Retrieval Enhanced Feedback via In-context Neural Error-book

Jongyeop Hyun
School of CSE
Chung-Ang University
jesussuej@cau.ac.kr[*]

Bumsoo Kim
School of CSE
Chung-Ang University
bumsoo@cau.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) have significantly improved reasoning capabilities, with in-context learning (ICL) emerging as a key technique for adaptation without retraining. While previous works have focused on leveraging correct examples, recent research highlights the importance of learning from errors to enhance performance. However, existing methods lack a structured framework for analyzing and mitigating errors, particularly in Multimodal Large Language Models (MLLMs), where integrating visual and textual inputs adds complexity. To address this issue, we propose REFINE: Retrieval-Enhanced Feedback via In-context Neural Error-book, a teacher-student framework that systematically structures errors and provides targeted feedback. REFINE introduces three systematic queries to construct structured feedback—Feed-Target, Feed-Check, and Feed-Path—to enhance multimodal reasoning by prioritizing relevant visual information, diagnosing critical failure points, and formulating corrective actions. Unlike prior approaches that rely on redundant retrievals, REFINE optimizes structured feedback retrieval, improving inference efficiency, token usage, and scalability. Our results demonstrate substantial speedup, reduced computational costs, and successful generalization, highlighting REFINE's potential for enhancing multimodal reasoning.

1 Introduction

"The only real mistake is the one from which we learn nothing." — Henry Ford

Recent LLM advancements show superior reasoning performance, with extensive research on in-context learning (ICL) to enhance human-like reasoning capabilities. Early works on ICL generated responses based on a few provided correct examples, enabling the models to adapt to new tasks without extensive retraining (Brown et al., 2020; Dong et al., 2024; Wei et al., 2023). Although ICL has demonstrated effectiveness by primarily leveraging *correct* examples (Min et al., 2022), subsequent works have re-examined the fact that human learning is also deeply rooted in learning from *errors* (Edmondson, 1996; Chialvo and Bak, 1999; Edmondson, 1999). Recent studies suggest that incorporating errors into the learning process can further improve LLM performance (Sun et al., 2024). These approaches typically identify recurring errors, extract underlying principles from them, and apply these insights to prevent similar errors in the future.

However, a critical limitation of these methodologies is the absence of a systematic framework for structuring errors, making it difficult to analyze and mitigate failure cases effectively. This challenge becomes even more pronounced in Multimodal Large Language Models (MLLMs) that jointly process multimodal inputs (Zhao et al., 2023). Unlike unimodal LLMs where errors in textual reasoning can often be traced and corrected through established interpretability techniques, MLLMs introduce additional complexity due to the integration of visual and textual modalities. Without a structured approach to diagnosing and addressing previous errors, failures in one modality can propagate through the system, making it more difficult to ensure reliable and interpretable outcomes (Lau et al., 2025). While recent works such as multimodal-CoT have shown that incorporating textual CoT reasoning and in-context learning can improve performance, a fundamental gap persists: the lack of structured error analysis for multimodal reasoning makes it unclear whether existing MLLMs can fully leverage CoT reasoning for visual understanding (Alayrac et al., 2022; Zhao et al., 2023). Furthermore, the interpretability of MLLMs remains a significant challenge, as current approaches do not adequately explain how

[*] contact: mldljyh@postech.ac.kr

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14083–14098
November 4-9, 2025 ©2025 Association for Computational Linguistics

visual information contributes to reasoning, highlighting the need for a more rigorous framework for structuring and mitigating errors in multimodal AI systems.

To address these challenges, we propose **REFINE: Retrieval-Enhanced Feedback via In-context Neural Error-book**. Our proposed REFINE is a teacher-student framework where the teacher generates a structured Error-book based on the student's observed errors, establishes question-level feedback based on this Error-book, and the student retrieves and applies proper feedback to prevent the recurrence of similar errors. Instead of expecting the MLLM to extract the optimal intuition to resolve the error from the initial response, we prompt it to structure what went wrong during the inference process. In MLLMs, effective reasoning requires a stronger focus on visual inputs. To enhance this process, we introduce three structured feedback mechanisms: Feed-Target, Feed-Check, and Feed-Path.

"Feed-Target" extracts high-level observations essential for accurate inference based on an image-question pair. For instance, when answering a question that requires counting pedestrians or vehicles, the model must prioritize proper object detection within the visual input. "Feed-Check" retrospectively analyzes errors, identifying the most critical failure points. For example, in an image-question-answer triplet, an error may stem from incorrect perception of the object *people*. Thirdly, "Feed-Path" formulates corrective actions by generating explicit instructions to refine the model's response and mitigate previously identified errors. We integrate these feedback mechanisms while excluding self-regulatory feedback, which our analysis shows introduces noise rather than improving response quality. Since our feedback structure is well-organized, storing multiple insights for similar questions is unnecessary. Unlike previous approaches that often rely on retrieving and processing multiple samples, REFINE employs a deterministic single-nearest-neighbor strategy for its structured feedback. This ensures consistency and low overhead at inference time, in stark contrast to the inefficiencies and stochastic behavior of traditional ICL approaches, thereby enabling a structured framework to infer the correct chain-of-thought reasoning without redundant retrievals in multimodal tasks. Additionally, we demonstrate that task-level insight retrieval offers no measurable benefit in multimodal question-answering benchmarks.

Besides accuracy, our method significantly outperformed baselines in terms of inference efficiency. Our structured feedback retrieval is substantially faster ($44.7 - 76.4\times$ speedup compared to the RICP baseline) and more token-efficient, improving spatial complexity (approximately 64.2% fewer tokens). The use of precomputed embeddings and the removal of clustering significantly reduce computational costs, demonstrating scalability and feasibility in real-time settings. Furthermore, successful generalization from smaller subsets (MME-RealWorld-Lite) to larger-scale tests (MME-RealWorld) clearly illustrates the practical scalability of our approach.

2 Related Work

2.1 Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning enhances Large Language Model (LLM) problem-solving by breaking down tasks into intermediate logical steps, akin to human cognition (Wei et al., 2022). This structured approach has improved performance in mathematical, commonsense, and multimodal tasks. However, CoT's efficacy depends on the model's inherent reasoning and prompt quality. A key challenge is error propagation from incorrect intermediate steps, necessitating refinements like retrieval or error-correction mechanisms for improved robustness (Cao et al., 2023).

2.2 Language Instruction Understanding in Multimodal Tasks

Language instructions are vital for AI-user interaction in multimodal tasks, traditionally requiring precision. While recent Multimodal Large Language Models (MLLMs) better handle complex instructions, ambiguity remains a challenge (Fu et al., 2023; Fei et al., 2022). Efforts like WAFFLE-CLIP (Roth et al., 2023) and FUDD (Esfandiarpoor and Bach, 2024) address polysemy in image classification, and REPHRASE (Prasad et al., 2024) uses iterative prompting for visual question answering. However, many existing methods demand extensive interactions or predefined rules, hindering generalization and scalability across diverse tasks and models.

2.3 Complex Reasoning with Multimodal Large Language Models

MLLMs increasingly integrate visual and textual information (Caffagni et al., 2024), yet comprehending complex language in visual contexts remains difficult (Zhao et al., 2023; Kil et al., 2024). Approaches to enhance MLLM reasoning are either training-based, aligning models with image-text data, sometimes using synthetic data (Davidson et al., 2025), or non-training-based, like CoT, which simulates step-by-step reasoning (Pramanick et al., 2024; Kil et al., 2024). Training-based methods can be data-intensive and costly (Davidson et al., 2025), while non-training methods often presume strong pre-existing reasoning capabilities (Zhang et al., 2024a), limiting general-purpose intelligence development.

2.4 Systematic Structuring of MLLMs

Systematically structuring MLLMs with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) is crucial for performance and scalability. Studies like Wang et al. (2024) and VisRAG (Yu et al., 2024) show RAG's benefits in multimodal contexts, while RAVEN (Rao et al., 2024) highlights gains in tasks like image captioning. This underscores the need for integrating retrieval modules for adaptability. Our work also advocates for refined feedback flows, drawing on structured techniques to filter unhelpful signals, akin to our concise "Error-book" concept. Existing methods, however, often face data curation challenges or domain-specific limitations. Our approach aims for a distilled feedback structure, enhancing MLLM resilience and extensibility with minimal overhead.

2.5 Advancements in Error-Driven Learning

Recent error-driven learning methods include LEAP (Zhang et al., 2024b), generating static principles from errors, and TRAN (Tong et al., 2024), maintaining rules to avoid past mistakes. RICP (Sun et al., 2024) clusters errors into task-level principles and retrieves question-level insights. These methods often rely on generalized principles or pre-clustered errors, potentially misaligning with immediate task goals. Our approach differs by restructuring feedback based on the **Feedback Model** (Hattie and Timperley, 2007), focusing on task/process-level guidance through an "Error-book" addressing: **Feed-Target** (goal), **Feed-Check** (progress), and **Feed-Path** (actions). This prioritizes task-specific guidance, enabling efficient, deterministic retrieval without teacher intervention during inference, thus reducing overhead and improving precision over methods with rigid clustering or insufficient task focus.

3 Method

The objective of this paper is to develop a structured Neural Error-book that systematically provides feedback to enhance model performance effectively. To achieve this, the Neural Error-book construction involves three distinct feedback formulation stages. The detailed pseudocode for our REFINE framework is presented in Appendix B.

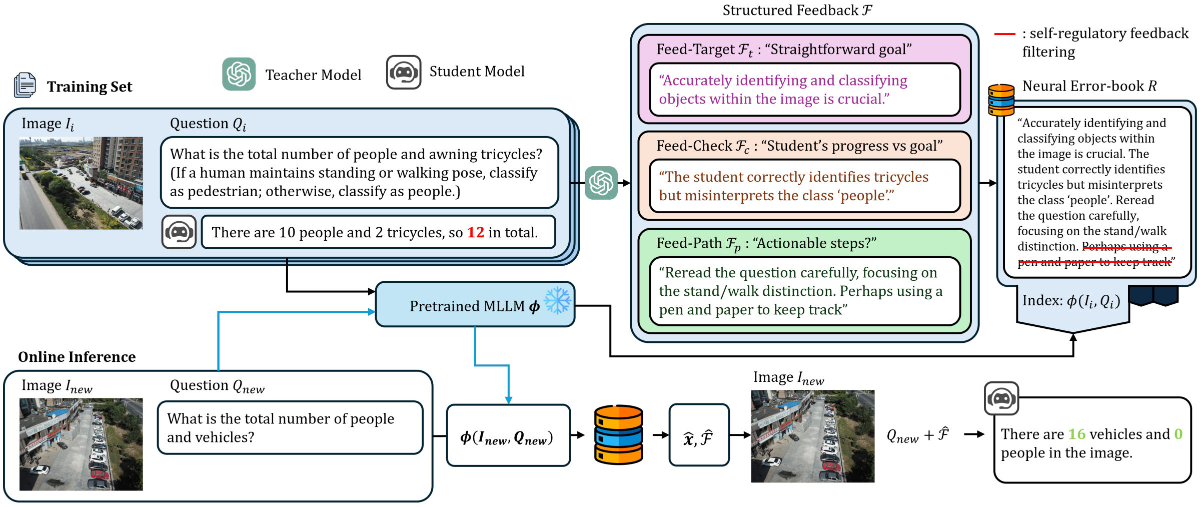
3.1 Structured Feedback Generation

REFINE is a teacher-student framework where the teacher model systematically analyzes errors made by the student model. We designed three guiding principles to construct a final structured feedback for each error (student misprediction) inspired by classic educational psychology (Hattie and Timperley, 2007): Feed-*Target*, Feed-*Check*, and Feed-*Path*. These principles guide structured error analysis and facilitate the generation of precise, actionable insights. Given a multimodal QA benchmark, our process begins by evaluating the student model on a training set of image-question pairs $\{(I_i, Q_i)\}_{i=1}^{N}$. Errors from student predictions $\hat{A}_i$ (incorrect answers) are aggregated and provided to the teacher model alongside their ground-truth reference A_i. The teacher model generates structured feedback through the threefold analysis:

1. **Feed-Target**: *"What is the straightforward goal of this task?"* Clarifies essential task requirements by extracting high-level observations necessary for accurate inference (e.g., "Proper object detection is essential for counting pedestrians and vehicles").

2. **Feed-Check**: *"How does the student's current progress align with the goal?"* Analyzes the student's mispredictions retrospectively, pinpointing critical failures in perception or reasoning (e.g., "Misclassification of 'people' due to overlooking pose criteria").

3. **Feed-Path**: *"What actionable steps bridge the gap to achieve the goal?"* Formulates explicit corrective instructions designed to help prevent error recurrence (e.g., "Re-analyze



Training Set
Teacher Model
Student Model
Image I_i
Question Q_i
What is the total number of people and awning tricycles? (If a human maintains standing or walking pose, classify as pedestrian; otherwise, classify as people.)
There are 10 people and 2 tricycles, so 12 in total.
Structured Feedback $\mathcal{F}$
Feed-Target $\mathcal{F}_t$: "Straightforward goal"
"Accurately identifying and classifying objects within the image is crucial."
Feed-Check $\mathcal{F}_c$: "Student's progress vs goal"
"The student correctly identifies tricycles but misinterprets the class 'people'."
Feed-Path $\mathcal{F}_p$: "Actionable steps?"
"Reread the question carefully, focusing on the stand/walk distinction. Perhaps using a pen and paper to keep track"
: self-regulatory feedback filtering
Neural Error-book R
"Accurately identifying and classifying objects within the image is crucial. The student correctly identifies tricycles but misinterprets the class 'people'. Reread the question carefully, focusing on the stand/walk distinction. ~~Perhaps using a pen and paper to keep track~~"
Index: $\phi(I_i, Q_i)$
Pretrained MLLM ϕ
Online Inference
Image I_{new}
Question Q_{new}
What is the total number of people and vehicles?
$\phi(I_{new}, Q_{new})$
$\hat{x}, \hat{\mathcal{F}}$
Image I_{new}
$Q_{new} + \hat{\mathcal{F}}$
There are 16 vehicles and 0 people in the image.

Figure 1: Overview of the REFINE. Given training set image-question pairs, REFINE extracts structured feedbacks under three systematic queries: Feed-Target, Feed-Check and Feed-Path. The self-regulatory questions are filtered out to construct the final feedback and our final Neural Error-book R is indexed under the multimodal embedding via the pre-trained model ϕ. During inference, the input image and question pairs are embedded to retrieve the most similar feedback within R under their multimodal embedding as the index. The retrieved feedback $\hat{\mathcal{F}}$ is used to enhance the query and obtain the final result.

image regions with sitting figures using the question's pose definitions").

3.2 Feedback Filtering

After Feed-Target, Feed-Check and Feed-Path, the resulting feedback instances are categorized. This classification is performed automatically by our teacher model. The model is provided with the definitions of "Task/Process-relative" feedback (that directly corrects task-specific errors or adds specifics for reasoning, e.g., "Adjust counts for occluded objects.") and "Self-Regulatory" feedback (that addresses metacognitive habits or personal traits, e.g., "To improve accuracy, try solving similar problems multiple times.") from Hattie and Timperley (2007) and then prompted to classify each generated instance. This automated method ensures the process is consistent and reproducible. Based on empirical observations suggesting that self-regulatory feedback tends to hinder final Chain-of-Thought (CoT) performance, we filter out feedback classified as self-regulatory.

3.3 Neural Error-book Construction

After filtering the generated feedback to retain only actionable task/process-level feedback (denoted as $\{\mathcal{F}_i\}_{i \in N_e}$, where N_e denotes error cases), we pair each feedback with the joint embedding for the corresponding image-question pairs to construct a structured Neural Error-book database R. Given ϕ as a pre-trained embedding model (e.g., voyage-multimodal-3 (VoyageAI, 2024)) that obtains a joint multimodal embedding from each image-question pair in the training set $x_i = (I_i, Q_i)$, we construct a Neural Error-book R as:

$$R = \{\phi(x_i), \mathcal{F}_i\}_{i \in N_e} \tag{1}$$

Since the Error-book is indexed with the joint embedding of the image-question pair, it enables efficient retrieval at inference time. Note that unlike previous work, we store only a single well-structured feedback instead of storing multiple redundant insights and clustering them afterwards for similar questions.

3.4 REFINE

During inference on unseen image-question pairs, we leverage the constructed Neural Error-book, which stores structured feedback indexed by the multimodal joint embeddings of the training set's image-question pairs. The retrieved feedback enhances the in-context feedback for the student model, improving precision and overall result quality. Given an unseen query $x_{\text{query}} = (I_{\text{query}}, Q_{\text{query}})$, our REFINE framework first computes the multimodal embedding of the query $\phi(x_{\text{query}})$. Then, from the Neural Error-book R, we retrieve the most relevant image-text sample in the training set to obtain the corresponding structured feedback $\hat{\mathcal{F}}$.

$$\hat{x}, \hat{\mathcal{F}} = \arg \max_{(\phi(x_i), \mathcal{F}_i) \in R} \frac{\phi(x_{\text{query}}) \cdot \phi(x_i)}{\|\phi(x_{\text{query}})\| \|\phi(x_i)\|} \quad (2)$$

The retrieved feedback $\hat{\mathcal{F}}$ is systematically integrated into the student model's prompt to guide its reasoning. Specifically, $\hat{\mathcal{F}}$ is appended to the original question Q_{query}, forming an enhanced prompt P_{enhanced} that combines task context with actionable corrective instructions:

$$P_{\text{enhanced}} = \langle Q_{\text{query}}, \hat{\mathcal{F}} \rangle \quad (3)$$

where $\langle \cdot \rangle$ denotes the structured prompt format. This design ensures the student model processes both the question and feedback in a unified context, directing attention to previously overlooked criteria or reasoning steps.

This deterministic single-nearest-neighbor strategy ensures consistency and low overhead at inference time, in stark contrast to the inefficiencies and stochastic behavior of traditional ICL approaches.

4 Experiment

4.1 Experimental Setup

We evaluated the effectiveness of our feedback mechanisms using three multimodal reasoning benchmarks: **MME-RealWorld** (Zhang et al., 2024c), **MMStar** (Chen et al., 2024), and **SEED-Bench-2-Plus** (Li et al., 2024). These benchmarks collectively emphasize diverse multimodal reasoning capabilities, including error diagnosis, procedural correction, and text-rich visual comprehension. The evaluation framework employed for these benchmarks was VLMEvalKit (Duan et al., 2024). All models in our experiments were used with a temperature setting of 0.0, and results were reported using the pass@1 metric. Specifically:

- **MME-RealWorld** covers complex visual reasoning tasks derived from realistic applications across diverse domains (e.g., Autonomous Driving, Diagrams, OCR).
- **MMStar** explicitly selects reasoning problems requiring multimodal integration, filtering out problems solvable without visual context.
- **SEED-Bench-2-Plus** evaluates text-rich visual comprehension through 2.3K multiple-choice questions spanning Charts, Maps, and Webs. Its scenarios simulate real-world complexity with embedded textual elements, assessing models' capacity to interpret visually grounded textual information.

To verify generalizability and scalability across different model sizes, we selected two representative multimodal models:

- **Pixtral-12B** (Agrawal et al., 2024) (a high-capacity model)
- **Qwen2.5-VL-3B-Instruct** (Bai et al., 2025) (a compact model)

Feedback generation employed Gemini-1.5-Pro (Team et al., 2024) for MME-RealWorld and MMStar, while SEED-Bench-2-Plus utilized Gemini-2.0-Flash (Pichai et al., 2024) due to its specialized text-rich processing capabilities.
We constructed our training and evaluation datasets as follows:

- **MME-RealWorld**: After creating an Errorbook based on MME-RealWorld-Lite (Reasoning) subset, we applied it directly to the remaining MME-RealWorld Reasoning subset, excluding the Lite portion to rigorously test generalization.
- **MMStar**: We divided the total items in the Instance Reasoning and Logical Reasoning categories into two equal parts, with one part forming the Train Set for creating the Errorbook and the remaining items forming a separate Test Set.
- **SEED-Bench-2-Plus**: We utilized the full benchmark dataset (2.3K items) and split it into equal halves, establishing distinct Train and Test sets to evaluate model adaptation to text-rich scenarios while preventing data leakage.

Our structured feedback approach was compared against the following baseline feedback methods:

- **Standard Prompting** (Brown et al., 2020): The LLM is asked to output the answer directly, without the intermediate reasoning process.
- **Chain of Thought** (Wei et al., 2022): The LLM is instructed to think step-by-step before providing the answer.

- **Direct Feedback** (Daheim et al., 2024): unstructured, open-ended feedback without explicit instructional framing.

- **RICP** (Sun et al., 2024): retrieved principles from errors clustered by error type, providing task- and question-specific guidance.

4.2 Results and Analysis

Structured Feedback Outperforms Clustering and Unstructured Approaches The superiority of our method over RICP and Direct Feedback highlights the importance of **task-specific granularity** over cluster-level generalizations. For instance, in MME-RealWorld(Reasoning)'s OCR with Complex Context, Pixtral-12B achieves a 24.25-point gain over Standard Prompting with our method, compared to RICP's marginal improvements. This suggests that cluster-level principles fail to address nuanced errors like misclassifying dynamic objects (e.g., distinguishing "people" vs. "pedestrians" based on pose). Our structured feedback, which explicitly defines task goals (Feed-Target) and actionable corrections (Feed-Path), bridges this gap by contextualizing errors within the specific reasoning process required for the task.

Domain-specific Feedback Efficacy Performance gains vary considerably across domains, revealing feedback-task alignment dynamics:

- **Diagram and Table Interpretation (MME-RealWorld)**: The significant improvement (+23.25 for Pixtral-12B) stems from feedback that clarifies hierarchical relationships (e.g., "Focus on nested chart labels first")—critical for parsing complex diagrams.

- **Logical Reasoning (MMStar)**: Smaller gains (+1.6 for Pixtral-12B) suggest feedback is less effective for abstract reasoning requiring implicit world knowledge (e.g., causality). Here, structured feedback aids factual corrections (e.g., misidentified object relationships) but struggles with higher-order logic gaps.

- **OCR Decline in Smaller Models**: Qwen2.5-VL-3B-Instruct's slight drop in OCR (-1.75) may reflect **feedback overload**: overly granular corrections (e.g., "Recount awning-tricycles after redefining 'people'") could confuse smaller models with limited reasoning depth, leading to overcorrection.

Model Capacity Dictates Feedback Utilization The contrast between Pixtral-12B and Qwen2.5-VL-3B-Instruct underscores **scaling laws for feedback internalization**:

- **Pixtral-12B** leverages structured feedback holistically, excelling in tasks requiring multi-step synthesis (e.g., OCR + counting). Its 58.25 score in OCR reflects an ability to chain corrections: first redefining terms (Feed-Target), then revising counts (Feed-Path).

- **Qwen2.5-VL-3B-Instruct** benefits most in **procedural tasks** (e.g., Instance Reasoning: +5.6) where feedback directly maps to executable steps. However, its performance plateaus in open-ended tasks (Logical Reasoning: +0.0), indicating limited capacity to generalize feedback beyond explicit instructions.

Feedback Type Impact Ablating feedback components would likely reveal hierarchical importance. Notably, Direct Feedback's inconsistent results—sometimes trailing Standard Prompting (e.g., MMStar Logical Reasoning 42.4 vs. 48.8) suggest unstructured feedback introduces noise, confusing the student model with irrelevant or contradictory advice.

Error-Type Correctability Our method excels in correcting **systematic procedural errors** (e.g., misapplying definitions) but is less effective for **knowledge gaps**. For example, in Diagram tasks, feedback like "Prioritize axis labels before interpreting trends" directly resolves a common student error, whereas Logical Reasoning errors (e.g., flawed causality chains) require external knowledge beyond feedback's scope.

Efficiency and Scalability of REFINE Besides accuracy, our method significantly outperformed baselines in terms of inference efficiency. Figure 2 shows that our structured feedback retrieval is substantially faster ($44.7 - 76.4\times$ speedup compared to RICP) and more token-efficient (approximately 64.2% fewer tokens than RICP on MMStar). Precomputed embeddings and the removal of clustering significantly reduce computational costs, demonstrating scalability and feasibility in real-time settings. Furthermore, successful generalization from smaller subsets (MME-RealWorld-Lite) to larger-scale tests (MME-RealWorld) clearly illustrates the practical scalability of our approach.

Method	**MME-RealWorld (Reasoning)**				
	Autonomous Driving	**Diagram & Table**	**Monitoring**	**OCR with Complex Context**	**Overall**
			Pixtral-12B		
Standard Prompting	29.66	27.75	15.80	34.00	27.82
CoT	27.86	28.00	22.70	44.25	30.16
Direct Feedback	33.05	32.00	24.71	40.50	32.89
RICP	29.87	28.25	23.56	44.25	31.26
REFINE	**35.06** (+5.40)	**51.00** (+23.25)	**31.32** (+15.52)	**58.25** (+24.25)	**41.92** (+14.10)
			Qwen2.5-VL-3B-Instruct		
Standard Prompting	21.19	21.75	14.94	40.25	23.90
CoT	27.54	25.50	19.25	**41.75**	28.49
Direct Feedback	29.98	19.75	21.55	35.00	27.58
RICP	19.60	23.00	13.79	39.75	23.14
REFINE	**34.43** (+13.24)	**26.75** (+5.00)	**24.71** (+9.77)	38.50 (-1.75)	**32.12** (+8.22)

Table 1: Performance (Accuracy %) of REFINE against baseline methods on MME-RealWorld (Reasoning) subtasks. Parentheses for REFINE indicate improvement over Standard Prompting.

Method	**MMStar (Reasoning)**			**SEEDBench-2-Plus**			
	Instance	**Logical**	**Overall**	**Chart**	**Map**	**Web**	**Total**
				Pixtral-12B			
Standard Prompting	59.2	48.8	54.0	50.86	50.00	56.67	52.23
CoT	59.2	**53.6**	56.4	49.14	47.04	54.55	49.96
Direct Feedback	60.8	42.4	51.6	47.65	47.29	50.61	48.38
RICP	59.2	48.0	53.6	49.63	46.55	47.27	47.85
REFINE	**64.8** (+5.6)	50.4 (+1.6)	**57.6** (+3.6)	**56.30** (+5.44)	**52.46** (+2.46)	**57.88** (+1.21)	**55.39** (+3.16)
				Qwen2.5-VL-3B-Instruct			
Standard Prompting	64.0	52.0	58.0	**63.57**	53.48	78.48	64.33
CoT	56.0	**54.4**	55.2	57.46	47.51	76.97	59.60
Direct Feedback	63.2	52.0	57.6	61.86	52.99	**78.79**	63.63
RICP	62.4	52.8	57.6	62.59	53.23	73.33	62.40
REFINE	**69.6** (+5.6)	52.0 (+0.0)	**60.8** (+2.8)	62.84 (-0.73)	**54.98** (+1.50)	78.18 (-0.30)	**64.50** (+0.17)

Table 2: Performance (Accuracy %) of REFINE against baseline methods on MMStar (Reasoning) and SEEDBench-2-Plus benchmarks. Parentheses for REFINE indicate improvement over Standard Prompting.

4.3 Ablation Study: Analyzing Feedback Component Contributions

Feedback Components We conducted an ablation study by adding Cluster-level feedback and CoT.

- **Cluster-level feedback**: To assess generalized feedback via clustering (cf. RICP (Sun et al., 2024)), we applied K-means (k=5) to the multimodal embeddings of training instances. For each data cluster, our teacher model generated a single generalized feedback from 20 samples of task/process feedback within that cluster. At inference, a query was assigned to its nearest cluster (by embedding similarity), combining this generalized feedback with REFINE's Task/Process feedback.
- **CoT** (Wei et al., 2022): Standard Chain-of-Thought prompting ('Let's think step by step.') was appended after REFINE's Task/Process feedback to elicit explicit reasoning from the student model, rather than altering the feedback content itself.
- **Self-Reg** (Hattie and Timperley, 2007): Self-regulatory feedback, normally filtered out (Section 3.2), was re-introduced alongside Task/Process feedback.

To assess the impact of individual feedback

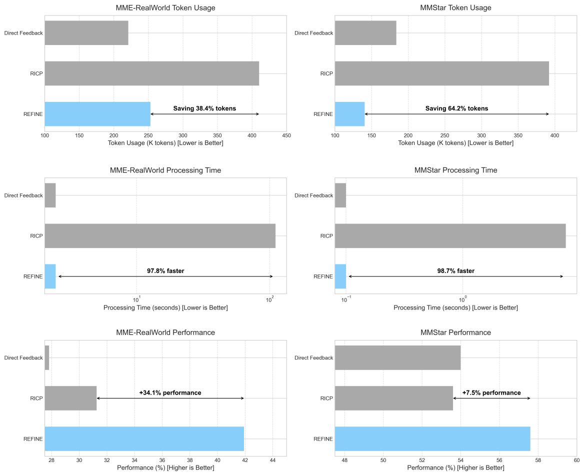


Figure 2: Performance (%) of REFINE and baseline methods versus Token Usage and Processing Time.

Configuration	**Overall Score**	**∆ from Baseline**
Task/Process (Baseline)	41.92	—
+ Self-Reg	32.50	−9.42 (−22.5%)
+ Cluster-Level	32.41	−9.51 (−22.7%)
+ CoT	32.17	−9.75 (−23.3%)
+ Self-Reg + Cluster-Level	32.07	−9.85 (−23.5%)

Table 3: Ablation study of REFINE's feedback components on MME-RealWorld (Overall Accuracy %, Pixtral-12B). Shows performance impact when adding components to REFINE's core Task/Process feedback.

types, we performed an ablation study (Table 3). The best overall performance (41.92) was achieved using Task/Process-level feedback exclusively, whereas the incorporation of additional feedback types consistently reduced model accuracy.

Specificity vs. Generalization Trade-off **Task/Process feedback** excels due to its direct alignment with the error context. For instance, advising the model to "recount standing/walking poses" directly addresses the miscounting error. But **Cluster-level feedback**, while intended to generalize insights, likely introduces noise. For example, feedback like "check definitions in counting tasks" may lack the precision needed for a **specific** question about "people vs pedestrians," leading to ambiguous guidance.

The 22.7% drop with Cluster-level feedback suggests that broad advice cannot substitute context-specific corrections.

Cognitive Overload in Multimodal Reasoning Adding **Self-Regulatory Feedback** (e.g., "Reflect on past errors") forces the model to split attention between executing the task and metacognitive monitoring, a challenge for vision-language models unoptimized for dual-task learning.

CoT exacerbates this by introducing open-ended reasoning steps (e.g., "Think about object definitions") that conflict with the structured corrective feedback. The 23.3% drop with CoT highlights incompatibility between exploratory and directive instructions.

Compounding Noise in Combined Feedback The worst performance (-23.5%) occurs when combining **Self-Regulatory + Cluster-level Feedback**. This suggests regulatory interference: the model receives vague Self-Regulatory cues *and* overgeneralized cluster advice, diluting the actionable signal. For example, a prompt mixing "Review past errors" (Self-Regulatory) and "Adjust counting strategies" (Cluster) provides no concrete steps to correct a specific miscount.

Embedding-Clustering Mismatch The reliance on embeddings for clustering raises questions: if the embedding space fails to capture fine-grained task nuances (e.g., subtle differences in "people" definitions), clusters may group dissimilar errors, leading to irrelevant feedback retrieval.

Why Task/Process Feedback Works The success of Task/Process feedback aligns with principles of *instructional alignment*:

- **Goal-Oriented**: Directly answers "What is needed to correct *this* error?"
- **Procedural Clarity**: Provides stepwise actions (e.g., "Reread the question, then recount").
- **Minimal Abstraction**: Avoids meta-commentary, reducing cognitive load.

This approach mirrors effective human tutoring, where immediate, task-focused corrections yield better learning outcomes than abstract or generalized advice, a finding now validated for multimodal AI systems.

5 Conclusion

We introduced REFINE, a teacher-student framework that systematically structures errors to deliver targeted feedback for multimodal reasoning. REFINE employs three structured queries—Feed-Target, Feed-Check, and Feed-Path—to prioritize visual information, diagnose failures, and guide corrective actions. By optimizing structured feedback retrieval, unlike methods reliant on redundant retrievals, REFINE significantly improves inference speed, token efficiency, and scalability. Empirical results confirm substantial performance gains and robust generalization, highlighting REFINE's effectiveness in advancing multimodal reasoning.

Limitations

The framework's effectiveness relies on the quality of teacher-generated feedback and the diversity of errors used in creating the Neural Error-book. Future research should explore its application to errors needing complex knowledge synthesis, its generalization to entirely new task domains, and further tailoring of feedback for varied model capacities.

Ethics Statement

This research adhered to the ACL Ethics Policy, utilizing only publicly available datasets and large language models for evaluation. The work's goal is to investigate methods for enhancing reasoning, and no negative ethical outcomes are anticipated.

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, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv: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, and 8 others. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. 2023. Probabilistic Tree-of-thought Reasoning for Answering Knowledge-intensive Complex Questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12541–12560, Singapore. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

D. R. Chialvo and P. Bak. 1999. Learning from mistakes. *Neuroscience*, 90(4):1137–1148.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.

Tim R. Davidson, Benoit Seguin, Enrico Bacis, Cesar Ilharco, and Hamza Harkous. 2025. Orchestrating Synthetic Data with Reasoning. In *Will Synthetic Data Finally Solve the Data Access Problem?*

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. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Amy C. Edmondson. 1996. Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *The Journal of Applied Behavioral Science*, 32:28 – 5.

Amy C. Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44:350 – 383.

Reza Esfandiarpoor and Stephen H Bach. 2024. Follow-up differential descriptions: Language models resolve ambiguities for image classification. In *ICLR*.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. 2022. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906, Dublin, Ireland. Association for Computational Linguistics.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2023. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *The Twelfth International Conference on Learning Representations*.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 28798–28827. Curran Associates, Inc.

Gregory Kang Ruey Lau, Hieu Dao, and Bryan Kian Hsiang Low. 2025. Uncertainty quantification for MLLMs. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. *Preprint*, arXiv:2404.16790.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. Introducing Gemini 2.0: Our new AI model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118807–118833. Curran Associates, Inc.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *ICLR*.

Varun Nagaraj Rao, Siddharth Choudhary, Aditya Deshpande, Ravi Kumar Satzoda, and Srikar Appalaraju. 2024. RAVEN: Multitask Retrieval Augmented Vision-Language Learning.

Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. 2023. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15746–15757.

Hao Sun, Yong Jiang, Bo Wang, Yingyan Hou, Yan Zhang, Pengjun Xie, and Fei Huang. 2024. Retrieved in-context principles from previous mistakes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8155–8169, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can LLMs Learn from Previous Mistakes? Investigating LLMs' Errors to Boost for Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3065–3080, Bangkok, Thailand. Association for Computational Linguistics.

VoyageAI. 2024. Voyage-multimodal-3: All-in-one embedding model for interleaved text, images, and screenshots. https://blog.voyageai.com/2024/11/12/voyage-multimodal-3/.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In *The Thirteenth International Conference on Learning Representations*.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2024a. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zhang, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. 2024b. In-Context Principle Learning from Mistakes. In *ICML 2024 Workshop on In-Context Learning*.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. 2024c. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? In *The Thirteenth International Conference on Learning Representations*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations*.

A Dataset Statistics

Benchmark	Error-book Size	Test Size
MME-RealWorld (Reasoning)	534	2,092
MMStar (Reasoning)	123	250
SEED-Bench-2-Plus	563	1,141

Table 4: Dataset split statistics for Error-book construction and testing.

The Neural Error-book for each benchmark was constructed using only the questions the student model answered incorrectly during the training phase.

For the MME-RealWorld benchmark, the Error-book was created based on errors from the MME-RealWorld-Lite subset.

For MMStar and SEED-Bench-2-Plus, the datasets were split into equal halves to create distinct train and test sets. The Error-book for these benchmarks was then built using the incorrectly answered questions from their respective 50% train sets.

B REFINE Algorithm Detail

Algorithm 1 REFINE

```
Require: Image-question pairs $D$, Teacher model $T$, Student model $S$
Ensure: Enhanced predictions
Note: $T$ is a Teacher model generating three-stage structured feedback (Feed-Target/Check/Path).
Note: FeedbackFilter filters to retain only actionable task/process-level feedbacks.
Stage 1: Error Recognition and Insight Generation
$D_{neg} \leftarrow \emptyset$
for each $(q, img, ans_{gt}, ans_{pred}) \in D$ where $ans_{pred} \neq ans_{gt}$ do
    $F_{target} \leftarrow T(q, img, ans_{pred},$ "Define learning goals"$)$
    $F_{check} \leftarrow T(q, img, ans_{pred},$ "Check current progress"$)$
    $F_{path} \leftarrow T(q, img, ans_{pred},$ "Plan next steps"$)$
    $F_{task/process} \leftarrow$ FeedbackFilter$(F_{target}, F_{check}, F_{path})$
    $D_{neg} \leftarrow D_{neg} \cup \{(q, img, F_{task/proces})\}$
end for
Stage 2: Feedback Construction
for each $(q, img, F_{task/process}) \in D_{neg}$ do
    $e_{q,img} \leftarrow$ Embedding$(q, img)$
    $R \leftarrow R \cup \{(e_{q,img}, F_{task/process})\}$
end for
Stage 3: Inference
for each test pair $(q_{test}, img_{test})$ do
    $e_{test} \leftarrow$ Embedding$(q_{test}, img_{test})$
    $F_{similar} \leftarrow$ NearestNeighbor$(e_{test}, R)$
    $ans \leftarrow S(q_{test} + F_{similar}, img_{test})$
end for
```

C Prompt for REFINE

Prompt For Feed-Target

Evaluate the student's response to a given question by reviewing their answer and providing feedback to reinforce the learning objective.

Consider the scope and importance of the concept or skill that the student should grasp. Emphasize the relevance of this understanding in the broader context of the subject matter.

# Steps

1. Review the original question and both the correct and incorrect answers provided by the student.
2. Identify the core learning objective or skill that the question aims to teach.
3. Clarify why this concept or skill is significant in the curriculum.
4. Relate the concept to the broader subject to illustrate its importance.

# Output Format

Respond in a single sentence that concisely reiterates the primary learning goal, emphasizes its significance, and indicates what successful mastery looks like.

# Notes

- Ensure the feedback is encouraging and constructive, aiding student comprehension and motivation.
- Tailor the response to align with educational standards and learning objectives specific to the curriculum.

IMPORTANT: DO NOT directly mention the correct answer or the corresponding options.

```
Question: {A specific question from the dataset}
Correct Answer: {The ground-truth answer for the question}
Incorrect student's Answer: {The model's incorrect answer}
```

Prompt For Feed-Check

Review the student's solution or explanation and evaluate the correctness of their approach by identifying correct elements, areas showing partial understanding, and any misunderstandings or errors.

Focus on pinpointing both accurate and inaccurate components, clearly articulating any misunderstandings such as conceptual errors, calculation slips, or instruction misinterpretations. Use precise language to highlight the analysis, such as noting specific steps where mistakes occur.

# Output Format

Provide your evaluation in a single, clear, and specific sentence that encompasses both strengths and misconceptions in the student's response.

# Notes

- Ensure to maintain the single-sentence structure for clarity and brevity.
- Clearly separate each element of the evaluation (correct parts, partial understanding, errors) within the sentence.
- Emphasize spotting both strengths and areas in need of improvement.

IMPORTANT: DO NOT directly mention the correct answer or the corresponding options.

```
Question: {A specific question from the dataset}
Correct Answer: {The ground-truth answer for the question}
Incorrect student's Answer: {The model's incorrect answer}
```

Prompt For Feed-Path

Review the question, the correct answer, and the incorrect answer provided by the model, then respond with actionable strategies to guide the student in correcting their misunderstanding and improving their understanding.

Offer both short-term and long-term steps, focusing on concrete, constructive actions that can help the student progress.

- Encourage actionable, bite-sized advice.
- Suggest multiple methods for approaching the subject.
- Maintain a supportive tone, emphasizing that mistakes are a learning opportunity.

# Output Format

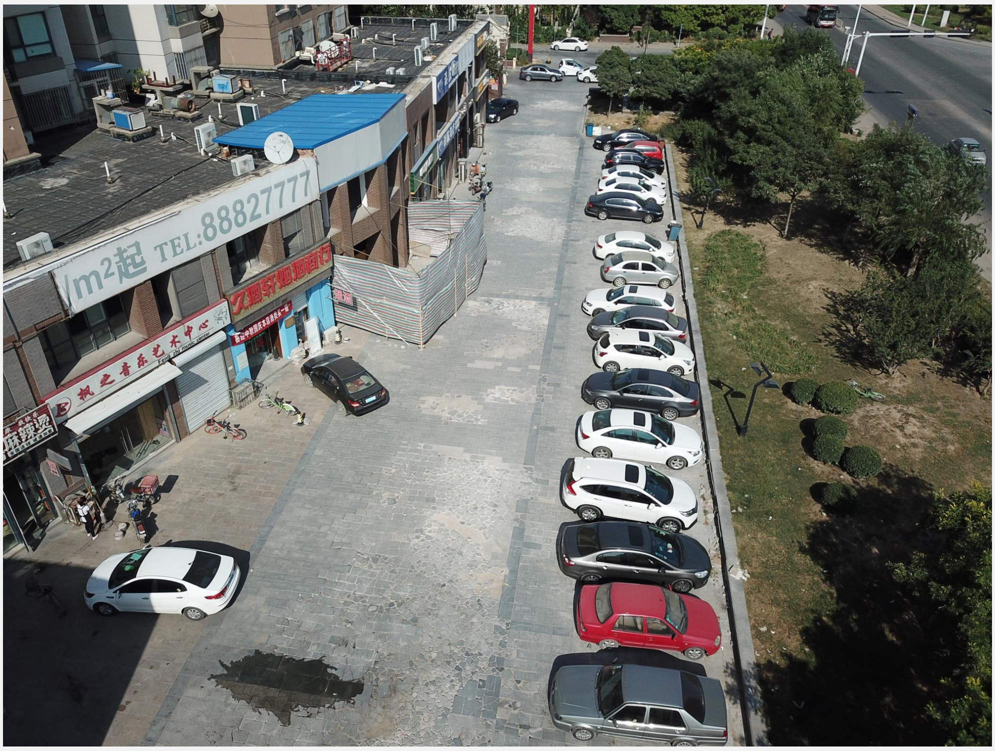
Provide your response in a single sentence, integrating specific actions the student can take to address their misunderstanding, and strategies to build on their correct reasoning and fix their errors.

IMPORTANT: DO NOT directly mention the correct answer or the corresponding options.

```
Question: {A specific question from the dataset}
Correct Answer: {The ground-truth answer for the question}
Incorrect student’s Answer: {The model’s incorrect answer}
```

D Detailed Example of REFINE

REFINE in Action



Question:

What is the total number of people and awning-tricycles in the image? (If a human maintains standing pose or walking, please classify it as pedestrian, otherwise, it is classified as a people.)
The choices are listed below:
(A) 15
(B) 4
(C) 0
(D) 3
(E) The image does not feature the objects

Model Response:

The image shows a street scene with parked cars, buildings, and a few pedestrians. There are no awning-tricycles visible in the image. To determine the total number of people and awning-tricycles, we need to count the number of people in the image. Upon examining the image, **we can see a few pedestrians walking on the sidewalk. Specifically, there are three people visible** in the image. Since there are no awning-tricycles in the image, the total number of people and awning-tricycles is simply the number of people, which is 3.
Final Answer: D [WRONG]

Retrieved Task/Process-level Feedback (from a similar past error):

During solving the solution, refer to the following information:
The student correctly identifies two awning-tricycles, but misinterprets the definition of "people," counting 10 individuals when only two meet the specified criteria of not being in a standing or walking pose, leading to an incorrect total of 12 instead of 2. **Reread the question carefully, noting the specific definition of "people" provided, then review the image focusing on individuals in standing or walking poses** and awning-tricycles, and finally recount these specific objects to arrive at the correct total.

REFINE Response:

The image shows a street scene with parked cars, buildings, and a few people. There are no awning-tricycles visible in the image.
Let's analyze the image step-by-step:

1. Identify the objects in the image: The image shows a street with parked cars, a few people, and buildings. There are no awning-tricycles visible.
2. Count the number of people: There are a few people visible in the image, **but we need to count only those who are not in a standing or walking pose.**
3. Count the number of awning-tricycles: There are no awning-tricycles visible in the image.

Based on the analysis, the total number of people and awning-tricycles in the image is 0.
Final Answer: C [CORRECT]