# Efficient and Effective In-context Demonstration Selection with Coreset

**Zihua Wang[1], Jiarui Wang[1], Haiyang Xu[2], Ming Yan[2], Fei Huang[2], Xu Yang[1], Xiu-Shen Wei[1], Siya Mi[3, 4], Yu Zhang[1]***

[1]School of Computer Science and Engineering and the Key Laboratory of New Generation Artificial Intelligence Technology and its Interdisciplinary Applications, Southeast University, Nanjing 210096, China.
[2]Tongyi Lab, Alibaba Group.
[3]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China.
[4]Purple Mountain Laboratories, Nanjing 210000, China.
zhang_yu@seu.edu.cn

## Abstract

In-context learning (ICL) has emerged as a powerful paradigm for Large Visual Language Models (LVLMs), enabling them to leverage a few examples directly from input contexts. However, the effectiveness of this approach is heavily reliant on the selection of demonstrations, a process that is NP-hard. Traditional strategies, including random, similarity-based sampling and infoscore-based sampling, often lead to inefficiencies or suboptimal performance, struggling to balance both efficiency and effectiveness in demonstration selection. In this paper, we propose a novel demonstration selection framework named Coreset-based Dual Retrieval (CoDR). We show that samples within a diverse subset achieve a higher expected mutual information. To implement this, we introduce a cluster-pruning method to construct a diverse coreset that aligns more effectively with the query while maintaining diversity. Additionally, we develop a dual retrieval mechanism that enhances the selection process by achieving global demonstration selection while preserving efficiency. Experimental results demonstrate that our method significantly improves the ICL performance compared to the existing strategies, providing a robust solution for effective and efficient demonstration selection.

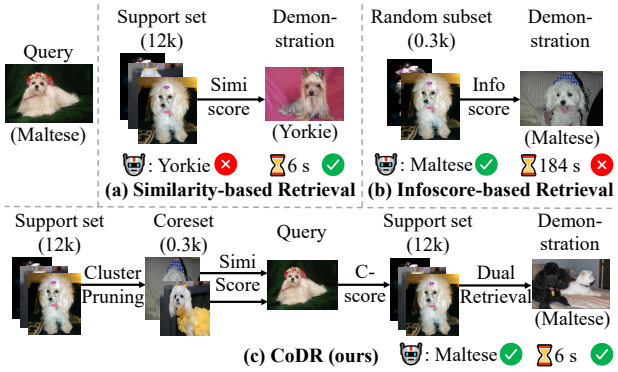


Figure 1: Similarity-based retrieval, infoscore-based retrieval and our proposed CoDR. CoDR introduces a dual retrieval mechanism: it first selects a coreset via cluster-pruning, then uses the similarity score between query and the samples in the coreset as a weighting coefficient, multiplying the pre-calculated C-score. The weighted scores guide the demonstration selection from the full support set.

## Introduction

In-context learning (ICL) is a groundbreaking approach that eliminates the need for conventional, data-intensive training methods. This innovative technique uses in-context demonstrations as prompts to enable few-shot learning, allowing models to perform tasks with minimal examples. Originally developed for Large Language Models (LLMs), ICL has recently gained traction in Large Vision-Language Models (LVLMs) (Awadalla et al. 2023; Laurençon et al. 2023) as well. It has shown remarkable effectiveness across various domains, including image captioning (IC), and visual question answering (VQA) tasks, underscoring its flexibility and potential across modalities.

One of the primary obstacles is to select appropriate demonstrations according to different queries. This selection process is crucial, as the performance of LLMs and LVLMs heavily depends on the quality and relevance of the provided demonstrations (Liu et al. 2022; Laurençon et al. 2024; Dong et al. 2024). However, selecting the appropriate demonstrations for each query is an NP-hard problem (Li and Qiu 2023). In other words, the quality of the selected demonstrations cannot be determined prior to the model execution of the inference process. Existing selection strategies can be likened to a form of speculation in the selection process. These strategies can be broadly categorized into two primary approaches. One focuses on efficiency, such as similarity-based selection (Liu et al. 2022). However, similarity-based approaches are prone to hallucination, as the outputs tend to overly adhere to the input demonstrations instead of adequately addressing the query (Liu et al. 2024). For example, as Figure 1 (a) shows, the image in the support set most similar to the query does not belong to the same dog breed, which is likely to lead to incorrect predictions when these samples are used as demonstrations. In instances where the model is presented with two highly similar images, and one is utilized as a demonstration, there is a propensity for the model to inaccurately clas-

*Yu Zhang is the corresponding author.

sify the query image and the demonstration as belonging to the same category, despite this not being the case. This issue is especially pronounced in tasks that require understanding of fine-grained image details, where superficial similarity fails to capture essential distinctions and leads to hallucinations (Wu and Yang 2024).

Another approach prioritizes effectiveness as the key property. For example, infoscore-based methods (Li and Qiu 2023) focus on selecting demonstrations that maximize the mutual information gain between the query and the samples in the support set. Infoscore is calculated by assessing each example's influence on the predicted probabilities of others, a process that requires traversing the entire support set, making it time-consuming. Consequently, existing infoscore-based approaches often limit their demonstration search to a subset of the support set, typically sampled randomly from the full support set (Li and Qiu 2023; Yang et al. 2024). Moreover, unlike traditional image–text retrieval methods (Cao et al. 2022, 2025; Xu et al. 2025), the ground truth for the query is unavailable, making it infeasible to compute the actual infoscore value. Instead, the sample with the highest probability is used to approximate the ground truth as an estimation (note that in the context of infoscore-based retrieval mentioned below, this refers to the estimated value). As shown in Figure 1 (b), infoscore-based methods limit the demonstration selection to a subset of 300 images which are randomly sampled from the support set. Despite this limitation, their computation time is significantly longer compared to similarity-based methods. This constraint results in sub-optimal selections, which ultimately reduces the overall performance by not fully exploring the diversity of the entire support set. Furthermore, the exponential growth of possible combinations further complicates the demonstration selection process, making it infeasible to enumerate all potential selections, especially as the number of demonstrations increases.

To address these challenges, we aim to develop a more effective and efficient demonstration selection strategy. First, to ensure the efficiency of the demonstration selection, we employ a subset of the support set, named coreset. Instead of being randomly sampled, the coreset is a subset that guarantees diversity. We mathematically prove that such a coreset leads to higher mutual information, enabling better transfer of knowledge from the demonstrations. To achieve diversity, we implement a cluster-pruning strategy to refine the supporting set into a coreset, where more complex clusters retain more samples after pruning.

Due to the absence of query ground-truth label, it becomes challenging to directly and qualitatively assess which demonstrations in the support set are most effective. However, the coreset, sampled from the support set, contains ground-truth labels. For each coreset sample, we treat it as a query and define the C-score to evaluate how well each sample in the support set fits the query. As Figure 1 (c) shows, for a given query, we first measure its similarity with each sample in the coreset. We then multiply this similarity score by the C-score of each coreset sample across the support set to obtain the accumulated scores. These accumulated scores provide a measure of the quality of each sample in the support set. Based on these scores, we select the top-k samples to form the k-shot demonstrations for the given query. As a result, it enhances performance by ensuring that the selected demonstrations are not limited to the coreset, thus having a broader and more effective selections while simultaneously maintaining efficiency. Compared to the similarity-based retrieval method, CoDR has almost the same retrieval time, but shows improvements of $7.01/5.18/5.78$ on the IC_CIDEr/VQA_ACC/FIC_ACC metrics, respectively. When compared with the infoscore-based method, CoDR demonstrates performance advantages of $2.40/1.34/1.79$, with nearly $4$ times faster retrieval speed.

Our contributions are concluded as follows:

- We demonstrate that a more diverse support set leads to higher mutual information expectation, enabling more effective knowledge transfer from the demonstrations.
- We introduce a coreset-based demonstration selection strategy that effectively balances diversity and relevance in visual-language in-context demonstration selection.
- Experimental results on image captioning, visual question answering, and fine-grained image classification tasks demonstrate that our approach significantly improves the performance of visual-text ICL on both the OpenFlamingo-v2 and Idefics-v2 models.

## Related Works

### Models with ICL Ability

ICL has emerged as a transformative paradigm in machine learning, with the potential to replace traditional few-shot training approaches by utilizing a few examples as the input (Mosbach et al. 2023). Inspired by GPT (Brown et al. 2020), ICL has found widespread application in textual generation tasks, *e.g.*, classification (Edwards and Camacho-Collados 2024), question answering (V, Bhattacharya, and Anand 2023; Peng et al. 2024b), table processing (Lu et al. 2025), and reading comprehension (Li et al. 2023a). Similarly, LVLMs like Flamingo (Alayrac et al. 2022), Idefics (Laurençon et al. 2023, 2024) leverage ICL to integrate image and text information, enabling richer Visual-Language (VL) tasks, *e.g.*, visual question answering (Li et al. 2024; Nie et al. 2024), image captioning (Yang et al. 2024, 2023; Ma et al. 2025), image classification (Zhang et al. 2024), and segmentation (Wang et al. 2023; Wen et al. 2025). Among the LVLMs with ICL capabilities, we employ OpenFlamingo-v2 (Awadalla et al. 2023) and Idefics-v2 (Laurençon et al. 2023) due to their robust ICL performance and open-source availability. This transparency ensures that neither model has been trained on the query or support sets, reducing potential bias.

### ICL Demonstration Strategy

To enhance the performance of ICL, researchers have explored a variety of strategies, primarily centered on optimizing the quality and structure of demonstrations. Early approaches emphasized improving demonstrations through formatting adjustments (Lu et al. 2022) or reordering (Kumar and Talukdar 2021), rather than selective curation of in-

stances. Recent research has increasingly emphasized selecting high-quality demonstrations to further optimize model performance. For example, metrics such as perplexity (Qin et al. 2024), BERTScore recall (Gupta, Gardner, and Singh 2023), and mutual information (Li and Qiu 2023; Liao, Zheng, and Yang 2022) are regarded as useful indicators of demonstration quality. However, some studies suggest that these metrics may be effective only for specific tasks (Wu et al. 2023; Li et al. 2023b). Demonstration selection strategies have also been found to be both data- and model-dependent (Peng et al. 2024a; Zhu, Ma, and Zhang 2025). Recent studies (Guo et al. 2023; Feng, Hong, and Zhang 2024) have explored the use of LLM-generated demonstrations to support ICL. While effective in some scenarios, these learned retrievers (Feng, Hong, and Zhang 2024; Shen et al. 2024; Askari, Poelitz, and Tang 2025) invoke LLMs for each query, leading to high computational overhead. Moreover, the quality of the retrieved demonstrations is highly sensitive to the initial prompt. Some methods employ calibration techniques, such as iterative refinement (Qin et al. 2024), while others enhance both the prompt and the demonstration (Yao et al. 2024; Jiang et al. 2025). A common limitation of these methods is their reliance on subsampling, as demonstrations are typically selected from a randomly drawn subset of the full support set, rather than the entire dataset. This constraint often arises due to computational or scalability challenges, yet it may lead to sub-optimal results.

## Coreset-based Dual Retrieval (CoDR)

In this section, we begin by mathematically demonstrating that a diverse coreset yields a higher expectation of mutual information. Subsequently, we describe a method for constructing a coreset with sufficient diversity. Lastly, we introduce dual retrieval, which leverages this diverse coreset to efficiently identify effective demonstrations from the entire support set using the C-score. The framework of CoDR is shown in Fig. 2.

### Preliminaries

Given a $n$-shot VL demonstrations $\langle x_i, y_i \rangle_{i=1}^{n}$, where $x_i$ represents the multimodal input (*e.g.*, an image and its associated question in VQA), and $y_i$ is the corresponding label (*e.g.*, the answer), along with a query input $x_q$, They are concatenated as $x' = \mathrm{concat}(\langle x_i, y_i \rangle_{i=1}^{n}, x_q)$ and fed into the LVLM, which then generates the predicted output $\hat{y}_q$.

### Coreset Construction

For a specific query $x_q$, selecting an appropriate set of demonstrations from the support set $\boldsymbol{S}$ is crucial to the ICL performance, yet remains a challenging task. Although existing strategies, such as similarity-based (Zebaze, Sagot, and Bawden 2025; An et al. 2023) and infoscore-based approaches (Li and Qiu 2023; Yang et al. 2024), offer practical heuristics, they often suffer from inefficiency in large-scale support sets and limited effectiveness in capturing semantic diversity relevant to the query. Our objective is to identify a compact subset $\boldsymbol{S}^* \subset \boldsymbol{S}$ that captures the most informative demonstrations for inference. We refer to this subset $\boldsymbol{S}^*$ as the coreset. Intuitively, a diverse subset offers broader coverage of the semantic space, potentially reducing the model's uncertainty when making predictions. To formalize this intuition, we employ mutual information as a metric to quantify the informativeness of the selected subset with respect to the query's target output. Next, we provide a theoretical justification for this assumption and introduce a practical method for constructing the coreset.

We define the mutual information to evaluate the input $x'$ and the model's target prediction $y_q$ as:

$$I(x', y_q) = H(y_q) - H(y_q \mid x'), \tag{1}$$

where $H$ represents the entropy. Since $H(y_q)$ is independent of the chosen demonstrations (assuming the same query $x_q$), maximizing $I(x', y_q)$ reduces to minimizing the conditional entropy $H(y_q \mid x')$.

We consider two subset selection scenarios: $\boldsymbol{S}_{\mathrm{div}}$, a diverse set of demonstrations sampled to maximize semantic variance, and $\boldsymbol{S}_{\mathrm{rand}}$, a randomly sampled set of demonstrations. To evaluate the quality of the selected subsets, we compare the expected conditional entropy between the query and the predicted label given the combined input: $E[H(x'_{\mathrm{div}}, y_q)]$ versus $E[H(x'_{\mathrm{rand}})]$, where $x'_{\mathrm{div}}$ and $x'_{\mathrm{rand}}$ denote the concatenation of the query $x_q$ with demonstrations from the $\boldsymbol{S}_{\mathrm{div}}$ and $\boldsymbol{S}_{\mathrm{rand}}$, respectively. The key idea is that the mutual information between a query and the support set increases when the support set provides better coverage of the query's latent class. To operationalize this notion of class coverage, we approximate the latent class of each sample. Since most of the VL tasks like VQA and image captioning lack category labels, we "assigned" categories (*e.g.*, through clustering the latent features).

Therefore, the conditional entropy (the second term of the Eq. 1) can be rewritten as:

$$H(y_q \mid x') = -\sum P(y_q \mid x') \log P(y_q \mid x'). \tag{2}$$

Since the function " $\cdot \log(\cdot)$" is convex, a diverse set—by evenly spreading probability mass across latent categories—leads to lower expected conditional entropy via Jensen's inequality. This directly implies that:

$$E[H(y_q \mid x'_{\mathrm{rand}})] = H\left[y_q \mid E(x'_{\mathrm{rand}})\right] \geq E[H(y_q \mid x'_{\mathrm{div}})], \tag{3}$$

where $E$ represents the mathematical expectation. Consequently, the subset with diverse demonstrations is expected to yield greater mutual information:

$$E[I(x'_{\mathrm{div}}, y_q)] \geq E[I(x'_{\mathrm{rand}}, y_q)]. \tag{4}$$

While applying greedy algorithms or K-means clustering can yield diverse subsets from the support set, these methods often fall short in ICL. This is because they prioritize more diverse sample features at the expense of considering the relevance and alignment between the demonstration samples and the query input. Thus, a more balanced approach that considers both diversity and relevance is essential to improve performance in ICL.

With the guidance of mutual information theory, we employ a cluster-based pruning method to obtain a subset $\boldsymbol{S}^*$

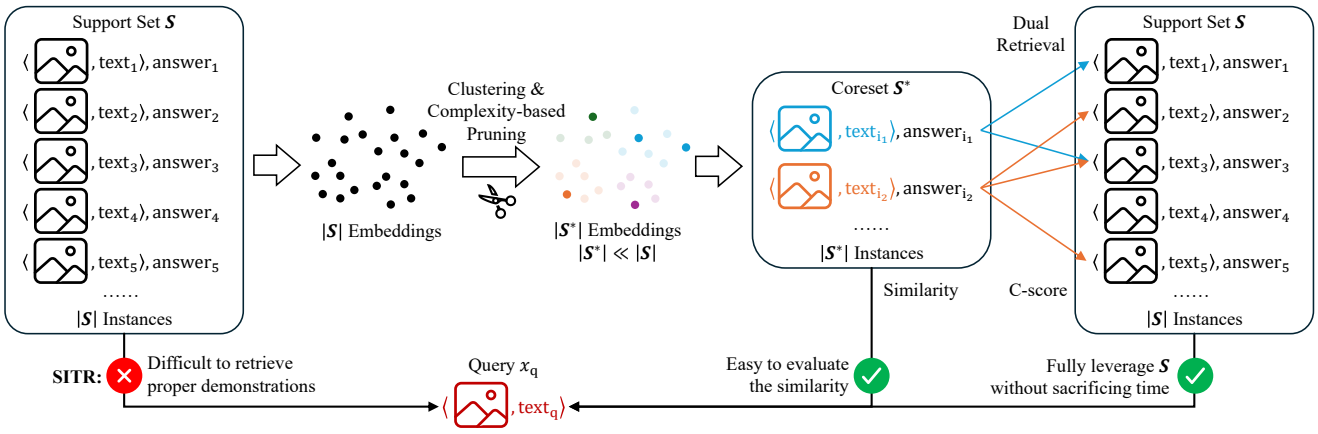


Figure 2: Architecture of CoDR. Our method construct a coreset $\boldsymbol{S}^*$ by clustering and pruning the support set $\boldsymbol{S}$. A dual retrieval module then performs global retrieval as follows: We precompute a C-score quantifying each support set sample's quality as a demonstration when queried with coreset samples. For an input query, we evaluate its similarity to each coreset sample, hypothesizing that higher similarity correlates with more similar demonstration selection. The final demonstration score is the product of this similarity and the precomputed C-score, enabling global retrieval across the entire support set.

of the support set $\boldsymbol{S}$. We notice that Density-Based Pruning (DBP) and self-supervised pruning (SSP-Pruning) are effective in pruning the training set in image classification (Abbas et al. 2024; Sorscher et al. 2022). They prune the cluster items based on similarity or diversity. Inspired by them, we use the LVLM to obtain the embedding of the support set $\boldsymbol{S}$, and then perform K-means clustering on this feature. However, we propose that pruning should be based on complexity $F$, which is defined as follows:

$$F_i = \overline{d(l,m)}_{l,m\in K_i} \cdot \overline{d(l,m)}_{l\in K_i, m\notin K_i}, \tag{5}$$

where $d$ represent the cosine distance between features. The first term of the formula denotes the intra-cluster distance for cluster $K_i$, measuring the average distance among points within the same cluster. The second term represents the average inter-cluster distance, calculating the mean distance between the points in cluster $K_i$ and the points in other clusters. Intuitively, a cluster with high intra-cluster distance requires more retained samples to preserve its internal variability. Meanwhile, high inter-cluster distance indicates uniqueness, necessitating sufficient representation to avoid losing distinct features. This dual perspective ensures that pruning retains informative clusters while reducing redundancy in over-represented ones. For clusters with higher complexity, more samples should be retained after pruning. Therefore, we preserve $\lfloor |K_i| \cdot \exp(F_i)/\sum_{j=1}^{n}\exp(F_j) \rfloor$ samples in cluster $K_i$. Since the coefficient $\exp(F_i)/\sum_{j=1}^{n}\exp(F_j)$ is strictly less than 1, this ensures the pruning process. We obtain the coreset $\boldsymbol{S}^*$ after cluster-pruning.

## Dual Retrieval

After obtaining the coreset $\boldsymbol{S}^*$, which ensures diversity, we note that due to the reduced sample size compared to the entire support set $\boldsymbol{S}$ after multiple pruning steps, selecting samples exclusively from the coreset may not yield a globally optimal solution. While the coreset effectively captures the most representative samples from the support set, it is limited in its ability to fully leverage the variety of available demonstrations on larger datasets. Therefore, we aim to use the coreset as a guide to retrieve samples globally.

We define a C-score $C(q|s)$ to quantify how well a demonstration $s \in \boldsymbol{S}$ supports the query $q = \langle x_q, y_q \rangle$.

$$C(q|s) = \text{Metric}(\text{LVLM}(\text{concat}(s, x_q)), y_q), \tag{6}$$

where $\text{Metric}(\hat{y}_q, y_q)$ is a metric function determined by the task, with larger values indicating more accurate predictions $\hat{y}_q$ (e.g., CIDEr for image captioning and ACC for VQA). However, with $q \in \boldsymbol{S_q}$, $y_q$ is unknown during inference.

Given the premise of ICL, where $x_q$ and $x$ belong to the same domain, we assume that $\boldsymbol{S_q}$ (query set) and $\boldsymbol{S}$ (support set) follow the same distribution. While $\boldsymbol{S_q}$ is unannotated, $\boldsymbol{S}$ contains ground-truth labels, allowing us to leverage labeled samples from $\boldsymbol{S}$ for ICL purposes. In this way, $C(q^*|s)$ can be easily obtained, where $s \in \boldsymbol{S}$ and $q^* \in \boldsymbol{S^*}$. Due to the characteristics of the clustering, the retained samples in $\boldsymbol{S}^*$ ensure diversity. Therefore, the C-score $C(q^*|s)$ can be regarded as kernels to represent the samples in $\boldsymbol{S_q}$:

$$\hat{C}(q|s) = \frac{\sum_{q^*\in \boldsymbol{S^*}}[d(q,q^*)\cdot C(q^*|s)]}{\sum_{q^*\in \boldsymbol{S^*}} d(q,q^*)}. \tag{7}$$

Based on the cosine similarity between the query sample $q$ and the coreset samples in $\boldsymbol{S}^*$, the contribution of each kernel can be calculated and thus obtain the score $\hat{C}(q|s)$. This process generates a score set indicating how effectively each demonstration $s \in \boldsymbol{S}$ supports the query sample $q \in \boldsymbol{S_q}$. Importantly, the C-score matrix $C(q^*|s)$ for each coreset sample $q$ can be precomputed once per task and reused across queries. Unlike traditional similarity-based retrieval

methods, which compute pairwise similarities across the entire support set, our approach merely calculates similarities between the input query and the coreset, which enables efficient large-scale deployment on demonstration selection.

From the perspective of ensemble learning, each 1-shot demonstration can be regarded as a weak learner. Aggregating multiple demonstrations is akin to combining weak learners to enhance overall performance. In the $n$-shot setting, selecting diverse and high-quality examples provides complementary information, helping to reduce the risk of overfitting to any single instance. Therefore, we construct the $n$-shot prompt by selecting the top-$n$ demonstrations based on their relevance scores, ensuring that the final representation is both informative and reliable for the query.

# Experiments

## Tasks and Datasets

Our approach is evaluated on 3 tasks: Image Captioning (IC), Visual Question Answering (VQA), and Fine-grained Image Classification (FIC). We evaluate on two mainstream open-source LVLM frameworks: OpenFlamingo-v2 (OFv2) (Awadalla et al. 2023) and Idefics-v2 (IDEv2) (Laurençon et al. 2023). We employ MSCOCO (Lin et al. 2014), VQAv2 (Antol et al. 2015), StanfordDogs (Khosla et al. 2011) for IC, VQA, FIC tasks, respectively.

## Implementation Details

**Shots number $n$.** Following the principle of diminishing marginal returns, our evaluation primarily focuses on 1-/2-/4-shot. Our ablation studies are conducted using a 4-shot configuration.

**Implementing ICL.** We set the maximum number of generated tokens to $16/5/5$ for IC/VQA/FIC tasks, respectively. For fluent sentence generation, we employ beam search with a beam size of 3. The features used for clustering and similarity-based retrieval are extracted by the encoder of the LVLMs (OFv2 and IDEv2). The inference process is conducted on an Nvidia A6000 GPU.

**Metrics.** To evaluate the effectiveness of VQA and FIC, we calculate accuracy scores (VQA_ACC and FIC_ACC), with higher scores indicating better performance of the VQA model. For the IC task, we observe that while the sentences generated through ICL may yield high scores, they can also contain factual errors. Therefore, in addition to traditional metrics CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), we further evaluate hallucination metrics, CHAIRs and CHAIRi (Rohrbach et al. 2018), as hallucinations are considered one of the severe challenges in LLM generations (Huang et al. 2025).

## Results and Analyses

**Methods for Comparison.** (1) **Random-Global (RG)**: RG randomly samples demonstrations from the whole support set $\boldsymbol{S}$. (2) **Random-Local (RL)**: RL first randomly samples a subset of the support set $\boldsymbol{S}$, then the demonstrations are randomly selected from this subset. (3) **Similarity-based Image-Text Retrieval-global (SITR)**: Benefiting from the capabilities of CLIP, the similarity between the features of image-text demonstrations and the query can be aligned in the feature domain. Searching for demonstrations in support set $S$ that are more similar to the query has been recognized as a potentially useful metric. (4) **Similarity-based Image-Text Retrieval-Local (SITRL)**: Similar to the RL strategy, we first randomly sample a subset of the support set and then conduct a search based on image-text similarity within this subset. (5) **Infoscore-based Retrieval-Local (IRL)**: Infoscore (Yang et al. 2024; Li and Qiu 2023) can be considered a greedy search strategy due to its high complexity and unsuitability for global search. Consequently, we utilize Infoscore to search for demonstrations within a randomly selected subset of 300 samples.

**Main Results.** The results of the various demonstration selecting strategies are listed in Table 1 (IC) and Table 2 (VQA and FIC). In IC task using the OFv2 model, our method significantly enhances the CIDEr while reducing the CHAIRi and CHAIRs compared to the random algorithms RG and RL. It outperforms both SITR and SITRL methods, particularly excelling in the CHAIR metrics, which indicates that CoDR reduces the hallucinations. In the IDEv2 model, we observe a similar trend, with our method attaining a CIDEr score of $89.57/115.06/117.54$ in the 1-/2-/4-shot scenarios. While other methods like SITR and IRL struggle to outperform RG on IDEv2 (likely due to the model's inherent complexity or feature distribution), CoDR's consistent performance highlights its robustness across different model architectures. This indicates that CoDR effectively balances relevance and diversity.

Table 1 also lists the average inference time required to process each query. Our method achieves average processing times of 3.95 seconds (OFv2) and 3.77 seconds (IDEv2), closely matching SITRL's performance (3.81 and 3.74 seconds, respectively). Compared to SITR, which is also a global search method, our method demonstrates a 1.42 second speed improvement for OFv2 and a 1.08 second improvement for IDEv2. These results demonstrate that CoDR achieves significantly better computational efficiency than conventional similarity-based retrieval methods.

Our method achieves near state-of-the-art performance in both VQA and FIC tasks compared to other strategies. In the IDEv2 4-shot VQA setting, CoDR attains an accuracy of 67.01, closely matching IDEv2-IRL (67.04), with a marginal difference of only 0.03. For the OFv2 model, our method demonstrates a significant improvement, achieving the highest VQA accuracy of $50.03/52.53/53.49$ under the 1-/2-/4-shot setting, compared to other strategies such as SITR and IRL, which yield accuracies of $40.17/43.58/47.50$ and $46.66/50.83/52.15$, respectively. In the FIC task, our approach outperforms the baseline methods, reaching $33.43/39.38/41.39$ for the 1-/2-/4-shot setting. When using the IDEv2 model, the performance advantage in VQA for our method is relatively smaller compared to FIC. Similarly, other approaches such as SITR and IRL do

| Model-Method | CIDEr ↑ | | | CHAIRs ↓ | | | CHAIRi ↓ | | | Time(s)↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1-shot | 2-shot | 4-shot | 1-shot | 2-shot | 4-shot | 1-shot | 2-shot | 4-shot | |
| OFv2-RG | 74.33 | 85.65 | 95.20 | 8.8 | 6.5 | 6.3 | 7.9 | 6.3 | 6.3 | - |
| OFv2-RL | 74.23 | 85.40 | 95.14 | 8.5 | 6.5 | 6.2 | 7.9 | 6.3 | 6.3 | - |
| OFv2-SITR | 73.10 | 82.99 | 97.29 | 19.32 | 8.3 | 6.2 | 14.35 | 6.6 | 5.9 | 5.37 |
| OFv2-SITRL | 73.40 | 82.94 | 97.22 | 17.10 | 8.9 | 6.2 | 13.1 | 6.7 | 6.0 | **3.81** |
| OFv2-IRL | 74.95 | 89.06 | 101.83 | 5.9 | 5.2 | 4.3 | 5.4 | 3.9 | 3.7 | 13.78 |
| OFv2-CoDR | **79.68** | **100.80** | **104.23** | **5.9** | **5.1** | **4.2** | **5.3** | **3.4** | **3.3** | 3.95 |
| IDEv2-RG | 75.59 | 100.33 | 112.70 | 7.4 | 4.9 | 5.2 | 7.3 | 3.9 | 3.9 | - |
| IDEv2-RL | 75.43 | 99.67 | 109.30 | 7.4 | 5.0 | 5.2 | 7.3 | 3.9 | 3.9 | - |
| IDEv2-SITR | 74.55 | 92.91 | 102.71 | 10.7 | 6.9 | 6.6 | 8.2 | 5.0 | 4.8 | 4.85 |
| IDEv2-SITRL | 74.59 | 92.70 | 101.76 | 13.9 | 6.8 | 5.3 | 7.9 | 5.0 | 4.5 | **3.74** |
| IDEv2-IRL | 79.61 | 105.01 | 111.29 | 5.6 | 5.2 | 4.8 | 5.6 | 5.1 | 4.6 | 13.35 |
| IDEv2-CoDR | **89.57** | **115.06** | **117.54** | **4.9** | **3.7** | **4.3** | **3.6** | **2.9** | **3.1** | 3.77 |

Table 1: 1-/2-/4-shot IC performance with various demonstration selection strategies.

| Model-Method | VQA_ACC ↑ | FIC_ACC ↑ |
|---|---|---|
| OFv2-RG | 41.97/45.92/48.95 | 17.83/28.51/34.36 |
| OFv2-RL | 42.12/46.07/48.99 | 18.56/29.29/33.16 |
| OFv2-SITR | 40.17/43.58/47.50 | 29.33/30.25/36.98 |
| OFv2-SITRL | 42.09/44.72/48.31 | 27.13/32.84/35.61 |
| OFv2-IRL | 46.66/50.83/52.15 | 29.88/33.79/39.60 |
| OFv2-CoDR | **50.03/52.53/53.49** | **33.43/39.38/41.39** |
| IDEv2-RG | 60.24/63.20/66.62 | 43.94/43.92/44.85 |
| IDEv2-RL | 60.03/64.23/66.71 | 41.70/41.97/43.02 |
| IDEv2-SITR | 58.65/62.50/66.93 | 47.14/53.10/54.42 |
| IDEv2-SITRL | 59.17/62.20/66.78 | 47.22/52.89/54.00 |
| IDEv2-IRL | 60.65/64.46/**67.04** | 43.85/49.88/50.17 |
| IDEv2-CoDR | **61.16/64.53**/67.01 | **52.74/58.64/66.28** |

Table 2: 1-/2-/4-shot VQA and FIC performance with various demonstration selection strategies.

not significantly outperform RG, likely due to limitations of the IDEv2 architecture. Nevertheless, CoDR maintains the strongest overall performance.

The performance of RG and RL is relatively close because the subset selected by the RL strategy is randomly sampled from the support set, which leads to a distribution that is expected to be similar to that of the support set, resulting in comparable performance. This similar pattern is also observed in SITR and SITRL. Furthermore, we find that similarity-based methods do not outperform random sampling in VQA, whereas they remain a relatively effective strategy in FIC. In the IC task, while similarity-based methods yield higher CIDEr scores, they also introduce excessive hallucinations, resulting in poorer CHAIRi and CHAIRs metrics. This occurs because the model tends to focus on the similar demonstrations while neglecting the query sample. In FIC, finding the most similar image may lead to selecting images from the same category, which are then adopted as the output by the model. However, due to the differences in labeling information between VQA and IC, this approach can inadvertently introduce hallucinations. In contrast, our method employs clustering to ensure the diversity of samples during the coreset selection process. Subsequently, the dual retrieval process effectively identifies the appropriate demonstrations based on the input query, thereby enhancing the performance of IC, VQA, and FIC tasks.

| Model-Task | Rand | Centroids | **CoDR** |
|---|---|---|---|
| OFv2-IC | 80.12 | 101.09 | **104.23** |
| OFv2-VQA | 46.86 | 52.24 | **53.49** |
| OFv2-FIC | 33.64 | 40.97 | **41.39** |
| IDEv2-IC | 113.93 | 117.22 | **117.54** |
| IDEv2-VQA | 64.37 | 66.38 | **67.01** |
| IDEv2-FIC | 50.25 | 63.56 | **66.28** |

Table 3: Ablations on different coreset selection methods.

| Model-Task | Rand | Simi | Div | **CoDR** |
|---|---|---|---|---|
| OFv2-IC | 78.30 | 97.83 | 79.05 | **104.23** |
| OFv2-VQA | 43.58 | 44.18 | 45.83 | **53.49** |
| OFv2-FIC | 29.53 | 36.47 | 24.44 | **41.39** |
| IDEv2-IC | 117.20 | 113.20 | 115.13 | **117.54** |
| IDEv2-VQA | 66.83 | 66.33 | 66.80 | **67.01** |
| IDEv2-FIC | 42.09 | 52.50 | 49.28 | **66.28** |

Table 4: Ablations on the effectiveness of dual retrieval.

## Ablations

**Coreset construction.** We evaluate the quality of coreset obtained using different methods. As shown in Table 3, we compare three approaches: random sampling (Rand), selecting samples closest to the cluster centroids (Centroids), and our proposed cluster-pruning method (CoDR). The results demonstrate that CoDR consistently achieves the best performance, which indicates that CoDR is effective in constructing a more representative and informative coreset. Moreover, CoDR also consistently outperforms the cluster-center baseline, which uses the centroids directly as coreset samples. This method captures coarse semantic structure, but fails to ensure optimal representativeness or diversity. CoDR, by contrast, applies additional pruning or normalization to select more globally informative samples. For

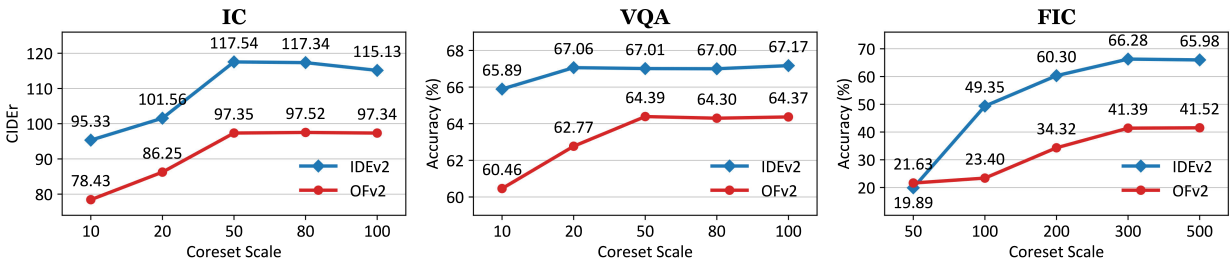


Figure 3: IC, VQA and FIC performance with different coreset scales.

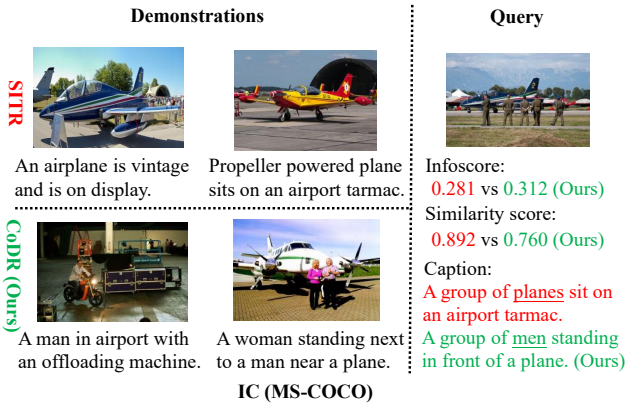


Figure 4: Visualization of SITR and CoDR on IC.

example, CoDR improves over the cluster-center method by +3.14 on OFv2-IC, +2.72 on IDEv2-FIC, and +1.13 on OFv2-VQA, showing that fine-grained coreset selection beyond simple clustering is crucial for performance.

**Coreset Scale.** The coreset size serves as the primary hyperparameter in the cluster-pruning process. We systematically investigate the performance of IC, VQA, and FIC under varying coreset scales. While larger coreset sizes expand the candidate sample pool for selection, this expansion also increases the probability of encountering redundant samples with high similarity, consequently elevating the computational overhead required to identify optimal demonstrations. The results in Fig. 3 indicate that a preferred coreset scale is approximately 50 for IC, 50 for VQA, and 300 for FIC. Due to the absence of category labeling for image-text pairs in VQA and IC tasks, a coreset scale of 50 samples is sufficient for effective sample selection. In contrast, FIC task has 120 categories (for the Stanford Dogs dataset), and when the coreset scale is less than 120, certain categories are absent, leading to performance degradation. However, with a coreset scale above 300, there is a higher likelihood that all 120 categories are represented in the coreset at least once, which indeed occurs in practice.

**Dual Retrieval.** We evaluate the effect of dual retrieval by comparing it with direct selection from the coreset using three alternatives: random (Rand), similarity-based (Simi), and diversity-based (Div) sampling, with results presented in Table 4. As shown in the table, the dual retrieval approach consistently outperforms random selection (Rand) from the coreset, as it adopts a more global selection strategy that provides richer candidate examples, whereas the coreset itself remains relatively small. We further observe that similarity-based sampling within the coreset generally surpasses random selection, since the coreset construction already ensures a certain degree of diversity. In addition, random sampling from the coreset yields better results than the RG strategy, which randomly selects from the full support set, and the RL strategy, which samples from a randomly chosen subset of it. Finally, diversity-based sampling consistently underperforms CoDR, particularly on OFv2-FIC, which demands strong semantic alignment. This indicates that maximizing diversity alone may introduce irrelevant examples, highlighting the necessity of jointly considering both relevance and diversity as in CoDR.

## Qualitative Results

Fig. 4 visualizes a representative query-demonstration pair, highlighting CoDR's advantages over SITR. Unlike SITR, CoDR selects diversified demonstrations that reduce hallucination while preserving relevance. For example, SITR erroneously predicts "airplanes" instead of "men", a clear error propagation from its second demonstration that deviates from ground-truth. This misalignment reveals SITR's over-reliance on demonstration content, undermining query fidelity. In contrast, CoDR yields more faithful predictions with more diverse and representative demonstrations.

## Conclusion

In this paper, we propose a novel demonstration selection approach, named CoDR. CoDR first prunes the support set into a compact yet diverse coreset, ensuring representational coverage across different data modes. With the guidance of mutual information, this diverse coreset facilitates richer information transfer from varied demonstrations, thereby improving generalization. Subsequently, our dual retrieval mechanism leverages this coreset as a guide to efficiently identify task-relevant examples from the full support set. Experimental evaluations validate the effectiveness of our approach, showing significant performance improvements over traditional selection strategies. In the future, we plan to extend CoDR to knowledge-intensive tasks, where identifying demonstrations that encode both factual and relational knowledge will be crucial in complex multimodal scenarios.

## Acknowledgments

This work was supported by National Key R&D Program of China (2021YFA1001100), National Natural Science Foundation of China under Grant (62576089, 62522602, 62576091), and the Fundamental Research Funds for the Central Universities (2242025K30024, 4009002401). This research work is supported by the Big Data Computing Center of Southeast University.

## References

Abbas, A.; Rusak, E.; Tirumala, K.; Brendel, W.; Chaudhuri, K.; and Morcos, A. S. 2024. Effective pruning of web-scale datasets based on complexity of concept clusters. In *ICLR 2024*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J. L.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Bińkowski, M. a.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Proceedings of NeurIPS 2022*, volume 35, 23716–23736. Curran Associates, Inc.

An, S.; Lin, Z.; Fu, Q.; Chen, B.; Zheng, N.; Lou, J.-G.; and Zhang, D. 2023. How Do In-Context Examples Affect Compositional Generalization? In *Proceedings of the ACL 2023*, 11027–11052. Association for Computational Linguistics.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of ICCV 2015*.

Askari, A.; Poelitz, C.; and Tang, X. 2025. Magic: Generating self-correction guideline for in-context text-to-sql. In *Proceedings of AAAI 2025*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Proceedings of NeurIPS 2020*, volume 33, 1877–1901. Curran Associates, Inc.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text Retrieval: A Survey on Recent Research and Development. arXiv:2203.14713.

Cao, M.; Zhou, X.; Jiang, D.; Du, B.; Ye, M.; and Zhang, M. 2025. Multilingual Text-to-Image Person Retrieval via Bidirectional Relation Reasoning and Aligning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A Survey on In-context Learning. In *Proceedings of EMNLP 2024*, 1107–1128. Association for Computational Linguistics.

Edwards, A.; and Camacho-Collados, J. 2024. Language Models for Text Classification: Is In-Context Learning Enough? In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10058–10072. Torino, Italia: ELRA and ICCL.

Feng, L.; Hong, M.; and Zhang, C. J. 2024. Auto-Demo Prompting: Leveraging Generated Outputs as Demonstrations for Enhanced Batch Prompting. arXiv:2410.01724.

Guo, J.; Li, J.; Li, D.; Tiong, A. M. H.; Li, B.; Tao, D.; and Hoi, S. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of CVPR 2023*.

Gupta, S.; Gardner, M.; and Singh, S. 2023. Coverage-based Example Selection for In-Context Learning. In *Findings of EMNLP 2023*, 13924–13950.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jiang, Y.; Fu, J.; Hao, C.; Hu, X.; Peng, Y.; Geng, X.; and Yang, X. 2025. Mimic In-Context Learning for Multimodal Tasks. In *Proceedings of CVPR 2025*, 29825–29835.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs. In *Proceedings of the CVPR 2011 Workshop on Fine-Grained Visual Categorization (FGVC)*, 1–6. IEEE.

Kumar, S.; and Talukdar, P. 2021. Reordering Examples Helps during Priming-based Few-Shot Learning. In *Findings of ACL-IJCNLP 2021*, 4507–4518. Online: Association for Computational Linguistics.

Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; Cord, M.; and Sanh, V. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Proceedings of NeurIPS 2023*, volume 36, 71683–71702. Curran Associates, Inc.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, 87874–87907. Curran Associates, Inc.

Li, L.; Peng, J.; Chen, H.; Gao, C.; and Yang, X. 2024. How to Configure Good In-Context Sequence for Visual Question Answering. In *Proceedings of CVPR 2024*, 26710–26720.

Li, L.; Zhang, H.; Li, C.; You, H.; and Cui, W. 2023a. Evaluation on ChatGPT for Chinese Language Understanding. *Data Intelligence*, 5(4): 885–903.

Li, X.; Lv, K.; Yan, H.; Lin, T.; Zhu, W.; Ni, Y.; Xie, G.; Wang, X.; and Qiu, X. 2023b. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of ACL 2023*, 4644–4668.

Li, X.; and Qiu, X. 2023. Finding Support Examples for In-Context Learning. In *Findings of EMNLP 2023*, 6219–6235.

Liao, C.; Zheng, Y.; and Yang, Z. 2022. Zero-Label Prompt Selection. arXiv:2211.04668.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022)*, 100–114. Dublin, Ireland and Online: Association for Computational Linguistics.

Lu, W.; Zhang, J.; Fan, J.; Fu, Z.; Chen, Y.; and Du, X. 2025. Large language model for table processing: A survey. *Frontiers of Computer Science*, 19(2): 192350.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity.

Ma, Z.; Wang, C.-X.; Ouyang, Y.-W.; Zhao, F.; Zhang, J.-B.; Huang, S.-J.; and Chen, J.-J. 2025. Hacking Reference-Free Image Captioning Metrics. *Frontiers of Computer Science*.

Mosbach, M.; Pimentel, T.; Ravfogel, S.; Klakow, D.; and Elazar, Y. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of ACL 2023*, 12284–12314. Toronto, Canada: Association for Computational Linguistics.

Nie, Z.; Zhang, R.; Wang, Z.; and Liu, X. 2024. Code-Style In-Context Learning for Knowledge-Based Question Answering. In *Proceedings of AAAI 2024*, 18833–18841.

Peng, K.; Ding, L.; Yuan, Y.; Liu, X.; Zhang, M.; Ouyang, Y.; and Tao, D. 2024a. Revisiting Demonstration Selection Strategies in In-Context Learning. In *Proceedings of ACL 2024*, 9090–9101. Association for Computational Linguistics.

Peng, Y.; Hao, C.; Yang, X.; Peng, J.; Hu, X.; and Geng, X. 2024b. Live: Learnable in-context vector for visual question answering. In *Proceedings of NeurIPS 2024*, volume 37, 9773–9800.

Qin, C.; Zhang, A.; Chen, C.; Dagar, A.; and Ye, W. 2024. In-Context Learning with Iterative Demonstration Selection. In *Findings of EMNLP 2024*, 7441–7455.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. Brussels, Belgium: Association for Computational Linguistics.

Shen, T.; Long, G.; Geng, X.; Tao, C.; Lei, Y.; Zhou, T.; Blumenstein, M.; and Jiang, D. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Proceedings of ACL 2024*, 15933–15946.

Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; and Morcos, A. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Proceedings of NeurIPS 2022*, volume 35, 19523–19536. Curran Associates, Inc.

V, V.; Bhattacharya, S.; and Anand, A. 2023. In-Context Ability Transfer for Question Decomposition in Complex QA. arXiv:2310.18371.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of CVPR 2015*.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. SegGPT: Segmenting Everything In Context. arXiv:2304.03284.

Wen, C.; Zhang, Y.; Fan, J.; Zhu, H.; Wei, X.-S.; Wang, Y.; Kou, Z.; and Sun, S. 2025. Object-level Correlation for Few-Shot Segmentation. In *Proceedings of ICCV 2025*, 23689–23699.

Wu, Y.; and Yang, X. 2024. A glance at in-context learning. *Frontiers of Computer Science*, 18(5): 185347.

Wu, Z.; Wang, Y.; Ye, J.; and Kong, L. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In *Proceedings of ACL 2023*, 1423–1436. Association for Computational Linguistics.

Xu, Y.; Wu, M.; Guo, Z.; Cao, M.; Ye, M.; and Laaksonen, J. 2025. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. *Visual Intelligence*, 3(1): 1–13.

Yang, X.; Peng, Y.; Ma, H.; Xu, S.; Zhang, C.; Han, Y.; and Zhang, H. 2024. Lever LM: Configuring In-Context Sequence to Lever Large Vision Language Models. In *Proceedings of NeurIPS 2024*, volume 37, 100341–100368. Curran Associates, Inc.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2023. Exploring Diverse In-Context Configurations for Image Captioning. In *Proceedings of NeurIPS 2023*, volume 36, 40924–40943. Curran Associates, Inc.

Yao, B.; Chen, G.; Zou, R.; Lu, Y.; Li, J.; Zhang, S.; Sang, Y.; Liu, S.; Hendler, J.; and Wang, D. 2024. More Samples or More Prompts? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling. In *Findings of NAACL 2024*, 1772–1790.

Zebaze, A. R.; Sagot, B.; and Bawden, R. 2025. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of NAACL 2025*, 1222–1252. Association for Computational Linguistics.

Zhang, J.; Wang, B.; Li, L.; Nakashima, Y.; and Nagahara, H. 2024. Instruct Me More! Random Prompting for Visual In-Context Learning. In *Proceedings of WACV 2024*, 2597–2606.

Zhu, Y.; Ma, H.; and Zhang, C. 2025. Exploring Task-Level Optimal Prompts for Visual In-Context Learning. In *Proceedings of AAAI 2025*.