MLLM Enriched Explainable Multiple Clustering

Shan Zhang[1], Liangrui Ren[1], Qiaoyu Tan[2,3], Carlotta Domeniconi[4], Wei Du[1], Jun Wang[1], Guoxian Yu[1]*

[1]School of Software and Joint SDU-NTU Centre for AI Research, Shandong University, Jinan, China
[2]NYU Shanghai Center for Data Science, Shanghai, China
[3]New York University Shanghai, Shanghai, China
[4]Department of Computer Science, George Mason University, VA, USA
{szhangs, renliangrui}@mail.sdu.edu.cn, qiaoyu.tan@nyu.edu, carlotta@cs.gmu.edu, {duwei, kingjun, gxyu}@sdu.edu.cn

Abstract

Multiple clustering aims to uncover diverse latent structures within the data, enabling a more comprehensive understanding of complex datasets. However, existing approaches either heavily rely on user-supplied keywords or disregard user-interested clustering types, limiting the ability to discover the full range of explainable clusterings of interests, particularly in high-dimensional settings. Furthermore, existing methods insufficiently leverage the rich textual semantics and fall short in fully integrating multi-modal information. To address these challenges, we propose MLLM enriched explainable Multiple Clustering ($\texttt{MLLM}_{MC}$), a novel framework that leverages multi-modal large language model (MLLM) to explore explainable non-redundant clustering. Specifically, $\texttt{MLLM}_{MC}$ first employs MLLM to generate sample descriptions, which serve as input for LLM to perform prompt-driven reasoning and infer latent clustering types, and then merges them with user-interested types to obtain diverse and explainable clustering types. For each selected type, $\texttt{MLLM}_{MC}$ utilizes MLLM to generate sample-level textual descriptions and aligns them with corresponding visual features through a cross-attention fusion module, which produces a semantically aligned and enriched representation for the target clustering type. Extensive experiments on six benchmark datasets from diverse domains demonstrate that $\texttt{MLLM}_{MC}$ achieves diverse, explainable, and high-quality clustering outcomes, outperforming state-of-the-art multiple clustering methods with a large margin.

Extended version —
https://www.sdu-idea.cn/codes.php?name=MLLMMC

Introduction

Clustering is a fundamental task that partitions unlabeled data into coherent groups, revealing the underlying structural patterns based on intrinsic similarity (Oyewole and Thopil 2023). Conventional clustering methods (Chang et al. 2017a; Park et al. 2021; Ren et al. 2020) often presume a single globally optimal partition, optimized by objectives, such as minimizing intra-cluster variance and maximizing inter-cluster separation. However, real-world data are inherently multi-faceted, exhibiting diverse semantic structures and potential grouping criteria that cannot be captured by a single partition (Ren et al. 2022). For instance, face images may be clustered by identity, pose, or emotion, each reflecting distinct semantics. This motivates the need for multiple clustering approaches that can uncover diverse, non-redundant grouping structures within the same dataset (Yu et al. 2024).

Existing multiple clustering methods can be broadly divided into two streams. The first type of methods focuses on generating diverse clustering within the original data space, using unsupervised (Bae and Bailey 2006; Chang et al. 2017b; Wang et al. 2021) or semi-supervised (Tokuda, Yamashita, and Yoshimoto 2021; Yao et al. 2023) strategies. The second type of methods discover alternative clusterings across different feature subspaces (Mautz et al. 2020; Wang et al. 2019), often leveraging deep neural networks (Miklautz et al. 2020; Ren et al. 2022; Yao and Hu 2024) to learn nonlinear projections that support clustering diversity and disentanglement.

More recently, the rise of multi-modal large language models (MLLMs) (Radford et al. 2021; Li et al. 2021, 2023) has enabled semantic-driven multiple clustering by aligning visual inputs with user-supplied keywords (Yao, Qian, and Hu 2024a,b; Kwon et al. 2024). In addition, Luo et al. (2024) proposed the concept of subpopulation structures, using large language models (LLMs) to reveal latent subpopulation distributions, which aligns with the idea of multiple clustering in capturing diverse data structures. Despite these advances, existing approaches face two critical limitations. First, these methods either overly rely on user-supplied keywords, which often reflect shallow prior knowledge and restrict the scope of semantic discovery, or solely depend on LLM reasoning, which may miss user-interested clustering types. In practice, clustering type selection is application-specific and should respect user interests for explainable and target usage. To address this, we propose using MLLMs and LLMs as knowledgeable assistants to support users in identifying clustering types, offering meaningful suggestions that may be overlooked based on user interests. Second, although MLLMs facilitate image-text alignment, the final clustering is typically made using the visual embedding alone or simple feature concatenation, which underutilize the enriched textual descriptions of samples. Prior work has shown that text features generated from images can significantly en-

*Corresponding author

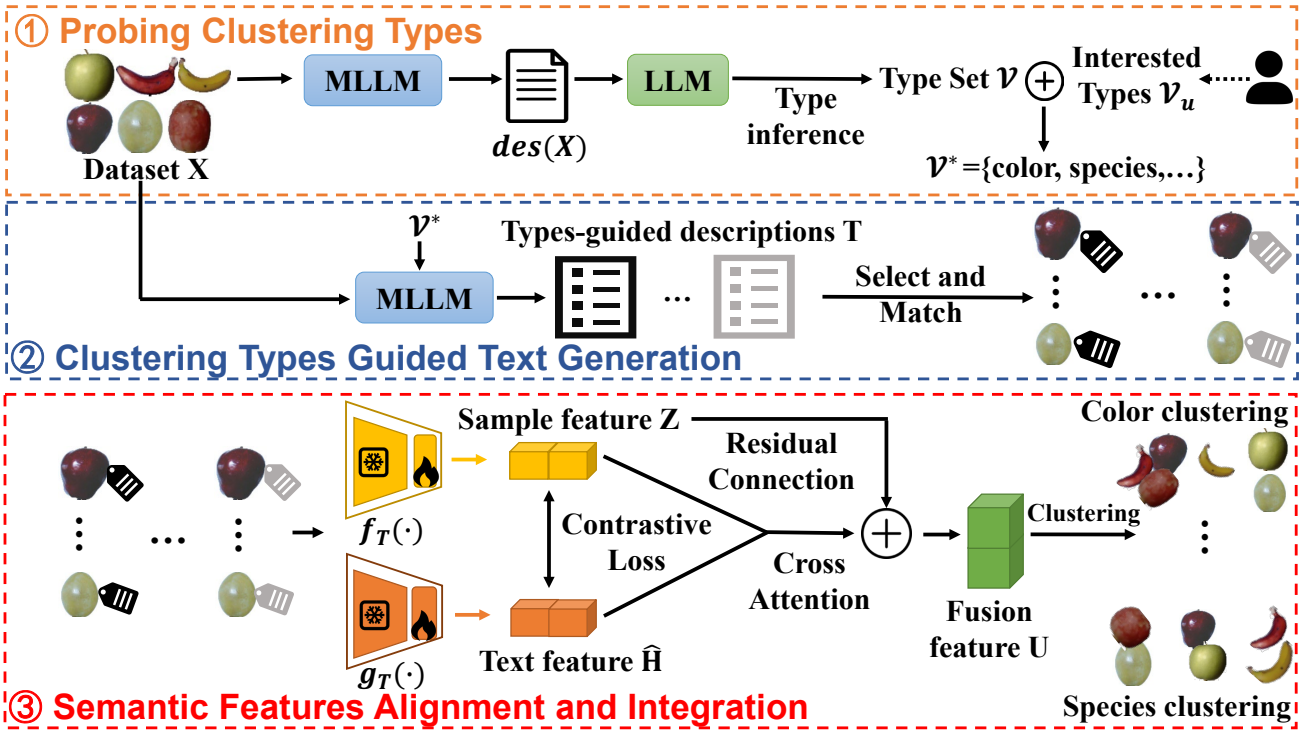


Figure 1: Framework overview of $\texttt{MLLM}_{MC}$. Given a dataset $\mathbf{X}$, we first use MLLM to generate dataset descriptions $des(\mathbf{X})$, which are then fed into LLM to infer a candidate clustering type set $\mathcal{V}$. Merging $\mathcal{V}$ with user-interested types $\mathcal{V}_u$ yields the augmented clustering type set $\mathcal{V}^*$. For each type, we select candidate semantic tags $\{\tilde{\mathbf{T}}_k\}_{k=1}^K$ from MLLM-generated descriptions $\{\mathbf{T}_k\}_{k=1}^K$, and match them with samples to form sample-text pairs $[\mathbf{x}_i, \hat{\mathbf{t}}_{k,i}]_{i=1}^N$, which are fused via cross-attention and residual connections into the representation $\{\mathbf{U}_k\}_{k=1}^K$. Finally, k-means is applied to $\{\mathbf{U}_k\}_{k=1}^K$ to yield multiple clusterings $\{\mathcal{C}_k\}_{k=1}^K$.

hance clustering performance by capturing latent semantics beyond visual similarity (Stephan et al. 2024). As such, we argue that integrating text representations into the clustering process rather than treating them as auxiliary cues, can provide a more comprehensive and semantically aligned understanding of clustering.

Motivated by these insights, we propose $\texttt{MLLM}_{MC}$, a novel framework that seamlessly integrates semantic textual representations generated by MLLMs with original visual features to achieve clustering from diverse perspectives, as illustrated in Fig. 1. $\texttt{MLLM}_{MC}$ addresses the dual challenges of limited user interests and under-exploited data semantics, enabling the discovery of richer clustering structures that align with both user interests and latent data characteristics. To address the ambiguity in the clustering objectives, $\texttt{MLLM}_{MC}$ first employs MLLM to produce sample descriptions, which subsequently serve as input to LLM to infer a set of candidate clustering types. If user-interested types are not among the generated candidates, they can be incorporated as additional guidance. These clustering types then guide the generation of sample-level textual descriptions via a multi-modal generative model, which conditions on the visual content and reflects the semantics of each clustering type (e.g., color-related descriptions for color-based clustering). Subsequently, $\texttt{MLLM}_{MC}$ adopts a cross-attention fusion mechanism to align and integrate visual and textual features, producing semantically enriched joint representations for clustering. This unified approach enables $\texttt{MLLM}_{MC}$ to produce diverse, explainable, and robust clustering results, offering a flexible yet principled way to uncover complex data structures from multiple semantic types.

The key contributions are summarized as follows: (i) $\texttt{MLLM}_{MC}$ leverages the reasoning capabilities of MLLMs and LLMs to assist users in probing latent and semantically meaningful clustering types, while allowing the incorporation of user interests for task-specific refinement. (ii) $\texttt{MLLM}_{MC}$ enhances the expressiveness of clustering by generating textual representations aligned with clustering types, and by fusing them with visual features via cross-attention to reveal deeper latent structures. (iii) We conduct extensive experiments on multiple benchmark datasets, showing that $\texttt{MLLM}_{MC}$ outperforms state-of-the-art baselines in producing diverse, high-quality, and explainable clustering results.

Related Work

Multiple Clustering

Multiple clustering aim to uncover diverse and semantically distinct data partitions beyond a single optimal structure. Early multiple clustering approaches primarily focused on generating alternative clustering within the original feature space (Bailey 2018; Yu et al. 2024). Some methods

(Yang and Zhang 2017; Chang et al. 2017b) adopt a sequential, semi-supervised strategy, where each new clustering is generated based on previous ones as reference points. Others aim to simultaneously identify multiple clusterings by jointly optimizing clustering quality and diversity, using criteria such as cluster centroids (Jain, Meka, and Dhillon 2008), mutual information (Dang and Bailey 2010), and diversity (Ren et al. 2023). Another line of work (Cui, Fern, and Dy 2007; Wang et al. 2019; Mautz et al. 2020) explores non-redundant clusterings across different feature subspaces, under the assumption that distinct subspaces encode different semantic views of the data. Deep learning models further extend this direction, where Miklautz et al. (2020) utilized autoencoders to learn multiple independent latent spaces, while Ren et al. (2022) introduced redundancy-controlled multi-head attention mechanisms to generate diverse nonlinear subspaces. Several studies (Yao et al. 2019; Wei et al. 2020, 2021) also expand this paradigm to multi-view or graph-structured data, demonstrating how different data modalities can support distinct clustering structures. More recent methods aim to combine automatic structure discovery with model-level diversity. For instance, Leiber et al. (2022) proposed a non-redundant clustering framework based on the principle of minimum description length (MDL). Metaxas, Tzimiropoulos, and Patras (2023) proposed a plug-and-play framework that incorporates a novel diversity-constrained loss into deep clustering models. These methods have promoted the development of clustering research towards deep representation learning and automatic structure discovery. Additionally, data augmentation strategies have been studied to generate distinct clusterings from multiple diversely augmented representations (Yao et al. 2023; Yao and Hu 2024; Zhang et al. 2025).

To improve the usability and adaptability, recent works begin to incorporate user interests through vision–language models. Yao, Qian, and Hu (2024a,b) leveraged MLLMs to align user-supplied keywords with image content, thereby generating clustering results that align with user interests. Kwon et al. (2024) used LLMs to cluster images based on user-supplied textual criteria. In addition, Luo et al. (2024) proposed the concept of subpopulations to capture multiple clustering structures, relying entirely on LLMs to infer the underlying subpopulations. However, such approaches either overly rely on subjective user inputs or solely depend on LLMs. In contrast, our $\texttt{MLLM}_{MC}$ integrates LLM reasoning and user interests to identify clustering types, where these types further guide the MLLM to generate the semantic representation, enabling the model to uncover richer underlying structures. In this way, $\texttt{MLLM}_{MC}$ empowers more comprehensive and explainable multiple clustering.

Multi-Modal Fusion

In multi-modal learning, effectively fusing visual and textual features is essential. Current fusion methods fall into three main categories: vector-based operations, deep learning architectures, and graph neural network (GNN) methods (Kuang et al. 2025).

Traditional fusion methods rely on straightforward mathematical operations, such as element-wise addition, element-wise multiplication, or vector concatenation, to combine features from different modalities. Due to their simplicity and computational efficiency, these methods are widely adopted in early multi-modal systems (Gandhi et al. 2023). Deep learning-based fusion approaches offer more expressive modeling capabilities. These can be further divided into non-attentional, attentional, and hierarchical strategies. Non-attentional methods typically employ CNNs (Ma, Lu, and Li 2016) and RNNs (Ren, Kiros, and Zemel 2015) to integrate aligned multi-modal features. The development of attention mechanisms has significantly advanced the precision of multi-modal fusion (Lu et al. 2023), where early attention models focus solely on salient visual regions, while co-attention (Ren et al. 2024), multi-level attention networks (Zhou et al. 2024; Xue et al. 2022) and symmetric attention (Sun et al. 2024) facilitate fine-grained alignment and reasoning across modalities. More recently, GNN-based methods leverage structural information by constructing scene, question, or semantic dependency graphs to model intra- and inter-modal relations (Zhang, Tsai, and Tsai 2024; Li et al. 2019).

In the context of clustering, recent studies have explored the use of external textual signals to guide clustering towards more semantically coherent structures (Cai et al. 2023; Qiu et al. 2024; Li et al. 2024). However, these methods focus on a single grouping, missing diverse structural insights. In contrast, $\texttt{MLLM}_{MC}$ leverages MLLMs to generate different perspective texts aligned with data, enabling the discovery of different explainable clusterings.

Methodology

In this section, we elaborate on the proposed $\texttt{MLLM}_{MC}$ framework, including clustering types exploration, clustering types informed text descriptions generation, and semantic features alignment and integration. The conceptual framework is presented in Fig. 1.

Probing Clustering Types

For datasets with high dimensionality or ambiguous semantics, it is often challenging for users to discern latent attributes and meaningful clustering types, thereby impeding the discovery of underlying data structures. In contrast, LLMs, with their broad knowledge and contextual understanding, offer an effective way to identify potential clustering types based on semantic and task-related cues.

However, LLMs require textual input and cannot directly process non-textual datasets. To address this, $\texttt{MLLM}_{MC}$ leverages the MLLM to convert each sample in the dataset into a textual description. These descriptions are then fed into the LLM to probe the dataset's potential clustering types. If the generated outputs do not align with the user's interests, user-interested types can be incorporated as supplementary input to refine and extend the set of candidate clustering types. This enables a task-oriented, knowledge-driven process for probing meaningful clusterings.

Specifically, given a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$, we first input its samples into the MLLM to generate textual descriptions $des(\mathbf{X})$, which then serves as input of LLM to probe

potential clustering types of $\mathbf{X}$. The probing prompt is denoted as $Pmt(\cdot)$ and follows a simple template, such as: `Based on the dataset descriptions, infer the possible clustering types for this dataset.`

In response, LLM generates a set of candidate clustering types, denoted as $\mathcal{V}$. Formally, we represent the prompt as $Pmt(des(\mathbf{X}))$, and probe the clustering type set $\mathcal{V}$ as:

$$\mathcal{V} = LLM(Pmt(des(\mathbf{X}))), \tag{1}$$

To ensure task-specific relevance, user-interested clustering types can be optionally included when $\mathcal{V}$ does not include user's interests, thereby giving a target set:

$$\mathcal{V}^* = \mathcal{V} \cup \mathcal{V}_u, \tag{2}$$

where $\mathcal{V}_u$ represents the set of user-interested clustering types, reflecting user interests and $\mathcal{V}^* = \{v_k\}_{k=1}^K$ is the final set of clustering types.

Clustering Types Guided Text Generation

After determining the target clustering type set $\mathcal{V}^*$, $\texttt{MLLM}_{MC}$ further leverages MLLM to generate sample-level textual descriptions $\mathbf{T}_k = \{\mathbf{t}_{k,i}\}_{i=1}^N$ relevant to the k-th clustering type v_k. These descriptions are distinct from the earlier descriptions used for clustering type probing. While the earlier descriptions aim to provide a coarse-grained and general overview of the dataset, the current ones focus on producing fine-grained and type-guided descriptions. This targeted generation aligns well with clustering types, leading to interpretable multiple clusterings. Specifically, we design prompts for each clustering type and input them into MLLM to guide the generation of sample-level descriptions. For example, when clustering based on color, the prompt `"What is the color of the image?"` can effectively elicit color-related information.

However, due to the subjectivity of generated descriptions, directly using them may result in noisy or inconsistent representations of sample semantics. To address this, we apply TF-IDF to $\mathbf{T}_k$ and extract the top M representative terms as candidate semantic tags:

$$\tilde{\mathbf{T}}_k = TopM(TF-IDF(\mathbf{T}_k)), \tag{3}$$

where $\tilde{\mathbf{T}}_k = \{\tilde{\mathbf{t}}_{k,j}\}_{j=1}^M$ are the candidate semantic tags for clustering v_k, and M is the number of clusters within it.

We then perform similarity-based alignment between samples and semantic tags in the feature space. The embedding of sample $\mathbf{x}_i$ is calculated as $\mathbf{e}_i = f_M(\mathbf{x}_i)$, and the embedding of the candidate semantic tag $\tilde{\mathbf{t}}_{k,j}$ is calculated as $\tilde{\mathbf{o}}_{k,j} = g_M(\tilde{\mathbf{t}}_{k,j})$, where $f_M(\cdot)$ and $g_M(\cdot)$ are feature extractors for original samples and candidate tags, respectively. The similarity between sample $\mathbf{x}_i$ and each candidate semantic tag $\tilde{\mathbf{t}}_{k,j}$ is computed as:

$$s_{i,k,j} = \mathbf{e}_i^T \tilde{\mathbf{o}}_{k,j}. \tag{4}$$

Then, the matched semantic tag for sample $\mathbf{x}_i$ is selected as:

$$\hat{\mathbf{t}}_{k,i} = arg\max_j s_{i,k,j}, \tag{5}$$

Finally, each sample $\mathbf{x}_i$ is paired with its most relevant semantic tag $\hat{\mathbf{t}}_{k,i}$, and the matched pairs $[\mathbf{x}_i, \hat{\mathbf{t}}_{k,i}]_{i=1}^N$ are used as enriched input for subsequent multi-modal fusion and multiple clustering tasks.

Semantic Features Alignment and Integration

Given the matched pairs $(\mathbf{x}_i, \hat{\mathbf{t}}_{k,i})$ of original samples and their matched semantic tags, we employ a frozen CLIP model augmented with a trainable Transformer layer, denoted as $f_T(\cdot)$ and $g_T(\cdot)$, to extract visual and textual embeddings, respectively. Notably, this setup is not fixed and can be flexibly replaced with alternative feature extractors. Specifically, the visual embedding $\mathbf{z}_i$ of sample $\mathbf{x}_i$ is obtained as $\mathbf{z}_i = f_T(\mathbf{x}_i)$, while the textual embedding $\hat{\mathbf{h}}_{k,i}$ of tag $\hat{\mathbf{t}}_{k,i}$ is computed as $\hat{\mathbf{h}}_{k,i} = g_T(\hat{\mathbf{t}}_{k,i})$. To ensure semantic consistency, we employ a contrastive objective that pulls aligned pairs together and pushes unaligned ones apart in the embedding spaces:

$$l_{k,i} = -log\frac{exp(\mathbf{z}_i^T \hat{\mathbf{h}}_{k,i}/\tau)}{\sum_{j=1}^N exp(\mathbf{z}_i^T \hat{\mathbf{h}}_{k,j}/\tau)}, \tag{6}$$

$$L_{con} = \frac{1}{KN}\sum_{k=1}^K \sum_{i=1}^N l_{k,i}, \tag{7}$$

where τ is the temperature parameter and L_{con} is the total contrastive loss computed over all samples. Guided by $\hat{\mathbf{h}}_{k,i}$, $\mathbf{z}_i$ is adapted into a clustering-specific representation $\mathbf{z}_{k,i}$, which is semantically aligned with the k-th clustering.

To integrate complementary information from both modalities, $\texttt{MLLM}_{MC}$ introduces a cross-attention fusion module that aligns the visual features toward semantically meaningful directions guided by texts. Specifically, in the cross-attention, $\mathbf{z}_{k,i}$ is treated as the query $\mathbf{Q}$, while $\hat{\mathbf{h}}_{k,i}$ serves as both the key $\mathbf{K}$ and value $\mathbf{V}$:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{8}$$

$$\mathbf{z}_{k,i}^* = Attention(\mathbf{z}_{k,i}, \hat{\mathbf{h}}_{k,i}, \hat{\mathbf{h}}_{k,i}), \tag{9}$$

where $\mathbf{z}_{k,i}^*$ denotes the text-guided visual representation after attention.

Instead of solely relying on the attended feature $\mathbf{z}_{k,i}^*$, $\texttt{MLLM}_{MC}$ employs a residual connection to combine it with the original input $\mathbf{z}_{k,i}$:

$$\mathbf{u}_{k,i} = \mathbf{z}_{k,i} + \mathbf{z}_{k,i}^*, \tag{10}$$

where $\mathbf{u}_{k,i}$ is the fused representation. Residual connection ensures that the original structural in the raw data is retained. Since the attention output $\mathbf{z}_{k,i}^*$ emphasizes semantic cues aligned with the textual modality, exclusively relying on it may suppress modality-specific features that are not directly referenced by the text. By incorporating the residual path, $\texttt{MLLM}_{MC}$ makes a balance between semantic alignment and structural integrity. Subsequently, $\texttt{MLLM}_{MC}$ applies k-means on the fused representations $\{\mathbf{U}_k\}_{k=1}^K$ to derive the multiple clustering outputs $\{\mathcal{C}_k\}_{k=1}^K$. We give the algorithmic procedure, time complexity and runtimes in the supplementary file.

Experiments

Experimental Setup

Baselines. We compare $\texttt{MLLM}_{MC}$ against representative and competitive multiple clustering methods: MNMF (Yang and Zhang 2017), ENRC (Miklautz et al. 2020), iMClusts (Ren et al. 2022), Multi-MaP (Yao, Qian, and Hu 2024b) and Multi-Sub (Yao, Qian, and Hu 2024a). These methods were discussed in related work and their configurations are provided in the supplementary files.

Datasets. Six benchmark datasets (Fruit, CMUFace, COIL, Cards, CIFAR and WebKB) are used to evaluate the performance of $\texttt{MLLM}_{MC}$ and other baselines. The first five are image datasets, the sixth is a text dataset. These datasets have been widely used to benchmark multiple clustering methods (Bailey 2018; Yu et al. 2024). Their statistical information is given in Table 1, with detailed information provided in the supplementary files.

Evaluation metrics. To quantitatively evaluate each method, we measure the performance using the Normalized Mutual Information (NMI) and Jaccard Index (JI) with reference to distinct ground-truth labels. Higher scores indicate better clustering quality and stronger alignment with targets.

Implementation. In our experiments, all MLLMs utilized are based on LLaVA-1.5-7b (Liu et al. 2023). The LLMs utilized, including clustering types inference and textual dataset evaluations, are based on LLaMA-3.1-8B-Instruct (Touvron et al. 2023). The feature extraction is based on CLIP-ViT-B/32. Notably, $\texttt{MLLM}_{MC}$ is highly flexible in design and does not rely on any specific vision or language backbone, making it broadly extensible and adaptable. All experiments are conducted on a Linux system equipped with an NVIDIA L40 GPU (46 GB memory) and PyTorch 2.4.1. Furthermore, the supplementary files include prompt templates for the used MLLM and LLM.

Datasets	Samples	Clustering Types	Clusters
COIL	648	color; shape	3; 3
Fruit	4856	color; species	4; 4
Cards	8029	suits; number	13; 4
CIFAR	50000	environment; type	3; 2
WebKB	1041	university; category	4; 4
CMUFace	640	identity; pose; emotion; glass	20; 4; 4; 2

Table 1: Used benchmark datasets, the last column indicates the number of clusters of respective clusterings.

Results and analysis

Table 2 provides the average clustering results of 10 independent runs of each method on benchmark datasets. The CLIP backbone used by Multi-MaP and Multi-Sub is designed for image-text alignment and is less suitable for textual datasets (e.g., WebKB). On textual datasets, $\texttt{MLLM}_{MC}$ uses LLM to extract keywords and fuses them with the original text features, demonstrating the effectiveness of feature integration within the text modality. We can find that $\texttt{MLLM}_{MC}$ achieves the best results in almost all cases, proving its effectiveness. In addition, we have the following important observations:

(i) **Deep vs. Shallow methods:** In most cases, deep multiple clustering methods, ENRC, iMClusts, Multi-MaP, Multi-Sub and $\texttt{MLLM}_{MC}$, perform better than the shallow MNMF. ENRC and iMClusts obtain satisfactory clustering performance on conventional datasets (i.e., Fruit), which are characterized by distinctive features and well-defined structures. However, their performance significantly declines when handling more complex datasets with larger semantic diversity and structural intricacy. This limitation is due to their reliance solely on non-redundant subspaces of the data, without the support of semantic guidance from text cues or the powerful reasoning capability offered by LLMs.

(ii) **MLLMs vs. DNN methods:** For large and structurally complex datasets such as Cards and CIFAR, the performance of iMClusts and ENRC is notably inferior to that of methods based on MLLMs, including Multi-Map, Multi-Sub, and $\texttt{MLLM}_{MC}$. This fact indicates that MLLMs demonstrate significant advantages in semantic reasoning and enriching features toward to target clusterings. Notably, $\texttt{MLLM}_{MC}$ underperforms Multi-MaP and Multi-Sub on CIFAR environmental clustering, because its generated environmental descriptions include irrelevant words, whereas Multi-MaP and Multi-Sub avoid this by direct LLM-generated references. $\texttt{MLLM}_{MC}$ consistently outperforms Multi-MaP and Multi-Sub across most datasets, demonstrating the efficacy of our cross-modal fusion mechanism in integrating information from different modalities and generating high-quality multiple clusterings. Specifically, Multi-MaP focuses on specific image features through textual prompts, but the final clusterings still entirely rely on image features. Multi-Sub simply concatenates image and text features and disregards the interactions between them. Unlike prior methods, $\texttt{MLLM}_{MC}$ employs cross-attention and residual connections to deeply fuse image-text features while preserving sample-specific characteristics. By better capturing semantic structures and data diversity, it produces superior multiple clusterings.

Ablation Study

To investigate the contribution components of $\texttt{MLLM}_{MC}$, we conduct an ablation study by introducing four variants: w/oContrast, w/oAttn, and w/oResidual, w/oMatch, which separately remove the contrastive loss, cross-attention, residual connection, and candidate tag matching. Figure 2 presents the NMI values of $\texttt{MLLM}_{MC}$ and its variants. We find that the full $\texttt{MLLM}_{MC}$ consistently surpasses its ablated variants, highlighting the importance of each component in achieving high-quality and diverse clustering outcomes. Similar JI trends are provided in the supplementary file.

Among all variants, w/oContrast suffers the most significant performance drop, confirming the crucial role of contrastive loss in promoting semantic alignment between visual and textual modalities. The contrastive objective explicitly guides the model to distinguish positive sample–text pairs from negative pairs, thereby reinforcing the intermodal consistency. In its absence, the model lacks suffi-

Dataset	Type	Metrics	MNMF	ENRC	iMClusts	Multi-MaP	Multi-Sub	MLLM$_{MC}$
COIL	color	NMI	.013±.001•	.084±.012•	.183±.001•	.153±.000•	.155±.001•	**.991±.002**
		JI	.215±.002•	.232±.000•	.284±.001•	.256±.001•	.213±.002•	**.994±.001**
	shape	NMI	.055±.003•	.125±.013•	.126±.002•	.249±.001•	.251±.000•	**.660±.008**
		JI	.225±.001•	.263±.004•	.273±.004•	.350±.001•	.352±.001•	**.612±.005**
Fruit	color	NMI	.019±.112•	.402±.003•	.421±.002•	.624±.000•	.660±.001•	**.694±.007**
		JI	.152±.022•	.301±.032•	.315±.041•	.442±.001•	.461±.001•	**.514±.008**
	species	NMI	.018±.010•	.380±.110•	.410±.002•	.528±.001•	.610±.003•	**.628±.005**
		JI	.163±.110•	.297±.131•	.311±.032•	.415±.002•	.401±.000•	**.490±.007**
CMUFace	identity	NMI	.228±.010•	.504±.001•	.527±.004•	.625±.001•	.543±.007•	**.658±.017**
		JI	.051±.023•	.164±.011•	.197±.002•	.245±.000•	.201±.005•	**.306±.011**
	pose	NMI	.017±.022•	.023±.207•	.026±.004•	.119±.001○	.032±.006•	**.120±.005**
		JI	.141±.112•	.153±.050•	.170±.001•	.202±.001•	.201±.001•	**.240±.002**
	emotion	NMI	.005±.002•	.004±.002•	.003±.000•	.007±.000•	.010±.000•	**.012±.000**
		JI	.142±.001•	.140±.001•	**.165±.000**○	.147±.001•	.149±.000•	.152±.000
	glass	NMI	.007±.002•	.007±.001•	.006±.010•	.029±.000•	.008±.001•	**.249±.011**
		JI	.401±.001•	.362±.000•	.360±.001•	.355±.002•	.352±.001•	**.480±.004**
Cards	number	NMI	.052±.031•	.100±.001•	.124±.003•	.129±.001•	.153±.000•	**.575±.004**
		JI	.071±.010•	.092±.002•	.103±.001•	.077±.000•	.121±.000•	**.345±.005**
	suits	NMI	.032±.010•	.090±.000•	.100±.001•	.181±.000•	.170±.001•	**.382±.021**
		JI	.211±.010•	.140±.001•	.205±.001•	.225±.001•	.191±.002•	**.355±.009**
CIFAR	environment	NMI	.097±.001•	.189±.000•	.192±.001•	.460±.001○	**.483±.001**○	.220±.011
		JI	.222±.011•	.272±.002•	.352±.001•	.609±.001○	**.617±.001**○	.463±.004
	type	NMI	.051±.010•	.183±.001•	.204±.002•	.497±.001•	.527±.001•	**.764±.033**
		JI	.142±.014•	.250±.002•	.371±.002•	.594±.000•	.662±.001•	**.866±.005**
WebKB	university	NMI	.012±.001•	.217±.000•	.451±.001•	-	-	**.793±.008**
		JI	.180±.011•	.221±.002•	.352±.001•	-	-	**.745±.014**
	category	NMI	.011±.010•	.133±.001•	.106±.002•	-	-	**.296±.023**
		JI	.142±.014•	.150±.002•	.137±.002•	-	-	**.317±.027**

Table 2: Performance of baselines on generating multiple clusterings. •/○ indicates whether MLLM$_{MC}$ is superior/inferior to the other method, with statistical significance checked by pairwise t-test at 95% level. The best results are highlighted in **bold** font.

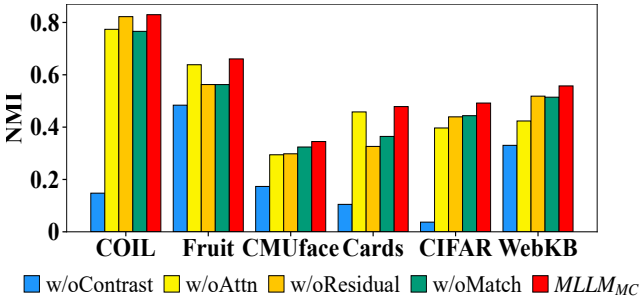


Figure 2: Results of MLLM$_{MC}$ and its variants.

cient semantic supervision, resulting in weaker feature fusion and diminished discriminative capacity, which ultimately degrades clustering performance. Furthermore, removing cross-attention, residual connections or candidate tags matching also causes performance degradation. The cross-attention mechanism enables the model to selectively attend to text features that are most relevant to the visual sample, facilitating targeted semantic fusion and enhancing representation alignment. The residual connection, on the other hand, ensures that critical information in the original visual features is preserved, particularly those elements that may not be fully captured by the text. The candidate tag matching enhances the validity and accuracy of the generated textual descriptions. In summary, these factors enable MLLM$_{MC}$ to achieve high-quality multiple clusterings.

Furthermore, to validate the effectiveness of integrating visual and textual modalities and the extensibility of MLLM$_{MC}$, we conducted three experiments. First, we compare MLLM$_{MC}$ with unimodal clustering using image-only and text-only inputs, as shown in Table 3. MLLM$_{MC}$ consistently outperforms unimodal baselines, proving that modality fusion enhances the representation for clustering. Notably, text features often achieve better performance than the image features, underscoring the critical role of semantic information. This observation again justifies that MLLMs can enrich semantic features to boost multiple clustering. Next, we evaluate MLLM$_{MC}$ by alternative MLLMs to generate textual descriptions for clustering and test on CIFAR-100, a dataset with a larger number of classes. The results and analyses are provided in the supplementary files.

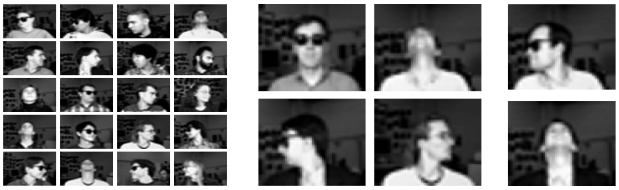
In additon, we examined the hyperparameters of MLLM$_{MC}$, including the number of transformer layers n_{layer}, learning rate lr, temperature parameter τ, and the number of candidate semantic tags M, with the results shown in the supplementary file. The results are stable within a certain hyperparameter range and fluctuate mod-

	COIL		Fruit		CMUFace				Cards		CIFAR	
	color	shape	color	species	identity	pose	emotion	glass	number	suits	environment	type
image	0.025	0.275	0.409	0.559	0.518	0.002	0.003	0.001	0.154	0.175	0.066	0.008
text	0.951	0.643	0.670	0.455	0.436	0.097	0.009	0.243	0.359	0.164	0.140	0.185
MLLM$_{MC}$	**0.990**	**0.662**	**0.694**	**0.623**	**0.658**	**0.120**	**0.012**	**0.258**	**0.575**	**0.382**	**0.213**	**0.764**

Table 3: Performance comparison of unimodal clustering and MLLM$_{MC}$. The best results are highlighted in **bold** font, and the second-best are marked with underlined.

erately beyond it, indicating MLLM$_{MC}$'s robustness and effectiveness across diverse conditions.

(a) **Color** and **Shape** clusterings on COIL.



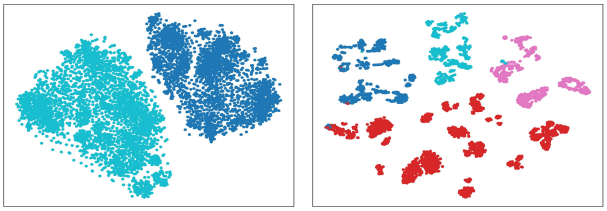
(b) **Identity**, **Pose** and **Glass** clusterings on CMUface.

Figure 3: Multiple clusterings on COIL and CMUface.

Visualization and Discovery of Novel Clusterings

To illustrate the interpretability of multiple clusterings generated by MLLM$_{MC}$, we visualize these for the COIL and CMUFace datasets. Figures 3(a) displays clustering results based on color and shape for COIL, where each row corresponds to a single cluster. Figures 3(b) presents identity-, pose-, and glasses-based clusterings on CMUFace. Across these cases, MLLM$_{MC}$ effectively distinguishes semantically distinct clustering types that align with the target clustering types set $\mathcal{V}^*$. These visualizations confirm the interpretability and practicality of MLLM$_{MC}$ in discovering and modeling diverse, meaningful clusterings within complex datasets.

To further check whether MLLM$_{MC}$ can leverage the prompting capabilities of MLLMs to discover novel clustering types beyond user interests, we further visualize the additional clustering types inferred by MLLMs that are not included in Table 2 and summarize the sources of these types in the supplementary files. Specifically, MLLM$_{MC}$ discovers *shape-based clustering* on the Fruit dataset, and *color-based clustering* on the Cards dataset. The shape of Fruit can be categorized into four types: irregular, heart, round, and banana, while Card suits can be divided into red and black. Since these clustering types lack ground-truth annotations, we apply t-SNE to visualize the clustering in a 2D scatter plot. Figure 4 reveals that new clusterings generated under these MLLM-suggested types exhibit clear structural separability, with well-defined inter-cluster boundaries and compact intra-cluster distributions, indicating strong semantic alignment and discriminative capability. This suggests that MLLMs can serve as effective tools in exploratory and heuristic clustering tasks, offering semantically coherent and structurally meaningful clustering types.



(a) Color clustering of Cards. (b) Shape clustering of Fruit.

Figure 4: 2D scatter plot of the **newly discovered clustering types** by MLLM$_{MC}$.

Conclusion and Limitations

In this paper, we propose a MLLM-enriched multiple clustering framework (MLLM$_{MC}$) that leverages the reasoning and knowledge representation capabilities of MLLMs and LLMs, and meticulously designed cross-attention based fusion on top of MLLMs, to address the key challenges of limited user interests and insufficient modality fusion in multiple clusterings. Experimental results on benchmark datasets demonstrate that MLLM$_{MC}$ outperforms existing methods in clustering diversity, quality and interpretability, and it can uncover novel clustering types.

Limitations and Broader impacts of MLLM$_{MC}$. Despite its effectiveness, MLLM$_{MC}$ still has several limitations. Its integration with MLLMs may introduce biases or hallucinations that affect clustering. Moreover, the performance of text generation depends on the generalization ability of MLLMs, which may vary across domains. Our work proves that MLLMs can serve as effective guidance tools in exploratory and heuristic clustering tasks, empower the discovery of semantically coherent and structurally meaningful clusterings. Future work will incorporate domain knowledge and improve robustness through adaptive prompting and alignment refinement.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2024YFF1206604), NSFC (62272276, 62432006 and 62531013), Shandong Provincial Natural Science Foundation (No. ZR2024JQ001), Taishan Scholars Program (No. tsqn202306007 and tsqn202408317), Postdoctoral Innovation Program of Shandong Province (No. SDCX-ZG-202501019), China Postdoctoral Science Foundation (No. 2025M771502).

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 *ICDM*, 53–62.

Bailey, J. 2018. Alternative clustering analysis: A review. *Data clustering*, 535–550.

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-enhanced image clustering. In *AAAI*, 6869–6878.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017a. Deep adaptive image clustering. In *ICCV*, 5879–5887.

Chang, Y.; Chen, J.; Cho, M. H.; Castaldi, P. J.; Silverman, E. K.; and Dy, J. G. 2017b. Multiple clustering views from multiple uncertain experts. In *ICML*, 674–683.

Cui, Y.; Fern, X. Z.; and Dy, J. G. 2007. Non-redundant multi-view clustering via orthogonalization. In *ICDM*, 133–142.

Dang, X. H.; and Bailey, J. 2010. Generation of alternative clusterings using the CAMI approach. In *SDM*, 118–129.

Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; and Hussain, A. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inform. Fus.*, 91: 424–444.

Jain, P.; Meka, R.; and Dhillon, I. S. 2008. Simultaneous unsupervised learning of disparate clusterings. *SADM*, 1(3): 195–210.

Kuang, J.; Shen, Y.; Xie, J.; Luo, H.; Xu, Z.; Li, R.; Li, Y.; Cheng, X.; Lin, X.; and Han, Y. 2025. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Comp. Surv.*, 57(8): 1–36.

Kwon, S.; Park, J.; Kim, M.; Cho, J.; Ryu, E. K.; and Lee, K. 2024. Image clustering conditioned on text criteria. In *ICLR*.

Leiber, C.; Mautz, D.; Plant, C.; and Böhm, C. 2022. Automatic parameter selection for non-redundant clustering. In *SDM*, 226–234.

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 *ICML*, 19730–19742.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 9694–9705.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-aware graph attention network for visual question answering. In *ICCV*, 10313–10322.

Li, Y.; Hu, P.; Peng, D.; Lv, J.; Fan, J.; and Peng, X. 2024. Image clustering with external guidance. In *ICML*, 27890–27902.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *NeurIPS*, 34892–34916.

Lu, S.; Liu, M.; Yin, L.; Yin, Z.; Liu, X.; and Zheng, W. 2023. The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Comp. Sci.*, 9: e1400.

Luo, Y.; An, R.; Zou, B.; Tang, Y.; Liu, J.; and Zhang, S. 2024. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *ECCV*, 235–252. Springer.

Ma, L.; Lu, Z.; and Li, H. 2016. Learning to answer questions from image using convolutional neural network. In *AAAI*, 3567–3573.

Mautz, D.; Ye, W.; Plant, C.; and Böhm, C. 2020. Non-redundant subspace clusterings with nr-kmeans and nr-dipmeans. *TKDD*, 14(5): 1–24.

Metaxas, I. M.; Tzimiropoulos, G.; and Patras, I. 2023. Divclust: controlling diversity in deep clustering. In *CVPR*, 3418–3428.

Miklautz, L.; Mautz, D.; Altinigneli, M. C.; Böhm, C.; and Plant, C. 2020. Deep embedded non-redundant clustering. In *AAAI*, 5174–5181.

Oyewole, G. J.; and Thopil, G. A. 2023. Data clustering: application and trends. *Artif. Intell. Rev.*, 56(7): 6439–6475.

Park, S.; Han, S.; Kim, S.; Kim, D.; Park, S.; Hong, S.; and Cha, M. 2021. Improving unsupervised image clustering with robust learning. In *CVPR*, 12278–12287.

Qiu, L.; Zhang, Q.; Chen, X.; and Cai, S. 2024. Multi-level cross-modal alignment for image clustering. In *AAAI*, volume 38, 14695–14703.

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 *ICML*, 8748–8763.

Ren, G.; Diao, L.; Guo, F.; and Hong, T. 2024. A co-attention based multi-modal fusion network for review helpfulness prediction. *IPM*, 61(1): 103573.

Ren, L.; Wang, J.; Li, Z.; Li, Q.; and Yu, G. 2023. scMCs: a framework for single-cell multi-omics data integration and multiple clusterings. *Bioinf.*, 39(4): btad133.

Ren, L.; Yu, G.; Wang, J.; Liu, L.; Domeniconi, C.; and Zhang, X. 2022. A diversified attention model for interpretable multiple clusterings. *TKDE*, 35(9): 8852–8864.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NeurIPS*, 2953–2961.

Ren, Y.; Wang, N.; Li, M.; and Xu, Z. 2020. Deep density-based image clustering. *KBS*, 197: 105841.

Stephan, A.; Miklautz, L.; Sidak, K.; Wahle, J. P.; Gipp, B.; Plant, C.; and Roth, B. 2024. Text-guided image clustering. In *EACL*, 2960–2976.

Sun, M.; Liu, X.; Wang, H.; and Liu, J. 2024. MixRGBX: Universal multi-modal tracking with symmetric mixed attention. *Neurocomp.*, 603: 128274.

Tokuda, T.; Yamashita, O.; and Yoshimoto, J. 2021. Multiple clustering for identifying subject clusters and brain sub-networks using functional connectivity matrices without vectorization. *Neural Netw.*, 142: 269–287.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, J.; Wang, X.; Yu, G.; Domeniconi, C.; Yu, Z.; and Zhang, Z. 2021. Discovering multiple co-clusterings with matrix factorization. *IEEE TCYB*, 51(7): 3576–3587.

Wang, X.; Wang, J.; Domeniconi, C.; Yu, G.; Xiao, G.; and Guo, M. 2019. Multiple independent subspace clusterings. In *AAAI*, 5353–5360.

Wei, S.; Wang, J.; Yu, G.; Domeniconi, C.; and Zhang, X. 2020. Multi-view multiple clusterings using deep matrix factorization. In *AAAI*, 6348–6355.

Wei, S.; Yu, G.; Wang, J.; Domeniconi, C.; and Zhang, X. 2021. Multiple clusterings of heterogeneous information networks. *Mach. Learn.*, 110(6): 1505–1526.

Xue, X.; Zhang, C.; Niu, Z.; and Wu, X. 2022. Multi-level attention map network for multimodal sentiment analysis. *TKDE*, 35(5): 5105–5118.

Yang, S.; and Zhang, L. 2017. Non-redundant multiple clustering by nonnegative matrix factorization. *Mach. Learn.*, 106(5): 695–712.

Yao, J.; and Hu, J. 2024. Dual-disentangled deep multiple clustering. In *SDM*, 679–687.

Yao, J.; Liu, E.; Rashid, M.; and Hu, J. 2023. Augdmc: Data augmentation guided deep multiple clustering. *Procedia Comp. Sci.*, 222: 571–580.

Yao, J.; Qian, Q.; and Hu, J. 2024a. Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning. In *NeurIPS*.

Yao, J.; Qian, Q.; and Hu, J. 2024b. Multi-modal proxy learning towards personalized visual multiple clustering. In *CVPR*, 14066–14075.

Yao, S.; Yu, G.; Wang, J.; Domeniconi, C.; and Zhang, X. 2019. Multi-view multiple clustering. In *IJCAI*, 4121–4127.

Yu, G.; Ren, L.; Wang, J.; Domeniconi, C.; and Zhang, X. 2024. Multiple clusterings: Recent advances and perspectives. *Comput. Sci. Rev.*, 52: 100621.

Zhang, J.; Tsai, P.-H.; and Tsai, M.-H. 2024. Semantic2Graph: graph-based multi-modal feature fusion for action segmentation in videos. *App. Intell.*, 54(2): 2084–2099.

Zhang, S.; Ren, L.; Wang, J.; Xu, Y.; Domeniconi, C.; and Yu, G. 2025. Aligning contrastive multiple clusterings with user interests. In *IJCAI*, 7011–7019.

Zhou, X.; Zhang, Y.; Wang, Z.; Lu, M.; and Liu, X. 2024. MAFN: multi-level attention fusion network for multimodal named entity recognition. *MTP*, 83(15): 45047–45058.