Improving Image Classification Tasks Using Fused Embeddings and Multimodal Models

Artur A. Oliveira[a], Mateus Espadoto[b], Roberto Hirata Jr.[c] and Roberto M. Cesar Jr.[d]
Institute of Mathematics and Statistics, University of São Paulo, Brazil
{*arturao, mespadot, hirata, cesar*}*@ime.usp.br*

Keywords: Prompt Engineering, Guided Embeddings, Multimodal Learning, Clustering, t-SNE Visualization, Zero-Shot Learning.

Abstract: In this paper, we address the challenge of flexible and scalable image classification by leveraging CLIP embeddings, a pre-trained multimodal model. Our novel strategy uses tailored textual prompts (e.g., "This is digit 9", "This is even/odd") to generate and fuse embeddings from both images and prompts, followed by clustering for classification. We present a prompt-guided embedding strategy that dynamically aligns multimodal representations to task-specific or grouped semantics, enhancing the utility of models like CLIP in clustering and constrained classification workflows. Additionally, we evaluate the embedding structures through clustering, classification, and t-SNE visualization, demonstrating the impact of prompts on embedding space separability and alignment. Our findings underscore CLIP's potential for flexible and scalable image classification, supporting zero-shot scenarios without the need for retraining.

1 INTRODUCTION

Pre-trained multimodal models, such as CLIP (Radford et al., 2021), have showcased exceptional generalization capabilities by aligning image and text representations within a shared embedding space. These models enable zero-shot learning, allowing for task adaptation without explicit retraining. However, their utility in scenarios such as unsupervised clustering and constrained classification, where novel or complex classification schemes arise, remains underexplored.

Constrained classification refers to workflows where the assignment of samples to categories must adhere to predefined semantic relationships. Unlike traditional classification methods that rely solely on static embeddings, constrained classification benefits from dynamic, task-driven structures within the embedding space. This paper introduces a framework that leverages task-specific and grouped prompts to guide embedding creation, aligning with such semantic constraints.

Task-specific prompts explicitly align embeddings with ground-truth classes, using descriptions like "This is digit 0", to emphasize precise class distinctions. Grouped prompts, in contrast, define higher-level semantic relationships, such as grouping "even" and "odd" digits, facilitating tasks where broader class groupings are sufficient or preferred. These prompt strategies enable us to structure the embedding space dynamically, providing a contrast to static image embeddings, which serve as a baseline in our analysis.

The contributions of this paper are as follows:

- We propose a novel prompt-guided embedding strategy that dynamically aligns multimodal representations to task-specific or grouped semantics, advancing the utility of models like CLIP in clustering and constrained classification workflows.
- We introduce a unified framework for evaluating embedding structures through clustering, classification, and visualization, highlighting the impact of prompts on embedding space separability and alignment.
- We conduct comprehensive experiments across three datasets: MNIST, CIFAR-10, and CIFAR-100 subsets, demonstrating that task-specific and grouped prompts significantly outperform image-only baselines in clustering and classification tasks.

[a] https://orcid.org/0000-0002-3606-1687
[b] https://orcid.org/0000-0002-1922-4309
[c] https://orcid.org/0000-0003-3861-7260
[d] https://orcid.org/0000-0003-2701-4288

Oliveira, A. A., Espadoto, M., Hirata Jr., R. and Cesar Jr., R. M.
Improving Image Classification Tasks Using Fused Embeddings and Multimodal Models.
DOI: 10.5220/0013365600003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 232-241
ISBN: 978-989-758-728-3; ISSN: 2184-4321

Through this work, we aim to bridge the gap between general-purpose multimodal models and task-specific workflows, showcasing how prompt-conditioned embeddings can enhance clustering quality and constrained classification accuracy. This research paves the way for exploring more flexible and adaptable embedding strategies in multimodal learning.

The rest of this paper is organized as follows: Section 2 reviews related work on multimodal learning, prompt design, and clustering techniques. Section 3 presents our methodology, detailing the design of task-specific and grouped prompts, the embedding framework, and the clustering and classification workflows. Section 4 describes the experimental setup, datasets, and results, showcasing the effectiveness of prompt-guided embeddings. Section 5 explores the implications of our findings, addressing limitations and potential opportunities for future work. Finally, Section 6 concludes with a summary of contributions and directions for further research.

2 RELATED WORK

Recent advances in multimodal learning have enabled models to effectively bridge visual and textual modalities, creating shared embedding spaces that capture semantic relationships across data types. These innovations have unlocked new capabilities, such as zero-shot generalization, allowing models to adapt to diverse tasks without additional fine-tuning. While much of the focus has been on leveraging these embeddings for classification and retrieval, their potential for unsupervised tasks like clustering and constrained classification remains underexplored. This section reviews advancements in multimodal learning, prompt design, and clustering techniques, highlighting key gaps in the current understanding of how prompting strategies shape embedding spaces.

2.1 Multimodal Learning with Natural Language Supervision

Advances in multimodal learning have introduced models capable of aligning visual and textual modalities in a shared embedding space. A prominent example is CLIP (Contrastive Language–Image Pretraining), which leverages natural language supervision to achieve zero-shot transfer across diverse tasks. CLIP's embedding space captures rich semantic relationships, enabling generalization without task-specific fine-tuning.

While CLIP's zero-shot performance is well-documented, less attention has been given to how its embeddings can be structured for unsupervised tasks like clustering and constrained classification. This presents an opportunity to understand and optimize the embedding space for these workflows.

2.2 Prompt Design in Multimodal Models

Prompt design plays a critical role in adapting general-purpose embeddings to task-specific needs. Textual prompts guide models like CLIP by aligning image embeddings with semantic concepts described in natural language (Li et al., 2024; Allingham et al., 2023; Huang et al., 2022). Well-crafted prompts have been shown to improve zero-shot classification by reducing the semantic gap between textual descriptions and image representations.

Recent studies have expanded the scope of prompt learning beyond task-specific classification. For instance, (Huang et al., 2022) introduced an unsupervised prompt learning framework for vision-language models, while (Li et al., 2024) proposed prompt-driven knowledge distillation to transfer knowledge between models.

2.3 Clustering and Classification in Embedding Spaces

Clustering is fundamental to understanding embedding spaces, providing insights into data organization and supporting classification tasks. Traditional clustering methods such as k-means (Lloyd, 1982), DBSCAN (Ester et al., 1996), and Spectral Clustering (Shi and Malik, 2000) have been primarily applied to unimodal embeddings derived from images or text alone. Their application to fused multimodal embeddings, where visual and textual features are integrated, to the best of our knowledge, remains limited.

Existing multimodal clustering methods, such as Multimodal Clustering Networks (MCN) (Chen et al., 2021), emphasize representation alignment across modalities but often rely on static embeddings. These approaches neglect the dynamic influence of prompts, which can lead to semantic overlap and misalignment between clusters and prompts. Methods like MoDE (Ma et al., 2024) and ModalPrompt (Zeng et al., 2024) incorporate prompts dynamically into clustering workflows but are restricted to specific use cases, leaving broader systematic approaches underdeveloped.

3 METHODOLOGY

This work explores how prompt-guided embeddings influence clustering and classification tasks in multimodal settings. By leveraging CLIP's ability to align visual and textual modalities, we design two prompting strategies: task-specific, and grouped, to guide embedding creation to reflect semantic relationships in the data. As baseline, we consider the case where no prompts are used, and clustering is performed solely using image embeddings. The image-only baseline evaluates clustering and classification performance without the influence of textual prompts, isolating the impact of semantic alignment introduced by task-specific and grouped prompts. These strategies are evaluated through a unified framework involving clustering, classification, and visualization.

3.1 Prompt-Guided Embedding Design

We use three types of prompts to structure the embedding space:

- **Task-Specific Prompts:** Class-level descriptions aligned with ground-truth labels (e.g., "This is a digit 0"). These prompts guide the embeddings to reflect precise semantic distinctions.
- **Grouped Prompts:** Higher-level groupings that capture relationships among multiple classes (e.g., "This is an even digit" for MNIST or "This is an animal" for CIFAR-10).
- **Swapped Prompts:** Intentionally misaligned prompts used to evaluate the robustness of clustering and classification.

For each prompt:

1. **Text Embeddings:** Prompts are tokenized and encoded using CLIP's text encoder.
2. **Image Embeddings:** Images are preprocessed and encoded through CLIP's image encoder.
3. **Fused Embeddings:** The final embeddings are the average of the image and text embeddings, creating a multimodal representation aligned with the semantic intent of the prompt.

For the baseline (image-only embeddings), clustering is performed solely on the image embeddings, without incorporating text features.

3.2 Clustering Framework

The training pipeline, illustrated in Fig. 1 (left), consists of the following steps:

1. **Generating Fused Embeddings:** Each training sample is processed through the CLIP model to produce image embeddings and paired with the corresponding prompt to generate text embeddings. The image and text embeddings are averaged to create fused multimodal representations, reflecting the semantic intent of the prompts.
2. **Applying Spectral Clustering:** The fused embeddings are used as input for Spectral Clustering, with the number of clusters set adaptively based on the dataset complexity. For datasets with high inter-class similarity or irregular cluster shapes, additional cluster centers were introduced to better capture the nuanced structure of the embedding space. The Spectral Clustering algorithm maps the fused embeddings into a lower-dimensional spectral space and identifies clusters based on their proximity in this space. This approach is chosen for its flexibility and ability to capture complex relationships within the embedding space.
3. **Cluster Label Assignment:** Once the clusters are formed, each cluster is assigned a representative label using majority voting. For every cluster:
   - The ground-truth labels of all samples within the cluster are counted.
   - The most frequent label is selected as the cluster's representative label.

   This step ensures alignment between the clusters and the dataset's semantic structure.
4. **Approximating Cluster Centroids:** Since Spectral Clustering does not provide explicit cluster centroids, these are approximated as the mean position of all fused embeddings within each cluster. Mathematically, for a cluster C_i containing n samples with embeddings $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, the centroid $\mathbf{c}_i$ is computed as:

$$\mathbf{c}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{e}_j$$

   These centroids are used during the classification phase to compute distances between test samples and clusters.

Evaluation of Clustering Quality:

- The effectiveness of the clustering is evaluated using metrics such as silhouette score (Rousseeuw, 1987), adjusted Rand index (ARI) (Hubert and Arabie, 1985), and adjusted normalized mutual information (ANMI) (Vinh et al., 2009; Scikit-Learn, 2024).

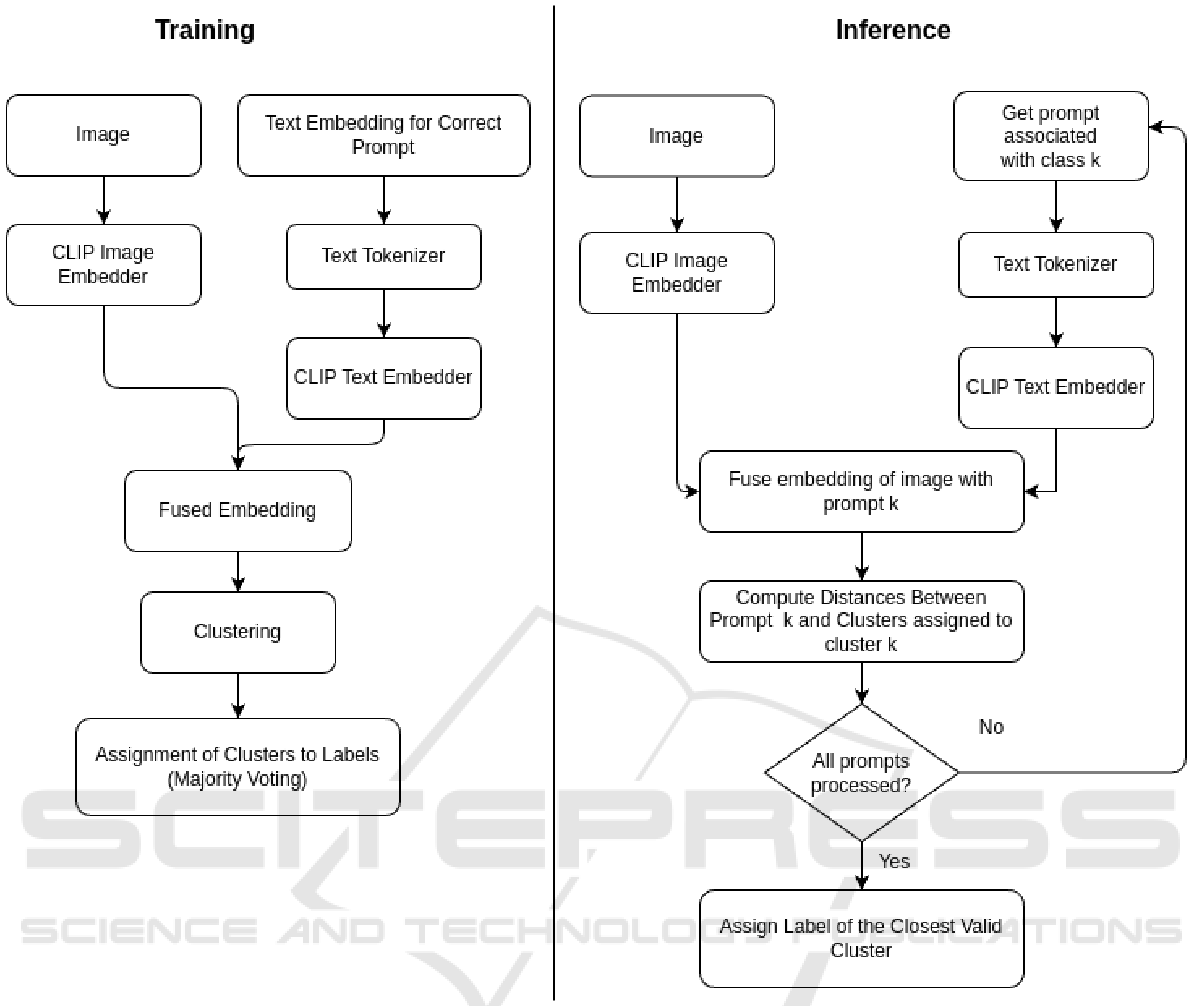

Training
Image
Text Embedding for Correct Prompt
CLIP Image Embedder
Text Tokenizer
CLIP Text Embedder
Fused Embedding
Clustering
Assignment of Clusters to Labels (Majority Voting)
Inference
Image
Get prompt associated with class k
CLIP Image Embedder
Text Tokenizer
CLIP Text Embedder
Fuse embedding of image with prompt k
Compute Distances Between Prompt k and Clusters assigned to cluster k
All prompts processed?
No
Yes
Assign Label of the Closest Valid Cluster

Figure 1: Diagrams showing the workflows for training (left) and inference (right). The training process involves generating fused embeddings of images and prompts, followed by clustering and label assignment. The inference process iteratively pairs an image with all class prompts to compute distances, determining the label of the closest cluster.

- These metrics provide insights into cluster separability, alignment with ground-truth labels, and the overall quality of the embedding space organization.

3.3 Classification Workflow

The inference workflow, shown in Fig. 1 (right), involves:

1. Generating fused embeddings for test samples paired iteratively with all prompts.
2. Filtering candidate clusters to only those corresponding to the prompt's intended class or grouping.
3. Assigning each test sample to the nearest cluster based on the filtered candidates.
4. Using the assigned cluster's label, determined during training, as the predicted class.

This filtering mechanism ensures semantic consistency between prompts and clusters, avoiding mismatches and improving classification reliability.

3.4 Visualization of Embedding Spaces

To qualitatively assess embedding space separability and alignment, we employ t-SNE (van der Maaten and Hinton, 2008) for dimensionality reduction. Visualizations compare the impact of different prompting strategies, color-coded by ground-truth and predicted labels. These plots highlight the influence of prompt design on clustering and classification outcomes.

4 EXPERIMENTS AND RESULTS

This section presents the evaluation of prompt-guided embeddings on clustering and classification tasks across three datasets of varying complexity. The experiments aim to assess the impact of task-specific, grouped, and swapped prompts on embedding alignment and downstream performance, using image-only embeddings as a baseline. We analyze the results both qualitatively, through visualizations of the embedding space, and quantitatively, using clustering and classification metrics.

4.1 Datasets and Experimental Setup

We conduct experiments using three datasets:

- **MNIST** (Deng, 2012), a simple dataset of grayscale handwritten digits, rendered as 28x28 images;
- **CIFAR-10** (Krizhevsky et al., a), which features 32x32 RGB images spanning 10 diverse classes;
- **CIFAR-100 Subsets** (Krizhevsky et al., b), with five randomly selected subsets of 10 classes each.

These datasets represent increasing levels of complexity, from clear class separability to greater inter-class similarity and diversity. Each dataset undergoes preprocessing for compatibility with the CLIP *ViT-B/32* backbone, including resizing images to 224x224 and normalizing them with CLIP's default mean and standard deviation. Textual prompts are tokenized and encoded using CLIP's text encoder.

4.1.1 Prompt Strategies

We evaluate the following prompting strategies:

- **Task-Specific Prompts:** Class-level descriptions such as "This is digit 0," aligned directly with ground-truth labels.
- **Grouped Prompts:** Semantic groupings, such as "This is a tool" or "This is a vehicle," reflecting broader relationships between classes.
- **Swapped Prompts:** Misaligned prompts used to evaluate the robustness of clustering against noisy semantic guidance.
- **Baseline:** Image-only embeddings, where no prompts are used, providing a benchmark for evaluating the added value of text guidance.

4.1.2 Clustering and Classification Workflow

Embeddings are generated by fusing image and text features, followed by clustering using Spectral Clustering with the number of clusters set to match the ground-truth classes. Cluster labels are assigned using majority voting over the training data. For the baseline, clustering is performed solely on image embeddings.

Test samples are projected into the embedding space, and classification is performed by assigning each sample to the nearest cluster center. For prompt-guided embeddings, filtering ensures alignment between test prompts and cluster labels.

4.2 Qualitative Analysis

To visualize the separability of the embedding space, we employ t-SNE (*t-distributed Stochastic Neighbor Embedding*). Figure 3 shows examples of CIFAR-10 embeddings under different prompting strategies.

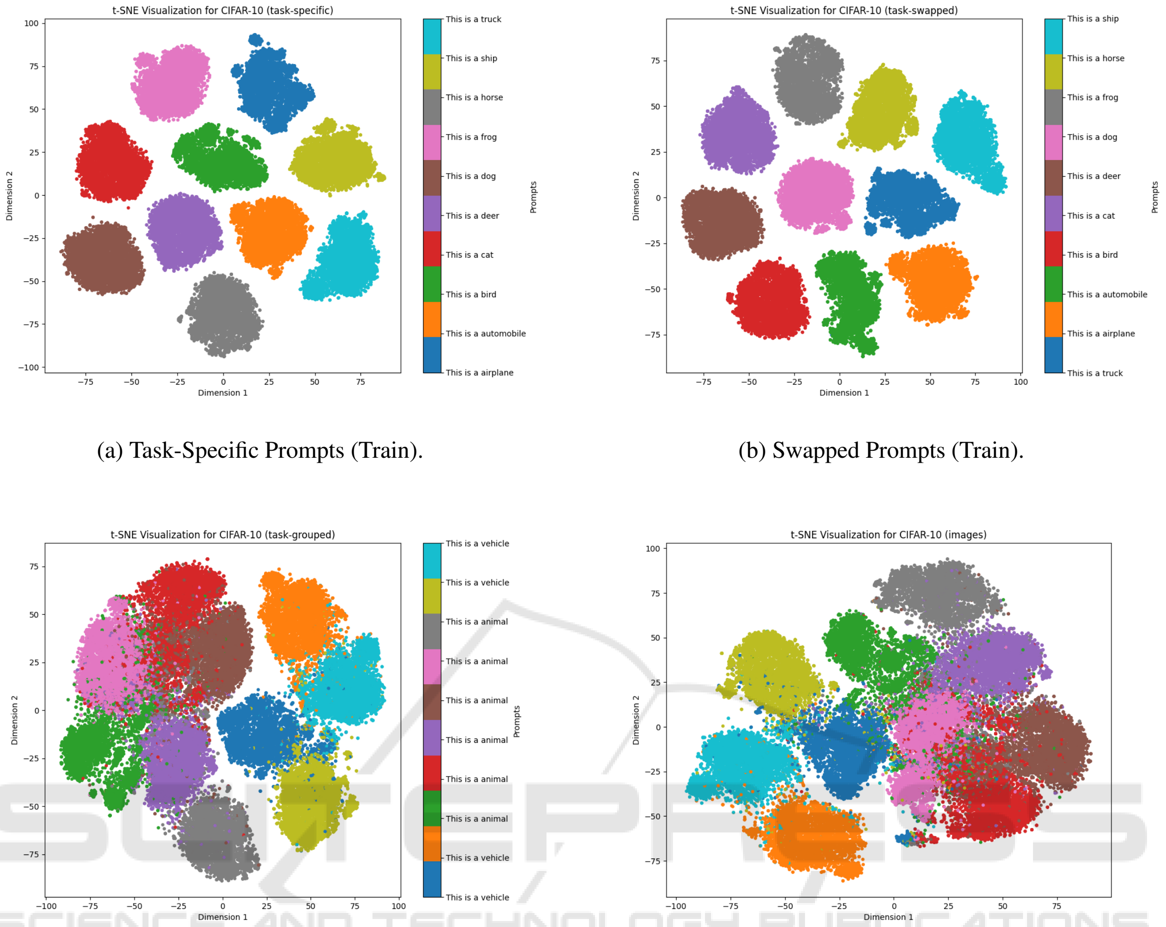
Task-specific prompts yield compact and well-separated clusters closely aligned with ground-truth labels, as seen in Figs. 2a and 3a. Grouped prompts, illustrated in Figs. 2c and 3c, produce meaningful separability but exhibit slight overlap within broader groupings. The swapped prompts, shown in Figs. 2b and 3b, highlight the robustness of the method, as clusters remain distinct despite noisy guidance. The baseline embeddings, shown in Figs. 2d and 3d, reveal significant overlap, underscoring the limitations of image-only embeddings. Figures. 3a, 3c, 3b and 3d are plots for the test embeddings and their colors represent the assignment performed by our test-time classification procedure described in Section 3.

Similar trends are observed in MNIST and CIFAR-100 visualizations. For MNIST, task-specific prompts achieve near-perfect separability, while CIFAR-100 datasets demonstrate the method's scalability despite increased complexity.

4.3 Metrics and Quantitative Setup

To quantitatively assess clustering performance, we evaluate the following metrics:

- **Silhouette Score:** Measures intra-cluster compactness relative to inter-cluster separation.
- **Adjusted Rand Index (ARI):** Evaluates the agreement between predicted clusters and ground-truth labels, adjusted for chance.
- **Adjusted Normalized Mutual Information (ANMI):** Quantifies the shared information between cluster assignments and ground-truth labels.



(a) Task-Specific Prompts (Train).

(b) Swapped Prompts (Train).

(c) Grouped Prompts (Train).

(d) Baseline: Image-Only Embeddings (Train).

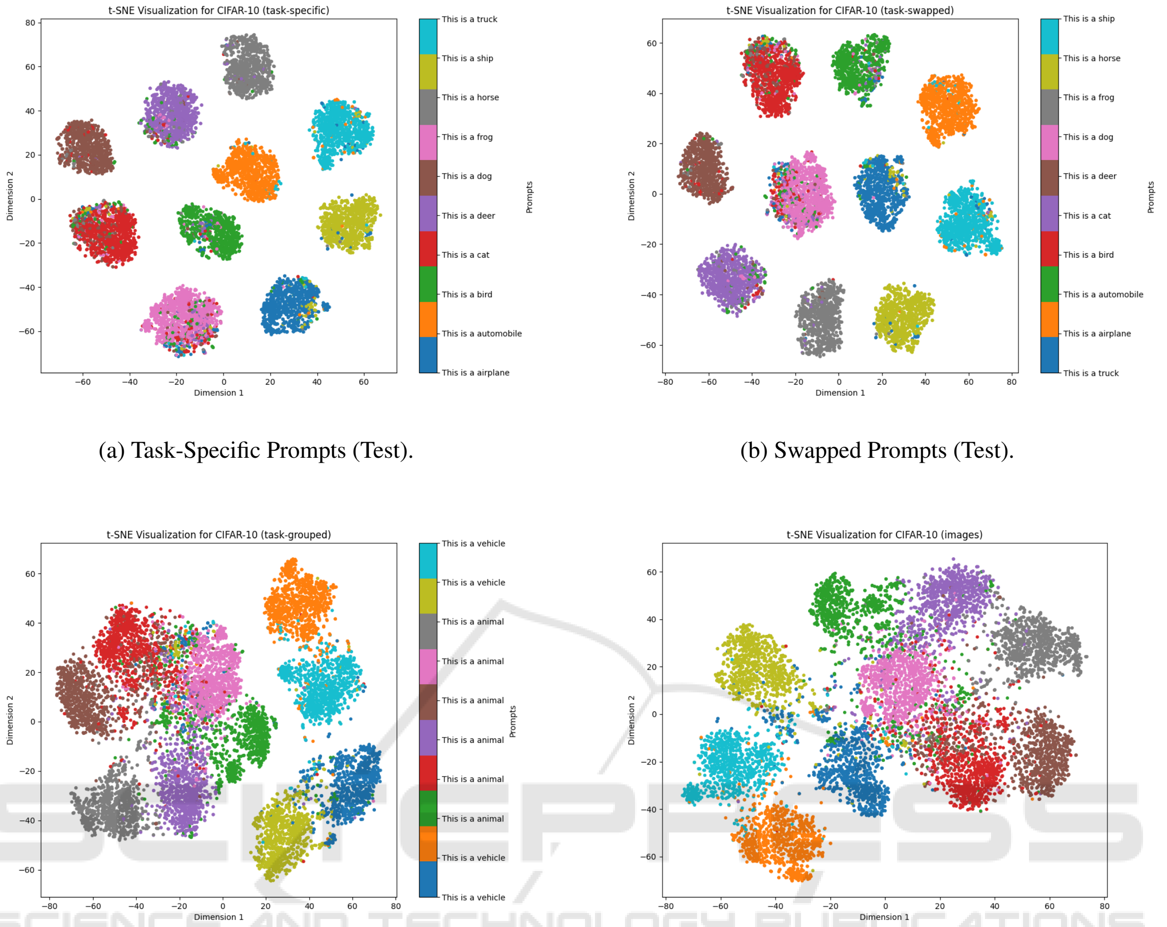
Figure 2: t-SNE visualizations for CIFAR-10 embeddings under different prompting strategies. Task-specific and grouped prompts show superior clustering, while generic and image-only embeddings suffer from significant overlap.

4.4 Quantitative Results

Table 1 presents the results across datasets and prompting strategies. For datasets with complex class structures, the number of cluster centers was adaptively increased to reflect the embedding space's complexity, ensuring robust clustering and improved classification outcomes. Key observations include:

1. **Task-Specific Prompts:** Consistently achieve the highest accuracy and clustering metrics, confirming their effectiveness in aligning embeddings with semantic intent.
2. **Grouped Prompts:** Perform well in binary-class tasks, where classes are grouped based on broader semantic definitions (e.g., ”even” vs. ”odd”). However, these results are not directly comparable to task-specific prompts due to the reduced number of classes and the binary nature of the task.
3. **Swapped Prompts:** Maintain robust performance, highlighting the resilience of prompt-guided embeddings to noisy or misaligned textual guidance.
4. **Baseline (Image-Only):** Achieves significantly lower metrics across all datasets, underscoring the importance of prompt-conditioned embeddings.

The results demonstrate the superiority of prompt-guided embeddings for clustering and classification tasks. Task-specific prompts consistently produce compact, well-separated clusters, enabling high classification accuracy. Grouped prompts provide a flexible alternative for binary or grouped-class definitions, while swapped prompts validate the robustness of the approach.



(a) Task-Specific Prompts (Test).

(b) Swapped Prompts (Test).

(c) Grouped Prompts (Test).

(d) Baseline: Image-Only Embeddings (Test).

Figure 3: t-SNE visualizations for CIFAR-10 test embeddings under different prompting strategies. Task-specific and grouped prompts show superior clustering, while generic and image-only embeddings suffer from significant overlap. The colors represent the assigned class using our classification strategy discussed in Section 3.

4.5 Summary of Findings

The experiments highlight the advantages of integrating textual prompts into embedding workflows, particularly for unsupervised clustering and constrained classification. Task-specific prompts are most effective in aligning embedding spaces with semantic intent, while grouped prompts offer a trade-off between granularity and flexibility. The baseline results emphasize the limitations of image-only embeddings, reinforcing the value of multimodal guidance.

5 DISCUSSION

Our study highlights the potential of prompt-guided embeddings to structure multimodal embedding spaces for clustering and classification tasks. By leveraging semantic cues encoded in task-specific and grouped prompts, our approach fosters the creation of compact and well-separated clusters. This alignment between semantic intent and embedding structure underpins the effectiveness of the proposed method, as evidenced by improved clustering metrics and classification accuracy compared to generic prompts.

5.1 Strengths of Prompt-Guided Embeddings

The experimental results validate our central hypothesis: prompt design significantly impacts the structure of embedding spaces. Task-specific prompts align closely with ground-truth labels, enabling precise class distinctions. Grouped prompts, on the other

Table 1: Experimental Results Across Datasets and Prompt Strategies. Note that results for grouped prompts are derived from binary-class tasks and are not directly comparable to task-specific or other multi-class prompt strategies.

Dataset	**Prompt Type**	**Accuracy ↑**	**Missing Classes**	**Silhouette Score ↑**	**ARI ↑**	**NMI ↑**
MNIST	Task-Specific	0.8500	None	**0.1747**	**0.7819**	**0.8964**
	Image Only	0.1073	None	0.0749	0.5885	0.7487
	Task-Swapped	**0.8667**	None	0.1551	0.7560	0.8912
	Task-Grouped	0.7487	None	0.1205	0.9995	0.9981
CIFAR-10	Task-Specific	**0.8792**	None	**0.1390**	**0.8591**	**0.9277**
	Image Only	0.0993	None	0.0637	0.5771	0.7306
	Task-Swapped	0.8790	None	0.0637	0.5768	0.7305
	Task-Grouped	0.6232	None	0.0915	0.6523	0.7138
CIFAR-100 (Subset 1)	Task-Specific	**0.702**	None	**0.1496**	**0.7815**	**0.8976**
	Image Only	0.144	None	0.0345	0.2042	0.4480
	Task-Swapped	0.7	None	0.1333	**0.7815**	**0.8976**
	Task-Grouped	0.52	None	0.0928	0.6749	0.7327
CIFAR-100 (Subset 2)	Task-Specific	**0.888**	None	**0.1164**	**0.7417**	**0.8892**
	Image Only	0.069	None	0.0563	0.4169	0.6735
	Task-Swapped	0.867	None	0.0998	0.6803	0.8640
	Task-Grouped	0.746	None	0.1619	0.7306	0.7506
CIFAR-100 (Subset 3)	Task-Specific	0.674	None	0.0929	0.6644	0.8583
	Image Only	0.1	None	0.0412	0.2190	0.4508
	Task-Swapped	**0.735**	None	**0.1105**	**0.7487**	**0.8884**
	Task-Grouped	0.504	None	0.0793	0.6338	0.7150
CIFAR-100 (Subset 4)	Task-Specific	**0.796**	None	0.1278	0.7944	**0.9022**
	Image Only	0.064	None	0.0631	0.3835	0.6122
	Task-Swapped	0.795	None	**0.1354**	**0.8006**	0.8985
	Task-Grouped	0.506	None	0.2377	0.7192	0.7443
CIFAR-100 (Subset 5)	Task-Specific	0.634	[96]	0.0819	0.6528	0.8332
	Image Only	0.133	None	0.0540	0.3174	0.5394
	Task-Swapped	**0.722**	None	**0.1167**	**0.7729**	**0.8976**
	Task-Grouped	0.537	None	0.0978	0.6697	0.7306

hand, capture broader semantic relationships, which are especially useful in cases where granular class distinctions are not necessary or desirable. Together, these strategies demonstrate the flexibility and efficacy of prompt-guided embeddings in enhancing representation quality for unsupervised and constrained tasks.

To address these challenges, we adapted our approach by increasing the number of cluster centers relative to the number of classes in datasets with higher complexity. This adjustment allowed the clustering process to capture more nuanced structures in the embedding space, improving classification performance. While this refinement demonstrates the method's flexibility, it also highlights the importance of considering cluster geometry in multimodal workflows.

5.2 Limitations and Challenges

5.2.1 Centroid-Based Classification Assumptions

A key component of our method is the centroid-based classification framework, which assumes that clusters formed by the fused embeddings are compact and well-separated. This assumption aligns with the observed improvements in clustering metrics, suggesting that prompt-guided embeddings indeed exhibit such properties. However, centroid-based classification may be suboptimal for scenarios where clusters are irregularly shaped or exhibit significant overlap.

Alternative classification schemes, such as nearest-neighbor methods or manifold-based approaches, could mitigate these issues. Nearest-neighbor methods are straightforward but would undermine the central premise that prompts structure the embedding space meaningfully. Manifold-based approaches, while theoretically robust, introduce additional complexity and computational overhead, diverging from the primary focus of this work. Addressing these scenarios presents an opportunity for future research.

5.2.2 Class Coverage in Clustering Assignments

One observation from our experiments, particularly with the image-only baseline embeddings, is the potential for some classes to remain unrepresented in the clustering process. This issue arises when the embedding space fails to separate certain classes effectively or when clustering algorithms struggle with ambiguous regions in the embedding space. However, rather than being a limitation of the method, this underscores the critical importance of prompt design.

Our results highlight that task-specific and grouped prompts introduce strong semantic cues, creating more meaningful and well-separated clusters that mitigate this issue. The absence of clusters for certain classes with image-only baseline embeddings validates our central hypothesis: specific and semantically aligned prompts play a pivotal role in structuring embedding spaces for effective clustering and classification.

This finding reinforces the necessity of prompt-guided approaches and provides a baseline for demonstrating the substantial improvements achieved with task-specific and grouped prompts. Future work may explore how to adapt or extend these prompts for more complex or imbalanced datasets, but the current study effectively demonstrates their advantages over generic baselines.

5.3 Unexplored Dynamics in Multimodal Alignment

Our findings reveal intriguing dynamics in CLIP's embedding space. The poor performance of image-only embeddings, even for simple datasets like MNIST, contrasts sharply with the effectiveness of text-guided embeddings, highlighting CLIP's reliance on multimodal alignment. The strong clustering performance under swapped prompts further emphasizes the dominant role of textual anchors in shaping semantic structures.

These dynamics raise questions about the intrinsic quality of image embeddings in CLIP and how textual prompts influence their semantic grounding. While this study focuses on demonstrating the utility of prompts for clustering and classification, future research could delve deeper into the interplay between multimodal alignment and individual modality performance.

5.4 Future Directions

Our findings create space for further exploration in several directions:

- **Exploration of Additional Prompt Strategies:** Beyond task-specific and grouped prompts, alternative designs such as adversarial or learned prompts may further enhance embedding space alignment.
- **Advanced Classification Techniques:** Investigating more sophisticated classification frameworks, such as manifold-based approaches, could provide insights into scenarios where centroid-based methods fall short.
- **Dynamic Prompt Adaptation:** Extending the method to dynamically adapt prompts based on dataset characteristics or clustering feedback could improve generalization to diverse tasks.
- **Class Coverage in Clustering:** Addressing the potential for unassigned clusters, particularly in challenging settings, through hybrid clustering methods or adaptive feedback mechanisms.

5.5 Broader Implications

The demonstrated impact of prompt design on embedding structures underscores the importance of integrating semantic guidance into multimodal models. This has implications beyond clustering and classification, potentially benefiting retrieval, generation, and other downstream tasks. As multimodal models continue to evolve, prompt-based strategies may serve as a critical tool for bridging the gap between general-purpose embeddings and task-specific needs.

By showcasing the interplay between prompts and embedding structures, this work contributes to the growing understanding of how natural language supervision can enhance multimodal representation learning. While challenges remain, the proposed method provides a foundation for further exploration and application in this rapidly advancing field.

6 CONCLUSION

This work introduced a novel strategy for leveraging CLIP to create guided embeddings for clustering and classification tasks for image data. By utilizing textual prompts, we demonstrated how embedding spaces could be shaped to align with semantic relationships in the data. Task-specific prompts enabled fine-grained separability for individual classes, while grouped prompts captured broader semantic groupings, offering flexibility for various application needs.

We showed that task-specific and grouped prompts significantly enhance clustering performance compared to image-only baselines, highlighting the critical role of prompt design in structuring embedding spaces. Furthermore, our method effectively adapts to zero-shot and constrained classification tasks, emphasizing the versatility of multimodal models in unsupervised workflows.

While the primary focus was on evaluating the influence of prompts on clustering and classification, our findings also underscore the potential for future work in prompt optimization, dynamic embedding structures, and applications to more complex datasets. This study contributes to a growing understanding of how natural language supervision can guide multimodal models, bridging the gap between zero-shot generalization and task-specific optimization.

ACKNOWLEDGEMENTS

This work was funded partially by FAPESP project 2022/15304-4 and MCTI (law 8.248, PPI-Softex - TIC 13 - 01245.010222/2022-44).

REFERENCES

Allingham, J. U., Ren, J., Dusenberry, M. W., Gu, X., Cui, Y., Tran, D., Liu, J. Z., and Lakshminarayanan, B. (2023). A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *International Conference on Machine Learning*, pages 547–568. PMLR.

Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A., Panda, R., Kingsbury, B., Feris, R., Harwath, D., et al. (2021). Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8012–8021.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Huang, T., Chu, J., and Wei, F. (2022). Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research).

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research).

Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., and Yang, J. (2024). Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Ma, J., Huang, P.-Y., Xie, S., Li, S.-W., Zettlemoyer, L., Chang, S.-F., Yih, W.-T., and Xu, H. (2024). Mode: Clip data experts via clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26354–26363.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Scikit-Learn (2024). *Adjusted Mutual Information Score - Scikit-learn 1.5.2 Documentation*. Scikit-learn. https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Zeng, F., Zhu, F., Guo, H., Zhang, X.-Y., and Liu, C.-L. (2024). Modalprompt: Dual-modality guided prompt for continual learning of large multimodal models. *arXiv preprint arXiv:2410.05849*.