ESMC: MLLM-Based Embedding Selection for Explainable Multiple Clustering

Xinyue Wang[1], **Yuheng Jia**[1,2,3*], **Hui Liu**[3], **Junhui Hou**[4]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]Department of Computing and Information Sciences, Saint Francis University, Hong Kong, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong, China
213222801@seu.edu.cn, yhjia@seu.edu.cn, h2liu@sfu.edu.hk, jh.hou@cityu.edu.hk

Abstract

Typical deep clustering methods, while achieving notable progress, can only provide one clustering result per dataset. This limitation arises from their assumption of a fixed underlying data distribution, which may fail to meet user needs and provide unsatisfactory clustering outcomes. Our work investigates how multi-modal large language models (MLLMs) can be leveraged to achieve user-driven clustering, emphasizing their adaptability to user-specified semantic requirements. However, directly using MLLM output for clustering has risks for producing unstructured and generic image descriptions instead of feature-specific and concrete ones. To address these issues, our method first discovers that MLLMs' hidden states of text tokens are strongly related to the corresponding features, and leverages these embeddings to perform clusterings from any user-defined criteria. We also employ a lightweight clustering head augmented with pseudo-label learning, significantly enhancing clustering accuracy. Extensive experiments demonstrate its competitive performance on diverse datasets and metrics.

Code — https://github.com/JCSTARS/Embedding-Selective-Multiple-Clustering

1 Introduction

Clustering is an unsupervised machine learning technique focused on grouping objects based on their specific patterns (Jain, Murty, and Flynn 1999; Xu and Wunsch 2005). Traditional clustering methods (MacQueen 1967; Ester et al. 1996) often rely on hand-crafted features and struggle with high-dimensional data. Deep clustering (Van Gansbeke et al. 2020; Ren et al. 2024; Caron et al. 2018) overcomes this limitation by leveraging embeddings from deep neural networks. Most of these methods group the dataset with a single standard. However, users may require clustering a set of images based on multiple criteria, as illustrated in Figure 1. This scenario highlights the challenge of multi-clustering, where diverse grouping objectives coexist for the same dataset.

Multiple clustering demonstrates that a single dataset can be partitioned into distinct clusters based on different underlying structures (Yu et al. 2024). While existing methods have offered solutions for obtaining diverse clusters (Bae and Bailey 2006; Dasgupta and Ng 2010; Ren et al. 2022; Yao et al. 2023), they often rely on inherent data characteristics and lack the flexibility to directly incorporate user-defined criteria.

Foundational models like CLIP (Radford et al. 2021) recently enable clustering tasks (Yao, Qian, and Hu 2024b,a) by aligning image and text features. However, CLIP's prompt-agnostic image embeddings limit clustering based on semantic features. On the other hand, MLLMs like LLaVA (Liu et al. 2023, 2024) leverage LLMs and visual instruction tuning, enabling user-defined clustering through modality alignment.

However, naively using MLLMs for multiple clustering can face several challenges. Firstly, MLLMs tend to produce unstructured data (Li et al. 2024), making it difficult to link semantically similar outputs to specific clustering labels. Furthermore, the model's response may offer a too general image description instead of strictly following provided instructions. A comprehensive discussion of these problems is presented in Section 4.3. To mitigate these issues, we propose Embedding Selective Multiple Clustering (ESMC). As shown in Figure 2, ESMC first derives internal embeddings of text prompt tokens to extract features. These embeddings are from hidden states of text tokens of the language model in MLLMs. As visualized in Figure 3, color-related features are strongly related to the 'color' prompt token, while prompt-unrelated or image-unrelated tokens present low logits. We also employ a clustering head with pseudo-label learning to group these embeddings into well-separated clusters.

Our contributions can be summarized as follows.

- We reveal that the internal embeddings of text prompt tokens can serve as condensed representations of corresponding features and can be utilized as an effective basis for multiple clustering.
- We demonstrate that a lightweight 2-layer MLP clustering head with pseudo label supervision achieves substantial improvements.
- Extensive experiments and ablation studies indicate that the proposed approach provides user-defined, diverse, and high-quality clustering outcomes.

*Yuheng Jia is the corresponding author.

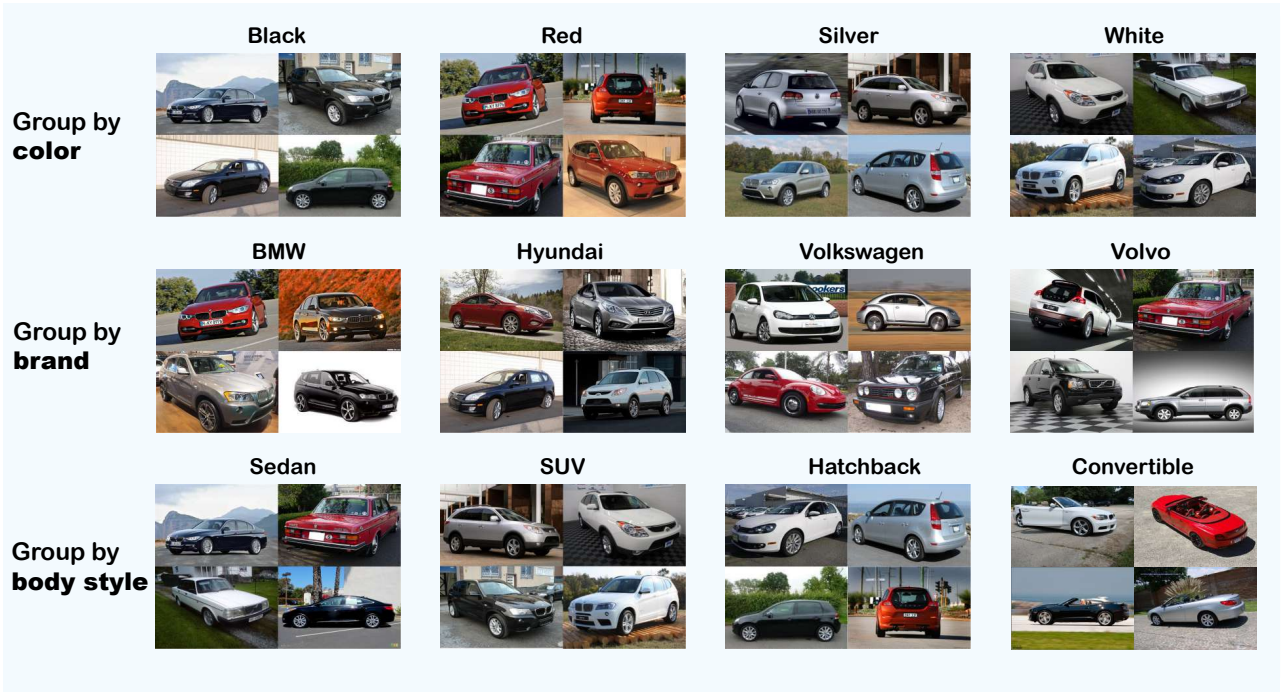

Figure 1: Multiple clustering of sample images on the Stanford_cars dataset with cluster criterion of color, manufacturer, and body style.

2 Related Work

Multiple clustering Multiple clustering aims to discover various meaningful ways to group a dataset (Yu et al. 2024). Traditional methods often achieve this by partitioning data based on its internal characteristics, such as constructing orthogonal subspaces (Cui, Fern, and Dy 2007; Niu, Dy, and Jordan 2013) or modeling independent mixture distributions (Jain, Meka, and Dhillon 2008; Tokuda et al. 2017), employing objective functions to improve both cluster quality and diversity (Yu et al. 2024). Deep learning methods further advance the field by learning richer and more diverse representations. For example, MCV (Guérin and Boots 2018) utilizes different pre-trained feature extractors as diverse views for the same data to construct multiple clusters. ENRC (Miklautz et al. 2020) finds multiple non-redundant clusters in the embedded space of the autoencoder. Multi-Map and Multi-Sub utilize CLIP (Radford et al. 2021) embeddings and proxy learning to generate multiple clusters. Most previous methods cannot provide clusters based on arbitrary user-defined criteria.

Multi-modal large language models MLLMs typically integrate three core components: a vision encoder for extracting visual features, a projection network to align visual and text modalities, and an autoregressive language backbone. Taking LLaVA-1.5 (Liu et al. 2024) as an example, its architecture comprises CLIP (Radford et al. 2021) as the vision encoder, a projection module, and Vicuna-v1.5 (Zheng et al. 2023) as the language foundation. Given an input image I, the CLIP image encoder $f_{vision\_encoder}$ will provide visual vectors. A mapping network f_{proj} is applied to project image features and text features to the same space and provides X_{visual} tokens:

$$\mathbf{X}_{visual} = f_{proj}(f_{vision\_enc}(I)) \in \mathbb{R}^{n \times d_{mllm}}, \quad (1)$$

where d_{mllm} is the dimension of MLLM embedding. The input prompt S is mapped into m tokens through the tokenizer $f_{tokenizer}$ to word embeddings from f_{embed}:

$$\mathbf{X}_{text} = f_{embed}(f_{tokenizer}(S)) \in \mathbb{R}^{m \times d_{mllm}}. \quad (2)$$

The concatenated modality embeddings $[\mathbf{X}_{vis}; \mathbf{X}_{text}]$ are fed into the transformer-based language model. Let $A_l(k) \in \mathbb{R}^{d_{model}}$ denote the k-th token's representation at the l-th transformer layer, and $A_0 = [\mathbf{X}_{vis}; \mathbf{X}_{text}]$. The final output is generated from $A_L \in \mathbb{R}^{(m+n+t) \times d_{model}}$ through an unembedding matrix $\mathbf{W}u \in \mathbb{R}^{dmodel \times |V|}$, where $|V|$ is the vocabulary size to map the embeddings to the vocabulary space, and t represents special tokens.

Cracking the Internal Mechanisms of MLLMs Detecting and understanding the internal workings of Large Language Models (LLMs) has long been recognized as crucial for trustworthy AI (Dai et al. 2022; Zou et al. 2024; Zhao et al. 2024). Inspired by these works, initial efforts have explored the interpretability of Multi-modal Large Language Models (MLLMs). For example, Vl-interp (Jiang et al. 2024) employs logit lens to identify and edit object hallucinations. LLaVA-interp (Neo et al. 2024) demonstrates a strong spatial correlation between object information and its original image location, indicating a refinement of visual input into interpretable language tokens. However, these investigations have primarily focused on the visual processing

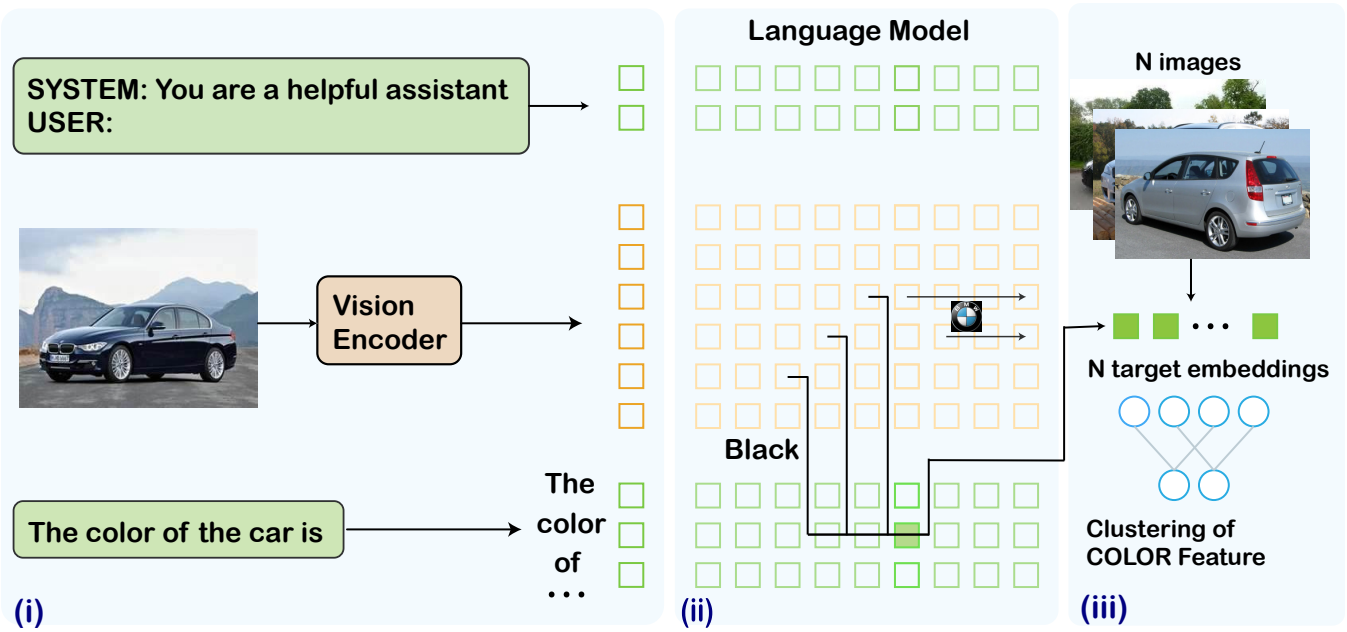

Figure 2: The ESMC framework consists of three key components: (i) the MLLM processes multimodal inputs (images and text prompts); (ii) target embeddings are selected from the language model's hidden states corresponding to text tokens; and (iii) a lightweight clustering head is trained to enhance cluster accuracy.

aspects of LVLMs, leaving the internal mechanisms related to prompt token embeddings unexplored, which motivates our work.

The logit lens (nostalgebraist 2020; Belrose et al. 2023; Pal et al. 2023) serves as an interpretability technique for analyzing language models by projecting intermediate layer representations through the unembedding matrix $\mathbf{W}_u$. Specifically, it computes linguistic probabilities via:

$$E_l(k) = \mathbf{W}_u \cdot A_l(k) \in \mathbb{R}^{|V|}, \tag{3}$$

where $E_l(k)$ denotes the k-th embedding feature of l-th layer in vocabulary space. While logit lens has been utilized as a tool to understand LLMs' hidden states, our proposed method first extends its application to explain prompt tokens and their semantic relationships with visual features in MLLMs.

3 Multiple Clustering with MLLM Hidden Embedding and Lightweight Clustering Head

We leverage user-defined prompts to specify clustering criteria and present our proposed pipeline as a two-stage framework:

Target embedding extraction Based on our novel observation of feature-related embeddings from text token embeddings, we sample a few images to get precise target embeddings, as described in detail in Section 3.1.

Clustering head training A lightweight clustering head is trained to enhance clustering performance. This network transforms raw embeddings into a structured latent space suitable for clustering, which is elaborated in Section 3.2.

3.1 Find Target Embeddings

We observe that specific prompt token shows strong correlations with semantically related visual features. The positions of these embeddings are determined by the input prompt tokens. For example, as shown in Figure 3, the ground-truth labels for color and brand features are "black" and "Volvo", respectively. The color-related feature of the input image (e.g., "black") correlates strongly with the embeddings at position $E_l(263)$, which corresponds to the "color" prompt token. Similarly, the brand-related feature (e.g., "Volvo") also shows high logits in the embeddings at position $E_l(263)$, aligning with the "brand" prompt token. Conversely, features not present in the image, such as "red" for color or "BMW" for brand, exhibit significantly lower logits in their respective embeddings, enabling clear distinctions between different features.

Though selecting embeddings from later layers corresponding to specific input tokens provides reasonable performance, we propose a refined approach to enhance clustering accuracy and semantic feature representation. Specifically, we recommend sampling a small set of images, like 10 examples for each feature, to identify embeddings with high contextual specificity, followed by 2 steps:

Keyword generation We use GPT-4 to generate feature-relevant keywords (e.g., asking "What are common car colors?" and GPT would provide tokens like "white", "black", "blue", "red", etc, as a response). These keywords are crucial to locate the embedding positions.

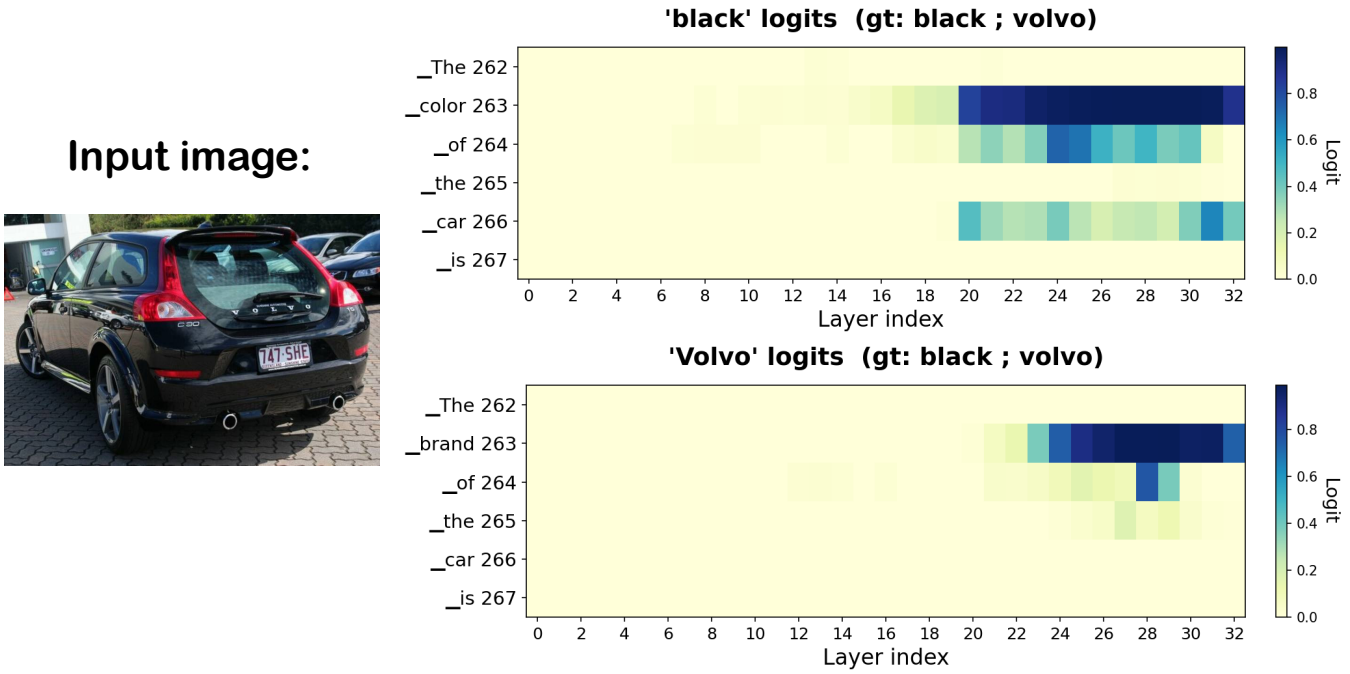

Figure 3: The relationship between vocabulary logits and text prompt tokens. The image is from the Stanford_cars dataset.

Embedding localization For each keyword, we track the embeddings of the sampled images with top-k logit and choose the final target embeddings from those with top-k logit shared by different keywords. For example, we choose the "black" keyword generated by GPT-4, and black cars are included in the sample of images, as indicated in Figure 3. The "black" token should present high logits in embeddings like $E_{27}(263)$, $E_{28}(263)$, $E_{29}(263)$ and $E_{30}(263)$. These embeddings are candidates for target embeddings. We define logit values exceeding 0.2 as high logits (Jiang et al. 2024).

3.2 Pseudo Label Clustering Head training

While typical clustering methods like K-means (MacQueen 1967) can be applied directly to target embeddings from user-defined prompt criteria, we observe that raw hidden state embeddings from LLMs/MLLMs often exhibit suboptimal performance in clustering tasks (Petukhova, Matos-Carvalho, and Fachada 2025). To mitigate this limitation, our approach incorporates a two-layer MLP clustering head to align high-dimensional embeddings with low-level semantic clusters.

The procedure is formalized as follows:

Initialization Clusters are initialized via K-means. Given $\mathcal{X} = x_1, x_2, \ldots, x_N$ where $x_i \in \mathbb{R}^{|V|}$, we initialize clusters using K-means:

$$C_1, C_2, \ldots, C_K = \arg\min_{C_k} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \quad (4)$$

where K is the given number of target clusters. N is the number of clustering samples. $\mu_k \in \mathbb{R}^d$ is the centroid of cluster C_k. x_i is the target embedding of $E_l(k)$

Pseudo-label generation For each cluster C_j with centroid μ_j, calculate the squared Euclidean distance between each data point $x_i \in C_j$ and its cluster centroid:

$$d_{ij} = ||x_i - \mu_j||^2. \quad (5)$$

For each cluster C_j, select the top α percent of data points that have the smallest distances d_{ij} to the cluster centroid μ_j. Let S_j be the set of these selected data points from cluster C_j, where α is a hyperparameter. Assign the pseudo-label $y_i = j$ to each selected data point $x_i \in S_j$. This creates a pseudo-labeled training set:

$$D_{pseudo} = (x_i, y_i) \mid x_i \in \bigcup_{j=1}^{k} Sj. \quad (6)$$

Network training The MLP is trained using these pseudo-labeled samples with cross-entropy loss to optimize clustering accuracy. The loss function can be denoted as:

$$L = -\frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}). \quad (7)$$

4 Experiments

4.1 Experiment Setup

Datasets We evaluate our method on seven multi-clustering benchmark datasets: Stanford_Cars (Yao, Qian, and Hu 2024a) includes two criteria: *color* (red, black, white, silver) and *type* (BMW, Hyundai, Volkswagen, Volvo). Flower dataset (Yao, Qian, and Hu 2024a) contains:

species (daisy, hyacinth, lily, rose) and *color* (pink, purple, red, white, yellow). Fruit (Hu et al. 2017) also contains two clustering criteria, *color* (red, green, yellow) and *species* (apple, banana, grape). Card (Yao et al. 2023) includes two criteria, *number* (ace to king) and *suits* (spades, clubs, hearts, diamonds). CMU_face (Günnemann et al. 2014) comprises three criteria, *emotion* (happy, sad, neutral, angry), *pose* (left, right, up, straight), and *glasses presence* (wearing sunglasses or not). Fruit360 (Yao et al. 2023) extends the Fruit dataset with additional samples: *color* (red, green, yellow, burgundy) and *species* (apple, banana, grape, cherry). CIFAR10 (Yao, Qian, and Hu 2024b)features two criteria, *environment* (sea, sky, land) and *object type* (animals, transportation). For the Flower and Stanford_Cars datasets, we extend prior work (Yao, Qian, and Hu 2024b) by constructing multi-criteria benchmarks based on dataset descriptions. The datasets are openly released with our codes.

Implementation details Target embeddings are extracted from the hidden states of LLaVA-1.5-7b (Liu et al. 2024) during inference, while user-defined criteria are encoded via prompts such as "The manufacturer of the car is", "The species of the fruit is", and "The color of the flower is." The pseudo-label ratio hyperparameter α is empirically set to 0.1 for the CMU_face and Card dataset, 0.2 for Fruit360 and CIFAR10 datasets, 0.3 for Stanford_cars and Flower datasets, and 0.4 for Fruit datasets. The clustering head is implemented as a two-layer MLP. We employ Normalized Mutual Information (NMI) (Meilă 2007) and the Rand Index (RI) (Rand 1971) as external metrics to measure the similarity between clustering results and ground truth. The experiments are conducted on one RTX3090 with 24GB of memory.

4.2 Results

As presented in Tables 1 and 2, the proposed ESMC method demonstrates overall superior performance compared to baselines across various clustering tasks and datasets, evidenced by its NMI and RI scores. On CIFAR10-Type, ESMC achieved the top NMI (0.8293) and RI (0.9520), marking a significant improvement over Multi-Sub (0.5271 for NMI, 0.7394 for RI) and Multi-Map (0.4967 for NMI, 0.7104 for RI). ESMC also exceeded baseline results for Fruit-Species, CMU_face-Glass, Flower dataset, CIFAR-10 dataset, and Stanford_Cars-Type datasets. It obtained perfect scores for Fruit-Species and achieved leading scores of NMI 0.8923 and RI 0.9726 on Flower-Species, surpassing Multi-Sub's NMI by 21.09%.

We note minor underperformance in the Fruit360-Color NMI and Fruit NMI scores. This may arise from the characteristics of the datasets, such as unclear images, and differences in the internal knowledge of the underlying models (e.g., LLaVA and CLIP).

4.3 Ablation Studies

Clustering of MLLM outputs yields suboptimal performance We compare our method with a baseline approach that clusters text outputs from LLaVA. Specifically, we first extract textual outputs from the model and generate semantic embeddings by feeding them into the CLIP text encoder (Radford et al. 2021). The inferior performance of the baseline (Table 3) arises from three key factors.

Firstly, the unstructured nature of MLLM outputs introduces inconsistent responses to inputs. As shown in Figure 4, when presented with two images of the same environment type described as "The background environment is a cloudy sky" and a much longer description, the CLIP text encoder may fail to produce semantically aligned embeddings. In contrast, MLLM embeddings from ESMC assign nearly identical logits to the "sky" token (0.8194 and 0.7436), which suggests consistency in semantic representation. These two images are highly likely to be clustered together due to their shared environmental features.

Secondly, MLLM tends to provide overly general descriptions instead of specific features. As illustrated in Figure 5, when prompted with "The species of the flower is", the model responds "pink rose" instead of focusing on the species feature, making it harder to cluster based on precise criteria.

Thirdly, language bias (He et al. 2025; Leng et al. 2024; Lee et al. 2024)—MLLM's tendency to prioritize prior linguistic knowledge over visual context—directly contributes to performance gaps. This bias creates a trade-off: excessive reliance on language priors may induce hallucinations, while underweighting linguistic guidance risks losing discriminative information encoded in the input prompts.

The proposed approach, ESMC, leverages intermediate hidden states from earlier layers of the MLLM, rather than the final output layer. This design not only preserves the text prompt's semantic guidance through embeddings at special positions encoded by quantized probability but also mitigates language bias by avoiding the dominance of higher-level of linguistic abstractions in later auto-regressive layers, which are prone to over-reliance on previously generated responses. By doing so, we ensure that our target embeddings are grounded in the initial, raw fusion of visual and linguistic information, while with less linguistic bias or language priors(Lin et al. 2024) from the luanguage model itself.

Clustering head improves performance Clustering head (Van Gansbeke et al. 2020; Niu, Shan, and Wang 2022; Jia et al. 2025) is a lightweight and effective technique for enhancing clustering performance. Our results, presented as NMI and RI gains in Table 3, show that the inclusion of a clustering head can lead to substantial improvements, reaching up to 12% on the Fruit360 dataset of species feature and the CMU_face dataset of Glass feature. This benefit is probably due to the high dimensionality of the MLLM embeddings, causing the curse of dimensionality(Indyk and Motwani 1998; Donoho et al. 2000). Notably, we observed that the clustering head did not enhance performance when applying CLIP embeddings.

5 Conclusion and Limitation

In conclusion, our proposed method, ESMC, introduces an explainable framework for multiple clustering tasks, addressing limitations of existing multimodal language mod-

Dataset	Criteria	MCV	ENRC	iMClusts	AugDMC	DDMC	MultiMap	MultiSub	**ESMC**
Stanford_Cars	Color	0.2103	0.2465	0.2336	0.2736	0.6899	0.7360	0.7533	**0.8138**
	Type	0.1650	0.2063	0.1963	0.2364	0.6045	0.6355	0.6616	**0.8817**
Flowers	Color	0.2938	0.3329	0.3169	0.3556	0.6327	0.6426	0.6940	**0.7283**
	Species	0.1326	0.1561	0.1894	0.1887	0.1996	0.6148	0.7580	**0.8923**
CIFAR-10	Type	0.1618	0.1826	0.2040	0.2855	0.3991	0.4967	0.5271	**0.8293**
	Environment	0.1379	0.1892	0.1920	0.2927	0.3782	0.4598	0.4828	**0.6075**
CMU_face	Emotion	0.1433	0.1592	0.0422	0.0161	0.1726	0.1786	0.2053	**0.2237**
	Glass	0.1201	0.1493	0.1299	0.1039	0.2261	0.3402	0.4870	**0.7665**
	Pose	0.3254	0.2290	0.4437	0.1320	0.4526	0.4693	0.5923	**0.6271**
Card	Order	0.0792	0.1225	0.1144	0.1440	0.1563	0.3633	0.3921	**0.4736**
	Suits	0.0430	0.0676	0.0716	0.0873	0.0933	0.2734	0.3104	**0.3586**
Fruit360	Color	0.3777	0.4264	0.4097	0.4594	0.4981	0.6239	0.6654	**0.6952**
	Species	0.2985	0.4142	0.3861	0.5139	0.5292	0.5284	**0.6123**	0.5036
Fruit	Color	0.6266	0.7103	0.7351	0.8517	0.8973	0.8619	**0.9693**	0.9308
	Species	0.2733	0.3187	0.3029	0.3546	0.3764	1.0000	1.0000	**1.0000**

Table 1: Quantitative comparison of 8 approaches using NMI. For methods involving k-means, the average result of 10 times is reported.

Dataset	Criteria	MCV	ENRC	iMClusts	AugDMC	DDMC	MultiMap	MultiSub	**ESMC**
Stanford_Cars	Color	0.5802	0.6779	0.6552	0.7525	0.8765	0.9193	**0.9387**	0.9235
	Type	0.5634	0.6217	0.5643	0.7356	0.7957	0.8399	0.8792	**0.9589**
Flowers	Color	0.5860	0.6214	0.6127	0.6931	0.7887	0.7984	0.8843	**0.8992**
	Species	0.5273	0.6065	0.6195	0.6077	0.6227	0.8321	0.8980	**0.9726**
CIFAR-10	Type	0.5634	0.6217	0.5643	0.7356	0.7957	0.8399	0.7394	**0.9589**
	Environment	0.3344	0.3599	0.3664	0.4689	0.5547	0.6737	0.7096	**0.8282**
CMU_face	Emotion	0.5268	0.6630	0.5932	0.5367	0.7593	0.7105	**0.8527**	0.6697
	Glass	0.4955	0.6209	0.5627	0.5361	0.7663	0.7068	0.8324	**0.9259**
	Pose	0.6028	0.5029	0.6114	0.5517	0.7904	0.6624	**0.8736**	0.8253
Card	Order	0.7128	0.7313	0.7658	0.8267	0.8326	0.8587	**0.8842**	0.8805
	Suits	0.3638	0.3801	0.3715	0.4228	0.6469	0.7039	**0.8504**	0.7285
Fruit360	Color	0.6791	0.6868	0.6841	0.7392	0.7472	0.8439	**0.8821**	0.8493
	Species	0.6176	0.6984	0.6732	0.7430	0.7703	0.7582	**0.8504**	0.7285
Fruit	Color	0.7685	0.8511	0.8632	0.9108	0.9383	0.9556	**0.9964**	0.9732
	Species	0.6597	0.6536	0.6743	0.7399	0.7621	1.0000	1.0000	**1.0000**

Table 2: Quantitative comparison of 8 approaches using RI. For methods involving k-means, the average result of 10 times is reported.

els (MLLMs). While MLLMs excel in visual feature extraction and contextual knowledge, they can yield unstructured or overly generic interpretations. To mitigate this, ESMC obtains target embeddings via sampling and employs a lightweight clustering head with pseudo-label learning for well-separated clusters. The method's transparency is grounded in the use of specific, interpretable text token embeddings as the clustering basis. By focusing the clustering process on these features, ESMC not only leverages the strengths of MLLMs but also enhances the interpretability of their internal representations. We also present the limitations and future directions.

Specific language model architecture Based on our study, MLLMs like LLaVA, built on LLaMA (Touvron et al. 2023), exhibit specific text prompt embedding properties. We found this architecture type fuses text and image features at a relatively late stage, with vision features injected as visual tokens.

Future work While our work focused on applying text prompt embeddings to multiple clustering tasks, we be-

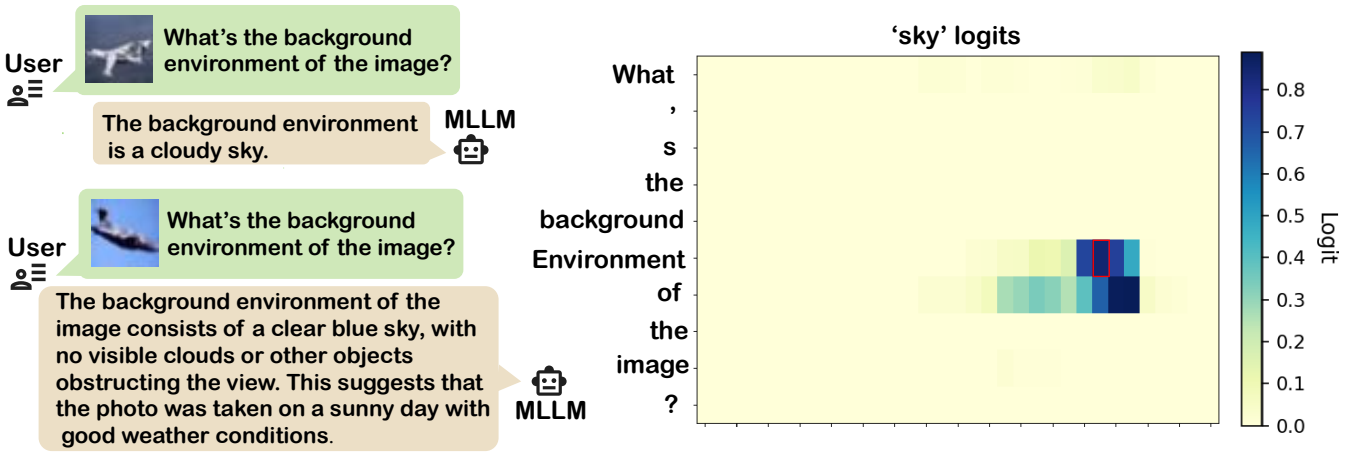

Figure 4: Left: Two MLLM outputs with shared semantic meaning but distinct formatting. Right: The target embedding(in red) shows higher consistency in "sky" logit, as the clustering label in the environment criterion.

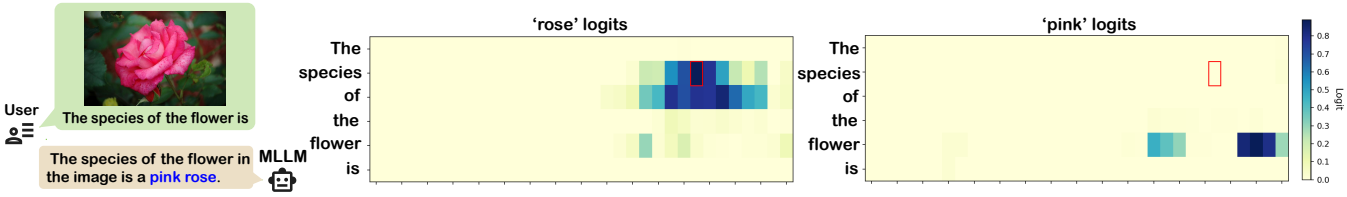

Figure 5: Left: Conversation reveals that the MLLM response can be too general. It responds as "pink rose" instead of "rose" directly, while we only care about the species feature instead of the color feature. Right: Our target embedding(in red) shows high logit in "rose" logit and low in "pink" logit, which could indicate the species feature.

Dataset	Criteria	NMI↑			RI↑		
		Output	w/o.MLP	**ESMC**	Output	w/o.MLP	**ESMC**
Stanford_Cars	Color	0.7871	0.7934	**0.8138**	0.9107	0.9129	**0.9235**
	Type	0.6462	0.8694	**0.8817**	0.8474	0.9517	**0.9589**
Flower	Color	0.6698	0.6835	**0.7283**	0.8736	0.8810	**0.8982**
	Species	0.8133	0.8462	**0.8923**	0.9329	0.9347	**0.9726**
CIFAR-10	Type	0.7505	0.7149	**0.8293**	0.8953	0.9137	**0.9520**
	Environment	0.0443	0.5484	**0.6075**	0.4941	0.7860	**0.8282**
CMU_face	Emotion	0.1498	0.1902	**0.2237**	0.5876	0.6346	**0.6697**
	Glass	0.6780	0.6445	**0.7665**	0.8279	0.7955	**0.9259**
	Pose	0.0913	0.5498	**0.6271**	0.5631	0.8140	**0.8253**
Card	Order	0.4387	0.3895	**0.4736**	0.8348	0.8712	**0.8805**
	Suits	0.2423	0.3334	**0.3586**	0.4258	0.6982	**0.7258**
Fruit360	Color	0.6071	0.6523	**0.6952**	0.7972	0.7584	**0.8493**
	Species	0.2199	0.3806	**0.5036**	0.5080	0.6824	**0.7285**
Fruit	Color	0.9308	0.9308	0.9308	0.9732	0.9732	0.9732
	Species	0.8441	1.0000	**1.0000**	0.9337	1.0000	**1.0000**

Table 3: (1) Output (direct clustering using the model's linguistic output), (2) Ours, w/o.MLP (clustering without an MLP-based clustering head), and (3) ESMC (The proposed method).

lieve MLLMs offer other effective ways for clustering or retrieval. Our research may also aid MLLM interpretability and grounding techniques. Ultimately, we aim for this work to foster more reliable and safer MLLM.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants U24A20322, 62576094 and 62422118. This work is also supported by Hong Kong UGC under grants UGC/FDS11/E03/24, UGC/FDS11/E03/25, and Hong Kong Research Grants Council under Grant 11219324.

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*, 53–62. IEEE.

Belrose, N.; Furman, Z.; Smith, L.; Halawi, D.; Ostrovsky, I.; McKinney, L.; Biderman, S.; and Steinhardt, J. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

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*, 132–149.

Cui, Y.; Fern, X. Z.; and Dy, J. G. 2007. Non-redundant multi-view clustering via orthogonalization. In *Seventh IEEE international conference on data mining*, 133–142. IEEE.

Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8493–8502.

Dasgupta, S.; and Ng, V. 2010. Mining clustering dimensions. In *Proceedings of the 27th International Conference on Machine Learning*, 263–270.

Donoho, D. L.; et al. 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000): 32.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Guérin, J.; and Boots, B. 2018. Improving image clustering with multiple pretrained cnn feature extractors. *arXiv preprint arXiv:1807.07760*.

Günnemann, 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.

He, J.; Zhu, K.; Guo, H.; Fang, J.; Hua, Z.; Jia, Y.; Tang, M.; Chua, T.-S.; and Wang, J. 2025. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3488–3501.

Hu, J.; Qian, Q.; Pei, J.; Jin, R.; and Zhu, S. 2017. Finding multiple stable clusterings. *Knowledge and Information Systems*, 51: 991–1021.

Indyk, P.; and Motwani, R. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 604–613.

Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data clustering: a review. *ACM computing surveys*, 31(3): 264–323.

Jain, P.; Meka, R.; and Dhillon, I. S. 2008. Simultaneous Unsupervised Learning of Disparate Clusterings. *Statistical Analysis and Data Mining*, 1(3): 195–210.

Jia, Y.; Cheng, J.; Liu, H.; and Hou, J. 2025. Towards Calibrated Deep Clustering Network. In *The Thirteenth International Conference on Learning Representations*.

Jiang, N.; Kachinthaya, A.; Petryk, S.; and Gandelsman, Y. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*.

Lee, S.; Park, S. H.; Jo, Y.; and Seo, M. 2024. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 391–404.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, D.; Zhao, Y.; Wang, Z.; Jung, C.; and Zhang, Z. 2024. Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. *ISPRS International Journal of Geo-Information*, 13(11): 405.

Lin, Z.; Chen, X.; Pathak, D.; Zhang, P.; and Ramanan, D. 2024. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 1205.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

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.

Meilă, M. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5): 873–895.

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.

Neo, C.; Ong, L.; Torr, P.; Geva, M.; Krueger, D.; and Barez, F. 2024. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*.

Niu, C.; Shan, H.; and Wang, G. 2022. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31: 7264–7278.

Niu, D.; Dy, J. G.; and Jordan, M. I. 2013. Iterative discovery of multiple alternativeclustering views. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1340–1353.

nostalgebraist. 2020. Interpreting GPT: The Logit Lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Pal, K.; Sun, J.; Yuan, A.; Wallace, B. C.; and Bau, D. 2023. Future Lens: Anticipating Subsequent Tokens from a Single Hidden State. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, 548–560.

Petukhova, A.; Matos-Carvalho, J. P.; and Fachada, N. 2025. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6: 100–108.

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.

Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P. S.; and He, L. 2024. Deep clustering: A comprehensive survey. *IEEE transactions on neural networks and learning systems*.

Tokuda, T.; Yoshimoto, J.; Shimizu, Y.; Okada, G.; Takamura, M.; Okamoto, Y.; Yamawaki, S.; and Doya, K. 2017. Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions. *PloS one*, 12(10): e0186566.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *European conference on computer vision*, 268–285. Springer.

Xu, R.; and Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3): 645–678.

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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 82705–82725.

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.

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

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37: 83345–83373.