Fallback-Enabled Closed-Set Classification: Cross-Modal Consistency in Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Vision-Language Models (VLMs) can describe and label images; however, this does not imply that they truly process what they are perceiving. Recent studies show that, despite their breadth of training, VLMs are surprisingly unreliable as classifiers, for either closed-world or open-world settings. In this work, we explore a deeper question: can a VLM recognize when an image falls outside the set of categories it is asked to choose from? Our results reveal a surprising failure mode: even when the notion of in-set versus out-of-set is explicitly defined, VLM models often assign plausible in-set labels to out-of-set images, violating the task's explicit constraint. Motivated by this, we propose a cross-modal consistency framework that reasons over both the visual and textual arms of the model and accepts an answer only when they agree. Experiments on three well-known datasets (DomainNet, VisDA and INaturalist-2021) demonstrate that this approach consistently improves balanced known *vs.* unknown detection over Source-Free Universal Domain Adaptation (SF-UniDA) baselines, showing that cross-modal consistency improves a VLM's ability to follow the task logic and distinguish when an image falls outside the intended label space. Our results suggest that, with strong VLMs, fallback behavior need not rely exclusively on specialized SF-UniDA adaptation pipelines: a lightweight cross-modal consistency decision rule can be competitive with representative SF-UniDA baselines on standard benchmarks.

1 Introduction

A long-standing challenge in (visual) recognition is to design systems that not only assign labels among known categories but also detect when an input does not belong to any of the known categories. This challenge has been studied under various names, such as open-world classification (Ding & Pang, 2024; Shu et al., 2017; Fei & Liu, 2016), universal domain adaptation (UniDA) (You et al., 2019b; Chang et al., 2022; Saito & Saenko, 2021; Choe et al., 2025). These formulations differ in scope: open-world classification typically assumes access to labeled training data and learns an unknown detector during training, while UniDA additionally addresses domain shift and partial label overlap. Source-free universal domain adaptation (SF-UniDA) (You et al., 2019a; Liang et al., 2021) further removes the dependency on source data, making adaptation feasible under privacy and efficiency constraints. This is so because it considers the case where one only has access to a pretrained model without access to source data and must adapt the model to a target domain.

The problem studied in these approaches has conceptual *and* practical implications. Epistemically, a system that "knows when it does not know" embodies a basic form of self-awareness long valued in philosophy, as Socrates taught, "wisdom begins by acknowledging the limits of one's knowledge". Practically, in safety-critical cases, e.g., if a medical condition does not belong to the known set of conditions in the diagnostic label set, the medical imaging recognition system should be able to detect it to prevent risky diagnoses.

The rapid development of Vision-Language Models (VLMs) (Bai et al., 2023; Achiam et al., 2023; Wei et al., 2022) has opened new possibilities to address this challenge (Yin et al., 2023; Kapoor et al., 2024). However, large foundation models suffer from serious hallucination issues Kalai et al. (2025); Huang et al. (2025); Xu et al. (2024). In fact, works such as Chen et al. (2023) found that large foundation models can produce correct answers, but lack internal self-awareness of their own output. Moreover, other works (Zhang et al.,

Figure 1: Illustration of the problem. From left to right, we showcase four scenarios, *i*) a known object gets wrongly recognized as unknown, *ii*) a known object gets recognized as known, *iii*) an unknown sample gets wrongly classified as one of the classes in the list because of the background, and lastly, *iv*) an unknown image gets successfully flagged as unknown. Over-rejection error (*i*), which unnecessarily triggers fallback will hurt system efficiency, while overconfidence error (*iii*), bypasses the fallback mechanism, risking unsafe downstream decisions.

2024; Mitra et al., 2025; Jiang et al., 2025) concluded that VLMs are *bad* at both open-world and closed-world image classification, with training data being the primary cause. Therefore, in order to deploy VLMs to meet the demands of a solution without relying on additional training, alternative procedures need to be taken to make their answers more reliable.

We first explore VLMs on a strict instruction-following setting as illustrated in Figure 1. Given an image and a closed-set label list, the model must assign a label from such a list, and either *i*) declare the image as "known", or *ii*) trigger a fallback decision of "unknown" when the image is judged to fall outside the intended label space. We define this problem as *fallback-enabled closed-set classification* with VLMs. Despite clear instructions and examples, our experiments on the DomainNet dataset expose a paradox: strong categorical labeling ability but weak logical boundary keeping, *i.e.*, very high true-negative rates for unknowns, yet poor true-positive rates for the known set, which results in many in-set images being mislabeled as "unknown". Importantly, clear instructions alone do not make the concept of "unknown" operational or sufficiently clear for these models.

Therefore, in order to use VLMs to address this problem, we propose a framework that tasks the VLM with reasoning along two arms and accepts a prediction only when they agree. Specifically, we make the following contributions.

- We propose a simple, model-agnostic framework that uses cross-modal consistency to perform closed-set classification with a fallback action. The method requires no fine-tuning or access to model internals. Moreover, it is easy to implement with existing VLM APIs.
- We define fallback-enabled closed-set classification with VLMs as a problem that is conceptually aligned with SF-UniDA. We explicitly bridge VLM reasoning with SF-UniDA, analyze their similarities and differences.
- We evaluate the performance of the proposed framework in comparison to the state-of-the-art SF-UniDA methods, not only using harmonic-mean accuracy and overall accuracy, but also the ratio of known *vs.* unknown accuracies as *balance* metric. This ratio reveals when a model is biased toward either over-accepting unknowns or over-rejecting knowns, which is critical for deploying fallback behavior in practice.
- Across six scenarios with three well-known datasets, we show that VLMs equipped with the proposed method can consistently outperform specialized SF-UniDA methods.

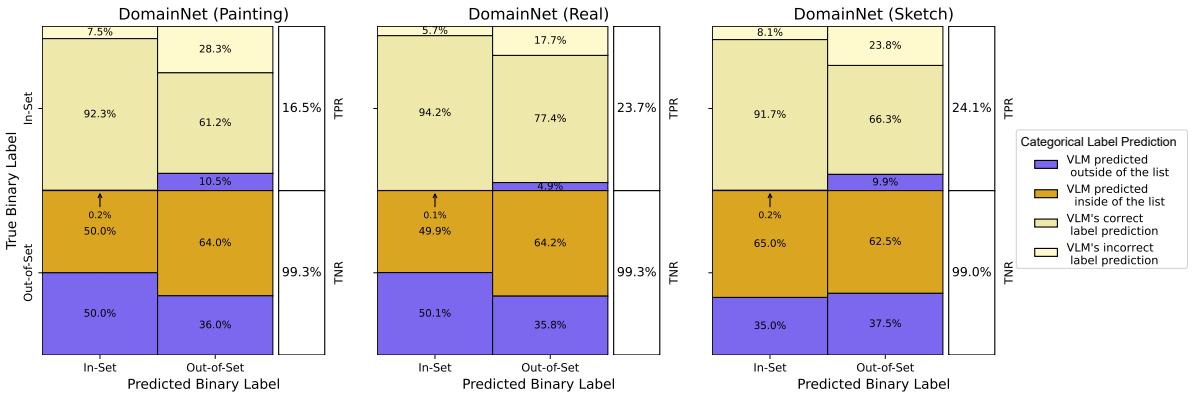


Figure 2: Preliminary analysis demonstrating the paradoxical behavior of VLMs on the DomainNet dataset. Even though the prompt explicitly defines the notions of in-set (known) and out-of-set (unknown) categories, the model fails to internalize this distinction, *i.e.*, it recognizes categorical labels accurately but struggles with binary discrimination between known and unknown samples. High TNR but low TPR indicates that while the model reliably rejects truly out-of-set images, it frequently misclassifies in-set ones as unknown, which reveals that clear instruction alone is insufficient for logical understanding of the concept "unknown".

2 A Paradoxical Finding in VLM Probing

We begin by investigating whether VLMs can accurately distinguish between known and unknown categories given a predefined list of labels. Specifically, we ask: *(i)* (Closed-set Classification) Can a VLM accurately determine whether an image belongs to a given list of labels? *(ii)* (Fallback Behavior) Can a VLM recognize when an image lies outside this list and thus should be classified as "unknown"?

We first evaluate the above using a direct prompting approach using GPT 4o-mini on three domains, specifically, Painting, Real, and Sketch, of DomainNet. The prompt we use is as follows, with {this image} and {known class list} being placeholders for the image query and the predefined list of class labels (the details are shown in Appendix A.2).

```
<System Prompt> You are an AI that classifies images based on a predefined list of categories.  If the image belongs to
a category in the GIVEN list (ONLY from the GIVEN list), then provide classname with the correct category name from
the given list and respond with unknown:  False; if the image does not belong to any category in the GIVEN list, then
select the closest possible match from the GIVEN list (DO NOT reply with labels outside of the list) as classname and
respond with unknown:  True.

<User Prompt> Does {this image} belong to one of the categories in the following list {known class list}?  Please format
the answer csv format with keys unknown and classname separated by ','
Example 1:
Image:  (picture of a aeroplane)
Response:  unknown:  False, classname:  'aeroplane'
Example 2:
Image:  (picture of a donkey)
Response:  unknown:  True, classname:  'horse'
```

This prompt explicitly articulates what constitutes an unknown image, encodes the task logic in a system prompt, clearly marks the predefined label set as a given list of classes to make these constraints less likely to be overlooked, and provides concrete formatting examples. These choices, *i.e.*, providing examples and setting the expected output format, follow structured prompting practices for multimodal LLMs to improve adherence to instructions (Sahoo et al., 2024; Zhang et al., 2023; Bsharat et al., 2023).

Figure 2 summarizes the empirical results of two prediction outputs from the VLM (GPT 4o-mini), the categorical label and the binary label "unknown", together in a confusion matrix, for all three domains (Painting, Real, and Sketch) of DomainNet. For each confusion matrix, the rows indicate the ground truth binary label: in-set (known) and out-of-set (unknown), while the columns indicate model predictions. We treat "known" (in-set) as the positive class, so the cells are true positives (TPs) (top-left), false negatives (FNs) (top-right), false positives (FPs) (bottom-left) and false negatives (FNs) (bottom-right). The per-

centage values in each cell denote the prediction accuracies of categorical labels within that subset. With the above definition, $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ measures how many known samples are correctly identified by a model, while $\text{TNR} = \frac{\text{TN}}{\text{TN+FP}}$ shows how often unknown images are correctly recognized. Our observations of the results in this setting revealed significant insights described below.

Weak known/unknown discrimination. When we analyze the predicted value of "unknown", the VLM is highly effective at identifying unknown images, correctly classifying these cases as "unknown" with up to 90% accuracy. However, the VLM shows limitations in identifying known images correctly, achieving accuracy as low as 20% in classifying known images as "known". In order to argue that a model is good at detecting what is "unknown", it should be able to have not only a good classification accuracy for "unknown", but also for "known" images. However, the high TNR but low TPR scores in Figure 2 indicate an imbalanced performance to correctly recognize "known" *vs.* "unknown", which does not match the desired behavior.

Strong label recognition. Despite poor known-sample detection, the high percentages in the top two cells of each of the confusion matrices in Figure 2 indicate that the VLM achieves high accuracy when assigning labels to known images, thus excelling at closed-set classification.

Paradoxical results. These findings suggest that the VLM has a solid label recognition ability (closed-set classification), but struggles with the discrimination of "known" *vs.* "unknown", *i.e.*, the desired fallback behavior, even with explicit instructions for them in the prompt.

Understanding the paradox. The results above suggest that the VLM's categorical reasoning and its binary known/unknown judgment are decoupled: the model can correctly identify what an object is, but cannot reliably determine whether that object belongs to the given label list. We hypothesize this occurs because the VLM's pretraining objective focuses on producing descriptive, plausible outputs rather than promoting logical constraint following. A more detailed analysis can be found in Appendix A.7.

3 A Solution using Cross-Modal Consistency

Section 2 illustrated that explicit instructions alone cannot reliably realize the fallback mechanism in closed-set classification that we seek to address. Specifically, VLMs may correctly find the category of the known labels, *i.e., the closed set*, but still misjudge in- *vs.* out-of-set membership, *i.e.*, the fallback behavior, even with strict task logic enforced in the prompt wording. A common practice for solving such a problem is to train a classifier with an additional class, thus extending the closed-set with an "unknown" class (Zhan et al., 2021; Shu et al., 2021). Other lines of work, such as UniDA (Liang et al., 2021) and SF-UniDA handle unknowns by thresholding uncertainty scores or synthesizing unknown samples (Bai et al., 2022) in open-world classification. Although effective when source data and retraining are available, these approaches are brittle under domain shift, sensitive to thresholding, and scale poorly as label lists change, especially in source-free settings. Here, we take a different route, by proposing a solution that leverages VLM's broad knowledge without additional training. Before providing the details, we first define the problem studied.

3.1 Problem Definition

To define what is "unknown", we first need to recognize what is "known". Given an image $I \in \mathcal{I}$, and a predefined and known list of labels (the closed set) $\mathcal{C} = \{c_1, c_2, \cdots, c_N\}$, with N being the total number of labels in the list, each image has two labels, one is its ground truth categorical label y which may or may not be in $\mathcal{C}$, and the other is a binary label δ denoting "known" *vs.* "unknown". Specifically, an image is defined as "known" if and only if its ground truth label is within the predefined known class list C. We define δ as

$$\delta = \begin{cases} 1 & \text{if } y \in \mathcal{C} \text{ (known)} \\ 0 & \text{if } y \notin \mathcal{C} \text{ (unknown)} \end{cases}.$$

We seek to understand whether a VLM model $f_{VLM}(\cdot)$ is able to determine whether the image I is "known" and can be appropriately labeled as any label in the set $\mathcal{C}$, or "unknown" and cannot be assigned any label in $\mathcal{C}$. To this end, we propose an approach that utilizes the cross-modal consistency of the VLM for unknown image identification. Specifically, query the VLM along two arms: i) prompt directly with the image; and

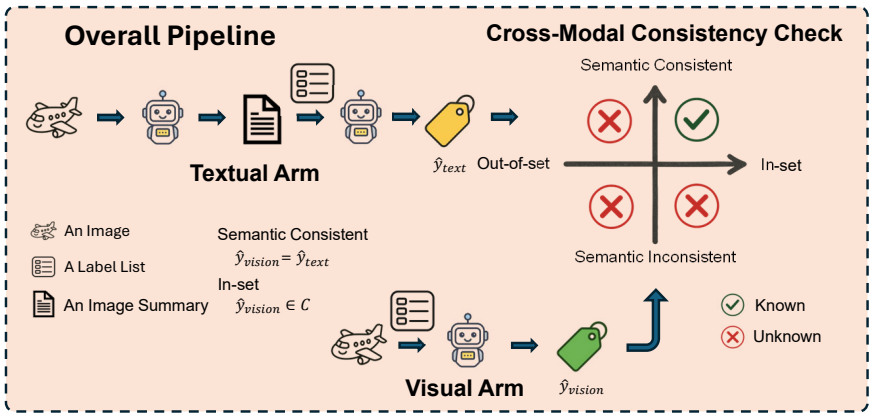


Figure 3: Illustration of the proposed method. The model makes predictions through a visual and a textual arm. A sample is marked known only when predictions from both arms agree on a label within the predefined set; otherwise, it is flagged as unknown.

ii) first summarize the image into text, then prompt the same question with such a summary. The details of the proposed framework are described below.

3.2 Direct Prompting (Visual Arm)

Given an image I, we first directly task the VLM model through prompting with giving a predicted label given the predefined list of known class labels $\mathcal{C}$. We show the partial prompt below for brevity and the full version is provided in the Appendix A.2.

```
"...  Does this image belong to one of the categories in the following list {known class list}?  Please format the answer
csv format with keys unknown and classname separated by ‘,’..."
```

Let $f_{VLM} : \mathcal{I} \times \mathcal{P} \to \mathcal{R}$, where $\mathcal{P}$ represents the prompt space and $\mathcal{R}$ is the unprocessed response space. We write

$$r_{vision} = f_{VLM}(I, p_{vision}(\mathcal{C})), \tag{1}$$

with $p_{vision}(\cdot)$ being a template for the prompt, which can be found in Appendix A.2.

We then define a parsing function $\phi(\cdot)$ (the detailed definition of which can be found in Appendix A.4) that extracts the categorical class label from the unprocessed response. Specifically, the two possible outcomes are $\hat{y}_{vision} = \phi(r_{vision}) \in \mathcal{C}$ if the parsing is successful and returns a categorical label from $\mathcal{C}$, while $\hat{y}_{vision} = \phi(r_{vision}) = \emptyset$, *i.e.*, the empty set, if the parsing fails or no valid categorical label is found because of failures following the instructions.

This process of generating $\hat{y}_{vision}$ constitutes the model's direct visual reasoning arm, which solely relies on its visual understanding and the label list $\mathcal{C}$.

3.3 Summary-Based Prompting (Textual Arm)

Then we task the VLM model to give a summary of the visual features of the image I with the following prompt. We denote the prompt by $p_{text}(\cdot)$.

```
"Can you give a summary of the image?"
```

The generated text $r_{summary} = f_{VLM}(I, p_{text}(\cdot))$ is then used as the context for a second classification prompt shown below (the full version of the prompt can be found in the Appendix A.2).

```
"...  Does this image belong to one of the categories in the following list {known class list} based on the following
summary {image summary}?  Please format the answer csv format with keys unknown and classname separated by ‘,’..."
```

The output of the VLM given this prompt produces a textual prediction

$$r_{text} = f_{VLM}(r_{summary}). \tag{2}$$

With the same parsing function $\phi(\cdot)$, we can obtain the predicted class label $\hat{y}_{text} = \phi(r_{text})$. Similarly to the direct prompting arm, we expect two outcomes, with $\hat{y}_{text}$ being a parsed string or an empty set.

3.4 Cross-modal Consistency of VLMs

Finally, we define the cross-modal consistency criterion to determine whether the model "knows what it knows", thus eliciting the desired fallback behavior. An image is predicted as known only when both reasoning arms agree on the class prediction and such class is within the label list $\mathcal{C}$:

$$\hat{\delta} = \mathbb{1}[\hat{y}_{vision} = \hat{y}_{text}] \cdot \mathbb{1}[\hat{y}_{vision} \in \mathcal{C}], \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function. If the two predictions disagree or the predicted label is not a member of $\mathcal{C}$, the image is labeled "unknown". This dual-arm validation filters out internally inconsistent predictions, reducing overconfidence from a single modality. Although prior work (Manakul et al., 2023; Wang et al., 2022) has considered multi-arm reasoning for reliability, we focus on the minimal and interpretable two-arm case due to the consideration of computational cost. Figure 3 illustrates the full workflow of the proposed cross-modal consistency approach. Given an image and a closed set of labels $\mathcal{C}$, it leverages the VLM by using two decision arms: a visual arm that takes $(I, \mathcal{C})$ and a textual arm that uses a generated description of that image, each of which produces a categorical label. An image is only accepted as "know" when the predicted labels from both arms agree and it is in the closed set $\mathcal{C}$; otherwise, it produces the fallback behavior, thus marking the prediction as "unknown". This procedure does not require training, source data (for the closed set $\mathcal{C}$) and is agnostic to the choice of VLM, thus specifically defining a decision rule for "unknown" labels rather than using separate classifiers (Qu et al., 2023) or thresholds on uncertainty estimation like in SF-UniDA approaches (Qu et al., 2023; Liang et al., 2021).

We also ask why the textual arm compensates for the alignment gap. Despite extensive training to align visual and textual representations, a residual misalignment persists between the two spaces - as evidenced by the paradox in Section 2, where categorical labeling and image summarization (which the alignment was directly optimized for) succeeds but logical set-membership judgment fails. The textual arm addresses this by converting the image into a text summary before classification, effectively collapsing the cross-modal problem into a uni-modal one. Once the visual content has been rendered as text, the comparison between the image description and the label list occurs entirely within the textual space, where language models are natively strong at logical reasoning. The summary thus acts as an explicit bridge that forces visual information through a linguistic bottleneck, narrowing the alignment gap at the point where the boundary decision is made. Agreement between the two arms then combines complementary strengths: the visual arm contributes direct perceptual credibility, while the textual arm contributes reliable within-modality logical reasoning, and disagreement signals that the cross-modal alignment has likely failed for the given image - precisely the cases where a conservative "uknown" fallback is appropriate.

4 Comparison to Source-free Universal Domain Adaptation

The fallback-enabled closed-set classification problem we study shows a close resemblance to Source-Free Universal Domain Adaptation (SF-UniDA) (You et al., 2019a; Liang et al., 2021; Qu et al., 2023; 2024). In SF-UniDA, one only has access to a pretrained source model for $\mathcal{C}$ (its training data are inaccessible), and the objective is to adapt this classification model to a new target domain so that samples from the known classes $\mathcal{C}$ are correctly classified, while unknown categories are reliably marked as "unknown". For example, in GLC (Qu et al., 2023), they leverage adaptive global one-vs-all clustering and local consensus clustering to effectively separate known and unknown samples. LEAD (Qu et al., 2024) proposes a decomposition-based learning framework that disentangles the source space into known and unknown components, thus

improving its robustness over general category-shift scenarios. These works underscore growing efforts to develop approaches that upcycle pre-trained models into new domains without accessing source data.

Large-scale VLMs share a similar trait: their training data is hidden from the user, yet they are expected to perform classification when prompted with a given set of labels. Furthermore, in the problem we study, since both VLMs and SF-UniDA models operate as predictors without access to source data, the nature of the problem is fundamentally similar in the sense that the objective is to elicit a fallback mechanism by distinguishing known versus unknown categories from model outputs.

Nevertheless, there are critical distinctions between these two settings. Specifically, SF-UniDA assumes a specific source-target domain relationship with partially overlapping label spaces (You et al., 2019a), whereas VLMs are foundation models trained on broad and multimodal corpora whose "source domain" is implicit and unobservable. Consequently, while SF-UniDA focuses on adapting decision boundaries between known and unknown target samples, our work analyzes how a large-scale VLM internally reasons about task-defined label boundaries when its pretraining data is inaccessible. These comparisons indicate that our cross-modal consistency framework can serve as an *effective alternative* to adaptation-based SF-UniDA pipelines for implementing fallback with strong VLMs. Later, we will show empirically that our rule matches or surpasses representative SF-UniDA baselines on the evaluated benchmarks in our setting.

Note on Capacity Difference: SF-UniDA baselines (GLC, LEAD) use ResNet-50 backbones ($\approx 25M$ parameters) trained on source data, while VLMs used in our experiments range from 7B (Qwen-2.5-7B-VL) to over 40B parameters (GPT-4o-mini, Gemini-2.0-flash). The comparison demonstrates that cross-modal consistency is an effective decision rule for the fallback task, but it should not be interpreted as a capacity-matched evaluation. We include SF-UniDA methods as structural reference points because the underlying task (classify knowns, reject unknowns, no source data access) is the same. To provide a more interpretable capacity-controlled reference, we additionally include open-set recognition with CLIP as a baseline.

A more extensive description of the **Additional Related Work** can be found in Appendix A.1.

5 Experiments

Datasets We test the proposed method on two well-known domain adaptation benchmarks, **DomainNet** (Peng et al., 2019) and **VisDA** (Peng et al., 2017), to compare against state-of-the-art UniDA approaches. **DomainNet** is a large-scale benchmark that contains 345 classes, with approximately 48K–172K images per domain. Following prior work (Qu et al., 2023; 2024), we conduct experiments with the Painting (P), Real (R), and Sketch (S) domains. **VisDA** is another challenging benchmark consisting of 12 object classes, where the source domain includes synthetic object renderings and the target domain consists of photo-realistic images.

In addition, we construct two datasets derived from **INaturalist-2021** (iNaturalist 2021 competition dataset) based on different taxonomic hierarchies, *phylum* and *class*. We explain the details of how the datasets are constructed in the Appendix A.9. The full dataset contains nearly 2.7M images across 10,000 species, with the phylum taxonomy yielding 13 classes and the class taxonomy yielding 51 classes. We selected **INaturalist-2021** for two main reasons. One is that it allows us to evaluate how well the models perform when the class imbalance issue is more severe, and the other is that it enables us to study how our method behaves when the labels are scientific terms instead of common words. To further illustrate the challenge of class imbalance, we visualize the distribution of sample sizes per class in Figure 5 in Appendix A.9.

Baselines *(1)* GLC (Qu et al., 2023): separates known and unknown data through adaptive one-vs-all clustering with Silhouette-based pseudo-labeling, which adaptively handles category shifts without prior knowledge. *(2)* LEAD (Qu et al., 2024): applies orthogonal decomposition to separate features associated with known and unknown source samples, thus allowing instance-level identification in target data. *(3)* Self-Consistency MV (Majority Vote)(Wang et al., 2022): queries the visual arm k times with temperature sampling; the final prediction is accepted only when a majority of the $k = 3$ runs agree on the same label in predefined list. Otherwise, the image is flagged as "unknown." This baseline is run on the open-source models only to control for API cost. *(4)* Self-Consistency w/ R (Rephrasing)(Khan & Fu, 2024): queries the visual arm $k = 3$ times with different rephrasings of the user prompt for visual and textual arms (see

Appendix A.3 for exact wording). *(5)* Open-set recognition with CLIP (ViT-L/14) (Miller et al., 2024): treats open-set recognition as a query-set problem: a VLM compares an image embedding to a finite set of label embeddings, and uses negative embeddings/words so the model can reject an input as unknown instead of forcing a wrong class. In the experiment, we show results with 1000 random words as negative embeddings for a more balanced known vs unknown results.

VLM Models We consider four VLMs, including two commercial APIs: GPT 4o-mini (Achiam et al., 2023) and Gemini-2.0-flash (Comanici et al., 2025), and two open-source models: LLaMA-3.2-Vision (Grattafiori et al., 2024) and Qwen-2.5-7B-VL (Bai et al., 2025). The implementation details can be found in Appendix A.5.1. Moreover, for reproducibility purposes, the source code is also included in the supplementary material and will be publicly available upon publication.

Table 1: Comparison to SoTA SF-UniDA methods on weighted accuracies for known (acc_{k-w}) and unknown (acc_{u-w}) classes. The results are formatted as acc_{k-w}/acc_{u-w}. (Best in bold and second best in underline)

Method	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	45.9/77.3	49.1/80.5	37.3/77.7	69.1/67.5	61.7/29.9	87.8/76.0	56.8/68.1
LEAD	41.1/68.7	51.9/73.3	34.8/69.7	72.6/87.1	77.6/65.6	63.5/64.6	56.9/71.5
OSR (w/ CLIP)	60.4/63.0	77.3/72.8	58.5/62.0	52.1/91.6	17.5/56.6	9.0/48.1	45.8/65.7
Self-Consistency MV (3V):							
Llama	69.4/54.4	71.4/56.4	60.3/52.0	65.6/50.8	75.0/15.4	70.8/46.7	68.8/46.0
Qwen	92.1/42.3	97.6/43.1	95.4/41.2	91.0/36.2	95.9/54.4	81.2/51.9	92.7/44.9
Self-Consistency w/ R (3V):							
Llama	66.7/60.4	67.5/62.9	63.1/46.2	66.6/46.8	96.7/10.6	86.1/47.0	74.5/45.7
Qwen	**93.7**/36.2	**98.2**/39.2	**96.2**/36.4	92.0/30.2	**99.9**/28.9	**99.0**/2.70	**96.5**/28.9
Visual-only (V):							
GPT 4o-mini	70.9/36.3	85.0/37.3	87.0/36.9	**96.8**/1.72	93.6/0.26	93.6/2.87	87.8/19.2
Gemini	72.2/21.9	78.3/35.4	74.0/26.6	91.7/6.70	98.4/0.07	74.9/17.5	81.6/18.0
Llama	60.0/69.2	63.5/78.7	46.9/79.4	60.8/53.8	93.6/10.8	56.5/51.5	63.5/57.2
Qwen	87.6/46.4	89.2/48.9	79.8/42.1	92.7/8.79	95.6/20.8	90.0/25.9	89.1/32.1
Texual-only (T):							
GPT 4o-mini	68.2/40.0	80.9/42.2	73.7/40.0	89.5/2.35	94.0/0.26	93.3/3.98	83.3/21.5
Gemini	65.6/35.2	71.3/47.9	62.4/37.0	85.2/6.70	96.9/0.07	90.2/18.5	78.6/24.2
Llama	47.9/70.1	55.3/79.8	40.1/79.9	64.6/53.8	83.6/64.6	57.0/51.5	58.1/66.6
Qwen	73.6/60.2	80.6/68.9	70.2/56.6	80.9/57.2	83.4/42.5	82.1/25.9	78.5/51.9
Cross-Modal Consistency (V+T)(Ours):							
GPT 4o-mini (w/ Ours)	73.9/62.1	85.0/62.8	81.2/62.1	94.3/37.7	93.8/42.4	86.9/49.0	85.9/52.7
Gemini (w/ Ours)	71.7/68.5	74.0/73.7	72.7/74.8	90.7/54.7	97.1/43.7	87.9/52.1	82.3/61.3
Llama (w/ Ours)	49.7/**88.9**	55.7/**91.8**	42.3/**91.3**	55.7/**77.0**	85.6/58.0	53.4/**94.4**	57.1/83.6
Qwen (w/ Ours)	72.4/85.6	83.3/87.4	76.0/84.0	84.2/73.8	83.8/**87.3**	80.3/94.3	80.0/**85.4**
Cross-Modal Consistency (with the best prompt design as reference): $*$: prompt var3, †: prompt original							
Qwen (w/ ours)	72.4/85.6*	83.3/87.4*	76.084.0*	84.2/73.8*	96.5/99.9†	80.9/92.3†	81.9/87.2

5.1 Metrics

***H* score**: $H = \frac{2 \cdot \overline{acc_k} \cdot acc_u}{\overline{acc_k} + acc_u}$ is the harmonic mean between the average of per-class accuracies ($\overline{acc_k}$) over the K classes in $\mathcal{C}$, and the binary accuracy of the unknown samples (acc_u). *Advantages* The $\overline{acc_k}$ focuses on maintaining discrimination within known classes. The average per-class accuracy prevents the dominant classes from masking the poor performance of rare known classes, while the binary unknown accuracy acc_u measures the task of unknown detection. When combined by harmonic mean, one intends to capture both aspects (closed-set classification and unknown detection) of the performance of UniDA methods.

Limitations Since $\overline{acc_k}$ is the unweighted mean average accuracy over K classes, class imbalance can distort its value, *i.e.*, the rare classes receive equal weights as the more frequent ones. If these infrequent classes have poor performance, then $\overline{acc_k}$ can be severely penalized by the poor performance of classes with few samples. Furthermore, since the harmonic mean is dominated by the component with the smaller value, the H score primarily reflects the worse metric, thus it cannot show whether the model is better with known or unknown classes. In practice, we often wish to evaluate not only the absolute performance on known and unknown samples separately (*e.g.*, their respective accuracies), but also the balance between them, that is, whether the model performs comparably well in recognizing known categories and rejecting unknown ones.

Table 2: Comparison to SoTA SF-UniDA methods on the balance ratio R. $R \to 1$ indicates balanced behavior for known and unknown recognition. (Best in bold and second best in underline. The complete table can be found in Table 11 in the Appendix.)

Method	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	0.599	0.614	0.481	**1.025**	2.063	<u>1.023</u>	<u>0.833</u>
LEAD	0.600	0.710	0.502	0.834	<u>1.182</u>	**0.983**	0.736
GPT 4o-mini (w/ Ours)	<u>1.190</u>	1.354	1.306	2.501	2.212	1.775	1.732
Gemini (w/ Ours)	**1.046**	**1.004**	**0.971**	1.660	2.224	1.687	1.432
Llama (w/ Ours)	0.560	0.611	0.463	0.722	1.476	0.565	0.733
Qwen (w/ Ours)	0.845	<u>0.953</u>	<u>0.905</u>	<u>1.141</u>	**0.960**	0.852	**0.943**

Table 3: H score (%) comparison with SoTA SF-UniDA methods. The first two rows show the SF-UniDA baseline results, cited from Qu et al. (2024). The last four rows show results with VLM models using our cross-modal consistency method (best in bold and second best is underlined.).

Method	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	59.1	50.5	50.7	73.1	32.4	47.8	53.8
LEAD	52.5	62.5	51.2	<u>76.8</u>	50.7	46.9	56.3
OSR w/ CLIP	61.0	<u>75.1</u>	60.8	68.7	38.9	12.0	52.8
Self-Consistency MV (Llama):	58.6	59.2	53.9	33.8	25.9	43.1	45.8
Self-Consistency MV (Qwen):	51.8	56.1	52.3	34.1	68.2	44.8	51.2
Self-Consistency w/ R (Llama):	61.1	65.8	62.9	30.1	18.9	57.5	49.4
Self-Consistency w/ R (Qwen):	51.3	56.1	50.2	45.1	44.8	5.24	42.1
GPT 4o-mini (w/ Ours)	67.2	72.4	70.2	54.0	56.0	50.5	61.7
Gemini (w/ Ours)	<u>69.9</u>	74.0	<u>73.3</u>	68.2	<u>57.5</u>	<u>53.3</u>	<u>66.0</u>
Llama (w/ Ours)	63.0	68.6	56.8	65.0	54.3	36.1	57.3
Qwen (w/ Ours)	**77.3**	**85.7**	**80.1**	**78.2**	**73.1**	**57.5**	**75.3**

In order to address the limitations of the H score, we also separately show the **weighted accuracy** for known and unknown classes. The weighted accuracies acc_{k-w} and acc_{u-w} incorporate class sample sizes to better reflect the influence of class imbalance. To further show whether each model has a balanced performance for known *vs.* unknown, we also show the **balance ratio** between weighted known accuracy and weighted unknown accuracy.

$$acc_{k-w} = \frac{\sum_K n_k \cdot acc_k}{\sum_K n_k} \quad (5), \qquad acc_{u-w} = \frac{\sum_M n_m \cdot acc_m}{\sum_M n_m} \quad (6) \qquad R = \frac{acc_{k-w}}{acc_{u-w}} \quad (7)$$

where k is the number of known classes in $\mathcal{C}$, M is the number of unknown classes (often called *private classes* in UniDA), which are assumed to be known only for evaluation purposes, acc_k, which is weighted by sample size n_k, is the accuracy of the known class k, while acc_m, the accuracy of the unknown class m, is weighted by sample size n_m. We expect a good model to have good and balanced performance for both known and unknown classes. Therefore, the closer the ratio R is to 1, the more balanced the model is. For comparison, we also show the mean accuracy for known classes *vs.* unknown classes in Appendix A.8. Furthermore, in the ablation study, we also evaluate the H score between the mean accuracy of the known classes and the accuracy of each private class, individually, specifically $H_{u-m} = \frac{2 \cdot \overline{acc_k} \cdot acc_{u-m}}{\overline{acc_k} + acc_{u-m}}$, where acc_{u-m} indicates the accuracy of the unknown samples from the unknown class m.

6 Results

We evaluate the proposed cross-modal consistency framework with the four VLMs across six scenarios in DomainNet, VisDA, and INaturalist-2021, as described above, to assess its ability to distinguish known and unknown categories compared to the state-of-the-art (SoTA) source-free UniDA (SF-UniDA) methods.

As shown in Table 3, across all benchmarks, our cross-modal consistency framework makes VLMs outperform the classical SoTA SF-UniDA baselines. Specifically, Qwen 2.5-7B-VL consistently achieves the highest H scores, surpassing classical UniDA approaches by up to 20% to 25%, while Gemini 2.0-flash consistently ranks second. The results demonstrate that VLMs equipped with cross-modal reasoning can rival or outperform

UniDA algorithms that are *specialized* for such tasks. Table 1 further illustrates how well are the weighted accuracies are for known and unknown samples, while table 2 shows their ratio R. The ideal behavior of a model is $R \approx 1$ with high scores for both acc_w and acc_w. Qwen 2.5-7B-VL achieves both the highest accuracy and the most balanced value for R. Gemini 2.0-flash also achieves strong absolute accuracies, but its R values shift away from 1 for certain datasets. In comparison, SF-UniDA methods exhibit a decent average balance ratio at times, yet their weighted accuracies are much lower in general. As a result, even though GLC performance looks "balanced" by ranking second in R value, its overall weighted accuracy cannot compete with Qwen 2.5-7B-VL and the other VLMs that are both more accurate and comparably (or better) balanced. The two self-consistency baselines consistently achieve higher known-class accuracy than our cross-modal method but substantially lower unknown-class accuracy, resulting in lower H scores overall. The cross-modal consistency method primarily improves unknown rejection (acc_{u-w}) relative to single-arm baselines, sometimes at a modest cost to known accuracy. This is the expected behavior of an agreement-based decision rule, disagreement between the arms makes mistakes on the side of caution by flagging the image as "unknown".

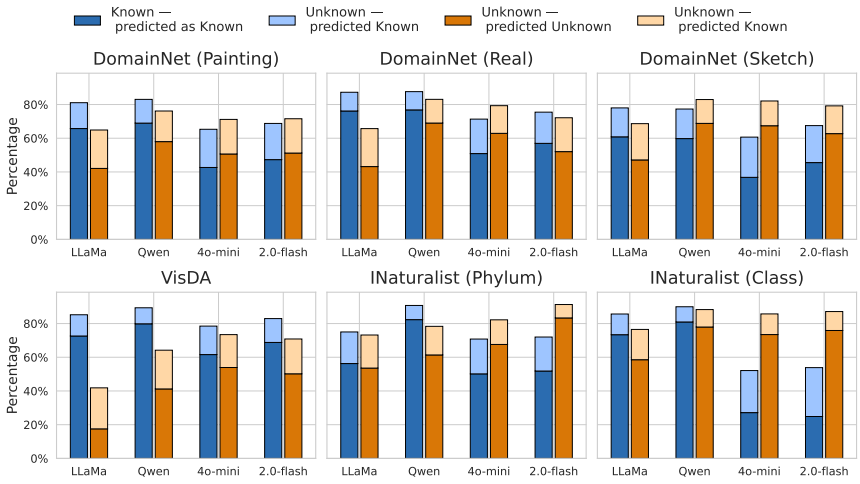


Figure 4: Positive Predictive Value (PPV, blue) and Negative Predictive Value (NPV, orange) of all four VLM models across six datasets. The lighter segments indicate the proportion of samples misclassified within that group for each dataset, while the darker bars correspond to correctly classified known / unknown samples.

Furthermore, we show positive predictive value (PPV) and negative predictive value (NPV) results in Figure 4 to show how reliable a VLM can be with our framework to behave like it "knows when it knows and when it does not", and observe that Qwen 2.5-7B-VL achieves the most reliable behavior. Overall, these results confirm that enforcing cross-modal consistency framework leads to more stable reasoning and balanced accuracy, reducing the bias between known vs unknown that characterizes classical UniDA baselines.

In the following section, we introduce our ablation studies to demonstrate the influence of prompt variations and removal of framework components.

6.1 Ablation Study

Influence of Prompt Variation To examine how image summary influences prediction with our framework, we tested three variations of $p_{text}(\cdot)$: *i*) Variation 1 (var 1) prompts with shorter length of image summaries to evaluate the effect of description verbosity. The motivation of this variation is to test whether reducing linguistic richness helps the models predict more accurately. *ii*) Variation 2 (var 2) "Describe the main object of the image". A prompt that explicitly directed the model to focus on the main object of the image. *iii*) Variation 3 (var 3) "Identify the primary object in the image, excluding any background elements". A prompt that went a step further by specifying that the background should be excluded, restricting attention solely to the primary object. This study allows us to disentangle whether richer contextual descriptions or

Table 4: Ablation Study on prompt variations with Qwen 2.5-7B-VL. We show H_{u-m} scores (%) for the individual target private classes in the VisDA validation dataset. Numbers in each cell represent the results on the validation set.

	Target Private Class			**Overall H score**
	skateboard	train	truck	
original	57.36	87.29	63.14	72.56
var 1	58.29	87.80	61.34	72.29
var 2	72.75	**90.56**	64.00	76.91
var 3	**84.45**	87.38	**65.21**	**78.24**

Table 5: H_{u-m} scores (%) for the individual target private classes in the VisDA validation set.

	Target Private Class			**Overall H score**
	skateboard	train	truck	
GLC	69.72	73.92	70.53	71.60
LEAD	75.56	57.92	**76.08**	77.10
Qwen	**84.45**	**87.38**	65.21	**78.24**

stronger object-centric guidance lead to more reliable recognition. As shown in Table 4, both the overall H score and the H score wrt. each individual target private class (which is class that do not show in the known label set but only exists in the target dataset) increase for VisDA as the prompts become more object-centric (from original to var 3, the view gets more object-centric). However, when we experiment with a similar ablation study for INaturalist, this trend reverses on INaturalist, for both phylum and class taxonomy, where variation 3 performs worse than the original prompt, the results of which are shown in the Appendix A.8. The explanation for the difference in performance trends lies in the fine-grained taxonomy of the dataset, in which many species can only be distinguished by subtle contextual cues such as texture, color tone, or typical habitat. When the prompt suppresses background information, these discriminative cues are lost and the model's alignment between visual features and (scientific) labels weakens. In addition, scientific names have limited pretraining data of such VLM models, which makes contextual signals even more crucial. As a whole, these results show that the decision of whether to exclude the background is image- and task-dependent. For example, for an image of "an arm with a dreamcatcher tattoo" in DomainNet, the ground-truth label "arm" competes with a plausible interpretation "tattoo", and the "right" choice of label depends on what the user intends to infer. Such image examples are discussed in the Appendix A.10. In summary, these observations highlight that prompt design should respect the semantic granularity of the task: use object-focused prompting enhances logical consistency in structured object domains, whereas prompt with context-aware phrasing for fine-grained classes or ecologically / environmentally dependent recognition. Importantly, these prompt variations represent explicit, controllable design choices for practitioners and not hidden degrees of freedom that undermine the generality of the proposed framework. By adapting the image summary prompt to match the semantic granularity of the task domain, users can systematically improve cross-modal consistency without requiring retraining or model modifications. This flexibility is a feature, not a limitation, as it enables practitioners to tailor the framework to their specific application context.

Discussion of VisDA results In Table 5, we observe that the VLMs achieve significantly lower H scores compared to the UniDA baselines. To better understand this gap, we examined the experimental setup in detail. The predefined label list contains nine known classes (aeroplane, bicycle, bus, car, horse, knife, motorcycle, person, plant), while the target private set consists of only three classes: skateboard, train, and truck. We calculate H_{u-m} scores for each target private class in VisDA. The results, which are shown in Table 5, reveal that the H_{u-m} score drops substantially when the private class is a truck. A closer inspection of the source data indicates why: the "car" category is represented exclusively by sedans, SUVs, hatchbacks, and sports cars, with no overlap with truck. For methods such as GLC and LEAD, which rely on pretrained source models, this restricted source definition of "car" enables them to more easily separate truck as an unknown class. In contrast, VLMs, trained on broader corpora, likely encode a richer concept of "car" that encompasses truck as a subtype. Consequently, the greater semantic knowledge of VLMs paradoxically makes it harder to avoid misclassifying a truck as a car, leading to a lower H_{u-m} score.

Influence of Cross-modal Consistency. To isolate the contribution of our cross-modal consistency mechanism, we ablate three configurations: (i) *visual-only* ($\hat{y}_{vision} \in \mathcal{C}$), (ii) *textual-only* ($\hat{y}_{text} \in \mathcal{C}$), and (iii) our full method combining both arms with consistency. As shown in Table 1, visual-only is prone

Table 6: Inference Cost and Latency Analysis

Model	Config	Calls	Latency (ms)	Throughput	Cost/100 Img	Total Tokens / Img
GPT 4o-mini	V	1	1069 ± 858	0.935	\$0.057	3,748
	T	2	1939 ± 1098	0.516	\$0.101	6,625
	V+T	3	3008 ± 1455	0.332	\$0.158	10,412
Gemini	V	1	1222 ± 505	0.819	\$0.021	2,108
	T	2	2329 ± 220	0.430	\$0.036	3,483
	V+T	3	3550 ± 264	0.282	\$0.057	5,591

to a *closed-set bias*: very high known-class weighted accuracy while severely under-rejecting unknowns. For example, on VisDA, Qwen's visual-only variant reaches 92.7/8.79 (acc_{k-w},acc_{u-w}), indicating many private-class images are incorrectly accepted as known. Introducing the textual arm helps mitigate this by enabling semantic mismatch reasoning, typically increasing unknown-class accuracy but sometimes at the cost of known accuracy. Our cross-modal consistency framework provides an additional, non-redundant check: by requiring agreement between what the model "sees" and what it can justify from its own summary, it increases recognition of "unknowns" while preserving strong known recognition. The large performance gain shows that by leveraging the consistency between visual and textual reasoning paths, we enhance the model's ability to recognize what it truly knows.

6.2 Inference Cost Analysis

For each commercial VLM we report five quantities: *(1)* number of API/model calls per image, *(2)* mean latency per image with standard deviation, *(3)* throughput in seconds per 100 images, *(4)* estimated monetary cost per 100 images, and *(5)* the total token usage per image. We measured latency, throughput, and monetary cost on a 100-image subset on DomainNet in Table 6. We note that the visual and textual arms are independent and can be dispatched in parallel, which would reduce the wall-clock time of V+T to approximately max(V, T) rather than V+T; we leave this engineering optimization to future work.

7 Conclusion

We present a cross-modal consistency framework that enables VLMs to better distinguish known from unknown images by requiring agreement between visual and textual classification arms. Experiments on DomainNet, VisDA, and INaturalist-2021 demonstrate that our approach for fallback-enabled closed-set classification consistently improved balanced known *vs.* unknown weighted accuracy and H score relative to SoTA SF-UniDA baselines. The results suggest that enforcing agreement enhances a VLM's logical understanding of "unknown", marking a step toward more self-aware and reliable vision-language reasoning. Moreover, our results suggest that the traditional SF-UniDA paradigm may no longer be the most effective way to obtain fallback behavior when strong VLMs are available. In our experiments, a frozen VLM equipped with a cross-modal consistency framework consistently matches or outperforms SF-UniDA methods, pointing toward a future in which open-world recognition relies more on VLM-based decision rules than on dedicated adapters on pretrained source classifiers.

8 Limitations and Future Works

Our study has several limitations. First, we focus on single-label image classification with fallback action, leaving multi-label classification and more complex vision tasks to interesting future work. Second, the proposed framework still requires multiple prompts per sample, which would increase latency and cost in large-scale commercial use. In the future, we plan to explore ways to reduce query cost. Third, because both reasoning arms rely on the same underlying VLM, systematic misunderstandings produce correlated errors that our agreement rule cannot detect. We included an analysis of this type of error in the Appendix A.6. Future work could explore using different VLMs for each arm or incorporating calibrated confidence scores to mitigate this.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Ke Bai, Guoyin Wang, Jiwei Li, Sunghyun Park, Sungjin Lee, Puyang Xu, Ricardo Henao, and Lawrence Carin. Open world classification with adaptive negative samples. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 4378–4392, 2022.

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*, 2025.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R Bowman, and Kyunghyun Cho. Two failures of self-consistency in the multi-step reasoning of llms. *arXiv preprint arXiv:2305.14279*, 2023.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024.

Seun-An Choe, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Universal domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4607–4617, June 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Choubo Ding and Guansong Pang. Improving open-world classification with disentangled foreground and background features. In *ACM Multimedia 2024*, 2024. URL `https://openreview.net/forum?id=6GLwA6jR3N`.

Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 506–514, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1061. URL `https://aclanthology.org/N16-1061/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mingfei Han, Haihong Hao, Jinxing Zhou, Zhihui Li, Yuhui Zheng, Xueqing Deng, Linjie Yang, and Xiaojun Chang. Self-consistency as a free lunch: Reducing hallucinations in vision-language models via self-reflection. *arXiv preprint arXiv:2509.23236*, 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025.

iNaturalist 2021 competition dataset. iNaturalist 2021 competition dataset. `https://github.com/visipedia/inat_comp/tree/master/2021`, 2021.

Yiwen Jiang, Deval Mehta, Siyuan Yan, Yaling Shen, Zimu Wang, and Zongyuan Ge. WISE: Weak-supervision-guided step-by-step explanations for multimodal LLMs in image classification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14674–14685, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.741. URL `https://aclanthology.org/2025.emnlp-main.741/`.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5830–5840, 2021.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL `https://arxiv.org/abs/2509.04664`.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972, 2024.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024.

Tom Kouwenhoven, Kiana Shahrasbi, and Tessa Verhoef. Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect. *arXiv preprint arXiv:2507.10013*, 2025.

Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Umad: Universal model adaptation under domain and category shift. *arXiv preprint arXiv:2112.08553*, 2021.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Enhancing few-shot vision-language classification with large multimodal model features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2760–2772, October 2025.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Sanqing Qu, Tianpei Zou, Florian Röhrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Upcycling models under domain and category shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20019–20028, 2023.

Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, and Changjun Jiang. Lead: Learning decomposition for source-free universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23334–23343, 2024.

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.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9000–9009, October 2021.

Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*, 2017.

Lei Shu, Yassine Benajiba, Saab Mansour, and Yi Zhang. Odist: Open world classification via distributionally shifted instances. 2021. URL `https://www.amazon.science/publications/odist-open-world-classification-via-distributionally-shifted-instances`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2720–2729, 2019b.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Albert YS Lam, and Xiao-Ming Wu. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3521–3532, 2021.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

A Appendix

A.1 Additional Related Work

Self-Consistency has recently emerged as an effective way to boost the reasoning abilities of large foundation models Wang et al. (2022). Khan & Fu (2024) explores a way to judge the reliability of the answer of a VLM by accessing the inconsistency of the model's answers over paraphrased versions of the original question, which are generated dynamically using a VQG model. Then, low consistency is used as a strong indicator of abstaining the model from answering. Han et al. (2025) harnesses the VLM's consistency between long and short answers to reduce hallucinations and improve the reliability of answers. In contrast, our method uses the cross-modal self-consistency of the outputs of a VLM as a primary signal to decide whether the system should output a categorical label prediction or trigger the "unknown" fallback behavior. Our method is different in the objective of the task and in that it is designed to operate under a constrained budget by limiting the number of VLM calls.

Open-World Classification is a classic natural language processing task Fei & Liu (2016) where, in an open environment, an ideal classifier must both assign inference examples to the correct classes seen during training and identify those examples that do not belong to any known classes (learned during training). For example, ODIST Shu et al. (2021) addresses the issue by creating distributionally shifted samples with the help of a pretrained BART model. ANS Bai et al. (2022) tackles this challenge by synthesizing negative examples in the feature space and pairing them with one-vs-all binary classifiers for each known class. While specific formulations vary, these methods generally assume access to labeled data for the in-scope classes during training, and learn the unknown detector jointly with or on top of such supervised representations. By contrast, the problem we consider forgoes using source samples when adapting or evaluating the method and instead focuses on the real-world deployment of a pretrained frozen model.

Open-Vocabulary Detection Recent methods such as YOLO-World Cheng et al. (2024), Grounding DINO Liu et al. (2024), and SAM 3 Carion et al. (2025) can detect objects from novel categories specified by text. However, these target object detection (localization + classification) rather than image-level classification with fallback behavior, making them architecturally different from our setting. Flamingo Alayrac et al. (2022) is a few-shot visual learner requiring in-context examples, which is a different paradigm from our zero-shot, inference-time-only setting. CLIP Miller et al. (2024); Radford et al. (2021) zero-shot classification is more directly comparable: it performs training-free image classification under a fixed label list using cosine similarity, and we include it as a baseline in Section 6. Methods like Joseph et al. (2021) use energy scores or confidence calibration for open-world detection. These typically require access to model logits or internal representations, which are unavailable when using VLMs through black-box APIs. Our framework is designed for the API-only setting and is complementary to logit-based approaches.

A.2 Prompt Templates

In all cases, curly-braced expressions such as {this image}, {known class list}, and {image summary} are placeholders that are filled programmatically before sending the query to the VLM; *classname* and *unknown* are expected variables.

Visual Arm Prompt For the visual arm, we pass the image directly together with the closed-set label list:

```
<System Prompt> You are an AI that classifies images based on a predefined list of categories. If the image belongs to
a category in the GIVEN list (ONLY from the GIVEN list), then provide classname with the correct category name from
the given list and respond with unknown: False; if the image does not belong to any category in the GIVEN list, then
select the closest possible match from the GIVEN list (DO NOT reply with labels outside of the list) as classname and
respond with unknown: True.

<User Prompt> Does {this image} belong to one of the categories in the following list {known class list}? Please format
the answer csv format with keys unknown and classname separated by ','
Example 1:
Image: (picture of a aeroplane)
Response: unknown: False, classname: 'aeroplane'
Example 2:
Image: (picture of a donkey)
Response: unknown: True, classname: 'horse'
```

Textual Arm Prompts The textual arm uses two prompts: one to obtain a textual summary of the image, and the other to perform a classification conditioned on that summary.

Image–Summary Prompt

```
<User Prompt>
Can you give a summary of the image?
```

Summary-Based Prompt

```
<System Prompt> You are an AI that classifies images based on a summary of the image. If the image belongs to a category
in the GIVEN list (ONLY from the GIVEN list), then provide classname with the correct category name from the given
list and respond with unknown: False; if the image does not belong to any category in the GIVEN list, then select the
closest possible match from the GIVEN list (DO NOT reply with labels outside of the list) as classname and respond with
unknown: True.
```

```
<User Prompt> Does {this image} belong to one of the categories in the following list {known class list} based on the
following summary: {image summary}? Please format the answer csv format with keys unknown and classname separated by
‘,’
Example 1:
Image: (picture of a aeroplane)
Response: unknown: False, classname: ’aeroplane’
Example 2:
Image: (picture of a donkey)
Response: unknown: True, classname: ’horse’
```

Placeholders

- {this image} is bound by attaching the actual image as input to the VLM.
- {known class list} is replaced with a label list in the closed set C.
- The model is expected to return a single line in the form

  ```
  unknown: True/False, classname: ’{one of the labels in {known class list}}’
  ```

{known class list} for Each Dataset We follow the SF-UniDA protocol for defining known vs. unknown classes, and we reuse the exact same class partitions for the source and target set when compared to GLC and LEAD on DomainNet and VisDA in their UniDA setting.

VisDA For VisDA, {known class list} (C in the paper notation) includes only 9 class labels (the first five above are in the tested target dataset), they are

```
’aeroplane’, ’bicycle’, ’bus’, ’car’, ’horse’, ’knife’, ’motorcycle’, ’person’, ’plant’
```

DomainNet (Painting, Real, Sketch) For each DomainNet scenario, {known class list} consists of 200 class labels as shown below. The first 150 classes are in the target dataset.

```
’The Eiffel Tower’, ’The Great Wall of China’, ’The Mona Lisa’, ’aircraft carrier’, ’airplane’, ’alarm clock’,
’ambulance’, ’angel’, ’animal migration’, ’ant’, ’anvil’, ’apple’, ’arm’, ’asparagus’, ’axe’, ’backpack’, ’banana’,
’bandage’, ’barn’, ’baseball’, ’baseball bat’, ’basket’, ’basketball’, ’bat’, ’bathtub’, ’beach’, ’bear’, ’beard’, ’bed’,
’bee’, ’belt’, ’bench’, ’bicycle’, ’binoculars’, ’bird’, ’birthday cake’, ’blackberry’, ’blueberry’, ’book’, ’boomerang’,
’bottlecap’, ’bowtie’, ’bracelet’, ’brain’, ’bread’, ’bridge’, ’broccoli’, ’broom’, ’bucket’, ’bulldozer’, ’bus’, ’bush’,
’butterfly’, ’cactus’, ’cake’, ’calculator’, ’calendar’, ’camel’, ’camera’, ’camouflage’, ’campfire’, ’candle’, ’cannon’,
’canoe’, ’car’, ’carrot’, ’castle’, ’cat’, ’ceiling fan’, ’cell phone’, ’cello’, ’chair’, ’chandelier’, ’church’,
’circle’,
```

```
'clarinet', 'clock', 'cloud', 'coffee cup', 'compass', 'computer', 'cookie', 'cooler', 'couch', 'cow', 'crab', 'crayon',
'crocodile', 'crown', 'cruise ship', 'cup', 'diamond', 'dishwasher', 'diving board', 'dog', 'dolphin', 'donut', 'door',
'dragon', 'dresser', 'drill', 'drums', 'duck', 'dumbbell', 'ear', 'elbow', 'elephant', 'envelope', 'eraser', 'eye',
'eyeglasses', 'face', 'fan', 'feather', 'fence', 'finger', 'fire hydrant', 'fireplace', 'firetruck', 'fish', 'flamingo',
'flashlight', 'flip flops', 'floor lamp', 'flower', 'flying saucer', 'foot', 'fork', 'frog', 'frying pan', 'garden',
'garden hose', 'giraffe', 'goatee', 'golf club', 'grapes', 'grass', 'guitar', 'hamburger', 'hammer', 'hand', 'harp',
'hat', 'headphones', 'hedgehog', 'helicopter', 'helmet', 'hexagon', 'hockey puck', 'hockey stick', 'horse', 'hospital',
'hot air balloon', 'hot dog', 'hot tub', 'hourglass', 'house', 'house plant', 'hurricane', 'ice cream', 'jacket', 'jail',
'kangaroo', 'key', 'keyboard', 'knee', 'knife', 'ladder', 'lantern', 'laptop', 'leaf', 'leg', 'light bulb', 'lighter',
'lighthouse', 'lightning', 'line', 'lion', 'lipstick', 'lobster', 'lollipop', 'mailbox', 'map', 'marker', 'matches',
'megaphone', 'mermaid', 'microphone', 'microwave', 'monkey', 'moon', 'mosquito', 'motorbike', 'mountain', 'mouse',
'moustache', 'mouth', 'mug', 'mushroom', 'nail'
```

INaturalist (phylum-level) For the phylum taxonomy, {known class list} (shown below) contains the subset of phyla designated as known in our split (9 classes in total), the first five labels appear in the target dataset. All other phyla that appear in the original INaturalist taxonomy but not in this subset are tested in target dataset and treated as "unknown" classes.

```
'Chlorophyta', 'Cnidaria', 'Ascomycota', 'Tracheophyta', 'Marchantiophyta', 'Basidiomycota', 'Bryophyta',
'Echinodermata', 'Chordata'
```

INaturalist (class-level) For the class taxonomy, {known class list} (shown below) contains 35 out of a total of 51 classes from INaturalist's class taxonomy, the first 20 of which are classes shared in the target dataset, and the rest of class taxonomy in INaturalist are treated as "unknown" classes.

```
'Arthoniomycetes', 'Aves', 'Marchantiopsida', 'Ulvophyceae', 'Jungermanniopsida', 'Malacostraca', 'Hexanauplia',
'Pucciniomycetes', 'Polychaeta', 'Pinopsida', 'Sphagnopsida', 'Clitellata', 'Diplopoda', 'Dacrymycetes', 'Echinoidea',
'Mammalia', 'Polypodiopsida', 'Dothideomycetes', 'Agaricomycetes', 'Florideophyceae', 'Polytrichopsida', 'Ascidiacea',
'Lecanoromycetes', 'Hydrozoa', 'Reptilia', 'Amphibia', 'Holothuroidea', 'Lycopodiopsida', 'Bivalvia', 'Chilopoda',
'Cycadopsida', 'Actinopterygii', 'Liliopsida', 'Insecta', 'Tremellomycetes'
```

A.3 Prompt Variants for Self-Consistency with Rephrasing

Below are the five semantically equivalent visual-arm prompt variants used for the V-rephrase baseline. All five share the same system prompt as the original visual arm (see A.2). For each sample, $k = 3$ prompts are selected out of the 8 randomly.

```
<User Prompt var 1>
Check if {this image} belongs to one category from this list:  {classlist}.
```

```
<User Prompt var 2>
Given {this image} and allowed labels {classlist}, decide if it is in-list or out-of-list.
```

```
<User Prompt var 3>
Classify {this image} using ONLY these categories::  {classlist}.
```

```
<User Prompt var 4>
Is {this image} an instance of any class in:  {classlist}.
```

```
<User Prompt var 5>
From the candidate labels {classlist}, pick the best class and report if none is exact.
```

```
<User Prompt var 6>
Determine whether this image matches one of the following categories:  {classlist}.
```

```
<User Prompt var 7>
Use this closed-set label list {classlist} to predict the closest category for {this image}.
```

```
<User Prompt var 8>
Analyze {this image} against this category vocabulary {classlist} and return the best label.
```

Algorithm 1: Parse VLM Response

Input: string r (raw VLM response)
Output: string *class_name*, string *unknown*
unknown $\leftarrow$ "";
class_name $\leftarrow$ "";
try:
parts $\leftarrow$ split(r, ',') ; // split on comma
foreach $p \in parts$ **do**
  if *contains*(p, "*unknown*") **then**
    $v \leftarrow$ split(p, "*unknown* : ")[-1];
    $v \leftarrow$ trimWhitespace(v);
    $v \leftarrow$ stripLeadingTrailingQuotes(v);
    $v \leftarrow$ removeAllSingleQuotes(v);
    unknown $\leftarrow v$;
  end
  if *contains*(p, "*class_name*") **then**
    $v \leftarrow$ split(p, "*class_name* : ")[-1];
    $v \leftarrow$ trimWhitespace(v);
    $v \leftarrow$ stripLeadingTrailingQuotes(v);
    $v \leftarrow$ removeAllSingleQuotes(v);
    class_name $\leftarrow v$;
  end
end
except Exception:
unknown $\leftarrow$ "";
class_name $\leftarrow$ "";
return (*class_name*, *unknown*) ;

A.4 Pseudo-code for Parsing Function

We next describe the parsing function $\phi(\cdot)$ that takes the raw VLM response string to recover the predicted label and unknown flag in Algorithm 1.

A.5 Experimental and Implementation Details

The source code of our paper can be found here `https://anonymous.4open.science/r/Fallback-Enabled-Closed-Set-Classification-8433/`.

A.5.1 Implementation Details

Input Preprocessing Each image is resized to 224×224 and normalized first. For VLMs, the preprocessed image is serialized as a base64-encoded data before sent to the model, as required by the APIs.

Randomness All experiments for DomainNet and VisDA are ran with a fixed random seed 2025 to make dataset sampling and baseline training reproducible. As for INaturalist, we use random seed 2 for phylum-level partition, and random seed 0 for class-level partition, to obtain a mixture of imbalanced classes for tested target dataset, the details of which can be found in Section A.9. For API-based VLMs (Gemini and GPT-4o-mini), we set the API seed to 2025. This makes repeated calls with the same input and prompt as deterministic as the API allows.

VLM Hyperparameter Configuration For both Gemini 2.0-flash and GPT4o-mini, we use the same configuration: Temperature = 1.0, maximum output token = 300 for all prompts, except for image summary prompt variation 1, for which we set maximum output token =150. Top-k and other sampling related parameters are left as default.

A.6 Correlated Error Analysis

The framework assumes that the visual arm and textual arm fail independently; however, given that both arms are instantiated from a single VLM, their errors are inherently correlated. When the model systematically misunderstands an object class, predicting on an identical incorrect label across both arms prevents the consistency mechanism from identifying such failures. To empirically quantify this concern, we analyze error correlation across all test datasets with Qwen 2.5-7B-VL in Table 7. While non-negligible, the analysis indicates that even though this type of error exists in practice, it enables the cross-modal consistency check to provide meaningful improvements in unknown detection.

Table 7: Error Correlation Analysis

	DomainNet			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
Error Ratio	3.51%	3.83%	4.67%	0.02%	3.75%	2.40%	3.03%

A.7 Understanding the paradox

A structural parallel in CLIP. The pattern of competence within individual modalities coexisting with failure at cross-modal binding is not unique to our setting. Kouwenhoven et al. (2025) demonstrates a structurally analogous failure in CLIP using the bouba-kiki effect, a well-established phenomenon in which humans reliably associate pseudowords like "bouba" with round shapes and "kiki" with jagged ones. Their experiments reveal that both CLIP's visual encoder and text encoder can individually separate curved from jagged shapes and their corresponding pseudowords in embedding space: the individual modalities are, in principle, separable. Yet when asked to *bind* these representations across modalities – matching pseudowords to the shapes they should evoke, the performance collapses to chance level. Neither the ResNet nor ViT variant of CLIP consistently maps both labels of a pseudoword pair to their intended shapes at above-chance rates, despite the bouba-kiki word pairs being among the most likely cross-modal associations to appear in training data.

Shared root cause: learning about ≠ mechanistically implementing We argue that both failures: CLIP's inability to bind phonetic form to visual shape, and our VLMs' inability to bind categorical knowledge to set-membership judgment, share a common root cause in their training objectives. CLIP's contrastive loss optimizes sentence-level alignment between images and text, but does not specifically promote learning fine-grained relations between sub-lexical phonetic features and visual geometry. Analogously, generative VLMs trained with next-token prediction learn *about* the concept of "unknown" from pretraining data, which means they can describe what it means for an object to fall outside a category set, but this semantic knowledge does not translate into a mechanistic procedure for computing set-membership against an arbitrary label list at inference time. In CLIP's case, the model has encountered the bouba-kiki association in its training distribution but lacks the internal mechanism to implement it; in our case, the VLM understands the logic of known/unknown boundaries but lacks the internal mechanism to enforce them. Neither training objective teaches the binding operation the task requires.

A.8 Extra Experimental Results

Comparing Table 1 and Table 8, we see that the VLMs, especially Qwen 2.5-7B-VL and Gemini-2.0-flash, with our framework, consistently outperform SF-UniDA baselines in both weighted and unweighted mean accuracies. In the two INaturalist scenarios, results in Table 8 show that mean accuracies give equal weight to each class and therefore amplifies the influence of rare tail classes, which leads to much lower mean accuracies. The results review that these tail classes remain harder even for the strongest VLMs.

Furthermore, comparing Table 1 and Table 9 on INaturalist, we show that changing the prompt for image summary generation mostly change the trade-off between known and unknown performance rather than giving a uniformly better solution. Only results on open-source VLMs are reported in Table 9 concerning cost of commercial APIs. The difference is that Table 1 uses image summaries that ignore the background / context, while Table 9 utilizes image summaries with more background description. On INaturalist, we see these variants nudging the balance in predictable ways: Variation 1, which produces more context for image summary, helps recover more known samples but at the cost of not recognizing more unknowns. These patterns are tightly linked to INaturalist's fine-grained and contextual nature: species-level distinctions often rely on subtle local cues.

Complementing these results, Table 10 reports H scores under the same original prompt and shows that VLMs with our framework consistently achieve higher H score performance than SF-UniDA baselines across DomainNet, VisDA, and INaturalist, confirming that the gains from cross-modal consistency is stable with different prompt designs.

Table 8: Comparison to SoTA SF-UniDA methods on mean accuracies for known (acc_{k-w}) and unknown (acc_{u-w}) classes. The results are formatted as acc_{k-w}/acc_{u-w}. The last four rows show results with VLM models using our cross-modal consistency method and **prompt variation 3 for generating image summary**. (Best in bold and second best in underline)

Methods	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	45.8/77.4	50.8/80.1	37.3/76.8	75.2/67.2	35.4/25.6	34.9/63.2	46.6/65.1
LEAD	37.6/71.5	48.5/72.0	34.0/70.7	68.4/**86.4**	41.3/44.1	**68.4**/**86.4**	49.7/71.9
Cross-Modal Consistency (V+T)(Ours):							
GPT 4o-mini	**73.3**/64.1	**85.4**/63.8	**80.6**/61.5	**95.1**/37.4	82.7/54.1	49.7/70.7	**77.8**/58.6
Gemini	71.2/69.3	74.4/73.8	71.9/74.1	90.9/60.5	**84.2**/49.2	54.7/72.1	74.6/66.5
Llama	48.8/**88.2**	55.0/**91.8**	41.2/**91.3**	56.5/76.5	51.0/**78.8**	22.3/85.9	45.8/**85.4**
Qwen	71.5/82.8	84.0/87.4	76.6/83.9	83.4/77.0	49.9/65.1	41.4/82.1	67.8/79.7

Table 9: Comparison to SoTA SF-UniDA methods on weighted accuracies for known (acc_{k-w}) and unknown (acc_{u-w}) classes. The results are formatted as acc_{k-w}/acc_{u-w}. The last two rows show results with VLM models using our cross-modal consistency method and **the original prompt for generating image summary**. (Best in bold and second best in underline)

Methods	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	45.9/77.3	49.1/80.5	37.3/77.7	69.1/67.5	61.7/29.9	77.8/76.0	56.8/68.2
LEAD	41.1/68.7	51.9/73.3	34.8/70.0	72.6/**87.1**	77.6/65.6	63.5/64.6	56.9/71.6
Cross-Modal Consistency (V+T)(Ours):							
Llama	48.0/**88.4**	55.3/**91.0**	41.9/**91.8**	55.3/74.9	88.5/54.6	54.4/**93.5**	57.2/82.4
Qwen	**63.1**/86.6	**81.2**/86.8	**71.3**/85.8	**82.1**/58.9	**96.5**/**99.9**	**80.9**/92.3	**79.2**/**85.1**

A.9 Dataset Details

For INaturalist, we construct two partitions that operate at different levels of the taxonomy. Both partitions inherit the severe class imbalance characteristic of INaturalist. As shown in Fig 5, the target-label bar plots follow a long-tailed distribution: a few head categories have thousands of samples, whereas many tail categories have only a handful of images. This imbalance makes the problem challenging for both SF-UniDA baselines and our VLM-based method, since rare classes contribute the same weights for estimating per-class performance as head classes with naive averaging. This motivates our use of balance-aware metrics (e.g., weighted accuracies and balance ratio) when comparing known vs. unknown performance in these INaturalist scenarios.

Table 10: H score (%) comparison with SoTA SF-UniDA methods. The first two rows show the SF-UniDA baseline results, cited from Qu et al. (2024). The last two rows show results with VLM models using our cross-modal consistency method and **the original prompt for generating image summary**. (best in bold and second best is underlined).

Methods	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	52.3	62.3	50.7	<u>73.1</u>	32.4	<u>47.8</u>	53.1
LEAD	52.1	62.4	51.2	**76.8**	50.7	46.9	56.7
Llama	<u>60.9</u>	<u>68.2</u>	<u>56.8</u>	63.9	<u>55.3</u>	38.0	<u>57.2</u>
Qwen	**73.2**	**83.9**	**77.9**	72.6	**79.2**	**61.3**	**74.7**

Table 11: Comparison to SoTA SF-UniDA methods on the balance ratio R. $R \rightarrow 1$ indicates balanced behavior for known and unknown recognition. (Best in bold and second best in underline)

Method	**DomainNet**			**VisDA**	**INaturalist**		**Avg**
	Painting	Real	Sketch		Phylum	Class	
GLC	0.599	0.614	0.481	**1.025**	2.063	<u>1.023</u>	<u>0.833</u>
LEAD	0.600	0.710	0.502	0.834	<u>1.182</u>	**0.983**	0.736
OSR w/ CLIP	0.959	1.062	0.944	0.569	0.309	0.187	0.672
Self-Consistency MV (Llama):	1.276	1.265	1.160	1.291	4.870	1.516	1.495
Self-Consistency MV (Qwen):	2.177	2.264	2.316	2.597	1.760	1.565	2.065
Self-Consistency w/ R (Llama):	1.104	1.073	1.366	1.423	9.123	1.832	1.630
Self-Consistency w/ R (Qwen):	2.588	2.505	2.643	3.046	3.457	36.67	3.339
GPT 4o-mini (w/ Ours)	<u>1.190</u>	1.354	1.306	2.501	2.212	1.775	1.732
Gemini (w/ Ours)	**1.046**	**1.004**	**0.971**	1.660	2.224	1.687	1.432
Llama (w/ Ours)	0.560	0.611	0.463	0.722	1.476	0.565	0.733
Qwen (w/ Ours)	0.845	<u>0.953</u>	<u>0.905</u>	<u>1.141</u>	**0.960**	0.852	**0.943**

A.10 Image Examples

Figure 6 first underscores that the underlying visual problem is more complicated than a single ground-truth label suggests. Each example image admits multiple reasonable interpretations: a kitchen scene that could be labeled as dishwasher, refrigerator, or microwave depending on which object the annotator or user focuses on, or an arm with a dreamcatcher tattoo that can be described as arm, tattoo, or even feather. Which label is “correct" is not intrinsic to the images alone but also depends on the annotation convention and on the downstream question when ask about the image. This highlights that the SF-UniDA setting is more complicated and challenging in practice.

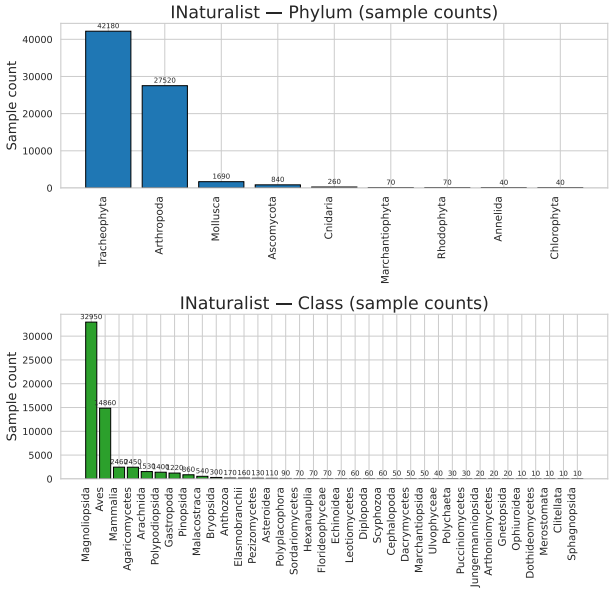
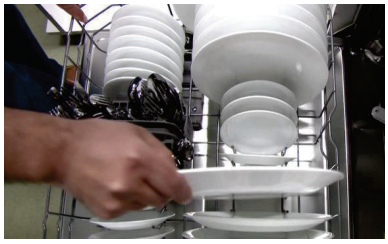
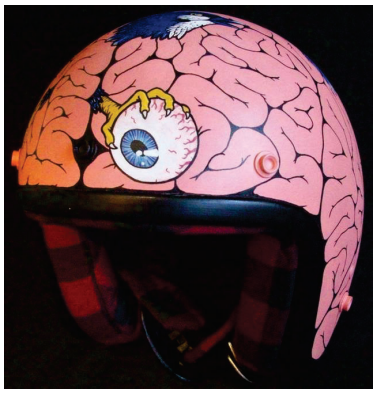


Figure 5: INaturalist Class Imbalance Problem



Ground truth: dishwasher (Real)
Visual Arm Prediction by Gemini-2.0-flash: dishwasher
Textual Arm Prediction by Gemini-2.0-flash: hand

Ground truth: beard (Real)
Visual Arm Prediction by Gemini-2.0-flash: face
Textual Arm Prediction by Gemini-2.0-flash: beard



Ground truth: brain (Painting)
Visual Arm Prediction by Gemini-2.0-flash: helmet
Textual Arm Prediction by Gemini-2.0-flash: helmet

Ground truth: dishwasher (Real)
Visual Arm Prediction by Gemini-2.0-flash: refrigerator
Textual Arm Prediction by Gemini-2.0-flash: refrigerator

Ground truth: arm (Painting)
Visual Arm Prediction by Gemini-2.0-flash: tattoo
Textual Arm Prediction by Gemini-2.0-flash: feather