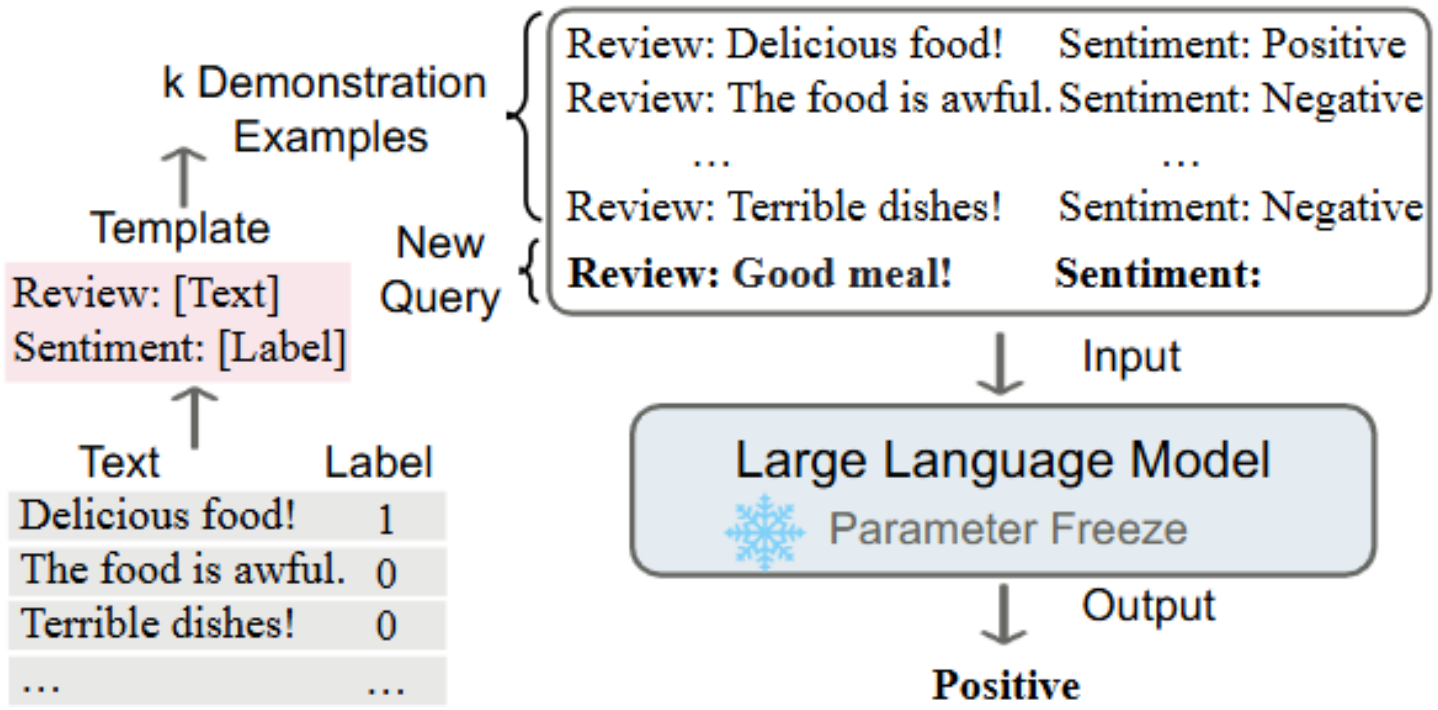
Incomplete In-context Learning

Wenqiang Wang
Sun Yat-sen University
wangwq69@mail2.sysu.edu.cn

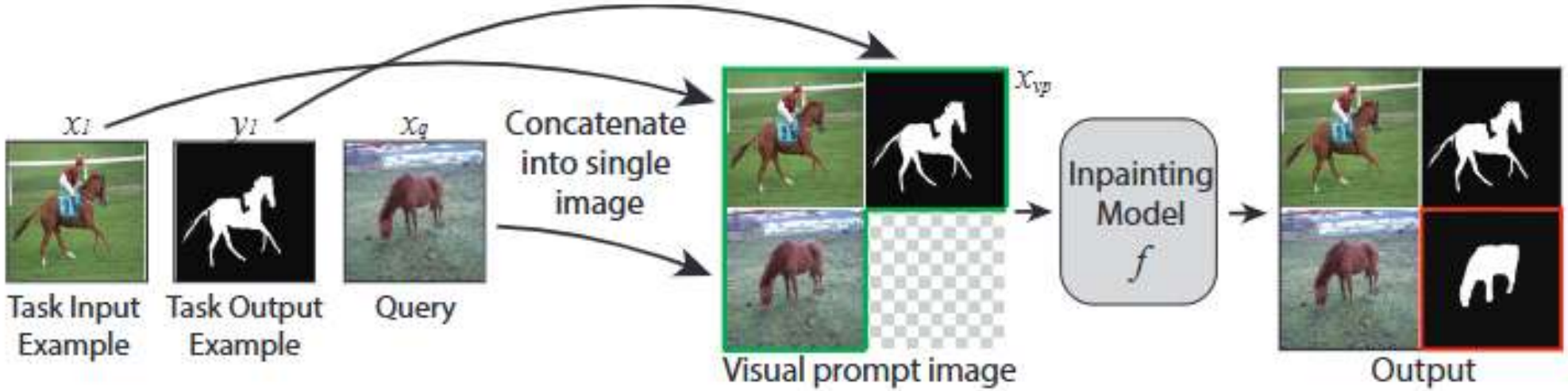
Yangshijie Zhang
Lanzhou University
zhangyshj2023@lzu.edu.cn

arXiv 2025

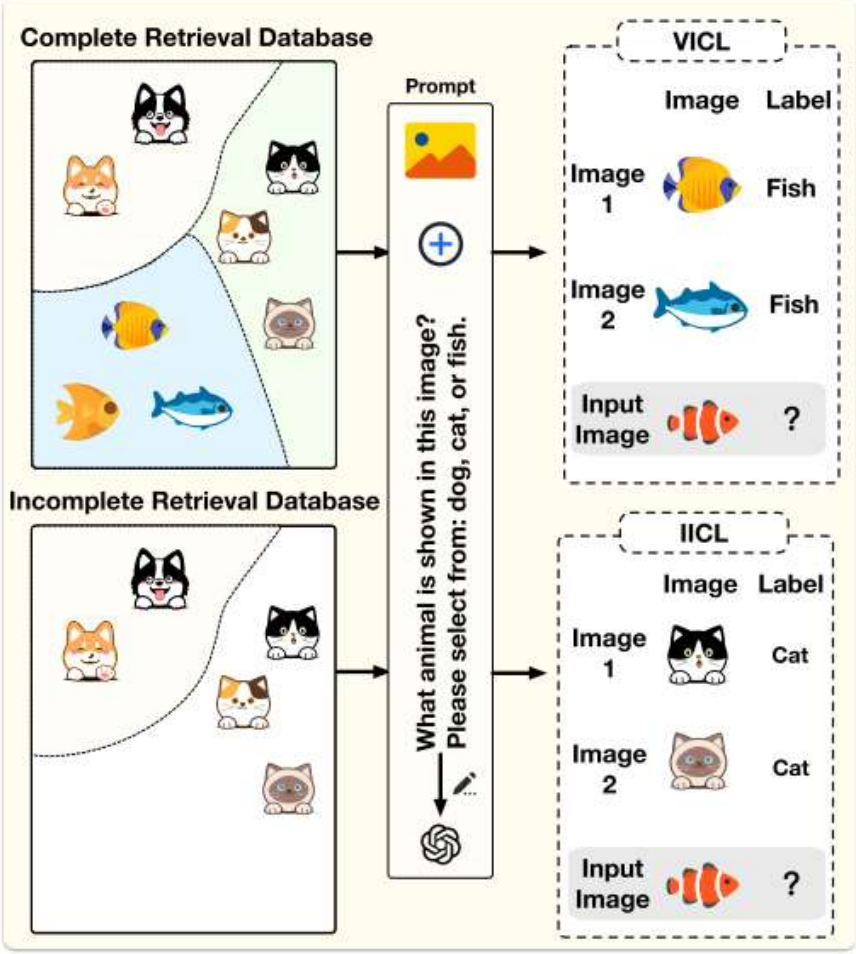
In-context Learning



Visual In-context Learning

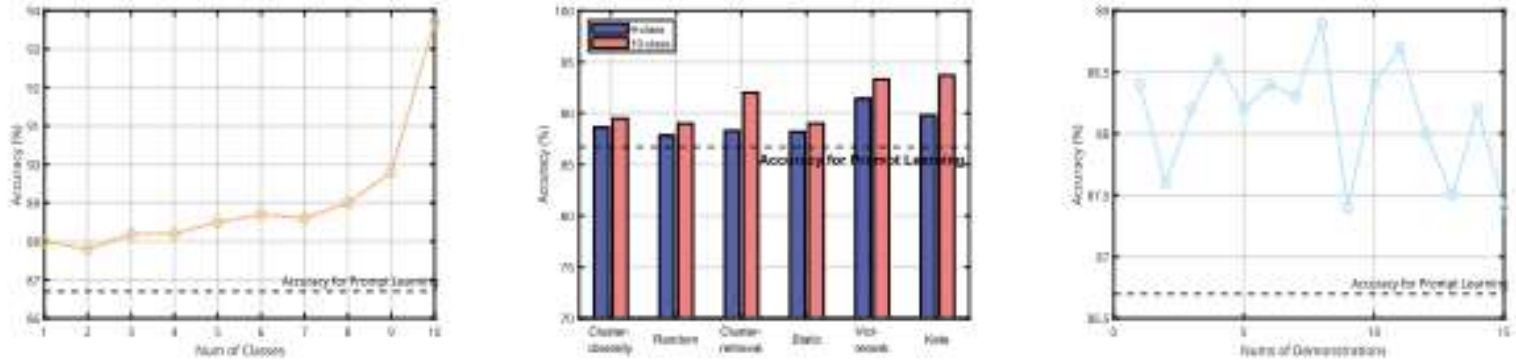


Incomplete In-context Learning



(a) Comparison of complete vs incomplete retrieval databases and VICL vs IICL scenarios. In the *incomplete retrieval database*, "fish" images are absent; thus, IICL fails to retrieve suitable demonstrations but retrieves "cat" images when the input is "fish", limiting IICL's performance.

Empirical Study of IICL

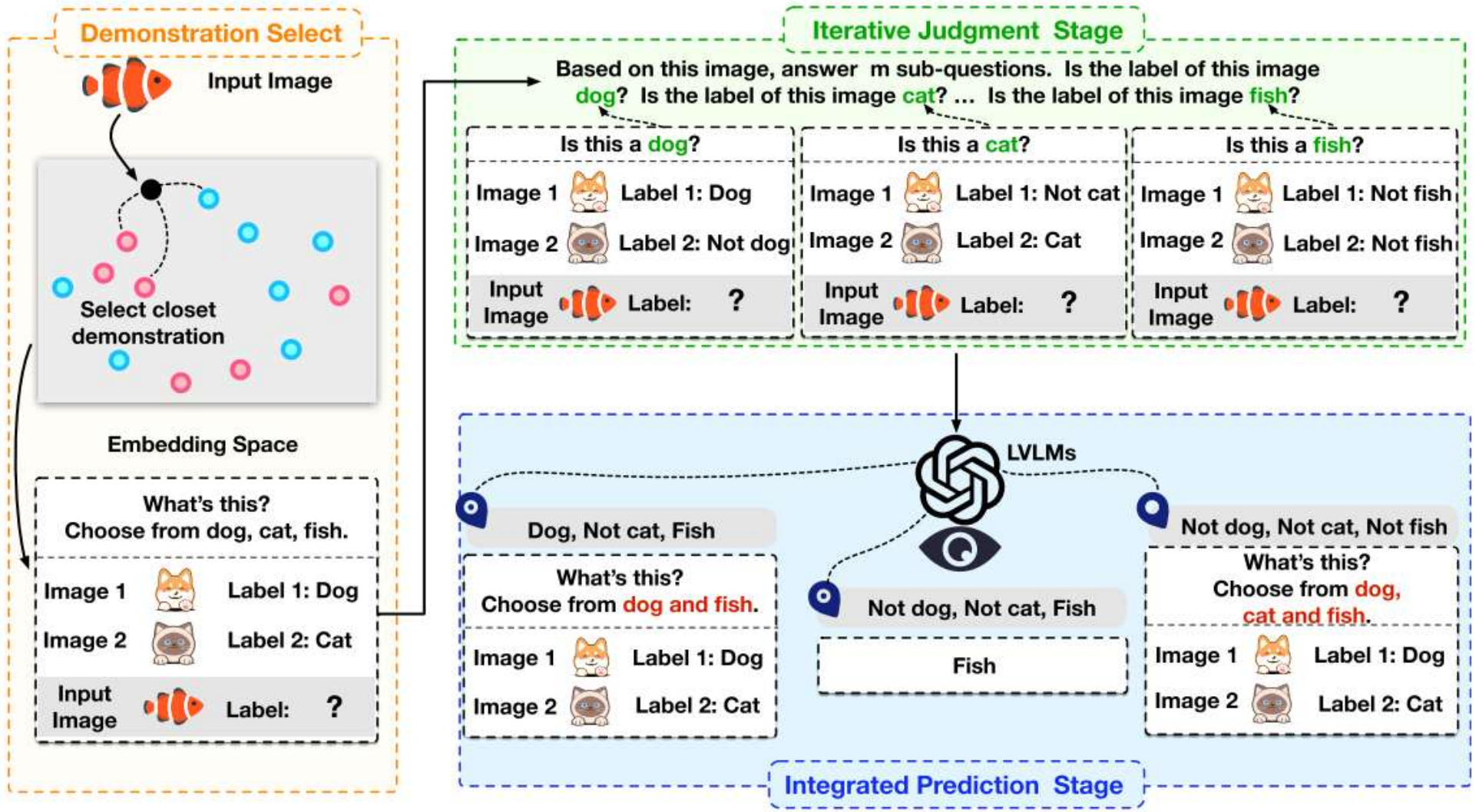


(a) (b) (c)

Figure 3. Subfigure (a) shows the empirical study of IICL with different missing label numbers. Subfigure (b) shows the empirical study of IICL with different VICL methods. Subfigure (c) shows the empirical study of IICL with different demonstration numbers.

Method

Overview



Iterative Judgment Stage

Retrieve k labeled images from the incomplete retrieval database that are most semantically similar to the input image and use them as demonstrations.

$$\mathbf{e}_{\mathbf{x}} = f_{\text{pre}}(\mathbf{x}), \mathbf{e}_i = f_{\text{pre}}(\mathbf{x}_i), \mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}. \quad (7)$$

$$s_i = \frac{\mathbf{e}_{\mathbf{x}} \cdot \mathbf{e}_i}{\|\mathbf{e}_{\mathbf{x}}\| \cdot \|\mathbf{e}_i\|}, \mathcal{S} = \{s_1, s_2, \cdots, s_n\}. \quad (8)$$

$$\begin{gathered} \mathcal{I} = \arg\operatorname{top}^k_{(\mathbf{x}_i, \mathbf{y}_i) \subseteq \mathcal{D}} \mathcal{S}, \\ \mathcal{D}_d = \{\mathbf{x}_i^d, y_i^d\}_{i=1}^k, y_i^d \in \{C_1, C_2, \cdots, C_w\}, \text{for } i \in \mathcal{I}, \end{gathered} \quad (9)$$

The m-class classification problem is transformed into a sequence of binary classification tasks, with each task corresponding to one of the m subclassification problems. In the j-th sub-classification task, the goal is to determine whether the label of the given image is C_j or $\overline{C_j}$.

Integrated Prediction Stage

Specifically, we first determine the number of predictions among $\{\hat{y}_{\mathbf{x}}^1, \hat{y}_{\mathbf{x}}^2, \cdots, \hat{y}_{\mathbf{x}}^m\}$

In the **Integrated Prediction Stage**, we refine the classification of the input image by leveraging both the image itself and the predictions obtained from the *Iterative Judgments Stage*. The final decision-making process falls into one of three distinct scenarios. ❶ All predictions are negative $(\overline{C_1}, \overline{C_2}, \ldots, \overline{C_m})$. In this case, no class has been positively identified. We directly perform m-label VICL for classification. ❷ Exactly one positive prediction is made (e.g., $\overline{C_1}, \overline{C_2}, \ldots, C_j, \ldots, \overline{C_m}$). The image is classified as C_j, and we assign C_j as the final predicted label. ❸ Multiple positive predictions are made (e.g., $\overline{C_1}, \overline{C_2}, \cdots, C_{j-1}, C_j, C_{j+1}, \cdots, \overline{C_m}$). Since multiple candidate labels exist, we perform an additional in-context learning classification among the positively predicted labels. For instance, if the predictions suggest the image could belong to C_{j-1}, C_j, or C_{j+1}, we refine the classification by treating it as a **three-class classification problem**.

$$\mathbb{I}_{\mathbf{x}}^j = \begin{cases} 1, & \text{if } \hat{y}^j = C_j, \\ 0, & \text{otherwise} \end{cases} \qquad \mathbb{I}_{\mathbf{x}} = \sum_{i=1}^{m} \mathbb{I}_{\mathbf{x}}^j, \tag{10}$$

$$\hat{y}_{\mathbf{x}} = \begin{cases} f_{\text{LVLM}}^m(\mathcal{D}_d, \mathbf{x}), & \text{if } \mathbb{I}_{\mathbf{x}} = 0, \\ C_j, s.t. \mathbb{I}_{\mathbf{x}}^j = 1 & \text{if } \mathbb{I}_{\mathbf{x}} = 1, \\ f_{\text{LVLM}}^u(\mathcal{D}_d, \mathbf{x}), & \text{if } \mathbb{I}_{\mathbf{x}} = u, 1 < u \leq m, \end{cases} \tag{11}$$

Experimental Setups

Datasets： conduct experiments using the CIFAR-10 and Fashion-MNIST datasets.

Metrics：employ accuracy as the metric, with higher accuracy reflecting greater performance.

Baselines：perform experiments using various ICL and VICL methods, including Static, Random, Clusteringretrieval, Kate, Cluster-Diversity, and Viclrerank.

Vision-language models and other setup：The Vision language model chosen for their experiments is InternVL2.5-8B and InternVL 2.5-4B. The pre-trained model employed in this study is CLIP (ViT-Base-Patch32) . Each experiment is conducted three times, and the average result is reported.

Comparsion IJIP with other VICL methods.

Table 2. Comparsion IJIP with other VICL methods. We evaluate performance primarily using accuracy(%)↑ under varying proportions of missing labels. The experiments are conducted with label missingness levels of 10%, 40%, and 90%. **Each experiment is conducted three times, and the average result is reported.**

Data	Methods	InternVL 2.5-8B			InternVL 2.5-4B		
		90%	40%	10%	90%	40%	10%
CIFAR-10	Cluster-diversity	86.8	88.7	88.3	78.7	80.4	83.2
	Random	87.4	88.9	88.6	79.8	81.7	83.7
	Cluster-retrieval	88.0	88.7	89.8	75.6	85.4	87.7
	Static	86.7	87.8	87.8	76.3	82.2	84.3
	Vicl-rerank	88.1	88.9	88.2	74.0	80.0	90.8
	Kate	88.0	88.7	89.8	75.0	80.3	90.1
	IJIP	**89.2**	**91.5**	**92.3**	**88.5**	**88.8**	**93.9**
Fashion-MNIST	Cluster-diversity	52.4	53.8	54.0	52.6	54.2	52.9
	Random	52.2	55.7	54.5	51.8	54.1	53.3
	Cluster-retrieval	46.6	56.6	57.8	46.7	51.6	49.6
	Static	49.9	53.1	54.6	51.4	52.8	50.3
	Vicl-rerank	43.2	60.9	75.7	41.1	61.3	72.4
	Kate	42.3	60.7	74.6	42.9	63.3	73.6
	IJIP	**52.8**	**68.9**	**78.9**	47.5	**64.1**	**77.4**

Effects of Different Demonstration Numbers

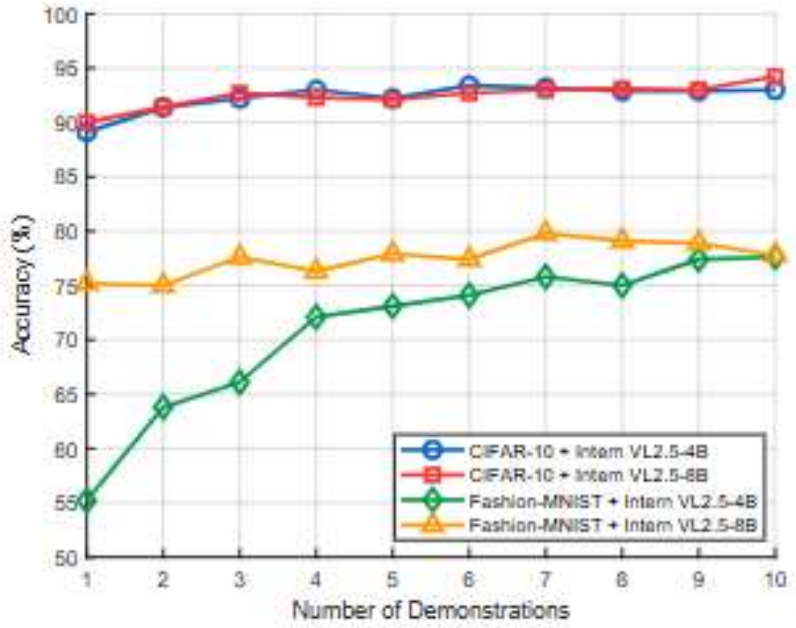


Figure 4. The accuracies(%)↑ of different demonstration numbers. Each experiment is conducted three times, and the average result is reported.

Effects of Different Missing labels

Table 6. Performance of IJIP using InternVL 2.5-4B and InternVL 2.5-8B on CIFAR-10 and Fashion-MNIST under varying proportions of missing labels.

Class Nums	CIFAR10		Fashion MNIST	
	4B	8B	4B	8B
90%	88.5	89.9	47.5	52.8
80%	88.5	88.5	48.0	55.2
70%	88.2	88.6	50.0	59.0
60%	89.4	89.0	52.4	61.0
50%	87.8	88.3	56.4	63.8
40%	88.8	91.5	64.1	68.9
30%	90.5	90.8	67.9	72.7
20%	90.0	92.6	71.9	72.2
10%	92.1	92.3	77.4	78.9

Effects of Different LVLMs' sizes

Table 7. The performance of IJIP with different LVLMs' sizes. We use InternVL 2.5-2B, InternVL 2.5-4B, InternVL 2.5-8B, and InternVL 2.5-26B. We evaluate performance primarily using accuracy(%)↑ under varying proportions of missing labels. The experiments are conducted with label missingness levels of 10%, 40%, and 90%.

LVLMs	90%	40%	10%	Average
2B	87.2	87.5	90.6	88.4
4B	88.5	88.8	93.9	90.4
8B	89.9	91.5	92.3	91.2
26B	95.4	98.1	98.3	97.3

Expanding IJIP from ICL to Prompt Learning

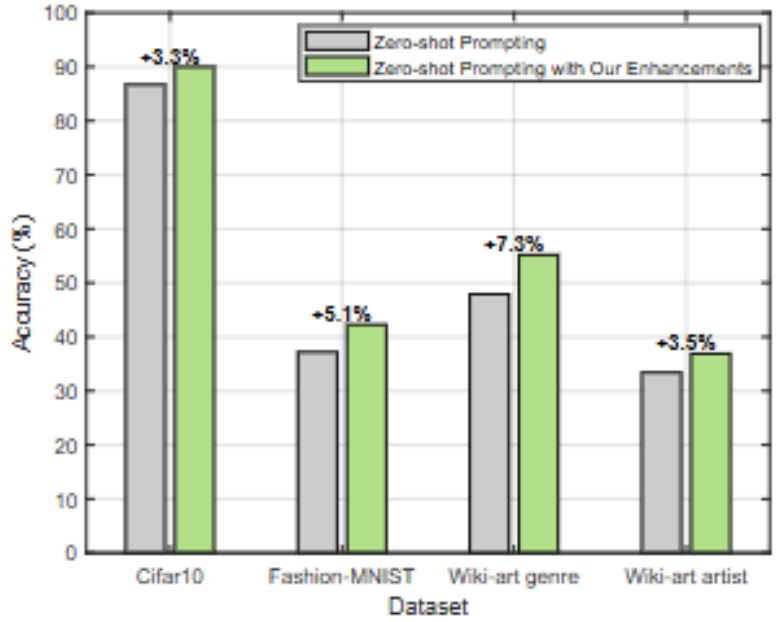


Figure 5. The effectiveness of IJIP in improving zero-shot prompt learning performance. Each experiment is conducted three times, and the average result is reported.

Thanks