Unified Multimodal Demonstration Retrieval for In-Context Learning in Large Vision-Language Models

Simon A. Aytes[a*], Jinheon Baek[a], Sung Ju Hwang[c]
[a]MLAI Lab, KAIST, Seoul, South Korea
[b]DeepAuto, Seoul, South Korea

[*]Corresponding author: Simon A. Aytes; saytes@kaist.ac.kr

Abstract

In-context learning (ICL) is a paradigm that enables models to make predictions based on examples provided within the input context, showing substantial promise for adapting vision-language models (VLMs) to diverse tasks without fine-tuning. However, extending ICL to multimodal settings—where models must dynamically retrieve and align both text and visual data—presents significant challenges. Existing ICL methods often rely on pre-defined (static) demonstrations and are primarily limited to unimodal text retrieval, reducing their adaptability and effectiveness in complex multimodal scenarios. We propose **AURA** (Adaptive Unified Retrieval and Alignment), a novel task-agnostic retrieval framework that enhances multimodal ICL by learning to retrieve and align relevant demonstrations across both unimodal and multimodal datasets. AURA operates within a unified multimodal embedding space structured through contrastive triplet learning, allowing it to dynamically adapt retrievals based on the task at hand. Furthermore, a VLM-guided feedback mechanism fine-tunes retrieval relevance and alignment, enabling AURA to respond effectively to varied task requirements. Experimental results across a broad range of vision-language benchmarks, including Visual Question Answering (VQA) and multimodal sentiment analysis, demonstrate that AURA not only improves retrieval precision and VLM response accuracy but also generalizes across disparate datasets, establishing a versatile framework for advancing multimodal in-context learning across varied tasks.

1 Introduction

Recent advancements in large language models (LLMs) and vision-language models (VLMs) have transformed natural language processing and multimodal tasks, achieving impressive results across applications such as image captioning (Vinyals et al., 2015), visual question answering (VQA) (Agrawal et al., 2016), and natural language understanding (Devlin et al., 2019). One of the key drivers of this success is in-context learning (ICL), where models perform tasks by conditioning on a few demonstrations without parameter updates (Brown et al., 2020). ICL has shown promise in enabling models to generalize to new tasks even with a few examples, particularly in unimodal contexts (Dong et al., 2024). However, extending ICL to complex multimodal scenarios, where both text and image data are essential, presents unique and largely unsolved challenges (Luo et al., 2024).

In the context of multimodal ICL, retrieval refers to selecting demonstrations relevant to the query from a large dataset, while alignment involves ensuring that these retrieved demonstrations are contextually appropriate for the query's modality and task requirements (Tsimpoukelli et al., 2021). In existing studies, retrieval-augmented generation models in text-only NLP primarily rely on static, unimodal retrievals, which are insufficient for multimodal tasks requiring dynamic, query-specific retrievals across image-text pairs (Lewis et al., 2021). This limitation often leads to issues with retrieval relevance and alignment, impacting performance on complex vision-language tasks like VQA, multimodal sentiment analysis, and cross-modal reasoning (J. Lu et al., 2022).

Overcoming these challenges requires a retrieval framework that can integrate diverse data types and align retrieved demonstrations with specific task contexts. Although models like CLIP (Radford et al., 2021) illustrate the potential of multimodal embedding spaces for cross-modal understanding, they lack the capability for dynamic, query-specific retrieval. Recent approaches, such as the Unified Demonstration

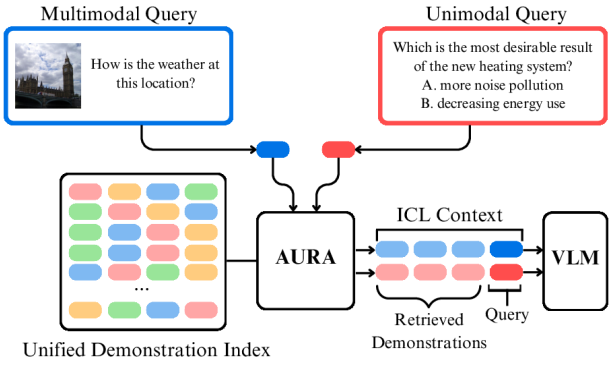

Figure 1: **Overview of AURA's unified multimodal demonstration retrieval process**. Given a query—either multimodal (image-text) or unimodal (text-only)—AURA retrieves task-relevant demonstrations from a unified multimodal demonstration index. The unified index and structured retrieval process enable AURA to dynamically adapt to diverse task types, improving the relevance and effectiveness of demonstrations for multimodal ICL.

Retriever (UDR) (X. Li et al., 2023) and Retrieval-based In-Context Learning (Ret-ICL) (Scarlatos & Lan, 2024), introduce task-aware retrieval mechanisms to enhance adaptability, but are often limited to static or unimodal contexts, limiting their flexibility in complex multimodal applications.

To address these limitations, we propose **AURA** (Adaptive Unified Retrieval and Alignment), a task-agnostic retrieval framework designed to enhance ICL by dynamically retrieving query-relevant demonstrations across both unimodal and multimodal datasets. AURA operates within a unified multimodal embedding space and utilizes contrastive triplet learning and an VLM-guided feedback mechanism to optimize retrieval relevance and alignment. During initial training, contrastive triplet learning organizes the embedding space by creating "triplets" consisting of a query, a positive sample (from the same dataset as the query), and a negative sample (from a different dataset). This phase encourages task-specific clustering and separation across datasets, supporting retrieval relevance. In the fine-tuning phase, VLM-guided feedback further refines this structure by evaluating the quality of retrieved demonstrations based on the VLM's prediction accuracy, promoting alignment with task requirements and enhancing intra-dataset clustering.

AURA's training process consists of two phases: an initial phase that establishes dataset-level clustering in the embedding space, and a fine-tuning phase that adjusts this clustering based on VLM feedback to improve retrieval precision within each dataset. During fine-tuning, the model adapts its embedding space to prioritize the proximity of effective demonstrations to each query, thereby enhancing relevance and reducing cross-dataset interference. AURA uses a FAISS-based indexing strategy (Douze et al., 2024; Johnson et al., 2019) at inference to efficiently retrieve demonstrations. FAISS optimizes for high-speed, low-latency retrieval by indexing embeddings in a high-dimensional space, which ensures scalability and fast response across diverse datasets.

Our experiments show that AURA successfully improves retrieval accuracy and task performance across vision-language benchmarks, including Visual Question Answering (VQA) and multimodal sentiment analysis. Furthermore, AURA demonstrates versatility by generalizing across disparate tasks, highlighting its potential as a flexible and adaptive retrieval framework for advancing multimodal ICL in real-world applications. Detailed analysis reveals that AURA's structured retrieval approach supports adaptation across tasks of varying complexity, highlighting its effectiveness in task-specific and cross-modal retrieval.

2 Related Works

In recent years, In-Context Learning (ICL) has garnered considerable attention for its ability to adapt large language models (LLMs) and vision-language models (VLMs) to diverse tasks without explicit fine-tuning. ICL achieves this by providing relevant demonstration examples alongside the input query, enabling models to infer task requirements from context. Early work, notably by Brown et al. (2020), demonstrated the capacity of autoregressive models like GPT-3 to perform various tasks using only a few in-context examples, albeit with limitations due to static, manually curated demonstrations that may lack adaptability for diverse inputs. More recent studies have proposed various approaches to enhance

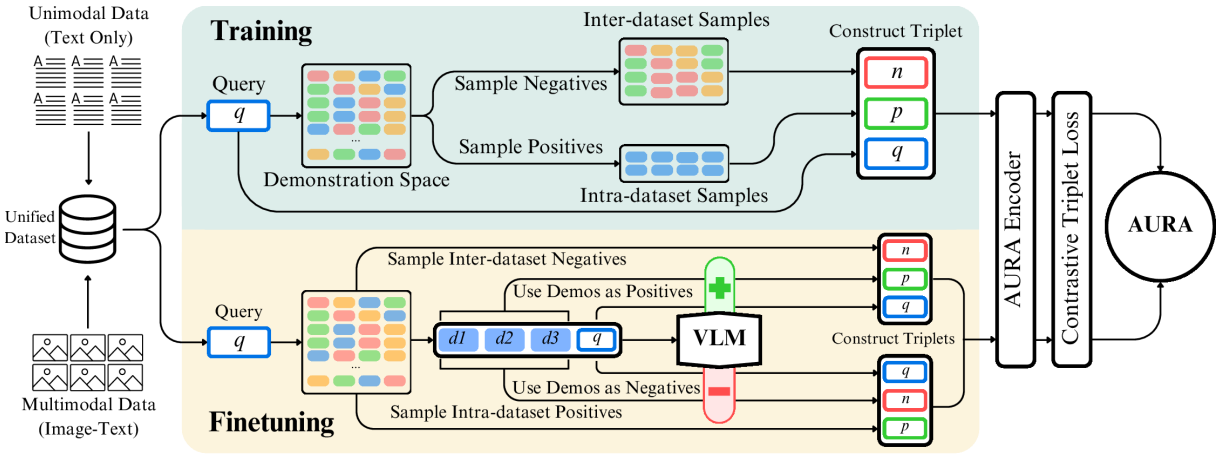

Figure 2: **Overview of AURA's training and fine-tuning**. In the training phase (top), intra-dataset samples (from the same dataset as the query) serve as positive examples, while inter-dataset samples (from different datasets) act as negatives, promoting dataset-level clustering and separation across unimodal (text) and multimodal (image-text) data. Each query (anchor) is contrasted with these samples to form triplets, optimized through contrastive triplet loss. In the fine-tuning phase (bottom), a VLM evaluates the quality of retrieved demonstrations; based on its performance, the demonstrations are labeled as positives or negatives to form task-specific triplets. This fine-tuning process refines intra-dataset clustering, improving retrieval accuracy by aligning demonstrations to task needs.

ICL by dynamically retrieving relevant demonstrations, a method known as Retrieval-based In-Context Learning (Ret-ICL) (Luo et al., 2024; Scarlatos & Lan, 2024; Sun et al., 2024).

Static and Manual Demonstration-Based ICL Traditional ICL methods rely heavily on fixed, manually selected examples, which are tailored for specific tasks but often fail to generalize across varied input distributions. Specifically, Brown et al. (2020) first showcased the utility of in-context demonstrations, setting a precedent for few-shot task generalization, while those demonstrations are fixed across queries. However, this static demonstration selection lacks adaptability, limiting performance when input queries deviate from the anticipated task structure. In other words, although effective in limited settings, these static approaches struggle in more complex, cross-modal contexts where fixed examples may inadequately capture essential modality-specific nuances (Devlin et al., 2019; Tsimpoukelli et al., 2021).

Retrieval-based In-Context Learning (Ret-ICL) The drawbacks of static ICL have led to the development of retrieval-based methods, where adaptive retrievers dynamically select demonstrations based on the query context, effectively tailoring examples to each input (Mosbach et al., 2023). Scarlatos and Lan (2024) propose Retrieval-based ICL (Ret-ICL), a sequential retrieval framework using reinforcement learning (RL) to optimize the order and selection of demonstrations, maximizing relevance to the task. Luo et al. (2024) extend the adaptability of Ret-ICL by investigating retrieval dynamics across varied input attributes, showing significant gains in alignment with input queries through selective retrieval.

Recent advancements like the Unified Demonstration Retriever (UDR) illustrate how multi-task retrieval can improve ICL's flexibility (X. Li et al., 2023). UDR employs a contrastive embedding approach to align task-agnostic demonstrations across a range of NLP applications, underscoring Ret-ICL's potential to handle diverse retrieval objectives. Nonetheless, many Ret-ICL models remain constrained to single-modality tasks, limiting their capacity to handle more complex, cross-modal requirements that are essential for multimodal applications (Sun et al., 2024; Tsimpoukelli et al., 2021).

Unified Cross-Modal and Cross-Task Models Cross-modal ICL models like UnifiedIO have emerged to address the challenges associated with handling diverse input modalities within a single framework (J. Lu et al., 2022). UnifiedIO leverages a sequence-to-sequence (Seq2Seq) architecture to process both visual and language data, enabling seamless integration across tasks and modalities. Models such as MURAL and CLIP further advance cross-modal learning through contrastive learning techniques, aligning image and text representations within a shared embedding space that facilitates zero-shot transfer (Jain

et al., 2021; Radford et al., 2021). These models have proven effective in zero-shot and few-shot scenarios, yet they typically lack retrieval-based adaptability, relying on pre-trained embeddings that may not optimally reflect query-specific contextual nuances (Zhou et al., 2022).

Unified retrieval frameworks like UDR introduce adaptability to these cross-modal models by allowing retrieval mechanisms to dynamically select demonstrations across tasks and modalities (X. Li et al., 2023). Despite these advancements, cross-modal models still require extensive pre-training on task-specific datasets, limiting their effectiveness in retrieval-based applications where input diversity is high and task-specific adaptation is challenging (Jain et al., 2021).

Multimodal Retrieval-Augmented Generation (RAG) Retrieval-augmented generation frameworks have been adapted for multimodal applications to bolster robustness and specificity in vision-language models (VLMs). Zhao et al. (2023) highlight the importance of adaptive retrieval in mitigating issues like hallucination and reasoning errors, supporting more contextually aligned responses by incorporating multimodal retrieval from dedicated databases. Models such as BLIP (J. Li et al., 2022) and PaLI (X. Chen et al., 2023) combine vision-language fusion with adaptive retrieval to address complex tasks like Visual Question Answering (VQA) and multimodal reasoning. These models generate task-specific outputs by retrieving relevant text-image pairs, thus enhancing response precision.

Despite these advancements, most multimodal retrieval-augmented systems still rely on fixed, predefined retrieval corpora, which may fail to capture nuanced task-specific or modality-specific requirements effectively. In contrast, emerging frameworks that attempt to integrate adaptive retrieval within a unified embedding space, although promising, remain limited in their ability to perform truly dynamic, query-specific retrieval (Guo et al., 2022; Sun et al., 2024). Our work builds upon these approaches by proposing a task-agnostic framework, AURA, which achieves a fully adaptive retrieval process that dynamically retrieves contextually relevant examples across diverse datasets and modalities in a unified embedding space. This approach enhances VLMs' capacity for retrieval precision and alignment, addressing limitations of prior fixed-corpus and partially adaptive models.

3 Methodology

Our proposed approach introduces a unified multimodal retriever model designed to enhance in-context learning through retrieval of task-relevant demonstrations across diverse unimodal and multimodal datasets. By leveraging both a contrastive triplet learning framework and a feedback loop from a vision-language model (VLM), our model dynamically aligns its embedding space to reflect the distinct requirements of each dataset. In this section, we detail the model architecture, training process, and the unique design choices aimed at optimizing retrieval performance for in-context learning.

3.1 Model Architecture

Our model extends the CLIP architecture (Radford et al., 2021), incorporating separate processing paths for text and image inputs through the use of modality-specific projection layers. The CLIP model's dual-modality encoder extracts unique features for each input type, projecting image features into a 1024-dimensional space and text features into a 512-dimensional space. These modality-specific features are then mapped to a unified 128-dimensional embedding space via linear projections:

$$z_{\text{image}} = f_{\text{image\_proj}}(g_{\text{image}}(x_{\text{image}})), \tag{1}$$

$$z_{\text{text}} = f_{\text{text\_proj}}(g_{\text{text}}(x_{\text{text}})), \tag{2}$$

where g_{image} and g_{text} represent the image and text encoders, respectively, and $f_{\text{image\_proj}}$ and $f_{\text{text\_proj}}$ are linear projections for each modality. This design enables the model to integrate visual and textual information within a shared 128-dimensional embedding space, an empirically chosen dimension that balances computational efficiency with retrieval accuracy.

3.2 Task-Specific Instructional Prompts

To differentiate datasets within the embedding space, we prepend each data query with a task-specific instructional prompt. This approach aligns with previous findings on unified retrieval models, where task-based prompts promote intra-dataset clustering while enhancing inter-dataset separation (X. Li et al., 2023). These prompts function as distinctive identifiers or "fingerprints" for each dataset, embedding

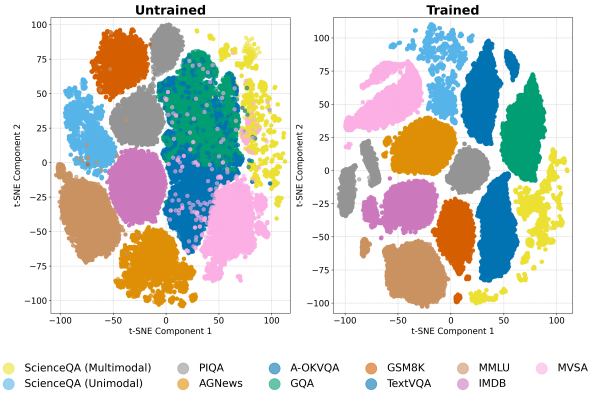

Figure 3: **t-SNE visualization of the unified embedding space**, showing distinct clusters for different dataset tasks. The plot at left shows the embedding space with the initial, untrained model. The plot at right reflects the trained model before finetuning. The clear boundaries between clusters highlight the model's ability to differentiate and separate tasks, facilitated by task-specific instructional prompts. This differentiation supports effective in-context learning by promoting intra-dataset clustering and minimizing cross-dataset interference.

dataset-specific semantics within the model's learned representations. For a query string x_{str} belonging to dataset d, the input text q is constructed as:

$$q = I[d] \cdot x_{str}, \tag{3}$$

where I is a dictionary of predefined task-specific prompt strings and $\cdot$ denotes the concatenation of two strings. This construction allows the model to distinguish datasets effectively, facilitating retrieval of demonstrations that are specific to each task.

3.3 Initial Training

The initial training phase is designed to establish clear distinctions between datasets and modalities within the embedding space, supporting retrieval precision for diverse tasks. The training objective is to position samples from the same dataset close to each other while creating broader separations between samples from different datasets or modalities. This setup ensures that the model can retrieve examples specific to a query's dataset and modality while avoiding irrelevant or incompatible demonstrations.

To achieve this, we employ a contrastive triplet loss with a focus on inter-dataset differentiation. For each query sample q, a positive sample p is selected from the same dataset, while a negative sample n is selected from a different dataset. By defining the triplets in this manner, we encourage each dataset to occupy a distinct locality within the embedding space, allowing the model to preserve dataset-specific and modality-specific clusters. This training objective is represented by the triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, d(q,p) - d(q,n) + \alpha\right), \tag{4}$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance and α is a margin parameter that enforces a minimum distance between positive and negative pairs.

The effectiveness of this training approach is illustrated in Figure 3, which shows a t-SNE visualization of the embedding space before and after initial training. In the untrained model (left), there is little structure or separation among the different datasets. After initial training (right), distinct clusters emerge, corresponding to different dataset tasks. This clustering effect, facilitated by the contrastive triplet learning framework, demonstrates the model's ability to organize the embedding space according to task and modality, enabling it to retrieve contextually relevant demonstrations within each dataset.

This structured approach to triplet selection mitigates modality leakage, where a unimodal query might retrieve multimodal demonstrations or vice versa, and ensures that task alignment is preserved at the dataset level. As a result, the initial training phase creates a well-defined embedding structure that supports the model's ability to retrieve contextually relevant demonstrations for each dataset.

Method	Multimodal (Image + Text)					Unimodal (Text Only)					
	TextVQA	MVSA	GQA	A-OKVQA	ScienceQA-MM	MMLU	IMDB	AGNews	PIQA	GSM8K	ScienceQA-UM
Zero-Shot	**63.40**	59.40	66.20	66.80	48.00	55.40	90.60	76.00	71.40	5.40	64.80
Random, All (n=3)	62.20	58.00	66.00	67.20	48.20	42.80	85.40	71.80	59.20	4.00	52.20
Random, All (n=5)	62.20	58.80	63.60	62.00	45.20	40.40	77.40	66.80	55.60	3.80	49.40
Random, Dataset (n=3)	59.20	60.40	58.20	66.60	48.20	54.40	86.80	78.60	72.40	6.20	65.60
Random, Dataset (n=5)	55.00	57.80	53.20	61.20	42.80	53.00	83.00	79.40	73.00	6.20	66.00
Ours (No FT)	59.40	59.20	58.40	66.20	49.00	53.80	89.00	80.20	72.60	5.20	**67.80**
Ours (Best)	61.40	**62.20**	**66.40**	**70.00**	**51.40**	**55.80**	**90.80**	**80.40**	**73.20**	**7.20**	**67.80**

Table 1: Accuracy performance comparison across datasets with scoring VLM. Results are shown for zero-shot, Random 5-shot and 3-shot, dataset-specific random 3-shot and 5-shot, and AURA configurations with and without fine-tuning. "Ours (best)" reflects the highest performance across different fine-tuning epochs (e=1, 3, 5). Best results are shown in **bold**, and the second-best are underlined.

3.4 Finetuning with VLM Feedback

In the finetuning phase, AURA shifts its focus from cross-dataset differentiation to refining intra-dataset alignment, optimizing the local structure of clusters to improve few-shot in-context learning (ICL) performance. The goal of this phase is to rearrange the members of each dataset's cluster in the embedding space so that the most effective ICL demonstrations—those that lead to correct outputs when used with a query—are positioned close to the query itself.

To construct triplets for finetuning, we use a demonstration selection process guided by VLM feedback. For each query in the finetuning dataset, we sample three random demonstrations from the same dataset as the query. These demonstrations are then used in a 3-shot ICL setup with the query, where the VLM's prediction is evaluated against the ground truth. If the VLM produces a correct prediction, the three demonstrations are designated as positive pairs with the query, while negative samples are selected from different datasets to further reinforce cross-dataset boundaries within the embedding space.

If the VLM's prediction is incorrect, the selected demonstrations are marked as negative pairs with the query, and positive samples are drawn randomly from within the same dataset. This setup introduces "noise" by incorporating less aligned, intra-dataset samples as positive pairs, which reduces the risk of overfitting to specific demonstration configurations. The contrastive triplet loss for finetuning remains consistent with the training phase as defined in (4), maintaining the objective of minimizing the distance between queries and effective ICL demonstrations while increasing the separation from less relevant examples.

Through this targeted refinement of intra-dataset clustering, finetuning with VLM feedback enables AURA to focus on task-level nuances within each dataset, facilitating more accurate retrievals. This process allows AURA to adaptively refine its embedding space for improved in-context learning, making it capable of retrieving contextually appropriate demonstrations tailored to the task requirements of each query.

4 Experiments

The experiments conducted are designed to evaluate the efficacy of AURA in retrieving contextually relevant demonstrations for in-context learning across a range of multimodal and unimodal datasets. Specifically, the experiments assess AURA's capacity for dataset-specific retrieval and in-context learning augmentation, using a broad selection of task types and domains across text and image modalities. These experiments were also structured to systematically investigate AURA's performance under various configurations, including both with and without fine-tuning on VLM feedback. This section provides a comprehensive overview of the datasets, baseline models, experimental setup, and the metrics used for evaluation.

4.1 Datasets

To rigorously test AURA's performance, we utilize a set of diverse and well-established benchmark datasets that encapsulate various multimodal (image-text) and unimodal (text only) learning challenges across domains including knowledge-based visual question answering (VQA), sentiment analysis, topic classification, physical commonsense reasoning, and scientific question answering. Specifically, we use the following benchmarks: **A-OKVQA** (Schwenk et al., 2022) for knowledge-based VQA, requiring the

	Multimodal (Image + Text)	Unimodal (Text Only)
Method	MMStar	MILU
Zero-Shot	31.60	52.80
Random, All (n=3)	32.80	42.00
Ours (Best)	**33.20**	**53.20**

Table 2: Accuracy performance comparison on two out-of-domain datasets; one unimodal and one multimodal. "Ours (best)" reflects the highest performance across different fine-tuning epochs (e=1, 3, 5). Best results are shown in **bold**.

model to integrate visual data with factual knowledge to respond accurately; **AGNews** (Zhang et al., 2015) as a topic classification dataset, assessing the model's ability to categorize textual news articles effectively; **GQA** (Hudson & Manning, 2019) for compositional reasoning in VQA, which requires understanding spatial relationships within images to accurately answer queries; **GSM8K** (Cobbe et al., 2021) a math problem-solving dataset to evaluate retrieval quality in structured reasoning; **IMDB** (Hendrycks, Burns, Basart, Critch, et al., 2021; Maas et al., 2011) a sentiment analysis task focused on text, requiring the model to discern the sentiment conveyed in movie reviews; **MMLU** (Hendrycks, Burns, Basart, Zou, et al., 2021), a general knowledge multiple-choice dataset, assessing AURA's ability to handle diverse factual topics without fine-tuning; **MVSA** (Niu et al., 2016) for multimodal sentiment analysis, integrating text and visual cues; **PIQA** (Bisk et al., 2020) a physical commonsense reasoning dataset to test everyday interaction knowledge; **ScienceQA-Multimodal** and **ScienceQA-Unimodal** (P. Lu et al., 2022) for scientific question answering, evaluating multimodal comprehension in domains that require logical reasoning; **TextVQA** (Singh et al., 2019) for text-focused visual question answering, emphasizing text comprehension within images.

4.2 Baselines and Model Configurations

We evaluate AURA against a range of baseline configurations to benchmark its retrieval capabilities. The zero-shot (ZS) baseline serves as a foundational measure of vision-language models (VLMs) without any demonstration retrieval, while random 5-shot and 3-shot baselines provide a lower bound for performance by selecting demonstrations arbitrarily from the unified dataset space, with minimal task alignment.

To introduce some degree of task relevance, we include dataset-specific random 5-shot and 3-shot baselines, which restrict retrieval to the query's parent dataset but lack the structured retrieval employed by AURA. These comparisons highlight the added value of AURA's context-aware and dynamically optimized retrieval approach.

AURA is evaluated in two configurations: AURA (No Finetune), which uses the pre-trained embedding space to assess the impact of structured retrieval alone, and AURA (Finetuned), which incorporates VLM-guided feedback during fine-tuning to refine retrieval precision. For the fine-tuned configuration, we report the best accuracy across multiple epochs, reflecting the model's peak performance.

4.3 Evaluation Metrics

Accuracy, defined as the percentage of correctly answered queries, serves as the primary metric for assessing retrieval quality and in-context learning performance. This metric provides a consistent basis for comparing the effectiveness of various retrieval configurations. Results are reported for all configurations, including zero-shot, random 3-shot and 5-shot, dataset-specific random 3-shot and 5-shot, and AURA's configurations. For AURA (Finetuned), the best accuracy achieved across fine-tuning epochs is highlighted as "Ours (best)," illustrating the impact of fine-tuning and VLM feedback on retrieval quality.

4.4 Experimental Setup

AURA operates within a 3-shot in-context learning framework, retrieving the top three most relevant demonstrations for each query to construct a context for VLM inference. Retrieval is facilitated by a FAISS-based indexing system, which efficiently stores and retrieves embeddings within a high-dimensional space, constructed by encoding all demonstration samples with AURA's pretrained model. This shared embedding space enables rapid retrievals optimized for task relevance. Additionally, we use LLaVA-1.6-7B (Liu et al., 2023) as the underlying VLM for scoring and inference tasks, providing a consistent platform for evaluating AURA's impact on in-context learning.

5 Results and Discussion

The results in Table 1 highlight AURA's robust performance across a diverse set of multimodal and unimodal tasks, showcasing reliable improvements over baseline retrieval methods. AURA consistently demonstrates the advantages of structured, context-aware retrieval, achieving strong accuracy on tasks that require fine-grained, task-specific alignment. By incorporating VLM-guided fine-tuning, AURA enhances its embedding space to optimize retrieval relevance for each query, improving contextual alignment and retrieval precision across a range of complex tasks.

AURA's strong performance across diverse tasks and modalities, particularly in multimodal contexts, highlights the adaptability of structured retrieval in in-context learning. The fine-tuning process enables the model to scale effectively across varying task requirements by capturing broad inter-dataset distinctions while refining intra-dataset relationships. These results demonstrate that AURA's retrieval adaptations not only enhance accuracy but also improve the flexibility and reliability of in-context learning frameworks. As a versatile solution for advancing LVLMs in real-world, multimodal applications, AURA's structured retrieval mechanism proves effective in achieving robust, task-specific alignment across complexities.

5.1 Effectiveness of Fine-Tuned Retrieval

AURA consistently surpasses random retrieval baselines, emphasizing the value of structured, task-aligned retrieval for enhancing performance. The fine-tuned configuration achieves the highest accuracy across most datasets, demonstrating that task-specific tuning can substantially improve retrieval precision. For example, AURA's fine-tuning particularly benefits tasks like MVSA (sentiment analysis) and A-OKVQA (knowledge-based VQA), where a deep semantic alignment between query and demonstration is essential for accurate predictions. The improvements in these datasets suggest that fine-tuning helps AURA to better differentiate between subtle, task-specific attributes within the shared embedding space, refining intra-dataset clustering and ensuring robust inter-dataset separation.

Interestingly, the dataset-specific random retrieval baselines (n=3 and n=5), which restrict retrieval to a query's parent dataset, provide a moderate performance boost over general random selection. This trend indicates that task-specific dataset filtering introduces some level of alignment, yet the lack of structured retrieval limits its efficacy. AURA's fine-tuned retrieval, in contrast, achieves a more precise alignment by dynamically selecting demonstrations based on both task relevance and semantic compatibility, which is particularly impactful in complex reasoning tasks like GQA and PIQA. This suggests that AURA's structured approach enables a nuanced representation of task requirements, capturing both high-level task structures and finer-grained contextual cues that are crucial for retrieval success.

5.2 Effects of Task Complexity and Modality

The impact of AURA's fine-tuning varies depending on task complexity and modality. In simpler tasks, such as AGNews (topic classification) and IMDB (sentiment analysis), AURA performs well in its pre-trained state, where broad semantic cues suffice for task success. Minimal fine-tuning is needed in these cases because the model's general embedding space already provides robust retrieval for straightforward tasks. This is reflected in the marginal gains achieved by fine-tuning, suggesting that retrieval alignment is less dependent on fine-tuning for tasks with well-defined semantic categories.

In contrast, complex tasks that require multimodal integration or nuanced reasoning, such as A-OKVQA, GQA, and ScienceQA-MM, see significant improvements from AURA's fine-tuning. These tasks involve intricate relationships between visual and textual information, demanding precise alignment of cross-modal cues. AURA's ability to dynamically adjust its embedding space during fine-tuning enhances retrieval for these tasks by refining intermodal dependencies and ensuring that retrieved demonstrations are contextually aligned with the query's modality and complexity level. The results indicate that fine-tuning is particularly beneficial for multimodal tasks, where visual and textual content must be integrated effectively. This trend suggests that fine-tuning adjusts AURA's embedding space to account for the complexities of multimodal tasks, enabling retrieval to capture the required cross-modal dependencies.

An interesting behavior observed in AURA's retrieval mechanism is its capacity to generalize across related tasks within similar meta-categories, even in the absence of direct task alignment. For instance, in the PIQA dataset, which involves physical commonsense reasoning, AURA's fine-tuned configuration achieves a notable improvement over random baselines. PIQA requires understanding common-sense physical interactions, a task that benefits from retrievals that align with real-world knowledge and logical

reasoning. This behavior suggests that AURA's embedding space captures the core structural features of each dataset, allowing it to generalize retrieval relevance based on task similarities.

5.3 Out-of-domain (OOD) Task Performance

To assess AURA's generalization to unseen domains, we evaluated it on two out-of-domain (OOD) datasets: MILU (Verma et al., 2024) (unimodal text-only) and MMStar (L. Chen et al., 2024) (multimodal image-text). For each dataset, we performed zero-shot, random 3-shot, and AURA 3-shot inference, with results shown in Table 2. AURA's retrievals were conducted without task-specific prompt strings, relying solely on its learned embeddings.

We evaluated AURA's OOD performance on three criteria: (1) modality alignment between the query and retrieved demonstrations, (2) task alignment with the query's meta-task (e.g., QA or sentiment analysis), and (3) overall ICL accuracy. Results show that AURA met all three conditions across both datasets, demonstrating strong generalization. AURA consistently retrieved demonstrations matching the modality of the query and achieved higher accuracy than both the random few-shot and zero-shot baselines.

Notably, AURA's retrieval selections showed insightful task-level alignment. For MILU, a multiple-choice QA dataset, AURA retrieved examples primarily from AGNews, a text classification dataset. While distinct in format, both datasets rely on high-context understanding, suggesting that AURA's embedding space effectively captures semantic similarities. Similarly, for MMStar, which involves complex image-text dependencies, AURA selected demonstrations from MVSA, a multimodal sentiment analysis dataset. Both tasks require nuanced integration of visual and textual information, highlighting AURA's ability to identify intermodal dependencies in OOD contexts.

6 Future Work

While this work establishes AURA as a versatile framework for unified multimodal retrieval, future research can expand its capabilities to address real-world applications and more specialized contexts. First, exploring AURA's performance in more complex multimodal scenarios, such as video-text retrieval or 3D spatial tasks, could push the boundaries of its retrieval precision and adaptability. These extensions would require further refinement of the embedding space to capture higher-dimensional dependencies, allowing AURA to handle increasingly sophisticated input modalities.

Another promising direction is to investigate the effects of dataset curation and demonstration selection on retrieval outcomes. Understanding the impact of selecting narrow or broad task distributions within the demonstration corpus could shed light on optimal dataset composition strategies. For instance, a broader selection could enhance cross-task flexibility, providing practical insights for deployment in dynamic, real-world settings where task types may vary widely.

Future work could also focus on developing more specialized training paradigms to enhance AURA's retrieval adaptability. This might include real-time retrieval updates, where the model adjusts its retrieval strategy based on live feedback, or task-specific fine-tuning processes that adapt AURA for specialized domains like medical imaging or autonomous navigation. By exploring these avenues, AURA could evolve into a robust, flexible tool tailored for high-stakes applications requiring adaptable, context-aware retrieval.

7 Conclusion

This work presented AURA (Adaptive Unified Retrieval and Alignment), a task-agnostic retrieval framework developed to advance in-context learning (ICL) in complex vision-language applications. AURA addresses the limitations of traditional retrieval methods that rely on static or unimodal demonstrations by dynamically retrieving and aligning query-relevant examples from both text and image modalities. By integrating a unified embedding space informed by contrastive triplet learning, AURA achieves flexible, query-specific retrieval, further refined through a vision-language model (VLM)-guided feedback mechanism to ensure retrieval relevance and alignment within each task's context.

Experimental evaluations across a diverse set of benchmarks, including Visual Question Answering (VQA) and multimodal sentiment analysis, demonstrate AURA's effectiveness in improving retrieval precision and enhancing task performance. The framework's task-agnostic adaptability allows it to operate effectively across varied datasets and task complexities, highlighting its potential as a robust solution

for multimodal in-context learning. AURA's architecture supports efficient retrieval within a unified multimodal structure, showcasing its scalability and flexibility in handling disparate datasets and task requirements.

AURA stands as a significant contribution to the development of adaptive multimodal retrieval frameworks, illustrating the feasibility of a unified, task-agnostic embedding space for in-context learning across complex, multimodal tasks. The structured, VLM-guided retrieval approach established in this work lays a foundation for future research in adaptable, contextually aware retrieval systems, and offers a promising pathway toward scalable multimodal applications capable of meeting the demands of real-world complexity.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2016). *Vqa: Visual question answering.* Retrieved from https://arxiv.org/abs/1505.00468

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., & Choi, Y. (2020). Piqa: Reasoning about physical commonsense in natural language. In *Thirty-fourth aaai conference on artificial intelligence.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020, July). *Language Models are Few-Shot Learners.* arXiv. Retrieved 2024-08-06, from http://arxiv.org/abs/2005.14165 (arXiv:2005.14165 [cs]) DOI: 10.48550/arXiv.2005.14165

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., ... others (2024). Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., ... Soricut, R. (2023). *Pali: A jointly-scaled multilingual language-image model.* Retrieved from https://arxiv.org/abs/2209.06794

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Retrieved from https://arxiv.org/abs/1810.04805

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., ... Sui, Z. (2024, November). A survey on in-context learning. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 1107–1128). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.64

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., ... Jégou, H. (2024). *The faiss library.* Retrieved from https://arxiv.org/abs/2401.08281

Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., & Kankanhalli, M. (2022). A unified end-to-end retriever-reader framework for knowledge-based vqa. In *Proceedings of the 30th acm international conference on multimedia* (p. 2061–2069). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3503161.3547870 DOI: 10.1145/3503161.3547870

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jain, A., Guo, M., Srinivasan, K., Chen, T., Kudugunta, S., Jia, C., ... Baldridge, J. (2021, September). *MURAL: Multimodal, Multitask Retrieval Across Languages.* arXiv. Retrieved 2024-08-06, from http://arxiv.org/abs/2109.05125 (arXiv:2109.05125 [cs]) DOI: 10.48550/arXiv.2109.05125

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive nlp tasks.* Retrieved from https://arxiv.org/abs/2005.11401

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.* Retrieved from https://arxiv.org/abs/2201.12086

Li, X., Lv, K., Yan, H., Lin, T., Zhu, W., Ni, Y., ... Qiu, X. (2023). *Unified demonstration retriever for in-context learning.* Retrieved from https://arxiv.org/abs/2305.04320

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). *Improved baselines with visual instruction tuning.*

Lu, J., Clark, C., Zellers, R., Mottaghi, R., & Kembhavi, A. (2022). *Unified-io: A unified model for vision, language, and multi-modal tasks.* Retrieved from https://arxiv.org/abs/2206.08916

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., ... Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th conference on neural information processing systems (neurips).*

Luo, M., Xu, X., Liu, Y., Pasupat, P., & Kazemi, M. (2024). *In-context learning with retrieved demonstrations for language models: A survey.* Retrieved from https://arxiv.org/abs/2401.11624

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1015

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., & Elazar, Y. (2023, May). *Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation.* arXiv. Retrieved 2024-07-05, from http://arxiv.org/abs/2305.16938 (arXiv:2305.16938 [cs])

Niu, T., Zhu, S., Pang, L., & El-Saddik, A. (2016). Sentiment analysis on multi-view social data. In *Multimedia modeling* (p. 15–27).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.* Retrieved from https://arxiv.org/abs/2103.00020

Scarlatos, A., & Lan, A. (2024). *Reticl: Sequential retrieval of in-context examples with reinforcement learning.* Retrieved from https://arxiv.org/abs/2305.14502

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022). *A-okvqa: A benchmark for visual question answering using world knowledge.* Retrieved from https://arxiv.org/abs/2206.01718

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., ... Rohrbach, M. (2019). Towards vqa models that can read. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8317–8326).

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., ... Wang, X. (2024). *Generative multimodal models are in-context learners.* Retrieved from https://arxiv.org/abs/2312.13286

Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., & Hill, F. (2021, July). *Multimodal Few-Shot Learning with Frozen Language Models.* arXiv. Retrieved 2024-08-06, from http://arxiv.org/abs/2106.13884 (arXiv:2106.13884 [cs]) DOI: 10.48550/arXiv.2106.13884

Verma, S., Khan, M. S. U. R., Kumar, V., Murthy, R., & Sen, J. (2024). *Milu: A multi-task indic language understanding benchmark.* Retrieved from https://arxiv.org/abs/2411.02538

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator.* Retrieved from https://arxiv.org/abs/1411.4555

Zhang, X., Zhao, J. J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Nips.*

Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... Joty, S. (2023). *Retrieving multimodal information for augmented generation: A survey.* Retrieved from https://arxiv.org/abs/2303.10868

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022, September). Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, *130*(9), 2337–2348. Retrieved 2024-08-06, from http://arxiv.org/abs/2109.01134 (arXiv:2109.01134 [cs]) DOI: 10.1007/s11263-022-01653-1

Appendix

A Datasets

A.1 Dataset Splitting

Task	Train	Fine-tune	Test
MMLU	10,000	600	500
IMDB	10,000	600	500
ScienceQA-UM	6,508	301	500
AGNews	10,000	600	500
PIQA	10,000	551	500
GSM8K	7,473	396	500
TextVQA	10,000	600	500
GQA	10,000	600	500
ScienceQA-MM	6,218	299	500
MVSA	10,000	600	500
A-OKVQA	10,000	344	500
MMStar	-	-	300
MILU	-	-	300
Total	100,199	5,491	5,900

Table 3: Dataset Splitting and Capping for Training, Fine-tuning, and Testing across all tasks.

For the core evaluation of AURA, each dataset is capped at a maximum of 10,000 samples for training, 500 for testing, and 600 for fine-tuning. Following established methods for ensuring computational efficiency, these sample-size limitations are meant to balance the trade-off between time cost and computation. Table 3 shows the specific data splitting and capping across all tasks. For out-of-domain (OOD) testing, we used 300 samples each from the MILU and MMStar datasets to evaluate AURA's generalization capabilities.

A.2 Modality Isolation

Some datasets contained both unimodal (text-only) and multimodal (text-image) data. To ensure a fair comparison of these modalities independent of one another, we split such datasets into modality-specific subsets to isolate these samples. Additionally, some multimodal datasets contain samples with multiple interleaved images and a single text query. To standardize task complexity, we omit these samples from the source datasets prior to splitting.

B Scoring LLM

For all scoring and inference experiments, we use the `llava-hf/llava-v1.6-mistral-7b-hf` checkpoint from Huggingface. This LLM configuration provides consistency across evaluation and fine-tuning, ensuring that all performance metrics directly reflect AURA's retrieval capabilities without variations introduced by different language models.

C ICL Prompts

Prompts play a crucial role in in-context learning (ICL) performance, as established by prior work. However, to minimize prompt-related variabilities, we adopted a simple, standardized prompt format for all tasks. The base prompt template is:

```
"Please respond with a single word or phrase."
```

Task	Instruction
MMLU	<MMLU> Provide concise and factual answers using relevant domain-specific knowledge.
IMDB	<IMDB> Analyze the given text and identify the underlying sentiment and emotions.
ScienceQA-Unimodal	<ScienceQA-Unimodal> Answer the question based on scientific knowledge and logical reasoning.
AGNews	<AGNews> Classify the news article into the correct category based on its content.
PIQA	<PIQA> Select the most plausible answer to the physical commonsense question.
GSM8K	<GSM8K> Solve the math problem step-by-step using logical reasoning and arithmetic.
TextVQA	<TextVQA> Interpret the image and extract relevant information to answer the given question.
GQA	<GQA> Analyze the visual elements logically to answer the question with precise reasoning.
ScienceQA-Multimodal	<ScienceQA-Multimodal> Answer the question based on the visual and scientific information provided.
MVSA	<MVSA> Analyze the image and text together to determine the sentiment expressed.
A-OKVQA	<A-OKVQA-Multimodal> Answer the open-ended question using visual content and external knowledge.

Table 4: Task Instructions for Various AI Benchmarks.

This approach reduces the risk of "yes-and" hallucinations, where the LLM might generate responses that deviate from the intended output format. It also limits extraneous factors that might affect ICL performance, allowing us to evaluate AURA's retrieval quality in a prompt-agnostic context.

For the IMDB sentiment analysis dataset, which does not include questions in its raw form, we use an additional task-specific prompt:

```
"Please classify this as either positive or negative."
```

For example, if the original input is `"I hate this movie"`, the full prompt becomes:

```
"I hate this movie.  <newline> Please classify this as either positive or negative.
Please respond with a single word or phrase."
```

D Task Instructions

Table 4 details the specific task instructions used during AURA's development. These instructions serve as "fingerprints" for each dataset, helping to differentiate tasks within the shared embedding space and facilitating more accurate retrieval by aligning with task-specific requirements.

E Model Architecture

The core architecture of AURA builds on the `openai/clip-vit-base-patch32` model, structured as a dual-encoder for processing text and image inputs. CLIP generates joint embeddings for text and image modalities within a shared semantic space, crucial for cross-modal tasks.

To adapt CLIP for task-specific retrieval, text and image embeddings are projected into a unified 128-dimensional embedding space using linear projection layers:

- Text features are encoded into 512 dimensions by CLIP's text encoder and projected to 128 dimensions.
- Combined text and image features are projected from 1024 to 128 dimensions.

This unified embedding space aligns text and image features with dataset-specific contexts, facilitating efficient multimodal and unimodal similarity searches during retrieval.

F AURA Code Overview

Text and Image Encoding: AURA uses the CLIP processor to tokenize text and preprocess images. Multimodal inputs are encoded, concatenated, and projected into the unified 128-dimensional embedding space via the `image_projection` layer. For unimodal (text-only) inputs, the `text_projection` layer processes the text embeddings.

Task-Specific Instructions: Task-specific instructions (e.g., "<TextVQA>") are prepended to input queries, providing contextual cues for better task differentiation.

Unified Embedding Space: Both text and multimodal inputs are projected into a compact 128-dimensional embedding space, balancing computational efficiency with retrieval accuracy.