\mathcal{L}_r$ .  $\mathcal{L}_p$  exploits user-provided positive feedback to pull samples from the same class closer, resulting in more compact intra-cluster structures. In contrast,  $\mathcal{L}_n$  leverages negative feedback to push samples from different classes apart, thereby enlarging inter-cluster separability. As exhibited in Table VIII, the positive and negative losses play complementary roles, and only their synergistic effect enables IPC to achieve substantially enhanced clustering performance. At the same time, the regularization loss  $\mathcal{L}_r$  helps prevent degenerate solutions and maintains stability during optimization. Overall, the ablation

study demonstrates the effective collaboration of these three losses, with a significant 13.6% average improvement.

Due to space limitations, several discussions are provided in the supplementary material, including detailed dataset descriptions, comparisons with pointwise queries and direct GPT querying, robustness to noisy user feedback, and experiments on two specialized-domain datasets.

## V. CONCLUSION

In this paper, we reveal that most existing clustering methods are inherently static. Specifically, this static paradigm blindly optimizes toward a single salient semantic criterion and is simultaneously bounded by the limited information of internal data. As a consequence, it inevitably suffers from the dual challenges of Rigid Criteria and Sample Ambiguity. To navigate beyond these two challenges, we propose an interactive personalized clustering paradigm (IPC), introducing user interaction to instruct the clustering process. To tackle the Rigid Criteria challenge, IPC constructs the text basis tailored to the user-specified criterion, aligning the image representation with the criterion via text basis transformation. Moreover, to address the Sample Ambiguity challenge, IPC mines highly valuable pairs based on their hardness, representativeness, and diversity, further leveraging user feedback to refine the cluster assignments. Extensive experiments demonstrate the superiority and robustness of IPC across diverse criteria. For future work, one potential direction is to investigate interactions across a broader range of modalities, not limited to image and text. Finally, we hope this work could stimulate new inspiration for the research community, facilitating a shift from the constrained static paradigm toward the more adaptive interactive paradigm.

## REFERENCES

- [1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, “Highly accurate protein structure prediction with alphafold,” *nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [2] G. Punj and D. W. Stewart, “Cluster analysis in marketing research: Review and suggestions for application,” *Journal of marketing research*, vol. 20, no. 2, pp. 134–148, 1983.
- [3] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, “An improved overlapping k-means clustering method for medical applications,” *Expert Systems with Applications*, vol. 67, pp. 12–18, 2017.
- [4] V. Mikuni and F. Canelli, “Unsupervised clustering for collider physics,” *Physical Review D*, vol. 103, no. 9, p. 092007, 2021.
- [5] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, “Model-based clustering for social networks,” *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 170, no. 2, pp. 301–354, 2007.
- [6] S. Lloyd, “Least squares quantization in pcm,” *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [7] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 14, 2001.
- [8] J. Yang, D. Parikh, and D. Batra, “Joint unsupervised learning of deep representations and image clusters,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5147–5156.
- [9] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *International conference on machine learning*. PMLR, 2016, pp. 478–487.
- [10] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, “Deep adaptive image clustering,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5879–5887.
- [11] X. Ji, J. F. Henriques, and A. Vedaldi, “Invariant information clustering for unsupervised image classification and segmentation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9865–9874.
- [12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [14] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 10, 2021, pp. 8547–8555.
- [15] H. Zhong, J. Wu, C. Chen, J. Huang, M. Deng, L. Nie, Z. Lin, and X.-S. Hua, “Graph contrastive clustering,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9224–9233.
- [16] Y. Shen, Z. Shen, M. Wang, J. Qin, P. Torr, and L. Shao, “You never cluster alone,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 734–27 746, 2021.
- [17] Z. Huang, J. Chen, J. Zhang, and H. Shan, “Learning representation for clustering via prototype scattering and positive sampling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7509–7524, 2022.
- [18] Z. Li and Y. Jia, “Conmix: Contrastive mixup at representation level for long-tailed deep clustering,” in *The Thirteenth International Conference on Learning Representations*, 2025.
- [19] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, “A closer look at memorization in deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 233–242.
- [20] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi, “Deep subspace clustering with sparsity prior,” in *Ijcai*, 2016, pp. 1925–1931.
- [21] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng, “Twin contrastive learning for online clustering,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2205–2221, 2022.
- [22] W. Van Gansbeke, S. Vandenheide, S. Georgoulis, M. Proesmans, and L. Van Gool, “Scan: Learning to classify images without labels,” in *European conference on computer vision*. Springer, 2020, pp. 268–285.
- [23] C. Niu, H. Shan, and G. Wang, “Spice: Semantic pseudo-labeling for image clustering,” *IEEE Transactions on Image Processing*, vol. 31, pp. 7264–7278, 2022.
- [24] H. Liu, P. Hu, C. Zhang, Y. Li, and X. Peng, “Interactive deep clustering via value mining,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 42 369–42 387, 2024.
- [25] R. Yang, P. Hu, X. Peng, X. Liu, and Y. Li, “Difc: Your diffusion model holds the secret to fine-grained clustering,” *arXiv preprint arXiv:2412.18838*, 2024.
- [26] J. Wu, Z. Li, W. Sun, J. Yin, L. Nie, and Z. Lin, “Clusmatch: Improving deep clustering by unified positive and negative pseudo-label learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [27] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, “Meta clustering,” in *Sixth International Conference on Data Mining (ICDM’06)*. IEEE, 2006, pp. 107–118.
- [28] E. Bae and J. Bailey, “Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity,” in *Sixth International Conference on Data Mining (ICDM’06)*. IEEE, 2006, pp. 53–62.
- [29] Y. Chang, J. Chen, M. H. Cho, P. J. Castaldi, E. K. Silverman, and J. G. Dy, “Multiple clustering views from multiple uncertain experts,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 674–683.
- [30] D. Niu, J. G. Dy, and M. I. Jordan, “Multiple non-redundant spectral clustering views,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 831–838.
- [31] L. Miklautz, D. Mautz, M. C. Altinigneli, C. Böhm, and C. Plant, “Deep embedded non-redundant clustering,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5174–5181.
- [32] L. Ren, G. Yu, J. Wang, L. Liu, C. Domeniconi, and X. Zhang, “A diversified attention model for interpretable multiple clusterings,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8852–8864, 2022.
- [33] J. Yao, E. Liu, M. Rashid, and J. Hu, “Augdmc: Data augmentation guided deep multiple clustering,” *Procedia Computer Science*, vol. 222, pp. 571–580, 2023.
- [34] S. Zhang, L. Ren, J. Wang, Y. Xu, C. Domeniconi, and G. Yu, “Aligning contrastive multiple clusterings with user interests,” in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 7011–7019.
- [35] S. Kwon, J. Park, M. Kim, J. Cho, E. K. Ryu, and K. Lee, “Image clustering conditioned on text criteria,” *arXiv preprint arXiv:2310.18297*, 2023.
- [36] M. Liu, Z. Zhong, J. Li, G. Franchi, S. Roy, and E. Ricci, “Organizing unstructured image collections using natural language,” *arXiv preprint arXiv:2410.05217*, 2024.
- [37] A. Stephan, L. Miklautz, C. Leiber, P. H. L. De Araujo, D. Repas, C. Plant, and B. Roth, “Text-guided alternative image clustering,” *arXiv preprint arXiv:2406.18589*, 2024.
- [38] H. Liu, C. Sun, P. Hu, Y. Li, and X. Peng, “Conditional representation learning for customized tasks,” *arXiv preprint arXiv:2510.04564*, 2025.
- [39] J. Yao, Q. Qian, and J. Hu, “Multi-modal proxy learning towards personalized visual multiple clustering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 066–14 075.
- [40] K. Wagstaff and C. Cardie, “Clustering with instance-level constraints,” *AAAI/IAAI*, vol. 1097, no. 577–584, p. 197, 2000.
- [41] H. Liu, Z. Tao, and Y. Fu, “Partition level constrained clustering,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2469–2483, 2017.
- [42] Z. Jiang, Y. Zhan, Q. Mao, and Y. Du, “Semi-supervised clustering under a ‘compact-cluster’ assumption,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5244–5256, 2022.
- [43] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, “Constrained k-means clustering with background knowledge,” in *Icml*, vol. 1, 2001, pp. 577–584.
- [44] L. Manduchi, K. Chin-Cheong, H. Michel, S. Wellmann, and J. Vogt, “Deep conditional gaussian mixture model for constrained clustering,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 303–11 314, 2021.
- [45] T. Nguyen, S. Ibrahim, and X. Fu, “Deep clustering with incomplete noisy pairwise annotations: A geometric regularization approach,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 25 980–26 007.
- [46] G. Ding, Y. Lu, P. Hu, M. Yang, Y. Lin, and X. Peng, “Visual abstraction: A plug-and-play approach for text-visual retrieval,” in *Forty-second International Conference on Machine Learning*, 2025.

- [47] Y. Lu, M. Yang, D. Peng, P. Hu, Y. Lin, and X. Peng, "Llava-reid: Selective multi-image questioner for interactive person re-identification," *arXiv preprint arXiv:2504.10174*, 2025.
- [48] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.
- [49] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *arXiv preprint arXiv:1705.02999*, 2017.
- [50] C. R. Gurbatri, N. Arpaia, and T. Danino, "Engineering bacteria as interactive cancer therapies," *Science*, vol. 378, no. 6622, pp. 858–864, 2022.
- [51] S. Wang, Z. Zhao, X. Ouyang, T. Liu, Q. Wang, and D. Shen, "Interactive computer-aided diagnosis on medical image using large language models," *Communications Engineering*, vol. 3, no. 1, p. 133, 2024.
- [52] S. Vaze, A. Vedaldi, and A. Zisserman, "No representation rules them all in category discovery," *Advances in Neural Information Processing Systems*, vol. 36, pp. 19 962–19 989, 2023.
- [53] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [54] X. Geng, S. Zhao, Y. Yu, B. Peng, P. Du, H. Chen, C. Li, and M. Wang, "Personalized clustering via targeted representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 16, 2025, pp. 16 790–16 798.
- [55] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- [56] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
- [57] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.
- [58] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [59] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [60] Y. Li, P. Hu, D. Peng, J. Lv, J. Fan, and X. Peng, "Image clustering with external guidance," in *International Conference on Machine Learning*. PMLR, 2024, pp. 27 890–27 902.
- [61] N. Adaloglou, F. Michels, H. Kalisch, and M. Kollmann, "Exploring the limits of deep image clustering using pretrained models," *arXiv preprint arXiv:2303.17896*, 2023.
- [62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [63] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.
- [64] G. Comanici, E. Bieber, M. Schaeckermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.
- [65] Anthropic, "Claude opus 4 & claude sonnet 4 system card," Anthropic, Tech. Rep., May 2025. [Online]. Available: <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>
- [66] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [67] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [68] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.
- [69] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, "Demystifying clip data," *arXiv preprint arXiv:2309.16671*, 2023.
- [70] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.



**Honglin Liu** received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2023. He is currently pursuing the Ph.D. degree in computer science at the College of Computer Science, Sichuan University. His research interests include deep clustering, representation learning, and multi-modal learning.



**Joey Tianyi Zhou** received the PhD degree in computer science from Nanyang Technological University, Singapore, in 2015. He is the deputy director and principal scientist with the A\*STAR Centre for Frontier AI Research (CFAR), Singapore. His research focuses on the efficiency and robustness of machine learning. He has published over 100 papers and received several awards at conferences such as ECCV, IJCAI, and ICCV. He serves as an Associate Editor or Area Chair for journals and conferences including IEEE TPAMI, IJCV, ICML, and NeurIPS.



**Peng Hu** received the Ph.D. degree in computer science and technology from Sichuan University, China, in 2019. He is currently a professor at the School of Artificial Intelligence, Sichuan University. His research interests mainly focus on multimodal learning, cross-modal retrieval, and network compression. In these areas, he has authored more than 50 articles in the top-tier conferences and journals.



**Yunfan Li** received his Ph.D. degree in 2025 from the College of Computer Science, Sichuan University, Chengdu, China, under the supervision of Professor Xi Peng. He is currently a postdoctoral researcher at the College of Computer Science, Sichuan University. His research interests include clustering, unsupervised learning, and AI for Science.



**Xi Peng** is currently the Cheung Kong distinguished professor at the School of Artificial Intelligence, Sichuan University. His current interests mainly focus on machine learning, multimedia analysis, and AI4Science. In these areas, he has co-authored around 100 articles in Nature Communications, JMLR, TPAMI, ICLR, ICML, NeurIPS, etc. Dr. Peng has served as an Associate Editor for five journals including IEEE TPAMI and IEEE TIP.

# Supplementary Material

Honglin Liu, Joey Tianyi Zhou, Peng Hu, Yunfan Li, Xi Peng

In the supplementary material, we provide additional information and analysis to complement the main paper, including the dataset description, more experiment details, and the rationality of the proposed value mining strategy.

## I. DATASET DESCRIPTION

In this section, we present samples from the datasets and give a concise description of the four extra criteria annotated in the widely used Stanford Cars [4] and CUB [8] datasets.

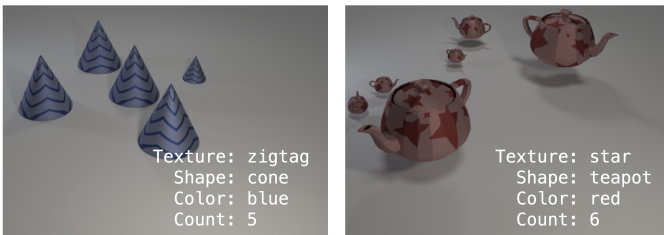
### A. Dataset Samples

In our experiments, we evaluate the proposed clustering framework on six datasets, i.e., Clevr4-10k [7], Cards [9], Fruit360 [9], CIFAR-10 [5], Stanford Cars, and CUB, which cover diverse visual contents and sixteen clustering criteria. These datasets provide a comprehensive testbed for assessing the effectiveness and robustness of our method under different data distributions and semantic grouping principles.

Derived from the CLEVR dataset [3], Clevr-4-10k is a synthetic dataset that consists of 10,531 samples designed with four distinct clustering criteria, as shown in Supplementary Table I. The dataset is generated through computer graphics tools, producing images that contain multiple objects arranged in fixed scenes, where example samples are illustrated in Supplementary Fig. 1.

SUPPLEMENTARY TABLE I  
CRITERIA AND VALUES OF THE CLEVR4-10K DATASET.

| Criterion | Value                                     | Number |
|-----------|-------------------------------------------|--------|
| Texture   | rubber, metal, checkered, ..., chessboard | 10     |
| Shape     | cube, sphere, monkey, ..., cylinder       | 10     |
| Color     | gray, red, blue, ..., orange              | 10     |
| Count     | one, two, three, ..., ten                 | 10     |



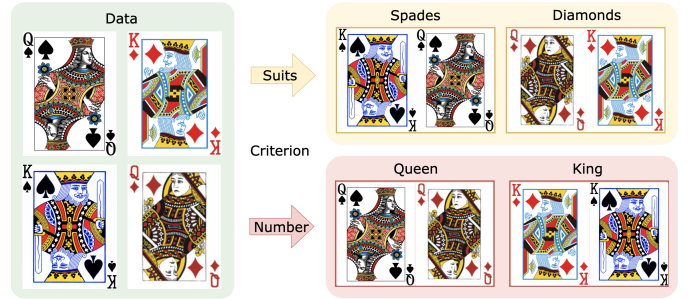
SUPPLEMENTARY FIG. 1. Examples of the Clevr4-10k dataset.

The Cards dataset contains 8,029 images of playing poker cards, which can be partitioned according to two independent dimensions, i.e., card number and suits, as exhibited in Supplementary Table II and Supplementary Fig. 2.

The Fruits360 dataset consists of 4,856 fruit images and supports two clustering criteria, namely color and species.

SUPPLEMENTARY TABLE II  
CRITERIA AND VALUES OF THE CARDS DATASET.

| Criterion | Value                           | Number |
|-----------|---------------------------------|--------|
| Number    | ace, two, three, ..., king      | 13     |
| Suits     | clubs, diamonds, hearts, spades | 4      |



SUPPLEMENTARY FIG. 2. Examples of the Cards dataset.

CIFAR-10 is a widely used image dataset comprising 60,000 low-resolution (32×32) images from 10 categories. In addition to the conventional object-based grouping, the dataset also supports an alternative scene-level criterion proposed in [1]. Supplementary Table III and Supplementary Table IV present example values of the corresponding datasets.

SUPPLEMENTARY TABLE III  
CRITERIA AND VALUES OF THE FRUIT360 DATASET.

| Criterion | Value                        | Number |
|-----------|------------------------------|--------|
| Color     | burgundy, green, red, yellow | 4      |
| Species   | apple, banana, cherry, grape | 4      |

SUPPLEMENTARY TABLE IV  
CRITERIA AND VALUES OF THE CIFAR-10 DATASET.

| Criterion | Value                                  | Number |
|-----------|----------------------------------------|--------|
| Object    | airplane, automobile, bird, ..., truck | 10     |
| Scene     | land, others                           | 2      |

Stanford Cars includes 16,185 fine-grained vehicle images with brand-level annotations. Beyond the original labels, we further annotate the dataset with color and type attributes to support clustering under different criteria, as shown in Supplementary Table V and Supplementary Fig. 3.

CUB contains 11,788 bird images with species-level labels. Beyond the original annotations, we manually annotate each image with action and scene information to enable multi-criterion clustering, as illustrated in Supplementary Table VI.

SUPPLEMENTARY TABLE V  
CRITERIA AND VALUES OF THE STANFORD CARS DATASET.

| Criterion | Value                                         | Number |
|-----------|-----------------------------------------------|--------|
| Brand     | 2012 Aston Martin V8 Vantage Convertible, ... | 196    |
| Color     | red, blue, black, ..., white                  | 8      |
| Type      | suv, sedan, coupe, ..., pickup                | 8      |



SUPPLEMENTARY FIG. 3. Examples of the Stanford Cars dataset.

## II. EXPERIMENT DETAILS

In this section, we provide details of the LLM prompt, LLM response, ground truth labels, and VLM prompt.

### A. LLM Prompt

Considering the space limitations in the main paper, we present a simplified prompt, while the complete version is provided here. We still take the “color” clustering criterion as an example, for which the complete prompt is as follows:

Generate common expressions to describe the **color** of objects, as many as possible, formatted as: [“...”, “...”, “...”, “...”]. Ensure all entries are unique and written in a single line, without any nested lists or additional formatting. You may describe the same **color** in different ways, such as “red”, “crimson”, or “scarlet”. Only generate the list, and do not include any additional information.

where the bold text indicates the clustering criterion, and the underlined portion consists of example synonyms that guide the model to produce more diverse descriptive texts.

### B. LLM Response and Ground Truth Labels

While the LLM-generated descriptive texts may initially seem close to the ground-truth labels, it would be misleading to interpret them as directly providing the missing labels for clustering. Nevertheless, there exists a significant discrepancy between the LLM-generated descriptive texts and the ground-truth labels.

To provide a comprehensive evaluation, we illustrate this discrepancy across multiple datasets and criteria, including the “color” criterion on Clevr4-10k, the “suits” criterion on the Cards dataset, and the “species” criterion on the Fruit360 dataset. The corresponding LLM-generated texts and ground truth labels are detailed below.

**Color Criterion on the Clevr4-10k dataset.** The LLM-generated texts are:

[“red”, “crimson”, “scarlet”, “ruby”, “cherry”, “rose”, “burgundy”, “cardinal”, “wine”, “firetruck red”, “ferrari red”,

SUPPLEMENTARY TABLE VI  
CRITERIA AND VALUES OF THE CUB DATASET.

| Criterion | Value                                            | Number |
|-----------|--------------------------------------------------|--------|
| Species   | black footed albatross, ..., common yellowthroat | 200    |
| Action    | swim, stand, fly, ..., walk                      | 6      |
| Scene     | sky, tree, river, ..., snow                      | 9      |

“carmine”, “sangria”, “apple red”, “poppy”, “raspberry”, “pink”, “blush”, “fuchsia”, “magenta”, “hot pink”, “baby pink”, “coral”, “peach”, “apricot”, “salmon”, “light pink”, “powder pink”, “lavender”, “violet”, “indigo”, “purple”, “amethyst”, “grape”, “orchid”, “eggplant”, “plum”, “lilac”, “mauve”, “periwinkle”, “blue”, “navy”, “azure”, “sky blue”, “cobalt”, “cerulean”, “sapphire”, “electric blue”, “royal blue”, “baby blue”, “powder blue”, “teal”, “turquoise”, “aqua”, “seafoam”, “green”, “emerald”, “forest green”, “mint”, “lime”, “olive”, “kelly green”, “sage”, “pea green”, “jade”, “chartreuse”, “yellow”, “gold”, “lemon”, “butter”, “canary”, “mustard”, “sunflower”, “cream”, “ivory”, “beige”, “tan”, “coffee”, “chocolate”, “brown”, “copper”, “rust”, “mahogany”, “walnut”, “espresso”, “gray”, “slate”, “charcoal”, “pewter”, “silver”, “platinum”, “black”, “coal”, “ebony”, “jet black”, “onyx”, “snow”, “ivory”, “white”, “pearl”, “bone”, “eggshell”, “vanilla”].

While the Ground Truth Labels are:

[“gray”, “red”, “blue”, “green”, “brown”, “purple”, “cyan”, “yellow”, “pink”, “orange”].

**Suits Criterion on the Cards dataset.** The LLM-generated texts are:

[“hearts”, “red hearts”, “heart suit”, “crimson hearts”, “spades”, “black spades”, “spade suit”, “dark spades”, “diamonds”, “red diamonds”, “diamond suit”, “sparkling diamonds”, “clubs”, “black clubs”, “club suit”, “green clubs”, “the heart suit”, “the spade suit”, “the diamond suit”, “the club suit”, “ace of hearts”, “ace of spades”, “ace of diamonds”, “ace of clubs”, “hearts suit”, “spade cards”, “diamond cards”, “club cards”, “the red suit”, “the black suit”, “the suit of hearts”, “the suit of spades”, “the suit of diamonds”, “the suit of clubs”, “card suit hearts”, “card suit spades”, “card suit diamonds”, “card suit clubs”, “hearts in poker”, “spades in poker”, “diamonds in poker”, “clubs in poker”, “hearts of playing cards”, “spades of playing cards”, “diamonds of playing cards”, “clubs of playing cards”].

While the Ground Truth Labels are:

[“clubs”, “diamonds”, “hearts”, “spades”].

**Species Criterion on the Fruit360 dataset.** The LLM-generated texts are:

[“apple”, “Fuji apple”, “Gala apple”, “Granny Smith apple”, “Red Delicious apple”, “Golden Delicious apple”, “Rome apple”, “Honeycrisp apple”, “McIntosh apple”, “Pink Lady apple”, “pear”, “Bartlett pear”, “Bosc pear”, “Anjou pear”, “Comice pear”, “Asian pear”, “Seckel pear”, “orange”, “Navel orange”, “Valencia orange”, “blood orange”, “Cara Cara orange”, “tangerine”, “mandarin”, “clementine”, “grapefruit”, “pomelo”, “lemon”, “Meyer lemon”, “lime”, “Key lime”, “Persian lime”, “Kaffir

lime", "grape", "Concord grape", "Thompson Seedless grape", "red grape", "green grape", "black grape", "muscadine grape", "scuppernong", "strawberry", "wild strawberry", "raspberry", "blackberry", "blueberry", "cranberry", "cherry", "Bing cherry", "Morello cherry", "peach", "white peach", "yellow peach", "donut peach", "nectarine", "white nectarine", "yellow nectarine", "apricot", "plum", "prune plum", "black plum", "green plum", "Damonson plum", "mirabelle plum", "banana", "Cavendish banana", "Manzano banana", "plantain", "kiwi", "kiwifruit", "golden kiwi", "red kiwi", "pineapple", "mango", "Alphonso mango", "Haden mango", "papaya", "Hawaiian papaya", "avocado", "Hass avocado", "Fuerte avocado", "pomegranate", "fig", "date", "Medjool date", "Deglet Noor date", "guava", "pineapple guava", "strawberry guava", "passion fruit", "purple passion fruit", "yellow passion fruit", "maracuja", "coconut", "cantaloupe", "honeydew melon", "watermelon", "muskmelon", "Galia melon", "persimmon", "Fuyu persimmon", "Hachiya persimmon", "quince", "star fruit", "carambola", "breadfruit", "jackfruit", "durian", "lychee", "rambutan", "satsuma", "kumquat", "mulberry", "gooseberry", "currant", "elderberry", "huckleberry", "boysenberry", "loganberry", "rhubarb", "medlar", "cherimoya", "soursop", "jujube", "longan", "mangosteen", "pawpaw", "salak", "snake fruit", "sloe", "ground cherry", "loquat", "sea buckthorn", "miracle berry", "salal berry", "serviceberry", "tamarind", "ice cream bean", "dragon fruit", "pitaya", "ackee", "akee", "jabuticaba", "sugar apple", "atemoya", "custard apple", "sapodilla", "mamey sapote", "black sapote", "white sapote", "cainito", "cupuaçu", "bael fruit", "noni fruit", "monstera deliciosa", "acai berry", "maqui berry", "chokeberry", "aronia berry", "lingonberry", "cloudberry", "salmonberry", "thimbleberry", "tamarillo", "ugli fruit", "yuzu", "Buddha's hand", "citron", "calamansi", "Sudachi" ].

While the Ground Truth Labels are:

["apple", "banana", "cherry", "grape"].

Though certain differences can be observed between the LLM-generated descriptions and the ground truth labels, they share a closely aligned semantic space. Consequently, these texts are sufficient to guide criterion alignment, highlighting the robustness of the proposed IPC.

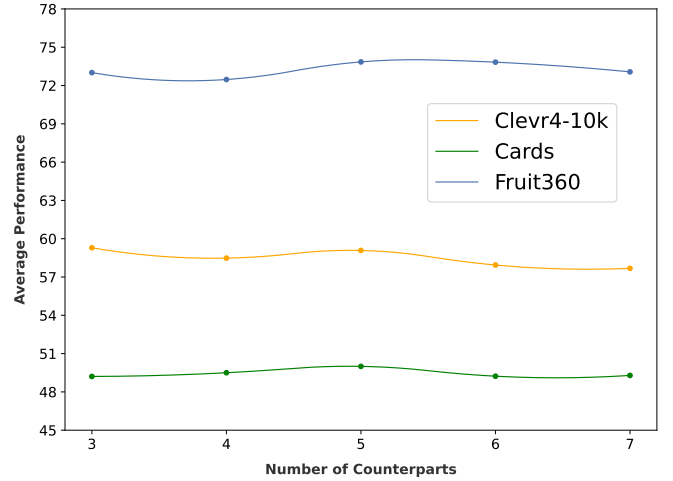
### C. The Number of Counterparts Per Sample

For each valuable sample, IPC constructs the pair by combining it with its  $B$ -th hardest counterparts to mitigate the accidental error. We set  $B$  to 5 in the experiments reported in the main paper. In fact, as shown in Supplementary Fig. 4, the performance exhibits insensitiveness with respect to this parameter  $B$ , which validates the stability of IPC. Within this stable range,  $B = 5$  performs slightly better, since a smaller  $B$  tends to be unstable, while a larger  $B$  leads to less ambiguous sample pairs.

## III. DISCUSSIONS

### A. Comparison of CLIP Zero-shot Classification and IPC Criterion Alignment

While the Criterion Alignment stage appears to be formally similar to CLIP's zero-shot classification, the two approaches



SUPPLEMENTARY FIG. 4. Performance with different numbers of counterparts per sample  $B$ .

are fundamentally different. We clarify the two core distinctions below to offer a clear perspective:

1) Different priors. CLIP's zero-shot classification requires pre-defined and explicit names for all candidate categories. In contrast, our Criterion Alignment stage does not require them. Notably, in a user-guided personalization scenario, expecting a user to provide explicit category labels is impractical. Users typically specify only a coarse-grained or high-level criterion (e.g., "cluster these images by color"), rather than listing every specific category name.

2) Distinct adaptability. CLIP's zero-shot classification acts as a classifier that directly outputs a probability distribution. In contrast, Criterion Alignment is a plug-and-play module that projects original representations into criterion-specific space. It can also be seamlessly integrated with the downstream clustering algorithm, such as k-means and our proposed Grouping Refinement.

### B. Comparison of Pairwise and Point Query

For the stage of Group Refinement, instead of the pairwise query we adopt, a more naive way is the pointwise query, i.e., directly query the affiliation between each sample and the cluster centroids. However, though it seems a simpler and more efficient alternative, we find that this pointwise query strategy leads to suboptimal performance due to error accumulation caused by imprecise cluster centroids, as shown in Supplementary Table VII. To be specific, since the initial and intermediate clusterings are inherently imperfect, the estimated centroids often deviate from the true semantic centers. Querying affiliations against these deviated centroids can introduce feedback bias, whereas our pairwise query strategy directly obtains the relationship between sample pairs and is free from this limitation.

### C. Comparison of IPC and GPT Querying

For the task of interactive personalized clustering, the most straightforward solution would be to directly query a powerful

SUPPLEMENTARY TABLE VII  
PERFORMANCE COMPARISON OF THE POINTWISE AND PAIRWISE QUERY STRATEGIES.

| Strategy                    | Clevr4-10k   |              |              | Cards        |              |              | Fruit360     |              |              | Average      |
|-----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                             | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |
| Pointwise Query             | 43.47        | 53.89        | 38.20        | 36.84        | 58.56        | 32.97        | 58.94        | 76.14        | 48.20        | 49.69        |
| <b>Pairwise Query (IPC)</b> | <b>57.98</b> | <b>65.10</b> | <b>54.16</b> | <b>44.90</b> | <b>62.90</b> | <b>42.19</b> | <b>70.69</b> | <b>84.14</b> | <b>66.74</b> | <b>60.98</b> |

SUPPLEMENTARY TABLE VIII  
PERFORMANCE AND OVERHEAD COMPARISON OF GPT QUERYING STRATEGIES AND OUR IPC.

| Strategy                 | Clevr4-10k   |              |              | Fruit360     |              |              | Average      | MLLM Queries | Tokens |
|--------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------|
|                          | NMI          | ACC          | ARI          | NMI          | ACC          | ARI          |              |              |        |
| CLIP                     | 21.43        | 29.34        | 17.61        | 46.23        | 56.65        | 30.31        | 33.59        | 0            | 0      |
| GPT <sub>Text</sub>      | 65.20        | 57.69        | 49.86        | 47.90        | 51.54        | 36.01        | 51.36        | 51,836       | 6M     |
| GPT <sub>Embedding</sub> | <b>68.57</b> | <b>68.26</b> | <b>56.50</b> | 42.49        | 57.32        | 33.79        | 54.49        | 51,836       | 6M     |
| <b>IPC</b>               | 57.98        | 65.10        | 54.16        | <b>70.69</b> | <b>84.14</b> | <b>66.74</b> | <b>66.47</b> | 6            | 1K     |

MLLM for the label of each image. Though it could introduce high computational costs, we complement this comparison to better demonstrate the effectiveness and necessity of IPC. To be specific, we include additional benchmarking experiments using GPT-4o, the MLLM used in the main text, under two distinct evaluation strategies:

1) GPT<sub>Text</sub>: For a direct evaluation, we query GPT-4o to directly output the cluster category for each sample. The generated text is then processed (e.g., case unification and text cleaning) and converted into prediction labels to compute the cluster metrics.

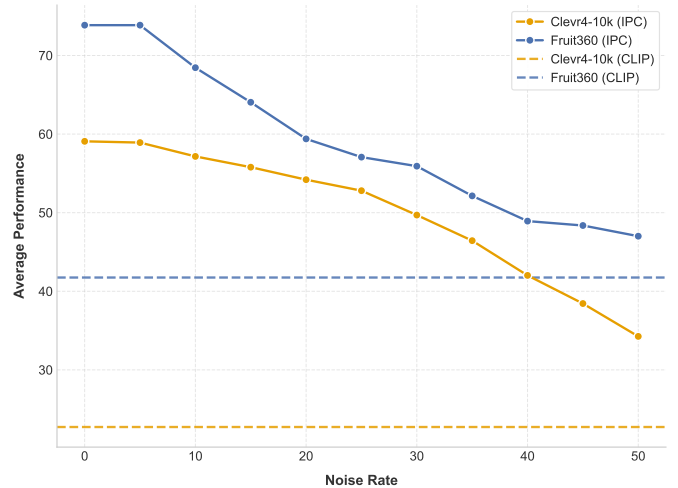
2) GPT<sub>Embedding</sub>: To further assess whether GPT-4o’s generated texts contain discriminative clustering information, we adopt an embedding-based evaluation. Specifically, we feed the generated text (also processed), together with the prompt used in IPC, into the CLIP model to extract textual representations. We then perform k-means on these embeddings to obtain clustering results.

It is worth noting the substantial computational overhead incurred by this GPT captioning approach. Specifically, for the Clevr4-10k dataset, GPT-4o required four separate queries per image to cover all four distinct criteria (i.e., texture, shape, color, and count), resulting in 42,124 queries for the 10,531 images. Similarly, for the Fruit360 dataset, two queries were executed for each of its 4,856 images. In total, this GPT querying process consumed over 6M tokens. In contrast, our IPC only needs to query GPT one time for each criterion.

As shown in Supplementary Table VIII, although the GPT<sub>Embedding</sub> strategy shows competitiveness on Clevr4-10k, our IPC achieves an average performance gain of near 12% with substantially lower costs, demonstrating its superior efficiency and effectiveness.

#### D. Influence of Noisy Annotation

For clarity, we simply use ground-truth labels to simulate the user feedback. While in real-world scenarios, users may provide noisy or inconsistent feedback. Therefore, to demonstrate the robustness of IPC, we evaluate its performance against varying degrees of user feedback noise. Specifically, we sim-



SUPPLEMENTARY FIG. 5. Performance under different noise rates on Clevr4-10k and Fruit360 datasets. The solid lines with markers indicate the clustering results of our IPC for the simulated interaction noises, and the dashed horizontal lines denote the performance of the vanilla CLIP model.

ulate noisy user feedback by randomly flipping the ground-truth pairwise relations, i.e., changing “the same cluster” to “different clusters” and vice versa, with noise rates ranging from 5% to 50%. To mitigate the impact of randomness, five independent trials are conducted for each noise rate, and the average performance is reported.

As exhibited in Supplementary Fig. 5, the performance of IPC is well-maintained when the noise rate is low. Even under extremely high noise rates, IPC still considerably improves the vanilla CLIP baseline. This demonstrates the robust error-tolerance of our IPC framework.

Notably, in real-world deployment, the actual user noise rate is expected to be relatively low. Compared to traditional multi-class annotation or direct cluster assignment, judging whether two samples belong to the same class under the specified criterion is significantly simpler. The former requires a global context and substantial time and effort, while the latter is much easier for users to execute, making high noise rates unlikely.

SUPPLEMENTARY TABLE IX  
CLUSTERING PERFORMANCE ON MORE SPECIALIZED DATASETS.

| Strategy   | GTSRB        |              |              | EuroSAT-Scene |              |              | EuroSAT-Human |              |              | Average      |
|------------|--------------|--------------|--------------|---------------|--------------|--------------|---------------|--------------|--------------|--------------|
|            | NMI          | ACC          | ARI          | NMI           | ACC          | ARI          | NMI           | ACC          | ARI          |              |
| CLIP       | 48.24        | 31.69        | 21.71        | 54.83         | 61.46        | 44.93        | 10.26         | 68.02        | 12.99        | 39.35        |
| Multi-Map  | 51.40        | 36.11        | 26.13        | 54.94         | 61.79        | 45.13        | 9.71          | 67.33        | 12.01        | 40.51        |
| DCGMM      | 51.80        | 35.79        | 25.63        | 60.56         | 68.24        | 53.60        | 12.32         | 70.24        | 16.39        | 43.84        |
| VolMax-DCC | 49.72        | 33.15        | 23.73        | 64.28         | 72.24        | 58.08        | 64.78         | 93.36        | 75.20        | 59.39        |
| <b>IPC</b> | <b>53.18</b> | <b>43.23</b> | <b>32.97</b> | <b>69.37</b>  | <b>78.33</b> | <b>64.49</b> | <b>70.08</b>  | <b>94.69</b> | <b>79.88</b> | <b>65.14</b> |

### E. Performance on More Specialized Datasets

In the main text, we include CUB (fine-grained birds) and Stanford Cars (fine-grained cars), both of which are widely recognized as challenging specialized benchmarks requiring expert-level discrimination. For instance, distinguishing specific categories, such as the “Nelson Sharp-tailed Sparrow” in CUB or the “2012 Aston Martin V8 Vantage Convertible” in Stanford Cars, necessitates capturing subtle visual nuances. Moreover, we also introduced non-salient human-annotated criteria for these two datasets, such as “action” and “scene”. The superior results on these criteria (Table II in the main text) clearly demonstrate that our IPC is inherently applicable to rare or domain-specific scenarios.

To further validate this point, we additionally conduct experiments on GTSRB [6] and EuroSAT [2], two datasets from specialized domains. Specifically, GTSRB contains traffic-sign categories with highly specialized semantics, such as “vehicles over 3.5 metric tons prohibited” and “road narrows on the right”. For EuroSAT, it consists of satellite remote-sensing images, which differ significantly from natural images in both visual characteristics and semantic content. Beyond the original scene criterion (EuroSAT-Scene), we further introduce a new criterion, the degree of human impact (EuroSAT-Human), which requires distinguishing the extent of human intervention and development. We evaluate CLIP, IPC, and three leading baselines from our main results (Table II in the main text) on these two specialized benchmarks, as presented in Supplementary Table IX.

Overall, these results, together with those reported in the original submission, provide comprehensive evidence that our IPC remains effective across both generic and specialized domains.

## IV. RATIONALITY OF THE VALUE MINING STRATEGY

As introduced in the main paper, our proposed IPC employs a two-step procedure to identify the top  $M$  high-value sample pairs. In the first step, it selects  $M/B$  samples according to their overall value  $\mathcal{V}$ , defined as the sum of representativeness  $\mathcal{R}$  and diversity  $\mathcal{D}$ . In the second step, for each of these selected samples, IPC chooses  $B$  counterparts from the entire dataset that have the largest hardness scores  $H$ , thus constructing the set of highly valuable sample pairs.

The hardness metric is explicitly designed to target hard ambiguous sample pairs, while it would lead to a prohibitive cost of  $N^2$ . To reduce the computational overhead, the representativeness and diversity metrics are introduced by pre-selecting high-value samples. Here we provide the justification

of the Theorem below, thereby validating the rationality of the proposed value mining strategy.

**Theorem 1.** *IPC sequentially selects samples in descending order of value, i.e.,*

$$\mathcal{V}_{s_m}^m \geq \mathcal{V}_{s_{m+1}}^{m+1}, \quad \forall m \in [1, M/B - 1], \quad (1)$$

where  $s_m$  denotes the index of the  $m$ -th selected sample, and  $\mathcal{V}_i^m$  denotes the value of the  $i$ -th sample in the  $m$ -th selection.

**Proof.** Let  $\mathbf{S}^m$  represent the indices of the samples selected after the  $m$ -th selection iteration, and let  $\mathbf{Y}^m$  denote the indices of the samples that remain unselected at this step.

Moreover, for a sample  $i$  that has not yet been selected at the  $m$ -th step (i.e.,  $i \in \mathbf{Y}^{m-1}$ ), we denote its representativeness and diversity by  $\mathcal{R}_i^m$  and  $\mathcal{D}_i^m$ , respectively.

By the definition of sample value, we have

$$\mathcal{V}_i^m = \mathcal{R}_i^m + \mathcal{D}_i^m, \quad (2)$$

Notably, because  $s_m$  is selected at the  $m$ -th step rather than  $s_{m+1}$ , it necessarily follows that

$$\mathcal{V}_{s_m}^m \geq \mathcal{V}_{s_{m+1}}^m. \quad (3)$$

By the definition of diversity, we have

$$\begin{aligned} \mathcal{D}_{s_{m+1}}^m &= \min_{i \in \mathbf{S}^{m-1}} \log(1 - r_{s_{m+1}} \cdot r_i) \\ &\geq \min_{i \in \mathbf{S}^m} \log(1 - r_{s_{m+1}} \cdot r_i) = \mathcal{D}_{s_{m+1}}^{m+1}, \end{aligned} \quad (4)$$

where  $r_i$  denotes the representation of the  $i$ -th sample. The inequality follows from the fact that  $\mathbf{S}^m = \mathbf{S}^{m-1} \cup s_j$ , which implies  $\mathbf{S}^{m-1} \subset \mathbf{S}^m$ .

Furthermore, as the representativeness score is irrelevant to the selection process, i.e.,  $\mathcal{R}_{s_{m+1}}^m = \mathcal{R}_{s_{m+1}}^{m+1}$ , we have

$$\mathcal{V}_{s_{m+1}}^m = \mathcal{R}_{s_{m+1}}^m + \mathcal{D}_{s_{m+1}}^m \geq \mathcal{R}_{s_{m+1}}^{m+1} + \mathcal{D}_{s_{m+1}}^{m+1} = \mathcal{V}_{s_{m+1}}^{m+1}. \quad (5)$$

Finally, by combining Eq. (3) and Eq. (5), we arrive at

$$\mathcal{V}_{s_m}^m \geq \mathcal{V}_{s_{m+1}}^m \geq \mathcal{V}_{s_{m+1}}^{m+1}, \quad (6)$$

which completes the proof.  $\square$

## REFERENCES

- [1] Xiwen Geng, Suyun Zhao, Yixin Yu, Borui Peng, Pan Du, Hong Chen, Cuiping Li, and Mengdie Wang. Personalized clustering via targeted representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16790–16798, 2025.
- [2] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

- [3] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- [4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [6] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.
- [7] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36:19962–19989, 2023.
- [8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [9] Jiawei Yao, Enbei Liu, Maham Rashid, and Juhua Hu. Augdmc: Data augmentation guided deep multiple clustering. *Procedia Computer Science*, 222:571–580, 2023.