ACD-RETRIEVAL: TOWARDS MULTIMODAL KNOWLEDGE TRANSFER VIA ALIGNED CROSS-DOMAIN DEMONSTRATION RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have achieved remarkable progress in complex reasoning tasks, yet existing reasoning enhancement methods still rely on high-quality in-domain annotations or expert demonstrations. We observe that, even when in-domain demonstrations are unavailable, demonstrations from other domains can provide transferable reasoning knowledge for multimodal ICL. However, existing multimodal retrieval methods mainly rely on simple embedding similarity, which may not effectively reflect how well a cross-domain demonstration transfers to the target reasoning task, potentially leading to unstable retrieval. To address this challenge, we propose **ACD-Retrieval**, a unified framework for multimodal cross-domain ICL. ACD-Retrieval performs domain-wise centering and SVD-based subspace projection to obtain more reliable cross-domain similarity estimation, while a self-selection mechanism suppresses negative transfer during inference. Extensive experiments demonstrate that our method consistently achieves positive transfer across nine MLLMs, while exhibiting strong generalization across diverse multimodal cross-domain settings.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently achieved remarkable progress in complex reasoning tasks, including visual understanding, mathematical reasoning, and logical reasoning (Bai et al., 2025; Team et al., 2025; Zhu et al., 2025b; Wang et al., 2024a). Despite these advances, current MLLMs still heavily rely on large amounts of manually annotated data or carefully curated expert demonstrations to achieve strong reasoning performance, such as supervised reasoning tuning with chain-of-thought annotations (Dong et al., 2025; Shao et al., 2024; Zhang et al., 2025; Ou & Yin, 2025).

In many real-world scenarios, obtaining high-quality in-domain annotations is expensive, scarce, or even entirely infeasible. To alleviate the dependence on large-scale annotations, in-context learning (ICL) (Brown et al., 2020; Radford et al., 2019; Dong et al., 2024; Mueller et al., 2024) provides a promising alternative by enabling models to perform reasoning from only a small number of demonstrations without parameter updates. Yet in many practical settings, even a small amount of high-quality in-domain demonstrations may be unavailable. This motivates the study of cross-domain ICL, where demonstrations from other domains are leveraged to provide transferable reasoning knowledge.

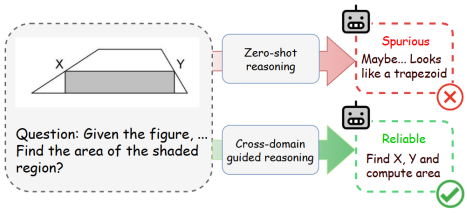


Figure 1: Cross-domain demonstrations improve reasoning reliability in multimodal ICL. Zero-shot reasoning often produces spurious predictions (top), while cross-domain guided reasoning leads to more reliable solutions (bottom).

Currently, multimodal cross-domain ICL remains largely underexplored. Although existing multimodal models provide effective representations for vision-language matching (Radford et al., 2021; Li et al., 2023; Microsoft et al., 2025; Agrawal et al., 2024; Li et al., 2026), most multimodal retrieval methods still rely on direct embedding similarity

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Baldassini et al., 2024), which may not capture transferable reasoning structures across domains. This leads to two major challenges for multimodal cross-domain ICL. The first is **cross-domain noise**, where domain discrepancies cause retrieved demonstrations to contain irrelevant or misleading information that interferes with reasoning. The second is **transfer uncertainty**, since cross-domain demonstrations do not consistently improve performance and may instead produce unstable or harmful reasoning behaviors.

To address these challenges, we propose **ACD-Retrieval**, a unified framework for multimodal cross-domain knowledge transfer. First, we encode multimodal queries using a unified vision-language embedding model and perform domain-wise centering to reduce domain-specific bias. Second, we introduce a Singular Value Decomposition (SVD)-based subspace projection mechanism that maps cross-domain representations into a shared low-dimensional space for more robust similarity estimation. Third, we design a self-selection mechanism that dynamically evaluates reasoning trajectories before and after incorporating demonstrations, thereby suppressing negative transfer. As illustrated in Figure 1, our framework improves reasoning reliability by leveraging properly selected cross-domain demonstrations. We conduct extensive experiments across 16 multimodal cross-domain transfer settings on MLLMs of different scales. Experimental results demonstrate that our method consistently achieves significant positive transfer and exhibits strong robustness and generalization abilities across different model families and scales. Our contributions are summarized as follows:

- We demonstrate that cross-domain demonstrations can provide transferable reasoning knowledge for multimodal reasoning, even when high-quality in-domain demonstrations are unavailable.
- We propose **ACD-Retrieval**, a unified framework that improves cross-domain demonstration retrieval through domain-wise centering and SVD-based subspace projection while mitigating negative transfer via self-selection.
- To the best of our knowledge, this is the first work to demonstrate effective multimodal cross-domain knowledge transfer for reasoning across diverse models and domains.

2 PRELIMINARIES

2.1 MULTIMODAL IN-CONTEXT LEARNING

ICL enables MLLMs to perform new tasks by conditioning on a small set of demonstrations without updating model parameters. Each demonstration is represented as $e_i = (x_i, y_i)$, where x_i denotes a multimodal input (e.g., an image-text pair) and y_i is the corresponding output. Given a demonstration set:

$$\mathcal{E} = \{(x_1, y_1), \ldots, (x_K, y_K)\},$$

and a test query x_{test}, the model prediction is obtained by sampling from the conditional distribution:

$$y_{\text{test}} \sim P_{\text{MLLM}}\left(y \mid x_1, y_1, \ldots, x_K, y_K, x_{\text{test}}\right).$$

2.2 DOMAIN ADAPTATION

Cross-domain ICL can also be viewed through the lens of domain adaptation, where the source demonstrations and target queries are sampled from related but non-identical distributions (Pan & Yang, 2010; Farahani et al., 2020). In this setting, the model must transfer task knowledge from a labeled source domain to an unseen target domain without parameter updates.

Formally, let the source and target domains be defined as:

$$\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}, \qquad \mathcal{D}_T = \{x_j^T\}_{j=1}^{N_T}.$$

A fundamental challenge is that their marginal input distributions satisfy a distinct discrepancy:

$$P_S(x) \neq P_T(x).$$

Such distribution shifts may arise from differences in visual appearance, language style, task format, or semantic composition across domains. As a result, demonstrations that are highly relevant in the source domain may become suboptimal when directly transferred to the target domain.

From a representation learning perspective, effective adaptation requires learning embeddings that preserve task-relevant semantics while reducing domain-specific variance. Given an encoder $f(\cdot)$, an ideal feature space should satisfy:

$$f(x^S) \approx f(x^T)$$

for semantically similar instances across domains, even when their raw inputs differ substantially. Therefore, improving cross-domain retrieval quality can be interpreted as reducing the discrepancy between source and target representations, enabling more informative demonstrations to be selected for in-context inference.

2.3 Cross-Domain Retrieval for In-Context Learning

In cross-domain ICL, demonstrations are retrieved from the source domain $\mathcal{D}_S$, where $x^S \sim P_S(x)$, while the test query x_{test} is drawn from the target domain $\mathcal{D}_T$, where $x^T \sim P_T(x)$. The goal of retrieval is to select a subset of demonstrations $\mathcal{E} \subset \mathcal{D}_S$ to construct the in-context prompt.

A common practice is to compute the similarity in the embedding space via the inner product:

$$\text{sim}(x^T, x^S) = \langle f(x^T), f(x^S) \rangle,$$

where $f(\cdot)$ denotes a multimodal encoder. This formulation enables the model to leverage contextual examples for rapid task adaptation, without requiring any parameter updates or additional training.

3 Method

Existing multimodal in-context learning methods mainly rely on direct embedding similarity for demonstration retrieval. However, under cross-domain settings, such approaches are highly vulnerable to domain-specific distribution shifts, often leading to unreliable retrieval and negative transfer. To address these limitations, we propose **ACD-Retrieval**, a training-free framework for multimodal cross-domain in-context learning. As illustrated in Figure 2, our framework consists of three stages: (1) Domain-wise centering, which removes global distribution shifts across domains; (2) Adaptive Subspace Projection, which constructs a shared low-dimensional subspace via singular value decomposition (SVD); and (3) Inference & Selection, which performs cross-domain retrieval for in-context construction, followed by in-context inference, independent zero-shot inference, and self-selection of the final prediction.

3.1 Domain-wise Centering

As shown in the first stage of Figure 2, source and target embeddings often exhibit domain-specific distribution shifts, which can distort cross-domain similarity estimation. To alleviate this issue, we perform domain-wise centering by independently subtracting the mean embedding within each domain.

For each domain $\mathcal{D}$, we compute its mean representation as $\mu_{\mathcal{D}} = \mathbb{E}_{x \sim \mathcal{D}}[f(x)]$. Given a source sample $x_i \in \mathcal{D}_i$ and a target query $x_{\text{test}} \in \mathcal{D}_T$, we then obtain their centered representations as

$$\tilde{z}_i = f(x_i) - \mu_{\mathcal{D}_i}, \quad \tilde{z}_{\text{test}} = f(x_{\text{test}}) - \mu_{\mathcal{D}_T}. \tag{1}$$

This operation reduces domain-specific global offsets and enhances the stability of cross-domain similarity estimation.

3.2 Adaptive Subspace Projection

As illustrated in the second stage of Figure 2, domain-wise centering reduces global distribution shifts but does not fully eliminate redundant domain-specific variations in high-dimensional embeddings. To further improve cross-domain structural alignment, we project centered representations into a compact subspace using SVD.

For each source domain $\mathcal{D}_i$, we construct an embedding matrix $Z_{\mathcal{D}_i} = [\tilde{z}_1, \ldots, \tilde{z}_{N_{\mathcal{D}_i}}]$, where each row corresponds to a centered representation from that domain, and perform SVD:

$$Z_{\mathcal{D}_i} = U_{\mathcal{D}_i} \Sigma_{\mathcal{D}_i} V_{\mathcal{D}_i}^{\top}. \tag{2}$$

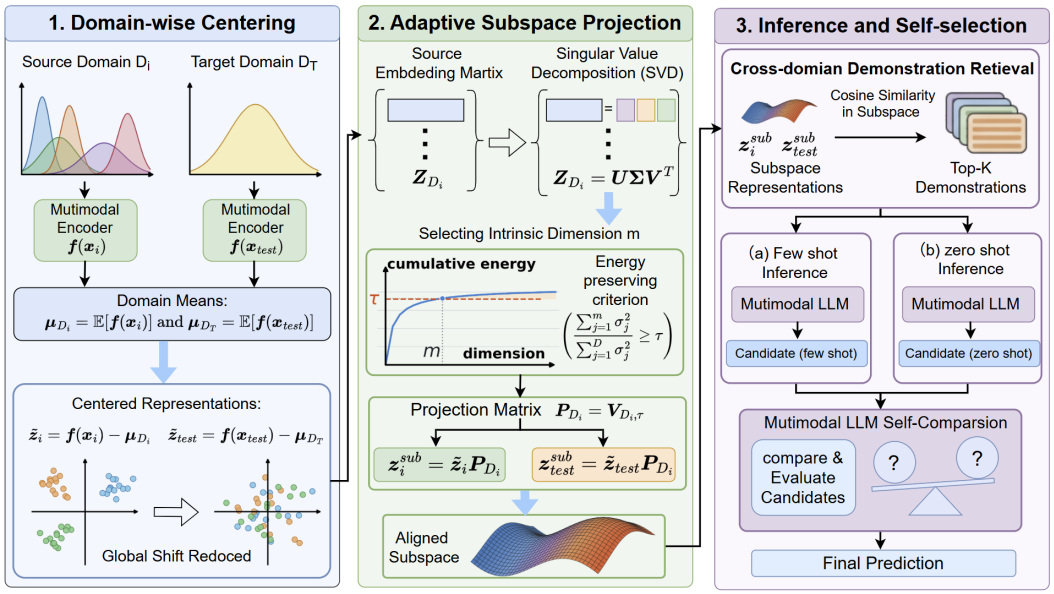


Figure 2: Overview of the proposed **ACD-Retrieval** framework. Our method consists of three stages: (1) **Domain-wise centering**, which aligns domain distributions by centering multimodal embeddings; (2) **Adaptive Subspace Projection**, which applies SVD with energy-based dimension selection to extract compact representations and construct a shared low-dimensional space; and (3) **Inference & Selection**, which first retrieves cross-domain demonstrations in the projected subspace and then performs self-selection between Zero-shot and in-context predictions to mitigate negative transfer.

We select the intrinsic dimensionality m using an energy-preserving criterion $\frac{\sum_{j=1}^{m}\sigma_j^2}{\sum_{j=1}^{D}\sigma_j^2} \geq \tau$, where τ controls the trade-off between compression and information preservation.

The resulting projection matrix is:

$$P_{\mathcal{D}_i} = V_{\mathcal{D}_i,\tau}. \tag{3}$$

We then project representations into source domain subspace:

$$z_i^{\text{sub}} = \tilde{z}_i P_{\mathcal{D}_i}, \quad z_{\text{test}}^{\text{sub}} = \tilde{z}_{\text{test}} P_{\mathcal{D}_i}. \tag{4}$$

We further normalize the projected representations to alleviate discrepancies across domain-specific subspaces, thereby making representations comparable for cross-domain similarity computation.

3.3 Inference and Self-selection

As shown in the third stage of Figure 2, we perform cross-domain demonstration retrieval using the projected subspace representations, followed by in-context inference. However, retrieved demonstrations may still introduce negative transfer under domain mismatch. To improve robustness, we further introduce a self-selection strategy that compares zero-shot and in-context predictions, and selects the more reliable output for final prediction.

Given a target query x_{test}, we compute cosine similarity in the projected space:

$$\text{Sim}(z_{\text{test}}^{\text{sub}}, z_i^{\text{sub}}) = \frac{z_{\text{test}}^{\text{sub}} \cdot z_i^{\text{sub}}}{\|z_{\text{test}}^{\text{sub}}\|_2 \, \|z_i^{\text{sub}}\|_2}. \tag{5}$$

Based on the similarity scores, we retrieve the top-K demonstrations to form the in-context set $\mathcal{E}$ for inference.

Using these demonstrations, we obtain $\hat{y}_{\text{icl}} \sim P_{\text{MLLM}}(\mathcal{E}, x_{\text{test}})$ and $\hat{y}_{\text{zs}} \sim P_{\text{MLLM}}(x_{\text{test}})$.

Since cross-domain demonstrations may introduce either beneficial transfer or negative interference, we further introduce a self-selection mechanism that prompts the MLLM itself to compare and evaluate both candidate predictions.

The final output is selected as:

$$\hat{y} = \text{Select}(\hat{y}_{\text{zs}}, \hat{y}_{\text{icl}}). \tag{6}$$

Specifically, we formulate this selection as a strict pairwise judgment task, where the MLLM evaluates both predictions under a zero-shot template. This self-selection mechanism allows the model to dynamically determine whether the retrieved demonstrations provide positive transfer or introduce negative interference, thereby enhancing robustness across target domains. The complete prompt template used for this self-selection phase is provided in Appendix B.

4 EXPERIMENTS

In this section, we systematically evaluate the effectiveness and underlying mechanism of **ACD-Retrieval** for multimodal cross-domain reasoning. Specifically, we aim to answer three research questions: **(RQ1)** Can ACD-Retrieval consistently improve cross-domain in-context reasoning across different MLLM families and model scales? **(RQ2)** Which components of ACD-Retrieval are primarily responsible for its performance gains under domain shifts? **(RQ3)** How does ACD-Retrieval repair reasoning failures caused by visual ambiguity and cross-domain retrieval noise? We first introduce the experimental setup, and then answer these questions through quantitative and qualitative analysis.

4.1 EXPERIMENTAL SETUP

4.1.1 MODELS.

We evaluate our approach on a diverse suite of multimodal large language models (MLLMs), covering multiple architectures and parameter scales. Specifically, we include **Qwen2.5-VL** (Bai et al., 2025) (3B, 7B, and 32B), **Gemma3** (Team et al., 2025) (4B, 12B, and 27B), and **InternVL3** (Zhu et al., 2025b) (2B, 8B, and 38B). This setup spans a wide range of model capacities (2B–38B), allowing us to systematically evaluate the effectiveness of our method across different architectures and scales. To ensure fair comparisons among retrieval methods, all retrieval-based methods evaluated with the same backbone use an identical prompting configuration and the same top-K setting; only the retrieval strategy differs. Detailed implementation settings are provided in Appendix A.1. The selection of key hyperparameters, including the energy threshold τ and the number of demonstrations top-K, is further analyzed in Appendix A.3.

4.1.2 DATASETS AND TASK.

We conduct experiments on **MATH-V** (Wang et al., 2024a), a challenging benchmark for multimodal mathematical reasoning. MATH-V contains 3,040 visual reasoning problems spanning 16 mathematical disciplines, including algebra, arithmetic, graph theory, topology, and counting.

To study cross-domain in-context learning, we treat each mathematical discipline as an independent domain. Given a target query from one domain, demonstrations are retrieved exclusively from other domains, introducing substantial distribution shifts in both visual and semantic spaces.

4.1.3 BASELINES.

We compare ACD-Retrieval with three baselines: **Zero-shot**, **RICES** (Baldassini et al., 2024), and **Embedding**. Implementation details for all baselines are provided in Appendix A.2.

4.2 RQ1: CAN ACD-RETRIEVAL IMPROVE CROSS-DOMAIN REASONING?

To evaluate whether ACD-Retrieval improves generalization under multimodal domain shifts, we compare our method with various baselines across multiple model families, parameter scales, and reasoning domains.

Table 1: Performance (%) across model families and prompting strategies on different mathematical reasoning domains. "Overall" denotes the sample-level accuracy pooled across all 16 target domains, computed as the total number of correct predictions divided by the total number of evaluation examples. The remaining columns report domain-specific accuracies. We report eight domains in this table, with full definitions provided in Appendix A.4; results for the remaining domains are provided in Appendix A.5. Red cells indicate performance improvement compared with the corresponding Zero-shot baseline, whereas blue cells indicate performance degradation.

Method	Model	Size	Alg.	AnaGeo	Arith.	CGeo	Comb.	Count	DGeo	Graph	Overall
Zero-shot	Gemma	4B	20.6	39.3	29.3	17.5	14.9	17.9	17.3	14.4	20.4
		12B	32.2	47.6	38.6	19.2	26.8	22.4	15.4	17.8	27.3
		27B	40.0	39.3	50.0	21.4	32.1	25.4	20.2	18.9	31.9
	Qwen	3B	19.7	28.6	31.4	16.9	12.5	14.9	18.3	14.4	22.7
		7B	27.5	29.8	32.9	20.4	17.9	13.4	19.2	12.2	25.7
		32B	41.5	40.5	53.6	28.2	30.4	19.4	23.1	24.4	36.2
	Intern	2B	13.3	26.2	16.4	13.3	12.5	10.4	17.3	15.6	15.0
		8B	24.9	42.9	33.6	16.6	19.6	16.4	16.4	17.8	22.6
		38B	39.7	48.8	47.9	28.6	31.6	29.9	22.1	28.9	32.4
RICES	Gemma	4B	19.1	33.3	19.3	20.1	13.1	7.5	21.1	12.2	19.6
		12B	33.0	33.3	36.4	24.4	27.4	17.9	21.1	21.1	27.9
		27B	41.7	41.7	48.6	26.0	33.3	16.4	26.9	24.4	33.4
	Qwen	3B	19.1	28.6	24.3	17.2	13.7	9.0	19.2	17.8	20.6
		7B	22.9	30.9	34.3	21.4	13.1	14.9	21.1	18.9	24.3
		32B	43.5	42.9	55.0	26.6	29.8	26.9	28.9	28.9	36.0
	Intern	2B	13.3	19.1	16.4	15.3	8.3	9.0	20.2	13.3	14.2
		8B	18.8	28.6	29.3	17.2	19.1	9.0	21.1	15.6	21.2
		38B	36.8	46.4	52.1	20.1	25.0	26.9	25.0	24.4	29.8
Embedding	Gemma	4B	18.6	28.6	27.1	15.3	13.7	9.0	12.5	15.6	19.1
		12B	33.3	36.9	42.1	21.8	25.0	19.4	18.3	22.2	27.7
		27B	43.2	35.7	45.7	26.0	29.2	20.9	22.1	28.9	33.7
	Qwen	3B	19.7	28.6	22.9	19.5	12.5	10.4	19.2	13.3	21.1
		7B	24.6	27.4	34.3	19.2	12.5	10.4	23.1	15.6	24.4
		32B	42.9	42.9	52.9	24.7	35.1	22.4	24.0	27.8	36.6
	Intern	2B	11.9	23.8	13.6	17.9	12.5	1.5	15.4	10.0	14.6
		8B	22.3	27.4	27.1	20.8	20.2	11.9	14.4	14.4	21.1
		38B	41.5	45.2	45.7	24.4	29.2	22.4	26.9	25.6	33.3
ACD-Retrieval (Ours)	Gemma	4B	21.4	36.9	26.4	17.5	14.9	16.4	22.1	16.7	**21.7**
		12B	35.9	45.2	41.4	21.1	28.0	14.9	20.2	24.4	**31.3**
		27B	44.6	46.4	48.6	26.6	33.9	26.9	21.1	25.6	**35.6**
	Qwen	3B	22.0	28.6	32.1	17.9	10.7	13.4	19.2	13.3	**24.0**
		7B	29.3	29.8	36.4	24.4	17.3	14.9	20.2	18.9	**27.6**
		32B	43.2	47.6	50.7	28.2	35.1	17.9	24.0	23.3	**38.4**
	Intern	2B	13.6	23.8	17.1	13.6	8.9	11.9	18.3	15.6	**15.7**
		8B	30.4	42.9	35.7	20.8	22.6	17.9	14.4	24.4	**25.9**
		38B	45.5	50.0	54.3	29.6	33.9	29.9	27.9	32.2	**36.6**

Table 1 reports the performance of different experimental settings across various models and scales under cross-domain transfer on MATH-V. Note that it only reports the overall accuracy and results on selected domains (e.g., Algebra (Alg.), Arithmetic (Arith.), Combinatorial Geometry (CGeo)); the cross-domain results for the remaining domains are provided in Appendix A.5.

Overall, ACD-Retrieval consistently achieves the strongest cross-domain reasoning performance across nearly all model families and scales, outperforming other baselines.

Compared with Zero-shot inference, our method yields stable improvements across almost all model scales. On the Gemma3 series, ACD-Retrieval improves overall accuracy from 20.43% to 21.74% (+1.31) on Gemma3-4B, and from 27.34% to 31.28% (+3.94) on Gemma3-12B. Similar gains are observed on Qwen2.5-VL and InternVL3, where improvements reach +1.91 on Qwen2.5-VL-7B and +3.29 on InternVL3-8B.

Interestingly, even larger models with stronger intrinsic reasoning capabilities continue to benefit from our retrieval strategy in Overall accuracy, achieving +2.20 on Qwen2.5-VL-32B and +4.24

Table 2: Ablation study across different models and scales. *Overall* accuracy (%) is the sample-level accuracy pooled across all 16 target domains, computed as the total number of correct predictions divided by the total number of evaluation examples. Zero-shot serves as the baseline without in-context demonstrations. RICES and Embedding represent different retrieval methods, while "w/o Self-selection" denotes the variant with the self-selection mechanism removed from our full framework. The **ACD-Retrieval (Full)** configuration is highlighted in bold.

Method	Gemma-4B	Gemma-12B	Gemma-27B	Qwen-3B	Qwen-7B	Qwen-32B	Intern-2B	Intern-8B	Intern-38B
Zero-shot	20.43	27.34	31.91	22.70	25.66	36.22	14.97	22.57	32.37
RICES	19.57	27.93	33.39	20.62	24.31	36.05	14.18	21.22	29.80
Embedding	19.11	27.70	33.72	21.12	24.38	36.61	14.61	21.15	33.29
w/o Self-selection	19.93	29.38	33.98	21.51	24.31	37.20	15.13	23.16	33.52
ACD-Retrieval (Full)	**21.74**	**31.28**	**35.59**	**23.98**	**27.57**	**38.42**	**15.66**	**25.86**	**36.61**

on InternVL3-38B. This suggests that our method remains effective even for strong MLLMs. In contrast, conventional retrieval methods such as RICES and Embedding often suffer from negative transfer and occasionally underperform Zero-shot inference. For example, on Gemma3-4B, RICES reduces Overall accuracy from 20.43% to 19.57%, while Embedding further drops to 19.11%. Similar degradation is observed on Qwen2.5-VL-3B, where RICES decreases performance from 22.70% to 20.62%, and Embedding only achieves 21.12%.

We further analyze domain-level performance and observe that conventional retrieval methods often suffer from negative transfer on reasoning-intensive tasks, whereas ACD-Retrieval achieves positive transfer in terms of overall accuracy, although its gains vary across individual domains. For example, on InternVL3-38B, our method improves Alg. from 39.71% to 45.51%, while RICES drops to 36.81%; on Arith., our method further improves performance to 54.29%, whereas Embedding only reaches 45.71%. Similar trends are observed on Gemma3-12B, where our method improves Alg. from 32.17% to 35.94%, while Embedding only reaches 33.33% and RICES achieves 33.04%.

Taken together, these results demonstrate that ACD-Retrieval provides robust and scalable improvements for multimodal cross-domain in-context reasoning across diverse architectures and parameter scales.

4.3 RQ2: What Makes ACD-Retrieval Effective?

To understand which components contribute to the performance gains of ACD-Retrieval, we conduct ablation studies across all model families and scales. Table 2 reports the effects of progressively removing (1) subspace alignment and (2) the self-selection mechanism.

We first observe that directly applying conventional retrieval methods often fails to improve performance under cross-domain settings. In several cases, these methods even underperform Zero-shot inference. For example, on Gemma3-4B, RICES and Embedding reduce performance from 20.43% to 19.57% and 19.11%, respectively. This observation suggests that naive embedding similarity is highly vulnerable to domain shifts, where visually similar demonstrations may still contain misleading reasoning patterns.

After introducing subspace alignment while removing the self-selection mechanism (*w/o Self-selection*), performance becomes more stable across model scales. Compared with Embedding, aligned retrieval improves overall accuracy in eight of the nine evaluated model configurations. For example, accuracy improves from 27.70% to 29.38% on Gemma3-12B, from 21.15% to 23.16% on InternVL3-8B, and from 36.61% to 37.20% on Qwen2.5-VL-32B. These results indicate that, compared with conventional retrieval methods, our retrieval strategy achieves more reliable cross-domain demonstration matching while reducing noise introduced by domain discrepancies.

Adding the self-selection mechanism leads to further consistent gains across all model families. Compared with the *w/o Self-selection* variant, incorporating the self-selection mechanism consistently improves performance across all evaluated models, demonstrating its effectiveness under cross-domain settings. Specifically, relative to the *w/o Self-selection* configuration, the complete framework achieves additional gains of +1.81 on Gemma3-4B, +1.90 on Gemma3-12B, +3.26 on Qwen2.5-VL-7B, and +3.09 on InternVL3-38B. These results suggest that even after retrieval alignment, cross-domain demonstrations may still introduce noisy or misleading reasoning patterns dur-

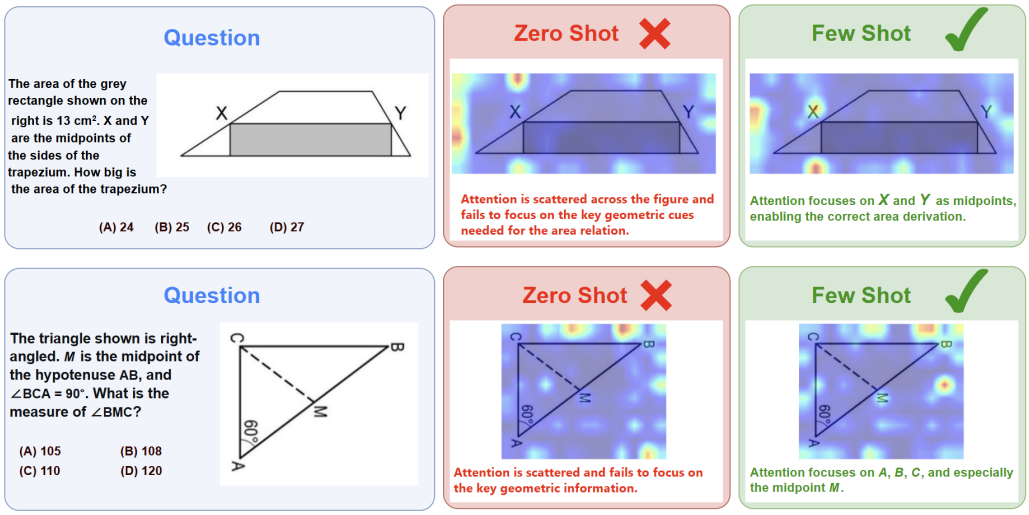


Figure 3: **Case study comparing Zero-shot inference and cross-domain few-shot reasoning.** *Top:* For the trapezium area problem, Zero-shot exhibits scattered attention and misses critical geometric cues, whereas ACD-Retrieval successfully attends to midpoints X and Y on the legs to guide correct reasoning. *Bottom:* For the right-triangle median problem, Zero-shot displays diffuse attention and generates an erroneous prediction, while ACD-Retrieval concentrates on essential keypoints (e.g., A, B, C, and midpoint M). Overall, ACD-Retrieval effectively shifts visual attention from task-irrelevant background regions to crucial geometric structures needed for accurate problem solving.

ing inference. The self-selection mechanism effectively mitigates such negative transfer through adaptive demonstration verification. Notably, the gains are consistently observed across different model families and parameter scales, indicating that self-selection broadly improves the robustness of multimodal cross-domain ICL.

Importantly, the two components address different sources of cross-domain uncertainty. Subspace alignment improves overall accuracy over conventional Embedding retrieval in eight of the nine evaluated model configurations. Self-selection further filters potentially misleading demonstrations by verifying whether they are beneficial to the current reasoning trajectory. Across the nine evaluated model configurations, self-selection yields an average additional gain of 2.07 percentage points over aligned retrieval without selection, suggesting that improved retrieval quality alone is insufficient to guarantee positive transfer. This complementary behavior indicates that ACD-Retrieval separates demonstration matching from demonstration utilization, reducing both retrieval noise and downstream negative transfer.

4.4 RQ3: How Does ACD-Retrieval Repair Reasoning Failures?

To better understand how ACD-Retrieval improves cross-domain reasoning, we compare representative Zero-shot and cross-domain few-shot cases in Figure 3. The first task requires the model to solve a geometric area problem: given that the gray rectangle has an area of $13\,\mathrm{cm}^2$, and points X and Y are the midpoints of the corresponding sides, the model must infer the area of the trapezium. The second asks for $\angle BMC$ in a right triangle with M the midpoint of the hypotenuse. Both problems require recovering midpoint structure rather than reading isolated labels.

Trapezium-area relation. Under Zero-shot prompting (top-middle panel), attention is scattered across the figure and fails to focus on the midpoint annotations. Without treating X and Y as the structural hinge of the construction, the model invents inconsistent height–base relations and predicts quantities such as 39 or 52, neither of which matches an option. With ACD-Retrieval (top-right panel), attention concentrates on X and Y together with the gray rectangle. This focus supports the midpoint-based area identity $\frac{(a+b)h}{4} = 13$, which yields $(a+b)h = 52$ and the trapezium area $\frac{(a+b)h}{2} = 26$, corresponding to the correct option C.

Right-triangle midpoint angle. The second example shows the same pattern in an angular setting. In the Zero-shot condition (bottom-middle panel), attention remains diffuse and does not lock onto the midpoint M. The model incorrectly assumes that the median bisects $\angle C$, sets $\angle ACM = \angle BCM = 45°$, and predicts $\angle BMC = 180° - 45° - 45° = 135°$, which is wrong. With ACD-Retrieval (bottom-right panel), attention focuses on A, B, C, and especially the midpoint M. Recognizing $CM = AM = MB$ and $\angle A = 60°$ gives $\angle B = 30°$; in isosceles $\triangle BMC$, $\angle BMC = 180° - 30° - 30° = 120°$, and the model selects the correct option D.

These observations suggest that ACD-Retrieval improves cross-domain geometric reasoning not only by retrieving semantically related examples, but more importantly by redirecting attention toward structurally critical midpoints and transferring aligned reasoning patterns that repair broken Zero-shot trajectories, leading to more stable cross-domain generalization.

5 Related Work

5.1 Multimodal In-Context Learning and Generalization

In-context learning enables large language models to adapt to new tasks without parameter updates (Brown et al., 2020). Its effectiveness relies on knowledge transfer from pretraining and the ability to model contextual patterns, but often degrades under domain shifts (Öncel et al., 2024; Oh et al., 2025). Existing approaches for improving generalization include data-centric adaptation, such as data augmentation and task mixture (Wang et al., 2024b); prompt optimization, including calibration, search, and rewriting (Honda & Oka, 2025; He et al., 2024; Zhao et al., 2021); and parameter-efficient tuning methods such as Adapter, Prefix-Tuning, and LoRA (Houlsby et al., 2019; Li & Liang, 2021; Hu et al., 2022). Long-context models further improve reasoning under high-resolution inputs and extended sequences (Chen et al., 2026). In addition, retrieval-augmented methods improve robustness in out-of-distribution reasoning (Zhu et al., 2025a).

With the development of vision-language models, ICL has extended to multimodal settings. Flamingo, LLaVA, Kosmos-1, Qwen2.5-VL, InternVL3, and Gemma3 show strong few-shot and Zero-shot transferability (Alayrac et al., 2022; Liu et al., 2023; Huang et al., 2023; Bai et al., 2025; Zhu et al., 2025b; Team et al., 2025). However, multimodal models remain vulnerable to visual shifts and cross-modal misalignment in cross-domain scenarios.

5.2 Example Retrieval Methods

Example selection is highly important for ICL performance (Luo et al., 2024; Lu et al., 2022). Existing unimodal retrieval methods mainly rely on nearest-neighbor retrieval, task matching, and dense retrieval frameworks (Rubin et al., 2022), often combined with uncertainty estimation, coverage constraints, diversity selection, and maximal marginal relevance (MMR) reranking to improve example quality (Ling et al., 2024; Huang et al., 2025).

In multimodal settings, CLIP is widely used for joint image-text representation learning and cross-modal retrieval (Radford et al., 2021). Building on this paradigm, BLIP-2, Pixtral, and Phi-4 further extend ICL to complex vision-language reasoning tasks (Li et al., 2023; Microsoft et al., 2025; Agrawal et al., 2024). Together, these advances establish the foundation for cross-domain multimodal retrieval. However, directly applying simple similarity computation (Baldassini et al., 2024; Li et al., 2026) in cross-domain settings may fail to retrieve useful cross-domain demonstrations, leading to unstable retrieval and negative transfer.

6 Conclusion

We propose a unified framework that improves structural matching across domains through representation centering and subspace alignment. In addition, we design a self-selection strategy to suppress harmful demonstrations. Extensive experiments across diverse multimodal cross-domain transfer settings show that our method consistently achieves significant positive transfer. Our study validates the feasibility of cross-domain knowledge transfer in multimodal ICL and points to a promising direction for developing more efficient and robust retrieval methods.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 provides the mathematical formulation of domain-wise centering, adaptive SVD-based subspace projection, cross-domain demonstration retrieval, and self-selection. Section 4 describes the evaluated MLLM families, the MATH-V evaluation protocol, the baselines, and the Overall metric. Appendix A reports the retrieval encoder and the inference and hyperparameter settings, while Appendix B provides the exact self-selection prompt. The anonymous supplementary package will include the implementation and evaluation scripts, data indices, demonstration-pool construction, configuration files, prompts, answer-parsing rules, and per-example predictions.

ETHICS STATEMENT

This work uses publicly available model checkpoints and benchmark datasets for research on multimodal mathematical reasoning. It does not involve human participants or the collection of personal, private, or sensitive data. The study does not introduce new datasets or deploy the proposed method in high-stakes settings. All models and datasets are used solely for academic research in accordance with their applicable licenses and usage terms. We identify no specific ethical concerns arising from this study.

AI USE STATEMENT

Generative AI tools were used to support manuscript organization, English language polishing, editorial refinement, and LaTeX formatting. All AI-assisted revisions were carefully reviewed and verified by the authors against the source code, experimental records, and relevant literature. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. URL https://arxiv.org/abs/2410.07073.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 23716–23736, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1539–1550, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2020.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *Advances in Neural Information Processing Systems*, pp. 91077–91100, 2026.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9062–9072, 2025.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. *arXiv preprint arXiv:2010.03978*, 2020. URL https://arxiv.org/abs/2010.03978.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 13477–13499, 2024.

Ukyo Honda and Tatsushi Oka. Exploring explanations improves the robustness of in-context learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 23693–23714, 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022.

Hsiu-Yuan Huang, Zichen Wu, Yutong Yang, Junzhao Zhang, and Yunfang Wu. Unlocking the power of llm uncertainty for active in-context example selection. *arXiv preprint arXiv:2408.09172*, 2025. URL https://arxiv.org/abs/2408.09172.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*, pp. 72096–72109, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742, 2023.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 3357–3370, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8086–8098, 2022.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *Transactions on Machine Learning Research*, 2024.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. In-context learning generalizes, but not always robustly: The case of syntax. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 4761–4779, 2024.

Changdae Oh, Zhen Fang, Shawn Im, Xuefeng Du, and Yixuan Li. Understanding multimodal LLMs under distribution shifts: An information-theoretic approach. In *Proceedings of the International Conference on Machine Learning*, 2025.

Fırat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. Adaptation odyssey in LLMs: Why does additional pretraining sometimes fail to improve? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 19834–19843, 2024.

Linyu Ou and YuYang Yin. Empowering lightweight mllms with reasoning via long cot sft. *arXiv preprint arXiv:2509.03321*, 2025. URL https://arxiv.org/abs/2509.03321.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 2655–2671, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2403.16999*, 2024. URL https://arxiv.org/abs/2403.16999.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-Yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems*, 2024a.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1752–1767, 2024b.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

Table 3: Sensitivity analysis of the SVD energy threshold τ on the **Overall** metric (%).

Model	$\tau = 0.8$	$\tau = 0.9$	$\tau = 1.0$
Intern-8B	25.30	**25.86**	25.66

Table 4: Sensitivity analysis of the number of retrieved demonstrations top-K on the **Overall** metric (%).

Model	top-$K = 1$	top-$K = 2$	top-$K = 3$
Qwen-7B	27.07	27.24	**27.57**
Qwen-32B	**38.42**	36.51	37.17

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning*, pp. 12697–12706, 2021.

Jiachen Zhu, Congmin Zheng, Jianghao Lin, Kounianhua Du, Ying Wen, Yong Yu, Wang Jun, and Weinan Zhang. Retrieval-augmented process reward model for generalizable mathematical reasoning. In *Findings of the Association for Computational Linguistics*, pp. 8453–8468, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b. URL https://arxiv.org/abs/2504.10479.

A Appendix

A.1 Implementation Details.

For retrieval, we utilize the pretrained Qwen3-VL-Embedding (Li et al., 2026) model as the multimodal encoder to extract unified representations for both target queries and candidate demonstrations. These embeddings are further used to compute cross-domain similarity and construct the aligned shared subspace for demonstration retrieval. The energy threshold is set to $\tau = 0.9$. The number of demonstrations is set to top-$K = 3$ for medium and smaller models ($\leq$ 12B) and top-$K = 1$ for large-scale models ($>$ 12B). Notably, all hyperparameters are kept consistent across various domains to evaluate cross-domain generalization. All evaluation with MLLMs (e.g., Qwen2.5-VL, Gemma3, and InternVL3) is conducted under a zero-temperature setting ($T = 0$) to ensure deterministic outputs and fair comparison across models.

A.2 Baseline Details

We compare our method against the following baselines:

- **Zero-shot**: The model performs direct inference without any in-context demonstrations.
- **RICES**: A retrieval-based ICL method that separately computes visual and textual similarity using pretrained multimodal embeddings, and ranks demonstrations based on the sum of the two similarity scores.
- **Embedding**: Demonstrations are retrieved based on similarity in the original multimodal embedding space.
- **ACD-Retrieval**: Demonstrations are retrieved using SVD-based similarity computed in a shared aligned subspace, followed by the same self-selection mechanism.

A.3 HYPERPARAMETER ANALYSIS

We further investigate the sensitivity of our framework to two important hyperparameters: the SVD energy threshold τ, which determines the retained subspace rank (Table 3), and the number of retrieved demonstrations top-K used for in-context prompting (Table 4). All reported results in this section use the **Overall** metric, defined as the sample-level accuracy pooled across all 16 target domains: the total number of correct predictions divided by the total number of evaluation examples.

We keep the SVD energy threshold τ fixed across all models, as it primarily influences the structure of the embedding space, which is shaped by the underlying data distribution rather than model scale. Therefore, fixing τ ensures consistent subspace construction and avoids confounding variations introduced by model-specific tuning. In contrast, the number of retrieved demonstrations top-K is a model-dependent hyperparameter. For medium and smaller models, we use a unified top-K setting, as their reasoning and representation capacities are relatively close, which provides stable and consistent external guidance. For larger models, we adopt a different top-K setting, as they possess stronger intrinsic reasoning ability and are more sensitive to contextual redundancy, making fewer demonstrations more effective.

Effect of SVD Energy Threshold. The energy threshold τ controls how many singular components are preserved during subspace construction. A smaller threshold may discard transferable task-relevant structure, while a larger threshold may retain excessive domain-specific noise. We evaluate $\tau \in \{0.8, 0.9, 1.0\}$ on InternVL3-8B.

As shown in Table 3, performance improves from 25.30% at $\tau = 0.8$ to 25.86% at $\tau = 0.9$, and then slightly decreases to 25.66% at $\tau = 1.0$. This trend suggests that $\tau = 0.9$ achieves the best balance between preserving shared information and suppressing noisy directions. Therefore, we adopt $\tau = 0.9$ in all main experiments.

Effect of Demonstration Number. We also analyze the impact of the number of retrieved demonstrations top-K. In general, medium and smaller models benefit from more demonstrations, which provide stronger reasoning guidance, while larger models tend to prefer fewer but more informative examples to avoid context redundancy.

As shown in Table 4, for Qwen2.5-VL-7B, performance consistently improves as more demonstrations are added, achieving the best result at top-$K = 3$ (27.57%). In contrast, Qwen2.5-VL-32B achieves its best performance at top-$K = 1$ (38.42%). Performance decreases to 36.51% at top-$K = 2$ and then partially recovers to 37.17% at top-$K = 3$, remaining below the top-$K = 1$ result.

These observations indicate that the optimal number of demonstrations strongly depends on model scale. Medium and smaller models rely more on external examples for reasoning support, whereas larger models already possess stronger internal reasoning ability and are more sensitive to noisy or redundant context.

Based on these results, we use top-$K = 3$ for models with at most 12B parameters and top-$K = 1$ for models larger than 12B. This simple heuristic yields stable performance across model families without requiring expensive per-model hyperparameter tuning.

A.4 DEFINITIONS OF MATHEMATICAL DOMAINS

To facilitate understanding, we provide the full names and definitions of each mathematical domain used in our analysis.

- **Algebra (Alg.):** Problems involving variables, symbols, and rules for manipulating mathematical expressions and equations.
- **Analytic Geometry (AnaGeo):** The study of geometry using a coordinate system and the principles of algebra and analysis.
- **Arithmetic (Arith.):** The elementary branch of mathematics focusing on numbers and traditional operations like addition and subtraction.

Table 5: Performance (%) across remaining evaluation domains.

Method	Model	Size	Logic	MG-Ang.	MG-Area	MG-Len.	Solid	Stats.	Topo.	TransGeo
Zero-shot	Gemma	4B	27.7	12.1	20.0	22.5	19.3	25.9	13.0	20.2
		12B	27.7	21.4	27.8	31.2	28.3	29.3	21.7	20.8
		27B	37.0	21.4	31.8	36.5	34.4	36.2	13.0	25.0
	Qwen	3B	19.3	27.8	25.2	27.8	20.9	41.4	13.0	23.2
		7B	23.5	29.5	29.2	28.3	22.1	37.9	21.7	28.6
		32B	34.5	39.9	38.0	40.3	35.2	62.1	21.7	26.2
	Intern	2B	18.5	17.3	13.0	13.1	15.6	27.6	13.0	17.9
		8B	22.7	16.2	25.6	22.1	23.8	25.9	13.0	18.4
		38B	29.4	21.4	32.8	34.1	29.1	39.7	30.4	23.2
RICES	Gemma	4B	19.3	15.6	21.4	20.3	25.0	20.7	17.4	16.1
		12B	31.1	18.5	28.6	31.4	25.8	27.6	17.4	27.4
		27B	37.0	24.3	33.2	38.1	34.0	31.0	4.3	27.4
	Qwen	3B	23.5	26.0	21.2	25.2	20.1	29.3	8.7	14.9
		7B	22.7	32.4	26.4	27.2	22.1	37.9	8.7	20.2
		32B	30.2	31.8	38.8	38.8	36.9	41.4	26.1	28.6
	Intern	2B	11.8	9.8	14.6	12.9	16.0	22.4	13.0	17.3
		8B	24.4	12.7	24.6	23.8	22.5	24.1	21.7	19.6
		38B	25.2	18.5	31.6	29.6	30.3	50.0	26.1	20.8
Embedding	Gemma	4B	21.0	14.4	20.8	20.5	22.1	24.1	17.4	20.2
		12B	26.1	15.6	28.2	29.8	32.0	31.0	8.7	26.8
		27B	33.6	24.3	34.4	39.9	34.8	34.5	17.4	28.6
	Qwen	3B	18.5	28.9	22.4	23.6	22.9	31.0	17.4	17.9
		7B	25.2	33.0	27.2	24.9	22.5	37.9	26.1	25.0
		32B	38.7	40.5	37.0	41.0	36.9	53.5	39.1	23.8
	Intern	2B	16.8	15.6	13.2	13.6	15.2	19.0	34.8	19.1
		8B	16.8	16.8	23.2	22.1	21.7	29.3	26.1	18.4
		38B	34.5	28.9	34.4	34.5	32.4	50.0	17.4	28.0
ACD-Retrieval (Ours)	Gemma	4B	27.7	14.4	22.0	26.1	21.7	19.0	17.4	22.6
		12B	31.1	32.4	32.8	36.5	31.1	27.6	21.7	28.6
		27B	37.8	27.2	36.8	39.6	36.9	31.0	21.7	30.9
	Qwen	3B	18.5	31.8	27.6	30.5	22.9	36.2	13.0	22.6
		7B	26.9	35.3	29.4	31.0	24.6	37.9	21.7	25.6
		32B	36.1	44.5	38.8	45.0	39.3	56.9	34.8	30.4
	Intern	2B	16.8	14.4	14.0	17.1	15.6	27.6	13.0	22.6
		8B	26.9	17.3	28.8	25.8	24.2	27.6	17.4	25.6
		38B	32.8	28.3	38.0	39.2	31.6	46.5	21.7	29.2

- **Combinatorial Geometry (CGeo):** Focuses on the discrete properties of geometric objects, such as the arrangement of points and lines.
- **Combinatorics (Comb.):** The study of finite or countable discrete structures, including permutations and combinations.
- **Counting (Count):** Tasks requiring the systematic enumeration of objects or configurations following specific constraints.
- **Descriptive Geometry (DGeo):** A branch of geometry providing methods to represent three-dimensional objects in two dimensions.
- **Graph Theory (Graph):** The study of graphs, which are mathematical structures used to model pairwise relations between objects.
- **Logic (Logic):** Formal systems of reasoning, including propositional and predicate logic applied to mathematical proofs.
- **Metric Geometry - Angle (MG-Ang.):** Geometry problems specifically focused on the measurement and properties of angles within figures.
- **Metric Geometry - Area (MG-Area):** Quantitative analysis of the two-dimensional extent (area) of closed planar shapes.
- **Metric Geometry - Length (MG-Len.):** Metric problems involving the calculation of distances, perimeters, and segments in Euclidean space.

- **Solid Geometry (Solid):** The geometry of three-dimensional figures such as polyhedra, cylinders, cones, and spheres.
- **Statistics (Stats.):** The collection, analysis, interpretation, and presentation of quantitative data and probability.
- **Topology (Topo.):** The study of spatial properties invariant under continuous deformations like stretching or twisting.
- **Transformation Geometry (TransGeo):** The study of geometry through groups of transformations such as symmetry, rotation, and scaling.

A.5 Additional Main Results

This section presents additional cross-domain results on the remaining domains not included in the main body, as shown in Table 5.

The additional results further support the effectiveness of ACD-Retrieval across diverse reasoning domains. For example, on Qwen2.5-VL-32B, our method improves metric Geometry - length (MG-Len.) performance from 40.31% to 44.99% and Transformation Geometry (TransGeo) performance from 26.19% to 30.36%. On InternVL3-38B, our method achieves substantial gains on Statistics (Stats.), improving accuracy from 39.66% to 46.55%. Similar improvements are also observed on Gemma3-27B, where Topology (Topo.) accuracy increases from 13.04% to 21.74%.

In contrast, conventional retrieval methods exhibit significant instability and frequently suffer from negative transfer, occasionally even underperforming Zero-shot inference. For instance, on Qwen2.5-VL-32B, RICES degrades Logic performance from 34.45% to 30.25%. More strikingly, on Gemma3-12B, Embedding-based retrieval causes a sharp decline in Topo. accuracy, plummeting from 21.74% to a mere 8.70%.

Collectively, these results show that ACD-Retrieval achieves positive transfer in terms of overall accuracy, although the magnitude and direction of its effects vary across individual model-domain combinations. Compared with the evaluated retrieval baselines, it also reduces the risk of severe performance degradation in cross-domain reasoning tasks.

B Prompt Templates

Prompt Template for Phase-III Self-Selection

```
You are a strict judge.  Given the following math question with
image(s), and two model responses (Response-1 and Response-2),
determine which response is correct.  The order of the responses
does not imply correctness.  Do not favor Response-1 simply because
it appears first.  Output the final answer directly.  Put your final
answer in the form \boxed{Response-1} or \boxed{Response-2}.

Image-1:  <image>

Question:
{question}
{options}

Response-1:
{response_a}

Response-2:
{response_b}

Output the final answer directly.  Put your final answer in the form
\boxed{Response-1} or \boxed{Response-2}.
```