

Multi-modal Dynamic Proxy Learning for Personalized Multiple Clustering

Jinfeng Xu¹, Zheyu Chen², Shuo Yang¹, Jinze Li¹, Ziyue Peng³, Zewei Liu¹, Hewei Wang⁴, Jiayi Zhang⁵, Edith C. H. Ngai^{1*}

¹The University of Hong Kong,

²Beijing Institute of Technology,

³University of Macau,

⁴Carnegie Mellon University,

⁵University of Nottingham,

{jinfeng, shuo.yang, lijinze-hku}@connect.hku.hk, {zewliu, chngai}@eee.hku.hk, zheyu.chen@bit.edu.cn, smyjz19@nottingham.edu.cn, heweiw@alumni.cmu.edu,

Abstract

Multiple clustering aims to discover diverse latent structures from different perspectives, yet existing methods generate exhaustive clusterings without discerning user interest, necessitating laborious manual screening. Current multi-modal solutions suffer from static semantic rigidity: predefined candidate words fail to adapt to dataset-specific concepts, and fixed fusion strategies ignore evolving feature interactions. To overcome these limitations, we propose Multi-DProxy, a novel multi-modal dynamic proxy learning framework that leverages cross-modal alignment through learnable textual proxies. Multi-DProxy introduces 1) gated cross-modal fusion that synthesizes discriminative joint representations by adaptively modeling feature interactions. 2) dual-constraint proxy optimization where user interest constraints enforce semantic consistency with domain concepts while concept constraints employ hard example mining to enhance cluster discrimination. 3) dynamic candidate management that refines textual proxies through iterative clustering feedback. Therefore, Multi-DProxy not only effectively captures a user’s interest through proxies but also enables the identification of relevant clusterings with greater precision. Extensive experiments demonstrate state-of-the-art performance with significant improvements over existing methods across a broad set of multi-clustering benchmarks.

Introduction

Clustering, a cornerstone of unsupervised learning, aims to uncover latent structures by grouping data based on intrinsic similarities. Traditional works rely on handcrafted features or monolithic representations (MacQueen 1967; Ng, Jordan, and Weiss 2001; Caron et al. 2018, 2020), often failing to capture the multifaceted nature of real-world data. While deep clustering works (Chu et al. 2024; Ouldoughi, Kuo, and Kira 2023; Qian 2023; Qian et al. 2022) have improved expressiveness, they typically produce a single partitioning, disregarding the inherent complexity of data that can be meaningfully grouped from diverse perspectives. This limitation spurred the development of multiple clustering (Miklautz et al. 2020; Ren et al. 2022; Yao et al.

2023), which seeks diverse partitions revealing complementary structures. However, existing works generate exhaustive clusterings without discerning user interest, necessitating laborious manual screening to identify relevant groupings—a significant practical bottleneck. Multimodal information is flooding the Internet (Xu et al. 2025a,b). Recent works leverage multi-modal models like CLIP (Radford et al. 2021) to align user interests (expressed as keywords, e.g., “color”) with visual representations. Recent works such as Multi-MaP (Yao, Qian, and Hu 2024b) and Multi-Sub (Yao, Qian, and Hu 2024a) employ proxy learning, where textual prompts guide the extraction of interest-biased embeddings. Despite promising results, these solutions exhibit critical limitations:

- **Static Semantic Rigidity:** Predefined candidate words (e.g., “red”, “blue”, “green” for “color”) fail to adapt to dataset-specific concepts, leading to misalignment when LLMs’ suggestions mismatch ground-truth categories.
- **Inflexible Feature Fusion:** Fixed fusion strategies (e.g., concatenation or simple averaging) ignore evolving feature interactions between modalities, yielding suboptimal joint representations.

To overcome these deficiencies, we introduce Multi-DProxy, a novel Multi-modal Dynamic Proxy Learning framework that synergizes gated cross-modal fusion, adaptive textual proxies, and dynamic candidates to generate personalized clusterings aligned with user interest. Our core innovations address the limitations head-on:

- **Gated Cross-Modal Fusion:** A hierarchical attention module with sigmoid-gated residuals dynamically recalibrates visual-textual interactions, prioritizing discriminative attributes through bidirectional feature modulation.
- **Dual-Constraint Proxy Optimization:** We enforce semantic consistency via user interest constraints (aligning proxies with concept centroids) while enhancing cluster discrimination via concept constraints using contrastive learning on fused features and relevant proxies. This replaces rigid candidate sets with learnable, semantically grounded proxies.
- **Dynamic Candidate Management:** An iterative feedback loop refines textual semantics by scoring candi-

*Corresponding authors

dates against evolving cluster centroids. This continuously adapts proxies to emergent data structures, mitigating static rigidity.

Multi-DProxy not only precisely captures user interests but also enables efficient identification of relevant clustering. Theoretical analysis proves proxy stability under dynamic updates and elucidates how visual features gate textual representations to prioritize salient attributes during fusion. Extensive experiments on a broad set of multi-clustering benchmarks demonstrate state-of-the-art performance. Our contribution can be summarized as:

- The first framework unifying learnable textual proxies, dynamic candidate refinement, and adaptive feature fusion for interest-aware multiple clustering.
- A theoretically grounded dual-constraint mechanism ensuring semantic coherence and cluster discrimination.
- We conduct extensive experiments on all publicly available multiple clustering tasks, which empirically demonstrate the superiority of the proposed Multi-DProxy in precisely capturing the user’s interest.

Related Work

Multiple clustering explores diverse data partitions from different perspectives, gaining increasing attention. Early methods rely on hand-crafted rules and representations. For example, COALA (Bae and Bailey 2006) generates new clusters using existing ones as a constraint, Hu et al. (Hu et al. 2017) maximized eigengap across subspaces, and Dang et al. (Dang and Bailey 2010) utilize an expectation-maximization framework to optimize mutual information. Recent approaches leverage learning-based techniques for better representations. For instance, ENRC (Miklautz et al. 2020) optimizes clustering objectives within a latent space learned by an auto-encoder, iMClusts (Ren et al. 2022) leverages auto-encoders and multi-head attention to learn diverse feature representations, and AugDMC (Yao et al. 2023) applies data augmentation to generate diverse image perspectives. However, it remains challenging to identify the clustering most relevant to user interests. Recently, Multi-MaP (Yao, Qian, and Hu 2024b) and Multi-Sub (Yao, Qian, and Hu 2024a) integrate CLIP embeddings with proxy learning to generate data representations aligned with user interests. While effective, these methods exhibit static semantic rigidity: predefined candidate words fail to adapt to dataset-specific concepts, fixed fusion strategies ignore evolving feature interactions, and CLIP inherently lacks deep contextual understanding for nuanced intent capture (Yao, Qian, and Hu 2024b,a). To address these limitations, we propose Multi-DProxy, a multi-modal dynamic proxy learning framework. Unlike static methods, Multi-DProxy leverages learnable textual proxies optimized via dual constraints—semantic consistency via concept centroid alignment and cluster discrimination via hard example mining.

Methodology

Multi-DProxy introduces a novel dynamic proxy learning framework that generates personalized clusterings aligned

with user intent through adaptive cross-modal alignment. Multi-DProxy transforms high-level concepts into learnable textual proxies that guide visual feature extraction. As illustrated in Figure 1.

Multi-modal Pre-training

First, we briefly review the training objective in CLIP as follows, and then describe the details of our Multi-DProxy method based on that. Given a set of image-text pairs as $\{v_i, t_i\}_{i=1}^D$, where D is the total number of datasets, and v_i is an image and t_i is the corresponding text description, their vision and text representations can be obtained by two encoders as $\mathbf{v}_i = f_v(v_i) \in \mathbb{R}^d$ and $\mathbf{t}_i = f_t(t_i) \in \mathbb{R}^d$, where \mathbf{v}_i and \mathbf{t}_i have the unit norm and d is latent dimension. Multi-DProxy employs frozen pre-trained CLIP encoders ($f_v(\cdot)$ for vision and $f_t(\cdot)$ for text). Moreover, a user-specified concept u (e.g., “color”) to refer to user interest.

Base Proxy Initialization

For each input image x_i , we generate an initial base proxy embedding by processing a unified placeholder token “*” using CLIP’s reference word embedding function: $\mathbf{w}_i' = f_t(“*”) \in \mathbb{R}^d$. We initialize and maintain D different proxies $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_D\}$. Multi-DProxy optimizes adaptive proxy embeddings \mathbf{w}_i , and facilitate identifying relevant clustering through three interconnected components:

- **Gated Cross-Modal Fusion:** synthesizes discriminative joint representations through adaptive feature interaction.
- **Dynamic Candidate Management:** iteratively refines textual semantics via clustering feedback.
- **Dual-Constraint Proxy Optimization:** ensures semantic consistency while enhancing cluster discrimination.

Gated Cross-Modal Fusion

We propose a Gated Cross-Modal Fusion module that dynamically synthesizes discriminative joint representations through hierarchical bidirectional attention and adaptive feature recalibration. Let $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_D\}$ and $\mathbf{T} = \{\mathbf{t}_1^*, \dots, \mathbf{t}_D^*\}$ denote visual and textual representations, respectively. Here $\mathbf{t}_i^* = [\mathbf{t}_i; \mathbf{w}_i]$. The component comprises core parts as following:

Bidirectional Cross-Attention For layer $l \in 1, 2, \dots, L$:

$$\begin{aligned}\mathbf{V}_{\text{attn}}^l &= \text{MultiHead}(\mathbf{V}^{l-1}, \mathbf{T}^{l-1}, \mathbf{T}^{l-1}), \\ \mathbf{T}_{\text{attn}}^l &= \text{MultiHead}(\mathbf{T}^{l-1}, \mathbf{V}^{l-1}, \mathbf{V}^{l-1}),\end{aligned}\quad (1)$$

where $\text{MultiHead}(\cdot)$ implements multi-head scaled dot-product attention.

Gated Residual Fusion Adaptive feature recalibration via sigmoid-gated residuals:

$$\begin{aligned}\mathbf{V}^l &= \mathbf{V}^{l-1} + \sigma(\mathbf{W}_g^{\mathbf{V}} [\mathbf{V}^{l-1}; \mathbf{V}_{\text{attn}}^l]) \odot \mathbf{V}_{\text{attn}}^l, \\ \mathbf{T}^l &= \mathbf{T}^{l-1} + \sigma(\mathbf{W}_g^{\mathbf{T}} [\mathbf{T}^{l-1}; \mathbf{T}_{\text{attn}}^l]) \odot \mathbf{T}_{\text{attn}}^l,\end{aligned}\quad (2)$$

$\sigma(\cdot)$ denotes the sigmoid function (distinct from γ in Eq.3). Projection matrices $\mathbf{W}_g^{\mathbf{V}} \in \mathbb{R}^{d \times 2d}$ and $\mathbf{W}_g^{\mathbf{T}} \in \mathbb{R}^{d \times 2d}$ transform concatenated features $[\cdot; \cdot]$.

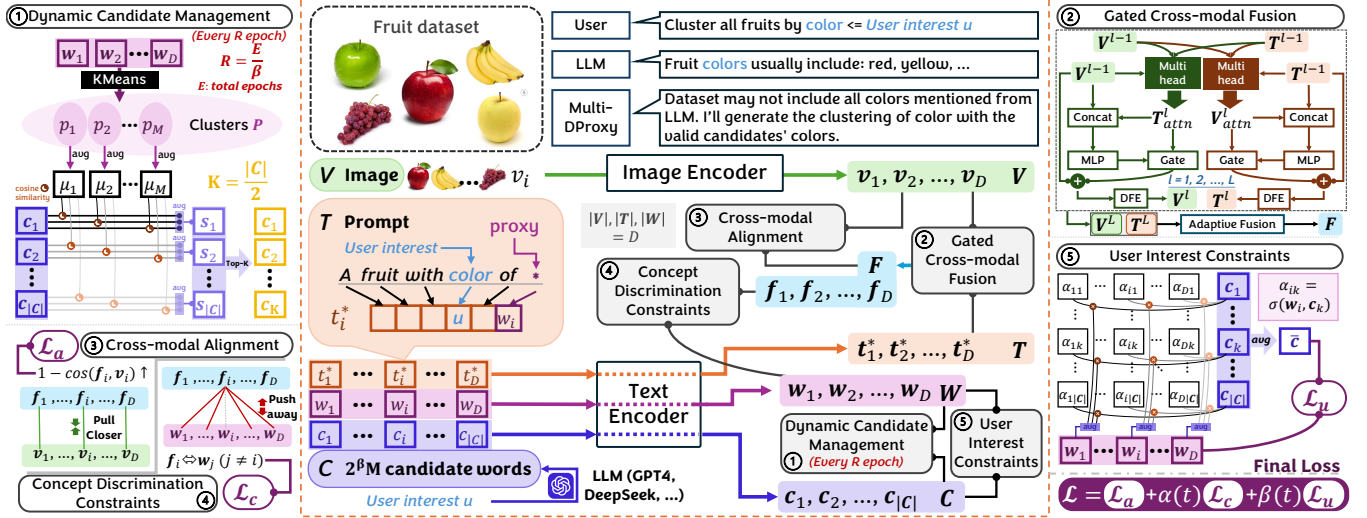


Figure 1: Overview of the Multi-DProxy framework. The central pipeline illustrates the overall architecture, while the key components are detailed on both sides: (1) Dynamic Candidate Management updates candidate words every R epochs; (2) Gated Cross-modal Fusion integrates visual and textual representations; (3) Cross-modal Alignment reduces modality discrepancies; (4) Concept Discrimination Constraints enhance cluster separability; and (5) User Interest Constraints ensure alignment with domain-specific concepts.

Discriminative Feature Enhancement (DFE) Post-attention refinement via LayerNorm and FFN:

$$\begin{aligned} \mathbf{V}^l &= \text{LayerNorm}(\mathbf{V}^l + \text{FFN}(\mathbf{V}^l)), \\ \mathbf{T}^l &= \text{LayerNorm}(\mathbf{T}^l + \text{FFN}(\mathbf{T}^l)). \end{aligned} \quad (3)$$

Adaptive Feature Fusion Final representation synthesis via temperature-scaled cosine similarity:

$$\mathbf{F} = \lambda \mathbf{T}^L + (1 - \lambda) \mathbf{V}^L, \quad \lambda = \sigma\left(\frac{\langle \mathbf{T}^L, \mathbf{V}^L \rangle}{\tau}\right), \quad (4)$$

where $\lambda \in [0, 1]$ is a learnable dynamic modality weight, τ is a learnable temperature parameter (initialized by 0.1), $\langle \cdot, \cdot \rangle$ is inner product, and $\sigma(\cdot)$ is the sigmoid function. This dynamically balances modal contributions based on inter-modal agreement.

Dynamic Candidate Management

To overcome static semantic rigidity, we introduce a Dynamic Candidate Set that evolves with the clustering structure through iterative refinement. The system maintains and dynamically updates candidate words based on their alignment with emerging cluster structures. The update process occurs every R epochs (where R is a configurable update interval hyperparameter) as follows:

- **Proxy Embedding Collection:** Collect all learnable proxy embeddings $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_D\}$ from the current training state.
- **Cluster Analysis:** Perform K-means clustering on the proxy embeddings to discover latent structures: $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_M\} = \text{KMeans}(\mathbf{W}, M)$, where M corresponds to the number of ground-truth classes.

- **Cluster Centroid Calculation:** Compute centroids for each discovered cluster: $\mu_i = \frac{1}{|\mathbf{P}_i|} \sum_{j \in \mathbf{P}_i} \mathbf{w}_j$.
- **Candidate Scoring:** Evaluate each candidate word c_i by measuring its average similarity to all cluster centroids: $s_i = \frac{1}{M} \sum_{j=1}^M \cos(\mathbf{c}_i, \mu_j)$, where cosine similarity serves as the alignment metric.
- **Candidate Selection:** Update the candidate set by retaining the top- K candidates with the highest alignment scores: $\mathbf{C}_{\text{new}} = \arg \text{top-K}_{c_i \in \mathbf{C}}(s_i)$, where $K = |\mathbf{C}|/2$.
- **Embedding Refresh:** Recompute embeddings for the new candidate set \mathbf{C}_{new} using CLIP’s reference word embedding function: $\mathbf{C}_{\text{new}} = f_t(\mathbf{C}_{\text{new}})$.

Here, the update cycle R is a hyperparameter. This closed-loop refinement strategy enables continuous adaptation to emergent data patterns. The candidate set evolves from generic initial wide range concepts (e.g., "red", "green", "blue", "burgundy", "emerald", "cyan", ... for color) to dataset-specific semantics (e.g., "green", "emerald", "cyan", ...) through iterative feedback from the clustering process.

Remark 1 Initially, the LLM generates $2^B M$ candidate words (refer to the **Dual-Constraint Proxy Optimization** section), where $\beta = E/R$, E represents the total number of training epochs, and R denotes the interval for updating candidates. After completing E epochs of training, the candidate words are reduced to M , aligning with the number of ground-truth classes. This ensures that clustering is not misled by erroneous guidance and effectively filters out dataset-irrelevant candidate words generated by the LLM throughout the process.

Dual-Constraint Proxy Optimization

User Interest Constraints To enforce proxies alignment with domain concepts, we initialize candidate words $\mathcal{C} = \{c_1, \dots, c_{2^\beta M}\}$ using GPT-4 (e.g., {"red", "blue", "green"} for user interest u "color") with embeddings $\mathbf{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_{2^\beta M}\} \in \mathbb{R}^{2^\beta M \times d}$, where M is the total number of ground-truth classes, embedding of each candidate is computed by $\mathbf{c}_i = f_t(c_i)$. Notably, β is a scalar parameter calculated by E/R , where E is the total training epochs and R is the configurable referring interval hyperparameter introduced in **Dynamic Candidate Management** section. These Candidate embeddings inject GPT-4’s domain knowledge as semantic priors. Each proxy is computed as a semantic-weighted combination:

$$\mathbf{w}_i = \sum_{k=1}^{|\mathcal{C}|} \alpha_{ik} \mathbf{c}_k, \quad \alpha_{ik} = \frac{\exp(\mathbf{w}_i'^\top \mathbf{c}_k / \tau_\alpha)}{\sum_j^{|\mathcal{C}|} \exp(\mathbf{w}_i'^\top \mathbf{c}_j / \tau_\alpha)}, \quad (5)$$

where \mathbf{w}_i' denotes the basic proxy and τ_α the temperature parameter. Proxies \mathbf{w}_i explicitly represent weighted combinations of domain concepts \mathbf{c}_k . The semantic consistency loss minimizes deviation from the concept centroid:

$$\mathcal{L}_u = \frac{1}{D} \sum_{i=1}^D \|\mathbf{w}_i - \bar{\mathbf{c}}\|_2^2, \quad \bar{\mathbf{c}} = \frac{1}{|\mathcal{C}|} \sum_{k=1}^{|\mathcal{C}|} \mathbf{c}_k, \quad (6)$$

where $\bar{\mathbf{c}} \in \mathbb{R}^d$ is the centroid of candidate embeddings. \mathcal{L}_u ensures semantic coherence with user-specified concept u .

Concept Discrimination Constraints To enhance cluster separability, we employ contrastive learning on fused features $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_B\}$ within a batch:

$$\mathcal{L}_c = \frac{1}{B} \sum_{i=1}^B \log \sum_{j \neq i} \exp(\mathbf{f}_i^\top \mathbf{w}_j / \sigma), \quad (7)$$

where σ controls negative sample hardness and B is training batch. The inner term $\sum_{j \neq i} \exp(\mathbf{f}_i^\top \mathbf{w}_j / \sigma)$ computes an exponential weighted sum of the similarities between sample i ’s fused feature \mathbf{f}_i and all proxy vectors \mathbf{w}_j associated with clusters other than its own. Minimizing the logarithm of this sum ($\log(\cdot)$) strongly penalizes high similarity scores between \mathbf{f}_i and incorrect proxies \mathbf{w}_j ($j \neq i$).

Optimization Framework

Let \mathbf{v}_i and \mathbf{f}_i denote *visual* and *fused* features of sample i , respectively. The unified loss combines:

$$\mathcal{L} = \underbrace{\frac{1}{D} \sum_{i=1}^D (1 - \cos(\mathbf{f}_i, \mathbf{v}_i))}_{\text{Cross-modal Alignment } \mathcal{L}_a} + \alpha(t) \mathcal{L}_u + \beta(t) \mathcal{L}_c, \quad (8)$$

where constraint weights following adaptive schedules: $\alpha(t) = \min(0.5, 0.1 + 0.4 \frac{t}{E})$, $\beta(t) = 0.1 \times (1 - \cos(\frac{\pi t}{E}))$, where t denotes current epoch and E total epochs. Empirical evidence demonstrates that this dynamic scheduling design reduces the sensitivity of pre-defined hyperparameters

to different datasets while achieving consistent performance advantages. This design progressively strengthens semantic constraints while maintaining stable cluster discrimination throughout training. Moreover, the cross-modal alignment term encourages the integration of multi-modal features for the same sample, thereby reducing discrepancies among different modalities. Notably, the final clusters are calculated by fused features \mathbf{F} .

We present pseudo-code in Algorithm 1 to offer a clearer and more comprehensive introduction to our Multi-DProxy. Additionally, an anonymous code repository is provided in the **Supplementary Material** for further reference.

Theoretical Analysis

Proposition 1 (Proxy Stability) *The dynamic candidate update reduces semantic drift by bounding proxy divergence:*

$$\|\mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)}\|_2 \leq \gamma \max_k \|\mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)}\|_2, \quad (9)$$

where $\gamma = \max_i \sum_k \alpha_{ik}$ is the maximum attention mass (bounded by 1), and $\mathbf{c}_k^{(t)}$ denotes candidate k at iteration t . The bound ensures proxy stability during candidate updates.

Proposition 1 quantifies how candidate updates control semantic drift and provides theoretical justification for dynamic refinement (Proof in **Supplementary Material**).

Theorem 1 (Cross-modal Attention Discriminability) *The gradient of the alignment loss \mathcal{L}_{align} with respect to the query projection matrix \mathbf{W}_Q satisfies:*

$$\frac{\partial \mathcal{L}_{align}}{\partial \mathbf{W}_Q} \propto \sum_{i=1}^B \mathbf{v}_i \mathbf{v}_i^\top \mathbf{t}_i \mathbf{t}_i^\top \mathbf{\Lambda}_i + \mathcal{O}(\epsilon) \quad (10)$$

where $\mathbf{t}_i = f_t(t_i)$ and $\mathbf{v}_i = f_v(v_i)$ denote text and visual features for the i -th sample. $\mathbf{\Lambda}_i = \frac{\partial \mathcal{L}_{align}}{\partial \cos(\mathbf{f}_i, \mathbf{v}_i)} \cdot \frac{1}{\|\mathbf{f}_i\| \|\mathbf{v}_i\|}$ is normalization factor, B is batch size, and $\mathcal{O}(\epsilon)$ is higher-order terms. This demonstrates that visual features \mathbf{v}_i modulate text representations proportionally to their discriminative power $\mathbf{v}_i \mathbf{v}_i^\top$, prioritizing semantically salient attributes during fusion.

Theorem 1 reveals visual features gate text representation learning and explains why discriminative attributes are prioritized (Proof in **Supplementary Material**).

Experiment

Dataset

To demonstrate the effectiveness of Multi-DProxy, we conduct extensive evaluations across a diverse array of publicly available visual datasets commonly adopted for multi-clustering benchmarks (Yao, Qian, and Hu 2024b). This comprehensive datasets includes: **Stanford Cars** (Yao, Qian, and Hu 2024b), **Card**, **CMUface** (Günemann et al. 2014), **Flowers** (Yao, Qian, and Hu 2024b), **Fruit** (Hu et al. 2017), **Fruit360** (Yao et al. 2023), and **CIFAR-10** (Yao, Qian, and Hu 2024a). Detailed introduction and statistical information are provided in the **Supplementary Material**.

Algorithm 1: Multi-DProxy Framework

Require: \mathcal{D} : Dataset $\{v_i, t_i\}_{i=1}^D$, $f_v(\cdot)$, $f_t(\cdot)$: Pre-trained CLIP encoders, u : User interest concept (e.g., "color"), M : Number of ground-truth classes, E : Total training epochs, R : Candidate update interval, K : Initial candidate size ($K = 2^\beta M$ where $\beta = E/R$).

- 1: **Initialize:**
- 2: Initialize $\mathbf{W} = \mathbf{w}_1, \dots, \mathbf{w}_D$.
- 3: Get $\mathbf{V} = \mathbf{v}_1, \dots, \mathbf{v}_D$, where $\mathbf{v}_i = f_v(v_i)$.
- 4: Get $\mathbf{T} = [\mathbf{t}_1; \mathbf{w}_1], \dots, [\mathbf{t}_D; \mathbf{w}_D]$, where $\mathbf{t}_i = f_t(t_i)$.
- 5: Generate candidates $\mathcal{C} \leftarrow \text{GPT-4}(u)$ with $|\mathcal{C}| = 2^\beta M$.
- 6: Get $\mathbf{C} \leftarrow f_t(\mathcal{C})$.
- 7: **for** epoch $t = 1$ to E **do**
- 8: *// Gated Cross-Modal Fusion:*
- 9: **for** $l = 1$ to L **do**
- 10: $\mathbf{V}_{\text{attn}}^l \leftarrow \text{MultiHead}(\mathbf{V}^{l-1}, \mathbf{T}^{l-1}, \mathbf{T}^{l-1})$.
- 11: $\mathbf{T}_{\text{attn}}^l \leftarrow \text{MultiHead}(\mathbf{T}^{l-1}, \mathbf{V}^{l-1}, \mathbf{V}^{l-1})$.
- 12: $\mathbf{V}^l \leftarrow \mathbf{V}^{l-1} + \sigma(\mathbf{W}_g^V[\mathbf{V}^{l-1}, \mathbf{V}_{\text{attn}}^l]) \odot \mathbf{V}_{\text{attn}}^l$.
- 13: $\mathbf{T}^l \leftarrow \mathbf{T}^{l-1} + \sigma(\mathbf{W}_g^T[\mathbf{T}^{l-1}, \mathbf{T}_{\text{attn}}^l]) \odot \mathbf{T}_{\text{attn}}^l$.
- 14: $\mathbf{V}^l \leftarrow \text{LayerNorm}(\mathbf{V}^l + \text{FFN}(\mathbf{V}^l))$.
- 15: $\mathbf{T}^l \leftarrow \text{LayerNorm}(\mathbf{T}^l + \text{FFN}(\mathbf{T}^l))$.
- 16: **end for**
- 17: $\lambda \leftarrow \sigma(\langle \mathbf{T}^L, \mathbf{V}^L \rangle / \tau)$.
- 18: $\mathbf{F} \leftarrow \lambda \mathbf{T}^L + (1 - \lambda) \mathbf{V}^L$.
- 19: *// Dual-Constraint Proxy Optimization:*
- 20: **for** $i = 1$ to D **do**
- 21: $\mathbf{w}_i \leftarrow \sum_{k=1}^{|\mathcal{C}|} \alpha_{ik} \mathbf{c}_k$.
- 22: **end for**
- 23: $\bar{\mathbf{c}} \leftarrow \frac{1}{|\mathcal{C}|} \sum_k \mathbf{c}_k$.
- 24: $\mathcal{L}_u \leftarrow \frac{1}{|\mathcal{B}|} \sum_i \|\mathbf{w}_i - \bar{\mathbf{c}}\|_2^2$.
- 25: $\mathcal{L}_c \leftarrow \frac{1}{|\mathcal{B}|} \sum_i \log \sum_{j \neq i} \exp(\mathbf{f}_i^\top \mathbf{w}_j / \sigma)$.
- 26: *// Optimization Framework:*
- 27: **for** batch $\mathcal{B} \in \mathcal{D}$ **do**
- 28: $\mathcal{L}_{\text{align}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_i (1 - \cos(\mathbf{f}_i, \mathbf{v}_i))$
- 29: $\mathcal{L} \leftarrow \mathcal{L}_{\text{align}} + \alpha(t) \mathcal{L}_u + \beta(t) \mathcal{L}_c$.
- 30: Update parameters via $\nabla \mathcal{L}$
- 31: **end for**
- 32: *// Dynamic Candidate Management:*
- 33: **if** $t \bmod R = 0$ **then**
- 34: $\mathbf{P} \leftarrow \text{KMeans}(\mathbf{W}, M)$.
- 35: Compute centroids $\boldsymbol{\mu}_i \leftarrow \frac{1}{|\mathbf{P}_i|} \sum_{j \in \mathbf{P}_i} \mathbf{w}_j$.
- 36: **for** each $\mathbf{c}_i \in \mathbf{C}$ **do**
- 37: $s_i \leftarrow \frac{1}{M} \sum_{j=1}^M \cos(\mathbf{c}_i, \boldsymbol{\mu}_j)$
- 38: **end for**
- 39: Get candidate words $\mathcal{C}_{\text{new}} \leftarrow \arg \text{top-K}_{\mathbf{c}_i \in \mathcal{C}}(s_i)$.
- 40: Update candidate embeddings $\mathbf{C} \leftarrow f_t(\mathcal{C}_{\text{new}})$.
- 41: **end if**
- 42: **end for**

Ensure: Fused features \mathbf{F} for clustering.

Baseline

We compare our proposed Multi-DProxy approach with eight state-of-the-art multiple clustering works. These works are as follows: **MSC** (Hu et al. 2017), **MCV** (Guérin and Boots 2018), **ENRC** (Miklautz et al. 2020), **iMClusts** (Ren et al. 2022), **AugDMC** (Yao et al. 2023), **DDMC** (Yao and Hu 2024), **Multi-DProxy** (Yao, Qian, and Hu 2024b), and **Multi-Sub** (Yao, Qian, and Hu 2024a). Details in the **Supplementary Material**.

Metric

To ensure the comparison’s fairness, we follow the widely used settings (Yao, Qian, and Hu 2024b,a) for evaluation. Specifically, we run k-means 10 times and report the average clustering performance using two metrics, namely, Normalized Mutual Information (NMI) (White, Steingold, and Fournelle 2004) and Rand index (RI) (Rand 1971). These metrics range from 0 to 1, with higher values indicating more accurate clustering results.

Hyperparameter

For each user’s preference, we train the model for $E = 1000$ epochs using the Adam optimizer with a momentum of 0.9. For all baselines, we conduct a comprehensive hyperparameter grid search based on their original papers. For our Multi-DProxy, we perform a grid search on learning rate from $\{1e^{-1}, 5e^{-2}, 1e^{-2}, 5e^{-3}, 1e^{-3}, 5e^{-4}\}$, weight decay from $\{5e^{-4}, 1e^{-4}, 5e^{-5}, 1e^{-5}, 0\}$ for all the experiments. Moreover, we tune candidate update interval R from $\{100, 200, 500\}$, temperature hyperparameters τ_α and σ from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, respectively. We follow most previous works in obtaining each clustering by applying KMeans (Lloyd 1982) to the learned representations. All experiments are performed on NVIDIA RTX 4090 GPUs.

Implementation Detail

To ensure fairness, we adopt the same settings as the baselines (Yao, Qian, and Hu 2024b,a) by selecting CLIP as the multi-modal model. Similarly, for LLMs, we follow the baselines (Yao, Qian, and Hu 2024b,a) to choose GPT-4. The impact of different multi-modal models and LLMs on performance is further explored and discussed in detail in the **Ablation Study** section.

Performance Comparison

Table 1 presents the clustering results, showing that Multi-DProxy consistently outperforms all baselines, highlighting its superiority. This demonstrates the strong generalization ability of the pre-trained multi-modal model, which effectively captures data features from diverse perspectives. Since our method employs the multi-modal encoder and LLM to generate clustering results, a natural question arises: **how would the performance compare if they were used directly in a zero-shot manner?**

To explore this, we introduce two zero-shot variants of CLIP: (1) **CLIP_{GPT}**, which uses GPT-4 to generate candidate labels and performs zero-shot classification with these labels as class names, and (2) **CLIP_{label}**, which directly uses

Dataset	Metric	Fruit		Fruit360		Card			CMUface			Stanford Cars		Flowers		CIFAR-10	
Clustering		Color	Species	Color	Species	Order	Suits	Emotion	Glass	Identity	Pose	Color	Type	Color	Species	Type	Environment
MSC	NMI↑	0.6886	0.1627	0.2544	0.2184	0.0807	0.0497	0.1284	0.1420	0.3892	0.3687	0.2331	0.1325	0.2561	0.1326	0.1547	0.1136
	RI↑	0.8051	0.6045	0.6054	0.5805	0.7805	0.3587	0.6736	0.5745	0.7326	0.6322	0.6158	0.5336	0.5965	0.5273	0.3296	0.3082
MCV	NMI↑	0.6266	0.2733	0.3776	0.2985	0.0792	0.0430	0.1433	0.1201	0.4637	0.3254	0.2103	0.1650	0.2938	0.1561	0.1618	0.1379
	RI↑	0.7685	0.6597	0.6791	0.6176	0.7128	0.3638	0.5268	0.4905	0.6247	0.6028	0.5802	0.5634	0.5860	0.6065	0.3305	0.3344
ENRC	NMI↑	0.7103	0.3187	0.4264	0.4142	0.1225	0.0676	0.1592	0.1493	0.5607	0.2290	0.2465	0.2063	0.3329	0.1894	0.1826	0.1892
	RI↑	0.8511	0.6536	0.6868	0.6984	0.7313	0.3801	0.6630	0.6209	0.7635	0.5029	0.6779	0.6217	0.6214	0.6195	0.3469	0.3599
iMClusts	NMI↑	0.7351	0.3029	0.4097	0.3861	0.1144	0.0716	0.0422	0.1929	0.5109	0.4437	0.2336	0.1963	0.3169	0.1887	0.2040	0.1920
	RI↑	0.8632	0.6743	0.6841	0.6732	0.7658	0.3715	0.5932	0.5627	0.8260	0.6114	0.6552	0.5643	0.6127	0.6077	0.3695	0.3664
AugDMC	NMI↑	0.8517	0.3546	0.4594	0.5139	0.1440	0.0873	0.0161	0.1039	0.5875	0.1320	0.2736	0.2364	0.3556	0.1996	0.2855	0.2927
	RI↑	0.9108	0.7399	0.7392	0.7430	0.8267	0.4228	0.5367	0.5361	0.8334	0.5517	0.7525	0.7356	0.6931	0.6227	0.4516	0.4689
DDMC	NMI↑	0.8973	0.3764	0.4981	0.5292	0.1563	0.0933	0.1726	0.2261	0.6360	0.4526	0.6899	0.6045	0.6327	0.6148	0.3991	0.3782
	RI↑	0.9383	0.7621	0.7472	0.7703	0.8326	0.6469	0.7593	0.7663	0.8907	0.7904	0.8765	0.7957	0.7887	0.8321	0.5827	0.5547
Multi-MaP	NMI↑	0.8619	1.0000	0.6239	0.5284	0.3653	0.2734	0.1786	0.3402	0.6625	0.4693	0.7360	0.6355	0.6426	0.6013	0.4969	0.4598
	RI↑	0.9526	1.0000	0.8243	0.7582	0.8587	0.7039	0.7105	0.7068	0.9496	0.6624	0.9193	0.8399	0.7984	0.8103	0.7104	0.6737
Multi-Sub	NMI↑	0.9693	1.0000	0.6654	0.6123	0.3921	0.3104	0.2053	0.4870	0.7441	0.5923	0.7533	0.6616	0.6940	0.6724	0.5271	0.4828
	RI↑	0.9964	1.0000	0.8821	0.8504	0.8842	0.7941	0.8527	0.8324	0.9834	0.8736	0.9387	0.8792	0.8843	0.8719	0.7394	0.7096
Multi-DProxy	NMI↑	1.0000	1.0000	0.7058	0.6490	0.5319	0.5008	0.2189	0.7739	0.7609	0.6646	0.7610	0.6829	0.7101	0.6888	0.5863	0.5431
	RI↑	1.0000	1.0000	0.8855	0.8537	0.9101	0.8848	0.8548	0.8381	0.9849	0.8991	0.9403	0.8901	0.8939	0.8897	0.7684	0.7204

Table 1: Comparison with state-of-the-art methods across multiple clustering benchmarks.

Dataset	Clustering	CLIP _{GPT}		CLIP _{label}		Multi-DProxy	
Metric		NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑
Fruit	Color	0.7912	0.9075	0.8629	0.9780	1.0000	1.0000
	Species	0.9793	0.9919	1.0000	1.0000	1.0000	1.0000
Fruit360	Color	0.5613	0.7305	0.5746	0.7673	0.7058	0.8855
	Species	0.4370	0.7552	0.5364	0.7631	0.6490	0.8537
Card	Order	0.3518	0.8458	0.3518	0.8458	0.5319	0.9101
	Suits	0.2711	0.6123	0.2711	0.6123	0.5008	0.8848
Card	Order	0.3518	0.8458	0.3518	0.8458	0.5319	0.9101
	Suits	0.2711	0.6123	0.2711	0.6123	0.5008	0.8848
CMUface	Emotion	0.1576	0.6532	0.1590	0.6619	0.2189	0.8548
	Glass	0.2905	0.6869	0.4686	0.7505	0.7739	0.8381
	Identity	0.1998	0.6388	0.2677	0.7545	0.7609	0.9849
	Pose	0.4088	0.6473	0.4691	0.6409	0.6646	0.8991
Stanford Cars	Color	0.6539	0.8237	0.6830	0.8642	0.7610	0.9403
	Type	0.6207	0.7931	0.6429	0.8456	0.6829	0.8901
Flowers	Color	0.5653	0.7629	0.5828	0.7836	0.7101	0.8939
	Species	0.5620	0.7553	0.6019	0.7996	0.6888	0.8897
CIFAR-10	Type	0.4935	0.6741	0.5087	0.7102	0.5863	0.7684
	Environment	0.4302	0.6507	0.4643	0.6801	0.5431	0.7204

Table 2: Zero-shot performance comparison.

the ground truth label set for zero-shot classification. Note that CLIP_{label} leverages an unfair setting, as the ground truth labels are known in advance, providing an upper bound for CLIP’s zero-shot performance. The results, shown in Table 2, align with expectations: CLIP_{label} generally outperforms CLIP_{GPT} due to the fixed and accurate ground truth labels, whereas CLIP_{GPT} relies on candidate labels that may introduce noise. Notably, both methods achieve identical performance on the Cards dataset, as GPT-4 generates candidate labels perfectly matching the ground truth.

Furthermore, Multi-DProxy outperforms CLIP_{GPT} in nearly all cases, indicating that the proposed method learns more effective features through its training process. Even when compared to CLIP_{label}, which benefits from access to the ground truth, Multi-DProxy achieves superior results in certain cases, such as clustering by color in the Fruit360 dataset. This is because CLIP tends to emphasize features from a single aspect, whereas Multi-DProxy learns a more comprehensive embedding of diverse features by leverag-

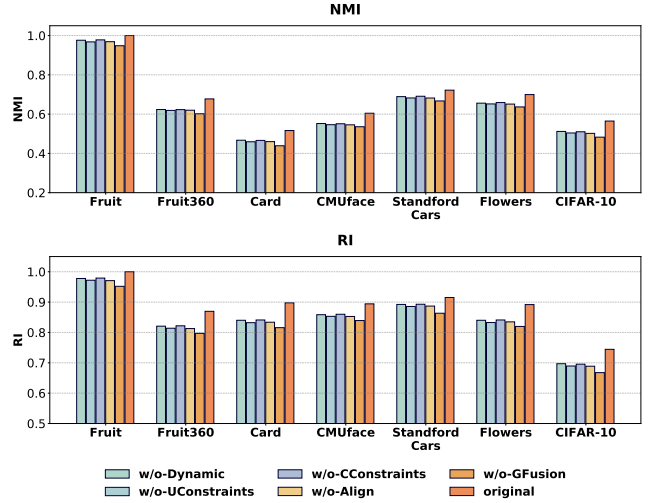


Figure 2: Ablation study. For each dataset, the average performance across all clustering objects is reported.

ing user-supervised training. Additionally, Multi-DProxy achieves competitive performance relative to CLIP_{label} in other cases, further validating the effectiveness of our proposed Multi-DProxy.

Ablation Study

To validate the effectiveness of our Multi-DProxy, we conduct experiments to justify the importance of key components. We design the following variants: 1) *w/o-Dynamic*, which removes Dynamic Candidate Management component and directly generates M candidate words via LLM. 2) *w/o-UConstraints*, which removes User Interest Constraints component. 3) *w/o-CConstraints*, which removes Concept Discrimination Constraints component. 4) *w/o-GFusion*, which replaces Gated Cross-Modal Fusion component by directly concatenating visual and textual repre-

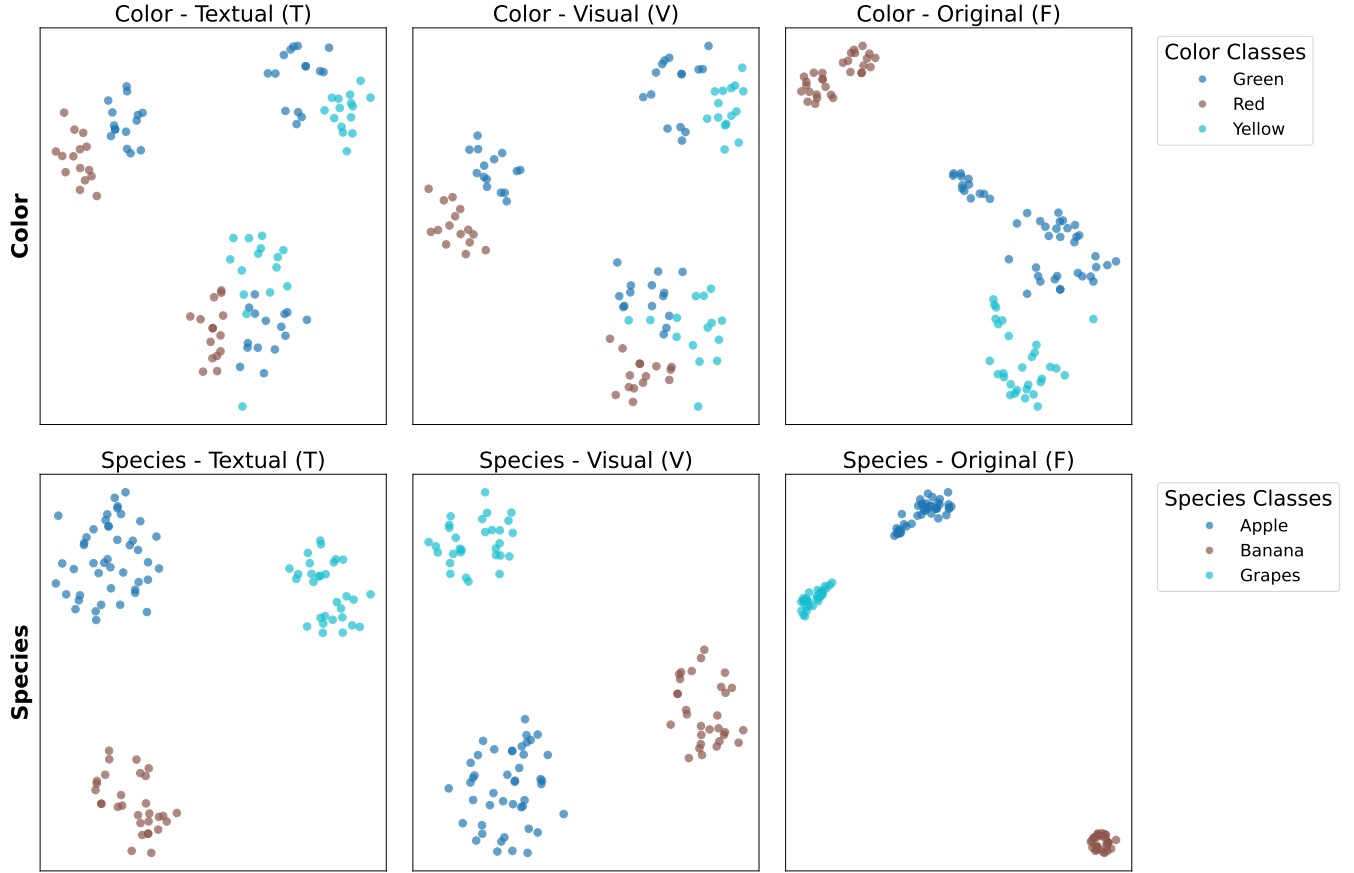


Figure 3: Visualization of textual, visual, and fused representations on the Fruit dataset.

sentations. For each dataset, the average performance across all clustering objects is reported. The results in Figure 2 show that the removal of any component results in a performance drop, demonstrating the effectiveness of all components. Furthermore, the Gated Cross-Modal Fusion component has a more significant impact on the model’s performance. Therefore, we conducted further exploration of the individual contributions of each modality.

We design the following variants: 1) $-T$, which only uses textual modality. 2) $-V$, which only uses visual modality¹. For each dataset, the average performance across all clustering objects is reported. The results in Figure 3 demonstrate that each modality possesses the capability to perform clustering independently. Moreover, the fused representation, enhanced by our tailored modality aggregation and alignment tasks, achieves significantly superior performance.

To further demonstrate the effectiveness of the fused representation, we visualize the representations obtained for $-T$, $-V$, and Original Multi-DProxy. The results are shown in Figure 3. Using visual information alone fails to achieve sat-

¹The ablation study evaluates the visual, textual, and fused representations of Multi-DProxy after 1000 training epochs. Thanks to the Cross-Modality Alignment component, the visual representation acquires clustering capabilities aligned with user interests.

Variant	Metric	$-T$	$-V$	Original
Fruit	NMI↑	0.7639	0.7421	1.0000
	RI↑	0.7719	0.7471	1.0000
Fruit360	NMI↑	0.5439	0.5326	0.6774
	RI↑	0.7410	0.7321	0.8696
Card	NMI↑	0.4439	0.4312	0.5164
	RI↑	0.8219	0.8138	0.8975
CMUface	NMI↑	0.5322	0.5233	0.6046
	RI↑	0.8231	0.8150	0.8942
Stanford Cars	NMI↑	0.6459	0.6378	0.7220
	RI↑	0.8199	0.8120	0.9152
Flowers	NMI↑	0.6369	0.6248	0.6995
	RI↑	0.8245	0.8129	0.8918
CIFAR-10	NMI↑	0.5030	0.4925	0.5647
	RI↑	0.6875	0.6789	0.7444

Table 3: Ablation study. For each dataset, the average performance across all clustering objects is reported.

isfactory clustering performance, whereas the fused representation, combining visual and textual information, better aligns the clustering results with user interests.

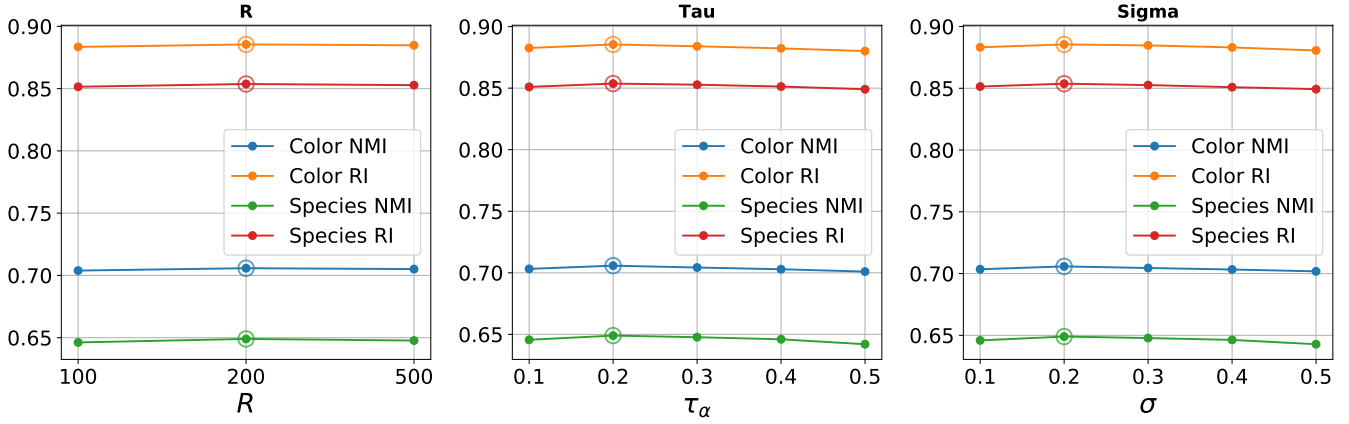


Figure 4: Hyperparameter analysis on the Fruit dataset.

In-depth Analysis

We investigate the impact of different multi-modal models and LLMs on Multi-DProxy’s performance. For multi-modal models, we select CLIP, ALIGN (Jia et al. 2021), and BLIP2 (Li et al. 2023), with ALIGN and BLIP2 offering larger datasets and stronger representation capabilities than CLIP. For LLMs, we choose GPT4, GPT4o (Yang et al. 2023), and DeepSeekV3 (Liu et al. 2024), with DeepSeekV3 and GPT4o providing richer knowledge than GPT4. We selected two representative datasets, and the experimental results in Table 4 show that different LLMs have a minimal impact on performance, as they are only used for generating candidate words. In contrast, stronger multi-modal models further enhance the performance of Multi-DProxy.

We further explored the efficiency advantages of Multi-DProxy over the sub-optimal baselines (Multi-Sub and Multi-MaP). We reported the average running time for clustering objects across two datasets. As shown in Figure 5, our method achieves significantly higher efficiency compared to these sub-optimal baselines.

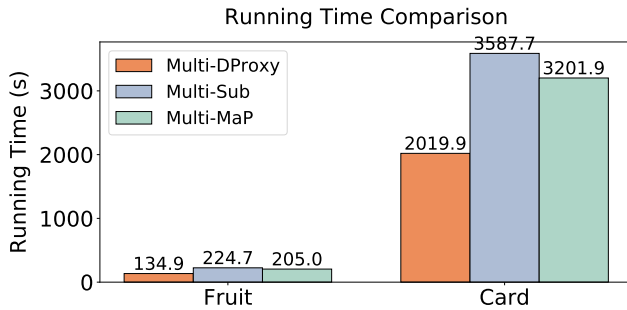


Figure 5: Efficiency study on the Fruit and Card datasets.

Hyperparameter Analysis

We further investigate the effect of candidate update interval R , temperature hyperparameter τ_α , and hyperparameter σ , respectively. We provide results on the Fruit360 dataset.

Dataset	Metric	Fruit360		Card	
Clustering		Color	Species	Order	Suits
CLIP-GPT4	NMI↑	0.7058	0.6490	0.5319	0.5008
	RI↑	0.8855	0.8537	0.9101	0.8848
CLIP-GPT4o	NMI↑	0.7059	0.6487	0.5323	0.5014
	RI↑	0.8862	0.8539	0.9109	0.8860
CLIP-DeepSeekV3	NMI↑	0.7048	0.6451	0.5322	0.5006
	RI↑	0.8850	0.8516	0.9088	0.8829
ALIGN-GPT4	NMI↑	0.7289	0.6647	0.5809	0.5215
	RI↑	0.8998	0.8717	0.9218	0.8901
ALIGN-GPT4o	NMI↑	0.7300	0.6655	0.5815	0.5224
	RI↑	0.9007	0.8728	0.9223	0.8911
ALIGN-DeepSeekV3	NMI↑	0.7292	0.6661	0.5820	0.5219
	RI↑	0.8995	0.8725	0.9219	0.8904
BLIP2-GPT4	NMI↑	0.7281	0.6597	0.5628	0.5178
	RI↑	0.8995	0.8699	0.9190	0.8872
BLIP2-GPT4o	NMI↑	0.7268	0.6600	0.5641	0.5183
	RI↑	0.9002	0.8711	0.9197	0.8881
BLIP2-DeepSeekV3	NMI↑	0.7285	0.6592	0.5629	0.5191
	RI↑	0.9000	0.8715	0.9188	0.8874

Table 4: Performance comparison across different multi-modal models and LLMs.

Figure 4 shows that the optimal choices for the temperature hyperparameters τ_α and σ are both 0.2, while the optimal choice for the candidate update interval R is 200. Notably, the model’s hyperparameters exhibit good performance across a reasonable range.

Conclusion

In this work, we presented Multi-DProxy, a dynamic proxy learning framework that overcomes the semantic rigidity of existing multi-modal clustering methods. By integrating learnable textual proxies refined through clustering feedback and gated cross-modal fusion that prioritizes discriminative features, our approach achieves precise alignment with user interests. Extensive validation across a board set of benchmarks shows state-of-the-art performance.

Acknowledgements

This work was supported by the Hong Kong UGC General Research Fund no. 17203320 and 17209822, and the project grants from the HKU-SCF FinTech Academy.

References

- Bae, E.; and Bailey, J. 2006. 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)*, 53–62. IEEE.
- Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.
- Chu, T. C.; Tong, S.; Ding, T.; Dai, X.; Haeffele, B.; Vidal, R.; and Ma, Y. 2024. Image clustering via the principle of rate reduction in the age of pretrained models. *International Conference on Learning Representations (ICLR)*.
- Dang, X. H.; and Bailey, J. 2010. Generation of alternative clusterings using the cami approach. In *Proceedings of the 2010 SIAM international conference on data mining*, 118–129. SIAM.
- Guérin, J.; and Boots, B. 2018. Improving image clustering with multiple pretrained cnn feature extractors. *arXiv preprint arXiv:1807.07760*.
- Günemann, S.; Färber, I.; Rüdiger, M.; and Seidl, T. 2014. Smvc: semi-supervised multi-view clustering in subspace projections. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 253–262.
- Hu, J.; Qian, Q.; Pei, J.; Jin, R.; and Zhu, S. 2017. Finding multiple stable clusterings. *Knowledge and Information Systems*, 51: 991–1021.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137.
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, 281–298. University of California press.
- Miklautz, L.; Mautz, D.; Altinigneli, M. C.; Böhm, C.; and Plant, C. 2020. Deep embedded non-redundant clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5174–5181.
- Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.
- Ouldnooghi, R.; Kuo, C.-W.; and Kira, Z. 2023. Clipgcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*.
- Poli, R.; Healy, M.; and Kameas, A. 2010. *Theory and applications of ontology: Computer applications*. Springer.
- Qian, Q. 2023. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16645–16654.
- Qian, Q.; Xu, Y.; Hu, J.; Li, H.; and Jin, R. 2022. Unsupervised visual representation learning by online constrained k-means. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16640–16649.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336): 846–850.
- Ren, L.; Yu, G.; Wang, J.; Liu, L.; Domeniconi, C.; and Zhang, X. 2022. A diversified attention model for interpretable multiple clusterings. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8852–8864.
- White, J.; Steingold, S.; and Fournelle, C. 2004. Performance metrics for group-detection algorithms. *Proceedings of Interface*, 2004: 5.
- Xu, J.; Chen, Z.; Yang, S.; Li, J.; Wang, H.; and Ngai, E. C. 2025a. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12908–12917.
- Xu, J.; Chen, Z.; Yang, S.; Li, J.; Wang, W.; Hu, X.; Hoi, S.; and Ngai, E. 2025b. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint arXiv:2502.15711*.
- Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.
- Yao, J.; and Hu, J. 2024. Dual-disentangled deep multiple clustering. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 679–687. SIAM.
- Yao, J.; Liu, E.; Rashid, M.; and Hu, J. 2023. Augdmc: Data augmentation guided deep multiple clustering. *Procedia Computer Science*, 222: 571–580.
- Yao, J.; Qian, Q.; and Hu, J. 2024a. Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yao, J.; Qian, Q.; and Hu, J. 2024b. Multi-modal proxy learning towards personalized visual multiple clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14066–14075.

Appendix

Proof of Proposition 1

Consider the proxy formulation at iteration t :

$$\mathbf{w}_i^{(t)} = \sum_{k=1}^K \alpha_{ik}^{(t)} \mathbf{c}_k^{(t)}. \quad (11)$$

After candidate update at $t+1$:

$$\mathbf{w}_i^{(t+1)} = \sum_{k=1}^K \alpha_{ik}^{(t+1)} \mathbf{c}_k^{(t+1)}. \quad (12)$$

The difference is bounded by:

$$\begin{aligned} \left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2 &= \left\| \sum_k \alpha_{ik}^{(t+1)} \mathbf{c}_k^{(t+1)} - \sum_k \alpha_{ik}^{(t)} \mathbf{c}_k^{(t)} \right\|_2 \\ &\leq \underbrace{\left\| \sum_k \alpha_{ik}^{(t+1)} (\mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)}) \right\|_2}_{\text{Term A}} + \underbrace{\left\| \sum_k (\alpha_{ik}^{(t+1)} - \alpha_{ik}^{(t)}) \mathbf{c}_k^{(t)} \right\|_2}_{\text{Term B}} \end{aligned} \quad (13)$$

Bounding Term A: Since α_{ik} are convex weights ($\sum_k \alpha_{ik} = 1, \alpha_{ik} \geq 0$):

$$\begin{aligned} \text{Term A} &\leq \sum_k \alpha_{ik}^{(t+1)} \left\| \mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)} \right\|_2 \\ &\leq \left(\sum_k \alpha_{ik}^{(t+1)} \right) \max_k \left\| \mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)} \right\|_2. \end{aligned} \quad (14)$$

Bounding Term B: From the softmax formulation:

$$\left| \alpha_{ik}^{(t+1)} - \alpha_{ik}^{(t)} \right| \leq L_\alpha \left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2, \quad (15)$$

where L_α is the Lipschitz constant of softmax. Since candidate embeddings are bounded ($\|\mathbf{c}_k\| \leq M$):

$$\text{Term B} \leq M \sum_k \left| \alpha_{ik}^{(t+1)} - \alpha_{ik}^{(t)} \right| \leq MKL_\alpha \left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2, \quad (16)$$

where K denotes the number of candidate words and M denotes the number of cores for K-means. Combining both bounds:

$$\left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2 \leq \gamma \max_k \left\| \mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)} \right\|_2 + \kappa \left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2, \quad (17)$$

where $\gamma = \sum_k \alpha_{ik}^{(t+1)}$ and $\kappa = MKL_\alpha$. For sufficiently small $\kappa < 1$ (ensured by normalization):

$$\left\| \mathbf{w}_i^{(t+1)} - \mathbf{w}_i^{(t)} \right\|_2 \leq \frac{\gamma}{1 - \kappa} \max_k \left\| \mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)} \right\|_2. \quad (18)$$

Taking $\gamma = \max_i \sum_k \alpha_{ik} = 1$ completes the proof.

Proof of Theorem 1

The alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \frac{1}{B} \sum_{i=1}^B (1 - \cos(\mathbf{F}_i, \mathbf{v}_i)), \quad (19)$$

where $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ and \mathbf{F}_i is the fused representation from gated cross-modal fusion (Eq. 4). The attention output (Eq. 1) is:

$$\mathbf{T}_{\text{attn}} = \text{MultiHead}(\mathbf{T}, \mathbf{V}, \mathbf{V}). \quad (20)$$

For single-head attention (generalizable to multi-head):

$$\begin{aligned} \mathbf{Q} &= \mathbf{T} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{V} \mathbf{W}_K, \quad \mathbf{V}_{\text{val}} = \mathbf{V} \mathbf{W}_V, \\ \mathbf{T}_{\text{attn}} &= \text{Softmax} \left(\frac{\mathbf{Q} \mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V}_{\text{val}}, \end{aligned} \quad (21)$$

The gradient decomposes as:

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \mathbf{W}_Q} = \sum_{i=1}^B \frac{\partial \mathcal{L}_{\text{align}}}{\partial \cos_i} \cdot \frac{\partial \cos_i}{\partial (\mathbf{F}_i^\top \mathbf{v}_i)} \cdot \frac{\partial (\mathbf{F}_i^\top \mathbf{v}_i)}{\partial \mathbf{W}_Q}, \quad (22)$$

where $\frac{\partial \mathcal{L}_{\text{align}}}{\partial \cos_i} = -\frac{1}{B}$, $\frac{\partial \cos_i}{\partial (\mathbf{F}_i^\top \mathbf{v}_i)} = \mathbf{\Lambda}_i$, and $\frac{\partial (\mathbf{F}_i^\top \mathbf{v}_i)}{\partial \mathbf{W}_Q} = \mathbf{v}_i^\top \frac{\partial \mathbf{F}_i}{\partial \mathbf{W}_Q}$.

Using residual connection (Eq. 2):

$$\begin{aligned} \mathbf{F}_i &\approx \mathbf{t}_i + \mathbf{T}_{\text{attn}, i}, \\ \frac{\partial \mathbf{F}_i}{\partial \mathbf{W}_Q} &\approx \frac{\partial}{\partial \mathbf{W}_Q} \left(\sum_{j=1}^B a_{ij} \mathbf{v}_j \mathbf{W}_V \right), \end{aligned} \quad (23)$$

where $a_{ij} = \text{Softmax} \left(\frac{\mathbf{t}_i \mathbf{W}_Q (\mathbf{v}_j \mathbf{W}_K)^\top}{\sqrt{d_k}} \right)$. Then, we got:

$$\frac{\partial a_{ij}}{\partial \mathbf{W}_Q} = \sum_{k=1}^B \frac{\partial a_{ij}}{\partial s_{ik}} \frac{\partial s_{ik}}{\partial \mathbf{W}_Q}, \quad (24)$$

where $s_{ik} = \frac{1}{\sqrt{d_k}} (\mathbf{t}_i \mathbf{W}_Q) (\mathbf{v}_k \mathbf{W}_K)^\top$, $\frac{\partial a_{ij}}{\partial s_{ik}} = a_{ij} (\delta_{jk} - a_{ik})$, and $\frac{\partial s_{ik}}{\partial \mathbf{W}_Q} = \frac{1}{\sqrt{d_k}} \mathbf{t}_i^\top \otimes (\mathbf{v}_k \mathbf{W}_K)$. Under diagonal-dominant attention ($a_{ii} \gg a_{ij}, j \neq i$):

$$\frac{\partial \mathbf{F}_i}{\partial \mathbf{W}_Q} \approx a_{ii} (1 - a_{ii}) \left[\frac{1}{\sqrt{d_k}} \mathbf{t}_i^\top \otimes (\mathbf{v}_i \mathbf{W}_V) \mathbf{W}_K^\top \mathbf{v}_i \right]. \quad (25)$$

Combining components:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{align}}}{\partial \mathbf{W}_Q} &\propto \sum_{i=1}^B \mathbf{v}_i^\top \left[a_{ii} (1 - a_{ii}) \cdot \frac{1}{\sqrt{d_k}} \mathbf{t}_i^\top \otimes (\mathbf{v}_i \mathbf{W}_V) \mathbf{W}_K^\top \mathbf{v}_i \right] \mathbf{\Lambda}_i \\ &\approx \sum_{i=1}^B \underbrace{\mathbf{v}_i \mathbf{v}_i^\top \mathbf{t}_i \mathbf{t}_i^\top \mathbf{\Lambda}_i \cdot a_{ii} (1 - a_{ii}) \frac{1}{\sqrt{d_k}} (\mathbf{W}_V \mathbf{W}_K^\top \mathbf{v}_i)}_{\text{scalar factor}}. \end{aligned} \quad (26)$$

Furthermore, the dominant term is:

$$\sum_{i=1}^B \mathbf{v}_i \mathbf{v}_i^\top \mathbf{t}_i \mathbf{t}_i^\top \mathbf{\Lambda}_i. \quad (27)$$

Dataset

To demonstrate the effectiveness of Multi-DProxy, we conduct extensive evaluations across a diverse array of publicly available visual datasets commonly adopted for multi-clustering benchmarks (Yao, Qian, and Hu 2024b). This

Dataset	# Samples	# Hand-crafted features	# Clusters
Fruit	105	shape descriptors; color histogram	3;3
Fruit360	4,856	shape descriptors; color histogram	4;4
Card	8,029	symbol shapes; color distribution	13;4
CMUface	640	HOG; edge maps	4;20;2;4
Stanford Cars	1,200	wheelbase length; body shape; color histogram	4;3
Flowers	1,600	petal shape; color histogram	4;4
CIFAR-10	60,000	edge detection; color histograms; shape descriptors	2;3

Table 5: Statistics of the experimental datasets.

comprehensive datasets includes: **Stanford Cars** (Yao, Qian, and Hu 2024b) (1,200 samples; clustering by color and vehicle type), **Card** (Yao et al. 2023) (8,029 samples; clustering by rank and suit), **CMUface** (Günnemann et al. 2014) (640 samples; clustering by pose, identity, glasses presence, and emotion), **Flowers** (Yao, Qian, and Hu 2024b) (1,600 samples; clustering by color and species), **Fruit** (Hu et al. 2017) (105 samples; clustering by species and color), **Fruit360** (Yao et al. 2023) (4,856 samples; clustering by species and color), and **CIFAR-10** (Yao, Qian, and Hu 2024a) (clustering by object type and environment). These datasets represent standard evaluations capturing varied multi-clustering challenges. Detailed statistical information regarding data size, feature representations, and cluster configurations is summarized in Table 5.

Notably, some data may encounter challenges in extracting meaningful candidate categories from LLMs, or their labels may lack semantic features. For instance, in identity clustering on the CMUface dataset (Günnemann et al. 2014), different identities represent distinct individuals, and the semantic meaning of names should not influence clustering outcomes. In such cases, following the widely used settings in previous works (Yao, Qian, and Hu 2024b,a), we randomly select candidate words from WordNet (Poli, Healy, and Kameas 2010) as reference categories.

Baselines

We compare our proposed Multi-DProxy approach with eight state-of-the-art multiple clustering works. These works are as follows:

- **MSC** (Hu et al. 2017), which utilizes hand-crafted features to automatically identify distinct feature subspaces for different clustering.
- **MCV** (Guérin and Boots 2018), which employs multiple pre-trained feature extractors to represent different views of the same data.
- **ENRC** (Miklautz et al. 2020), which integrates autoencoders with clustering objectives to generate diverse clustering.
- **iMClusters** (Ren et al. 2022), which leverages the representational power of deep autoencoders and multi-head attention to produce multiple salient embedding matrices and corresponding clustering.
- **AugDMC** (Yao et al. 2023), which uses data augmentations to automatically extract features corresponding

to various aspects of the data through a self-supervised prototype-based representation learning approach.

- **DDMC** (Yao and Hu 2024), which combines disentangled representation learning with a variational Expectation-Maximization (EM) framework.
- **Multi-DProxy** (Yao, Qian, and Hu 2024b), which relies on contrastive user-defined concepts to learn proxies tailored to user-specific interests.
- **Multi-Sub** (Yao, Qian, and Hu 2024a), which incorporates multi-modal subspace proxy learning and leverages the synergistic capabilities of CLIP and GPT-4 to better capture user preferences.

In our experiments, we include both traditional and deep learning-based baselines. Traditional methods rely on hand-crafted features, whereas deep learning methods directly process the original images as input.