IIR-VLM: IN-CONTEXT INSTANCE-LEVEL RECOGNITION FOR LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instance-level recognition (ILR) concerns distinguishing individual instances from one another, with person re-identification as a prominent example. Despite the impressive visual perception capabilities of modern VLMs, we find their performance on ILR unsatisfactory, often dramatically underperforming domain-specific ILR models. This limitation hinders many practical applications of VLMs, e.g., where recognizing familiar people and objects is crucial for effective visual understanding. We attribute this limitation in part to the VLM's general-purpose visual representation, which is not explicitly trained to distinguish visually similar instances. To overcome this challenge, we propose IIR-VLM, a VLM enhanced for In-context Instance-level Recognition. We integrate pretrained ILR expert models as auxiliary visual encoders to provide highly instance-discriminative features complementary to those of the VLM's general-purpose encoder. Through a lightweight two-stage training process, IIR-VLM aligns the expert encoder with the base model, learns to identify new instances in-context, and leverages this knowledge for instance-aware visual understanding. Experiments confirm the benefit of instance-discriminative visual pretraining, particularly for visually similar instances. On IIR-Bench, our challenging new benchmark spanning varying difficulty levels and diverse categories including persons, faces, pets, and general objects, IIR-VLM substantially outperforms strong VLM baselines and prior specialized methods.

1 INTRODUCTION

Instance-level Recognition (ILR) (Ypsilantis et al., 2021; Radenović et al., 2018; Weyand et al., 2020; Kordopatis-Zilos et al., 2025; Ypsilantis et al., 2023; Oh Song et al., 2016; Ye et al., 2021) requires models to distinguish specific instances rather than general categories. A prominent example is person re-identification (Ye et al., 2021; He et al., 2021; Niu et al., 2025), which focuses on identifying the same person across images from different cameras or in new environments. This challenge of fine-grained matching naturally extends to other categories, such as recognizing specific pets (Shinoda & Shiohara, 2024), vehicles (He et al., 2021), faces (Cao et al., 2018), or general objects (Kordopatis-Zilos et al., 2025; Ypsilantis et al., 2023; Oh Song et al., 2016). Together, these tasks illustrate a fundamental demand of visual perception, which is to differentiate specific instances across diverse domains.

Despite the impressive visual understanding capabilities of large vision-language models (VLMs), we find that they struggle with instance-level recognition. When prompted to identify the same instance from a small gallery of images, even state-of-the-art VLMs often fail to select the correct instance, especially when the gallery images exhibit high visual similarity. A wide range of real-world applications can be affected by such an inability. For example, users of a smart-home camera may expect their VLM to reliably distinguish between family members, or tell their own pet apart from their neighbor's similar-looking pet. Instance-level recognition capabilities must serve as a cornerstone of such personalization features.

We attribute the failure of VLMs on ILR to limitations in visual perception. Prior work shows that the pretext task of a visual encoder determines its suitability for different VLM tasks, *e.g.*, language-supervised encoders perform far better on OCR (Tong et al., 2024a). Yet the VLM's general-purpose visual encoder, often trained by language-image contrastive learning (Radford et al., 2021), is not

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.

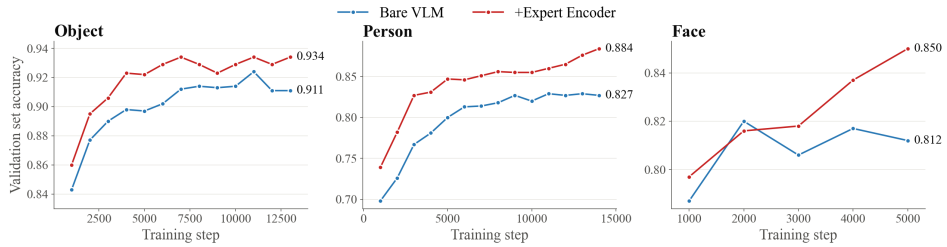

Figure 1: **Adding an instance-discriminative expert encoder consistently lifts VLM performance on instance-level recognition tasks.** Validation curves during training on instance-level recognition tasks, with and without the expert encoder.

optimized to separate visually similar instances. We argue that ILR, much like OCR, is a highly specialized task that relies on fine-grained, instance-specific visual cues, and therefore calls for an encoder explicitly trained on instance discrimination.

To meet this need, we create In-context Instance-level Recognition VLM, or IIR-VLM, which learns and recognizes new instances on-the-fly. Our framework achieves this with two primary components. First, we elevate the VLM's perception by integrating an instance-level recognition expert model as an auxiliary visual encoder. Through an attention-based adapter, we fuse the highly instance-discriminative features from the expert with those from the VLM's original, general-purpose visual encoder, which consistently helps the VLM learn instance-level recognition faster and converge to higher accuracy (Fig. 1).

Second, we introduce a two-stage, lightweight training process. We repurpose existing ILR datasets of diverse categories into an instruction-tuning format, where each task consists of a query and a small image gallery, essentially asking the model to learn instances from the gallery in-context. In Stage 1, the model is trained simply to identify the matching instance from the gallery, which aligns the expert features with the VLM. While Stage 1 establishes instance matching, practical VLM applications require the model to leverage recognized identities in broader visual understanding tasks. Stage 2 therefore trains the model to leverage this instance-level knowledge for visual understanding, with captioning as the central task. By supervising the instance label and the caption separately, Stage 2 preserves the matching capability of Stage 1 while resuming the natural language outputs of the VLM. This paradigm allows the VLM to identify and reason about any new instance provided in-context without fine-tuning for individual instances (Alaluf et al., 2024; Nguyen et al., 2024).

Extensive experiments validate our IIR-VLM's capabilities. On IIR-Bench, our benchmark of challenging ILR tasks spanning multiple difficulty levels and categories, IIR-VLM outperforms all open-source baselines and prior specialized methods, and surpasses Gemini 2.5 Pro and Gemini 3.1 Pro with a 3B base model. It also achieves the highest average accuracy on MM-ID (Ji et al., 2024), an unseen domain of instance identification in movie scenes. Through a controlled comparison with the general-purpose encoder our expert is initialized from, we show that the gain comes from the instance-discriminating training objective rather than additional features or parameters. The advantage of the expert also widens as the tasks involve more visually similar instances. Together, these results demonstrate that IIR-VLM can reliably recognize new instances in-context and leverage their identities for personalized visual understanding.

We summarize our contributions as follows:

(1) We show that an encoder pre-trained on instance-discriminating objectives complements the VLM's general-purpose encoder. We integrate it as an auxiliary visual encoder with a simple and effective method for elevating the instance-level recognition capabilities of VLMs.

(2) Based on the new architecture, we train IIR-VLM to perform In-context Instance-level Recognition through a lightweight, two-stage fine-tuning process, allowing VLMs to reliably learn and identify new instances on-the-fly.

(3) We establish IIR-Bench, a comprehensive instance-level recognition benchmark for VLMs, spanning varying difficulty and categories. On IIR-Bench, our IIR-VLM outperforms strong baselines and prior specialized methods in both instance matching and instance-aware visual understanding.

2 Related Work

Instance-level recognition and retrieval. Instance-level recognition (ILR) tasks (Weyand et al., 2020; Ypsilantis et al., 2021; Radenović et al., 2018; Kordopatis-Zilos et al., 2025; Ypsilantis et al., 2023) require models to identify specific instances from visually similar candidates. The most prominent example is person re-identification (Re-ID) (Ye et al., 2021; He et al., 2021; Zheng et al., 2015; Yang et al., 2020; Zhu et al., 2021), and similar tasks are developed for pets (Shinoda & Shiohara, 2024), vehicles (He et al., 2021), landmarks (Radenović et al., 2018), artwork (Ypsilantis et al., 2021), and online products (Oh Song et al., 2016). In terms of method, most work leverages comparison of instance embeddings, which are obtained by training models on instance classification, and are often jointly optimized with metric learning objectives to enforce separation between features of different instances. More recently, the field has explored non-category-specific ILR (Kordopatis-Zilos et al., 2025; Ypsilantis et al., 2023). UnED (Ypsilantis et al., 2023) proposes to learn a single embedding model for diverse instance retrieval tasks. ILIAS (Kordopatis-Zilos et al., 2025) verifies that pre-training on ILR tasks significantly improves the performance of general-purpose visual encoders. This finding directly motivates our work.

Vision-language models and visual encoders. In a typical VLM (Liu et al., 2023; Bai et al., 2025b; Lu et al., 2024), a visual encoder maps features to the token space of a large language model (LLM) via a projector. Tong et al. (2024a) reveal that an encoder's pretext task, such as vision-language alignment (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) or self-supervised learning (Caron et al., 2021; Oquab et al., 2023), determines its suitability for different VLM tasks. For example, VLMs based on language-supervised encoders perform far better on OCR and chart-related tasks (Tong et al., 2024a). Furthermore, features from different encoders can be aggregated to complement each other. Eyes Wide Shut (Tong et al., 2024b) finds that CLIP-based VLMs fail on image pairs that CLIP perceives as similar, and mixes in self-supervised DINOv2 features to mitigate this. Cambrian-1 (Tong et al., 2024a) similarly aggregates features from multiple vision encoders. Unlike these combinations of general-purpose encoders, we investigate the effect of explicitly leveraging specialized encoders. We argue that ILR, much like OCR, is a highly specialized task, and demonstrate that a VLM's ILR capabilities can be enhanced by complementing the general encoder with an encoder trained on instance-discriminating objectives.

VLM personalization. In training VLMs to understand specific instances, existing work mainly relies on per-instance training (Alaluf et al., 2024; Nguyen et al., 2024; An et al., 2024). This approach fine-tunes the VLM to associate a small image set of each new instance with a unique instance token. Specifically, MyVLM (Alaluf et al., 2024) trains individual concept heads that indicate the existence of instances, before deciding to include the instance token as part of its response. Yo'LLaVA (Nguyen et al., 2024) trains the instance token as well as a set of description tokens to help the VLM encode crucial information of the instance. While effective, it requires the curation of per-instance datasets and additional training for every new instance. A few studies pioneer instance understanding without per-instance training. PLVM (Pham et al., 2025) uses additional general-purpose encoders for gallery image encoding. IDA-VLM (Ji et al., 2024) applies a specifically designed encoder for face-related movie understanding. TWIN (Marsili et al., 2026) teaches VLMs instance identification tasks through reinforcement learning, where the model reasons about instance similarity before answering. Our work generalizes these approaches to diverse and challenging ILR tasks, identifying dedicated ILR expert encoders as the critical component.

3 Method

This paper proposes IIR-VLM, which advances large VLMs with an In-context Instance-level Recognition capability. We achieve this through a dual-component framework. We first integrate instance-level recognition models as auxiliary visual encoders, which provide instance-discriminating knowledge (Section 3.2) in addition to general-purpose visual representations. This enhanced VLM is then efficiently fine-tuned in a two-stage process (Section 3.3), aligning the feature spaces through challenging in-context instance matching in Stage 1 and enhancing the model with instance-aware understanding tasks in Stage 2. The pipeline is illustrated in Fig. 2 (Right).

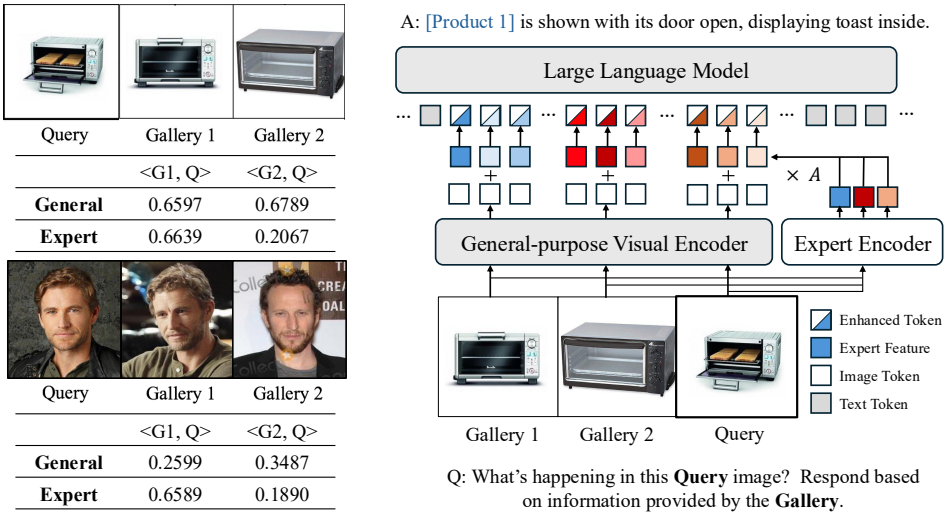

Figure 2: **Motivation and pipeline for expert encoder integration in IIR-VLM.** (Left) General-purpose visual encoders often struggle to distinguish visually similar instances, as demonstrated by high embedding similarity. Instance-level recognition (ILR) expert encoders maintain discriminative feature spaces. (Right) IIR-VLM performs in-context learning by leveraging gallery examples to execute visual understanding tasks on the query. Input images are processed by both a general-purpose encoder and an expert. These expert features are re-weighted by an attention score A and fused with the original tokens to provide the fine-grained discriminative cues necessary for ILR.

3.1 PRELIMINARY

A large Vision-Language Model (VLM) combines a visual encoder with a Large Language Model (LLM) through a projector, and works to process interleaved visual and text inputs to perform various visual perception tasks. Specifically, given a series of input text $X = (x_1, x_2, \cdots, x_{n-1})$ and image inputs c, VLMs predict

$$P(x_n|f(c), x_1, x_2, \cdots, x_{n-1}), \tag{1}$$

where $f(\cdot)$ denotes visual encoders, in order to decode responses based on inputs.

This paper investigates and advances the capabilities of VLMs on the in-context instance-level recognition (ILR) task (Radenović et al., 2018; Ypsilantis et al., 2021; Weyand et al., 2020). ILR can be described as matching images that depict the same instances across diverse categories, including persons, faces, and arbitrary objects. In a typical formulation, a query image c_q is provided along with a set of gallery images $(c_g^1, c_g^2, \cdots, c_g^M)$, where the goal is to identify which gallery image c_g^i shares the same "identity", *i.e.*, involves the same instance, as c_q. Traditional solutions involve ILR model f_e that are trained to extract instance-discriminating features, and images are matched through feature similarity:

$$y = \text{argmax}_i \, Sim(f_e(c_q), f_e(c_g^i)). \tag{2}$$

Instance-level recognition, and subsequently visual understanding based on instance knowledge, are invaluable for VLMs to master, though we show that VLMs struggle on the task when gallery images are highly similar. We refer to these traditional models f_e as *experts*, and show how they can be utilized to efficiently enhance VLMs on the task.

3.2 INTEGRATING ILR EXPERTS

Motivation. While general-purpose visual encoders adopted by VLMs demonstrate broad-domain understanding, they struggle with the fine-grained discrimination required for Instance-Level Recog-

nition (ILR). We illustrate this limitation in Fig. 2 (Left), where a standard general encoder DINOv2 (Oquab et al., 2023) consistently assigns higher similarity scores to incorrect distractors than to true matches. For example, when evaluating visually similar toaster ovens, the general encoder incorrectly prefers the distractor (0.6789) over the identical instance (0.6597). In the facial domain, the distractor again wins over the true match (0.3487 vs. 0.2599). In contrast, an ILR expert encoder trained explicitly on instance discrimination resolves this ambiguity. By maintaining a discriminative feature space, the expert correctly isolates the true matches across both domains. This stark contrast directly motivates our architectural design, demonstrating that general VLMs must benefit from expert cues to succeed at ILR tasks.

Integration mechanism. Our method enhances the VLM's image features by injecting instance-aware features from ILR models. For each image c containing a target instance of the expert, the VLM's original, general-purpose encoder provides a multi-token feature $f(c) = [f_1(c), f_2(c), \cdots f_N(c)]$, while the expert encoder extracts a single, highly discriminative one-dimensional identity feature, $f_e(c)$. To integrate the features, $f_e(c)$ is first projected through a trainable MLP into $f'_e(c)$, and then used as a query to calculate an attention score A against every token in $f(c)$:

$$A_i = \text{Softmax}_N \left(Sim(f_i(c), f'_e(c)) \right). \tag{3}$$

Inspired by the positional encoding mechanism (Vaswani et al., 2017), which injects spatial information into Transformer inputs through additive modulation, we treat the attention-enhanced feature $f'_e(c)$ as an "identity embedding". This embedding is added back to the original token features, effectively infusing fine-grained instance identity knowledge into the VLM's visual representation:

$$F_i(c) = f_i(c) + A_i \cdot f'_e(c). \tag{4}$$

The original encoder's multi-token features $f(c)$ are then replaced by the instance-aware feature sequence $F(c) = [F_1(c), F_2(c), \cdots, F_N(c)]$ for all subsequent VLM processing.

Selection of expert encoders. We empirically show that the additional visual representation *requires* a model explicitly trained on instance discrimination, instead of general visual encoders, for the best performance (Section 4.3). We therefore use expert encoders trained with instance classification objectives on large-scale instance-level recognition data. This step ensures that the expert possesses the instance-discriminative, fine-grained feature that effectively supplements the original visual encoder. We detail our choice of experts in Section 4.1.

3.3 Data engine and two-stage training

Dataset formulation. The training of VLMs requires datasets in the format of instruction tuning. To that end, we repurpose traditional instance-level recognition datasets (Ye et al., 2021; Yang et al., 2020; Li et al., 2014; Cao et al., 2018; Oh Song et al., 2016), which feature a large number of distinct instances and multiple images per instance. These images are grouped and embedded into user-assistant conversations, where every conversation includes a query image (c_q) and a gallery of candidate images ($G = \{c_g^1, c_g^2, \ldots, c_g^K\}$) that potentially contains matches for the queried instance. At inference time, the VLM is provided with the gallery in its context, and required to understand the query image based on matching results.

While some instances can be straightforwardly identified based on their visual appearance, instance-level recognition often requires more fine-grained attention to subtle visual cues, particularly when dealing with highly similar instances. In this work, we emphasize the capability of performing challenging fine-grained tasks, which are not immediately solvable by general visual perception. Leveraging general visual encoders, we modulate the difficulty level of the dataset by filling the gallery with strong distractors, selected based on visual similarity to the query. Specifically, we require $Sim(f(c_g^i), f(c_q)) > \tau$ for some pre-determined similarity threshold τ.

Two-stage fine-tuning with ILR awareness. With the dataset and auxiliary expert encoders, we introduce a lightweight, two-stage fine-tuning process to embed Instance-Level Recognition (ILR) awareness into the VLM. This process minimizes computational cost and mitigates overfitting by freezing the parameters of both the LLM and the visual encoders, and training only the projectors.

Stage 1 trains the VLM to identify which image in the gallery G matches the identity of the query image c_q, essentially answering a multiple-choice question. This stage aligns the expert features with the VLM. **Stage 2** trains the VLM to generate a detailed caption of c_q, leveraging the identity information in G. Following Stage 1, the model first outputs the label of the matching gallery image, and then follows with a caption that refers to the instance by a bracketed identity. For example, if Person 3 of the gallery matches, the model responds with `3`, followed by `[Person 3] walks into the room wearing a blue shirt.` The bracketed token can readily be replaced by an instance's name in practical applications.

Label-caption separate supervision. A standard next-token prediction (NTP) loss averages over all response tokens, which dilutes instance-identifying tokens and leads to lower matching performance in Stage 2. We instead compute the NTP loss separately on the label tokens $\mathcal{Y}_l$ and the caption tokens $\mathcal{Y}_c$:

$$\mathcal{L} = -\frac{1}{|\mathcal{Y}_l|}\sum_{t\in\mathcal{Y}_l}\log P(x_t \mid x_{<t}, F(c)) - \frac{1}{|\mathcal{Y}_c|}\sum_{t\in\mathcal{Y}_c}\log P(x_t \mid x_{<t}, F(c)). \tag{5}$$

The label term keeps the same weight regardless of caption length, so the matching capability learned in Stage 1 is preserved.

4 Experiments

We perform a series of experiments to verify the validity of IIR-VLM to learn instances in-context. In Section 4.1, we outline our experimental setup, including benchmark construction, training configuration, and sources of expert encoders. We evaluate our model on IIR-Bench and compare with existing methods in Section 4.2. Finally, through ablations in Section 4.3, we quantify the impact of expert encoders, task difficulty, and two-stage training on overall performance.

4.1 Setup

Datasets. We build our instruction-tuning training and test data by re-purposing instance-level recognition datasets of different domain. We put together datasets spanning four categories, , person, face, pet, and general objects. For the person category, we utilize a set of public person Re-ID datasets (Zheng et al., 2015; Yang et al., 2020; Zhu et al., 2021; Song et al., 2018; Li et al., 2014), which combines into 76k images. We use the VGGFace2 (Cao et al., 2018) dataset for the face category. Around 60k images are collected for the pet category. The Stanford Online Product (Oh Song et al., 2016) dataset, which contains 120k images representing 23k instances across 12 household categories, is used for learning general object ILR.

For instruction tuning, we select gallery length $K = 5$ and similarity threshold $\tau = 0.5$ by default to provide sufficiently challenging data. All instruction tuning datasets are split into training and test sets such that there are no overlapping identities between the two splits. The test set, which we name IIR-Bench, features 500 conversations per instance category. It serves as the cornerstone of our evaluation and will be released.

Finally, we generate captions with bracketed subjects (see Section 3.3) for all query images using Gemini 2.5 Flash and train models to perform captioning in the second stage.

Implementation details. We use Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) as our base model. During training, the LLM and the original visual encoders are both kept frozen, and only projectors between all encoders and the LLM are updated. Image inputs are presented as is to the VLM's original visual encoder.

We use a unified instance recognition model as the expert for all categories except faces. Inspired by UnED (Ypsilantis et al., 2023), we train this model ourselves on the training splits of instance-level recognition datasets covering products (Oh Song et al., 2016; Liu et al., 2016), persons (Zheng et al., 2015; Yang et al., 2020; Zhu et al., 2021; Song et al., 2018; Li et al., 2014), landmarks (Weyand et al., 2020), vehicles (He et al., 2021), and animals (Van Horn et al., 2018). The model is initialized from DINOv2 (Oquab et al., 2023) and trained with instance classification objectives. For face tasks, we use the ArcFace (Deng et al., 2019) face recognition model as it is trained on millions of identities with face-specific alignment and losses, which makes it much stronger on faces than general instance encoders. Face tasks are evaluated separately in IIR-Bench.

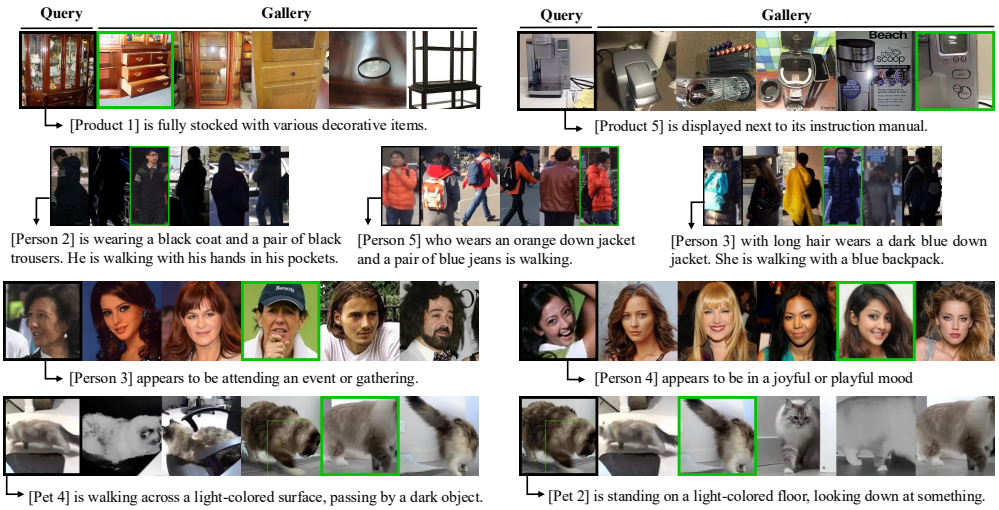

Figure 3: **Samples from IIR-Bench and corresponding model outputs.** We cover household items (object), persons, faces, and pets. In each row, the leftmost image is the **query**, followed by visually similar **gallery** images. Models select the gallery image with the matching instance (in green), and generate captions of the query referring to the correct instance.

4.2 Challenging instance-level recognition

Fig. 3 provides representative examples from IIR-Bench along with outputs from our model. Many test cases are challenging even to the human eye at first glance, as galleries contain highly similar negative samples. However, correct decisions can be made by focusing on subtle, fine-grained details. Our VLM consistently identifies the correct matching instance (marked in green), and produces correct captions for the query images that cater to the instance's immediate status, context, and background. The bracketed subject (*e.g.*, `[Person 3]`) allows for fast replacement of real, user-specified names, such as family members or pets in a smart-home visual assistant.

We systematically evaluate the performance of VLMs on IIR-Bench, our four-category ILR benchmark. We compare our IIR-VLM with state-of-the-art VLMs Gemini 2.5 Pro, Gemini 3.1 Pro, Gemini 3.8 Flash, Qwen2.5-VL-3B (Bai et al., 2025b) (our base model), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a) and Qwen3.5-4B (Team, 2026), as well as IDA-VLM (Ji et al., 2024), a recent specialized VLM proposed for instance learning and movie understanding. We also compare with TWIN (Marsili et al., 2026), which teaches VLMs instance identification tasks through reinforcement learning. Finally, we include Vanilla FT, which applies the same training to the base model without the expert encoder. All fine-tuned models are evaluated after both training stages. We report the accuracy of finding the correct instance in Table 1.

Among the general state-of-the-art models, Gemini 3.8 Flash scores the highest average baseline performance, demonstrating particular strength in the object category, reaching 98.0%. The open-source Qwen models improve steadily with model size and newer versions. Notably, Qwen3.5-4B reaches 85.4%, securing the highest rank among open-source baselines.

IDA-VLM (Ji et al., 2024) showcases its ILR capability by performing promisingly on the face category, scoring 85.6%. However, it is evidently less capable of generalizing its ILR capability to other categories, resulting in a lower overall average. TWIN (Marsili et al., 2026) scores 54.7% on average. It trains the model to reason about instance similarity before answering, and our results suggest that such reasoning does not transfer to galleries of highly similar distractors.

Our results demonstrate the necessity and value of the expert encoder. Vanilla FT shares the same base model, data, and training procedure as IIR-VLM, and only differs in the expert encoder. IIR-VLM consistently outperforms Vanilla FT across all categories. Quantitatively, the expert boosts the average accuracy from 90.3% to 92.4%, with the largest gain on the person category (+4.6%). By supplementing the general-purpose visual encoder with instance-discriminative features, the expert

Table 1: Accuracy (%) of VLMs on IIR-Bench, which tests models' capability on challenging instance-level recognition (ILR) tasks across four categories. Proprietary Gemini models are shown in gray for reference. Our models are built on Qwen2.5-VL-3B and trained with both stages. Bold marks the best result among non-proprietary models.

	Object	Person	Face	Pet	Avg
Gemini 2.5 Pro	91.4	82.4	88.8	81.6	86.1
Gemini 3.1 Pro	94.2	87.2	97.4	85.4	91.1
Gemini 3.8 Flash	98.0	94.2	97.2	87.2	94.2
IDA-VLM (Ji et al., 2024)	62.6	47.8	85.6	54.4	62.6
TWIN (Marsili et al., 2026)	60.8	53.8	58.0	46.2	54.7
Qwen2.5-VL-3B (Bai et al., 2025b)	54.0	51.6	57.4	47.8	52.7
Qwen2.5-VL-7B (Bai et al., 2025b)	73.2	70.2	85.8	67.6	74.2
Qwen3-VL-4B (Bai et al., 2025a)	77.6	77.6	89.0	77.0	80.3
Qwen3.5-4B (Team, 2026)	80.8	85.4	92.4	83.0	85.4
Vanilla FT	92.6	85.6	92.2	90.9	90.3
IIR-VLM (Ours)	**94.2**	**90.2**	**93.0**	**92.2**	**92.4**

Table 2: Instance-level recognition accuracy (%) on the matching tasks of MM-ID, proposed by IDA-VLM (Ji et al., 2024). Base denotes our base model Qwen2.5-VL-3B (Bai et al., 2025b). Avg is weighted by the number of samples per category.

	Face	Animal	Building	Vehicle	Avg
Qwen2.5-VL-3B (Bai et al., 2025b)	25.0	33.3	88.2	50.0	46.3
TWIN (Marsili et al., 2026)	42.9	33.3	88.2	40.0	52.2
IDA-VLM (Ji et al., 2024)	78.6	83.3	100	30.0	77.6
IIR-VLM (Ours)	92.9	66.7	100	60.0	**85.1**

encoder allows for a strong instance-discriminating capability. Our full model surpasses our base Qwen2.5-VL-3B model by 39.7%, leading all open-source VLMs tested, and surpassing Gemini 2.5 Pro and Gemini 3.1 Pro. Across categories, this advantage also creates visible gaps between the validation curves of Vanilla FT and IIR-VLM (Fig. 1), corroborating higher accuracy and faster convergence.

We further evaluate on the matching task of MM-ID (Ji et al., 2024) in Table 2. MM-ID evaluates VLMs on memorizing and utilizing identity information of visual inputs, with a particular focus on face understanding in movie scenes, and serves as an unseen domain for our model. Our method achieves the highest average accuracy of 85.1%, improving over our base model by 38.8%, demonstrating a promising generalization capability to unseen domains and diverse instance-level recognition problems.

4.3 Ablation Study

We study the effect of both critical components of our framework, namely introducing an ILR expert as an auxiliary encoder and two-stage training. We strengthen the argument of applying experts by testing the framework on varying difficulty levels of IIR-Bench.

VLM with expert encoders. We support the necessity of our dual-encoder design with ablation study in Table 3. We study the Stage 1 performance of VLMs with varying auxiliary encoder design on the object, person, and face categories. We compare three designs: no auxiliary encoder, an additional general-purpose DINOv2 (Oquab et al., 2023) encoder, and our expert encoder. With the exception of the face category, our expert encoder starts from the same DINOv2 weights and is further trained on instance-discriminating objectives. On these categories, the comparison between DINOv2 and our expert therefore isolates the effect of the instance-discriminating training objective.

Table 3: Stage 1 accuracy (%) of VLMs with different auxiliary encoders.

Aux. encoder	Object	Person	Face	Avg
None	91.8	87.4	91.4	90.2
DINOv2	93.4	88.8	91.8	91.3
Expert (Ours)	**94.4**	**90.4**	**92.4**	**92.4**

Table 4: Matching accuracy (%) after Stage 1 (S1) and Stage 1→2 (S1→2). Vanilla FT is without expert.

Stage	Model	Object	Person	Face	Pet	Avg
S1	Vanilla FT	91.8	87.4	91.4	86.7	89.3
	Ours	**94.4**	**90.4**	92.4	88.3	91.4
S1→2	Vanilla FT	92.6	85.6	92.2	90.9	90.3
	Ours	94.2	90.2	**93.0**	**92.2**	**92.4**

We demonstrate that adding just another general-purpose visual encoder is not sufficient. Compared with the original encoder alone, the DINOv2 variant brings alternative features and additional parameters to the model. Adding DINOv2 improves the average accuracy from 90.2% to 91.3%. This suggests that feature fusion is beneficial by providing supplementary information. However, it does not bring the performance to the level of our model.

Our proposed IIR-VLM, *i.e.*, the VLM with expert model, achieves the highest performance at 92.4% on average. It outperforms the DINO-supported VLM in every category, with the largest gap on person (1.6%). Since both models benefit from alternative features and additional parameters, this gain comes from the instance-discriminating training of the encoder.

Expert advantage on challenging tasks. Our similarity threshold parameter τ determines the difficulty of ILR tasks. VLMs are trained and tested on datasets with $\tau = 0.2$, $\tau = 0.5$, and $\tau = 0.8$. As the threshold grows, gallery images are increasingly similar to the query image, thus providing more challenges to ILR. As shown in Fig. 4, the advantage of the expert widens with difficulty. At the lowest difficulty ($\tau = 0.2$), the expert provides a modest gain of +0.8%. As the task becomes more challenging ($\tau = 0.5$), the advantage widens to +1.6%, and it reaches +3.1% at the highest difficulty ($\tau = 0.8$).

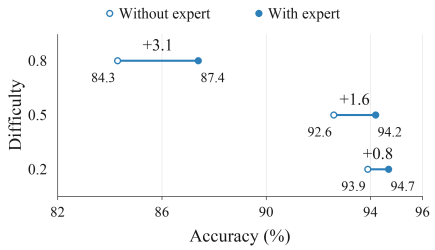

Figure 4: Performance gap between with and without expert encoder widens as difficulty increases.

This quantitatively validates that the features provided by the expert become increasingly critical as visual similarity rises. Unlike features provided by general-purpose visual encoders, the expert features possess the precise discriminative power required to isolate subtle differences within groups of visually similar instances.

Two-stage training. As shown in Table 4, training on captions in Stage 2 does not degrade instance matching. For IIR-VLM, the average accuracy moves from 91.4% after Stage 1 to 92.4% after Stage 2. It matches Stage 1 on object and person, and improves on face and pet. The model without the expert shows a similar trend, from 89.3% to 90.3%. We find that separately supervising the label and the caption in Stage 2 (Section 3.3) is key to this result. Stage 2 therefore preserves the instance matching capability learned in Stage 1, and also resumes the natural language outputs of the VLM. The advantage of the expert is maintained across both stages at 2.1%.

5 Conclusion

We propose IIR-VLM, which achieves strong in-context instance-level recognition by complementing the VLM's general-purpose encoder with an expert encoder pre-trained on instance-discriminating objectives. Our controlled comparison shows that the gain comes from this training objective rather than additional features or parameters, and the advantage of the expert widens as task difficulty increases. With a lightweight two-stage training, IIR-VLM outperforms open-source and specialized baselines on IIR-Bench, a diverse new benchmark. Future work should build on these results by extending IIR-VLM to instance-aware visual grounding, enabling the model to spatially localize recognized objects within images and videos.

REFERENCES

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2024.

Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022, 2021.

Yatai Ji, Shilong Zhang, Jie Wu, Peize Sun, Weifeng Chen, Xuefeng Xiao, Sidi Yang, Yujiu Yang, and Ping Luo. Ida-vlm: Towards movie understanding via id-aware large vision-language model. *arXiv preprint arXiv:2407.07577*, 2024.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14777–14787, 2025.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Damiano Marsili, Aditya Mehta, Ryan Y Lin, and Georgia Gkioxari. Same or not? enhancing visual perception in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17303–17315, 2026.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2024.

Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. *arXiv preprint arXiv:2502.19958*, 2025.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Plvm: A tuning-free approach for personalized large vision-language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3632–3641, 2025.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5706–5715, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2577–2591, 2020. doi: 10.1109/TPAMI.2019.2960509.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovskỳ, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 11290–11301, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

Yanlin Zhu, Xinyu Wu, Mingjia Su, Zhiming Wang, Zhicheng Zhang, and Guojun Yin. Rstpreid: A large-scale dataset for text-based person re-identification. In *International Conference on Pattern Recognition (ICPR)*, 2021.