Delineating Knowledge Boundaries for Honest Large Vision-Language Models

Junru Song[1], Yimeng Hu[1], Yijing Chen[1], Huining Li[1], Qian Li[1], Lizhen Cui[1], and Yuntao Du[1,2⋆]

[1] Joint SDU-NTU Centre for Artificial Intelligence Research & School of Software, Shandong University, P.R. China
[2] State Key Lab. for Novel Software Technology, Nanjing University, P.R. China

Abstract. Large Vision-Language Models (VLMs) have achieved remarkable multimodal performance yet remain prone to factual hallucinations, particularly in long-tail or specialized domains. Moreover, current models exhibit a weak capacity to refuse queries that exceed their parametric knowledge. In this paper, we propose a systematic framework to enhance the refusal capability of VLMs when facing such unknown questions. We first curate a model-specific "Visual-Idk" (Visual-I don't know) dataset, leveraging multi-sample consistency probing to distinguish between known and unknown facts. We then align the model using supervised fine-tuning followed by preference-aware optimization (e.g., DPO, ORPO) to effectively delineate its knowledge boundaries. Results on the Visual-Idk dataset show our method improves the Truthful Rate from 57.9% to 67.3%. Additionally, internal probing also demonstrates that the model genuinely recognizes its boundaries instead of just memorizing refusal patterns. Our framework further generalizes to out-of-distribution medical and perceptual domains, providing a robust path toward more trustworthy and prudent visual assistants.

Keywords: Large Vision-Language Models · Knowledge Boundaries · Preference Optimization

1 Introduction

The emergence of Large Vision-Language Models (VLMs) has marked a transformative milestone in multimodal artificial intelligence, enabling AI assistants to engage in complex reasoning and open-ended dialogue based on visual inputs [19]. Despite their impressive performance, these models are still highly prone to factual hallucinations [16]. In many cases, when confronted with questions involving images that contain long-tail entities or specialized domains beyond the model's parametric knowledge, such as rare biological species or complex medical radiographs, VLMs tend to produce confident yet entirely incorrect responses rather than acknowledging their knowledge limitations [4]. Such "blind

⋆ Corresponding authors.

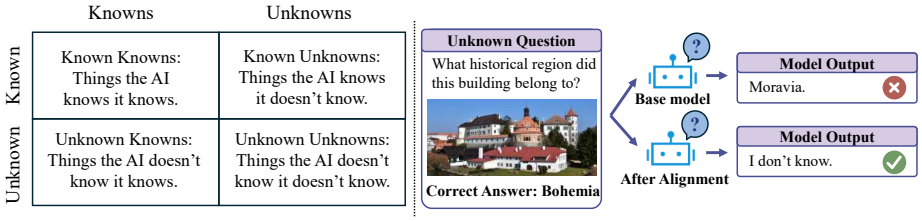

Fig. 1: Left: The four knowledge quadrants of a VLM, adapted from [6]. Our goal is to transform Unknown Unknowns (hallucinations) into Known Unknowns (standard refusals). **Right:** Comparison of model behaviors on an unknown question. The model after alignment correctly identifies its internal knowledge gap and provides a truthful refusal ("I don't know"), while the base model generates an incorrect response ("Moravia").

confidence" poses significant risks in high-stakes applications like medical diagnosis or legal document analysis, where the cost of misinformation is prohibitively high [13]. Consequently, a fundamental question arises for the more responsible AI: **Can visual assistants discern their own knowledge boundaries and articulate this awareness through natural language?**

Existing research on mitigating VLM hallucinations has primarily focused on reducing object-level misidentifications [28] or improving cross-modal alignment through reinforcement learning from human feedback (RLHF) [30]. While some studies [6] in the LLM domain have explored the "I don't know" (Idk) response for pure text, the multimodal landscape remains largely unaddressed. Current VLM "refusal" studies often focus on perceptual uncertainty, where images are intentionally blurred or corrupted [9]. However, we argue that the more insidious challenge lies in **epistemic uncertainty**, where the visual input is perfectly clear, but the model lacks the internal parametric knowledge to interpret it [27]. As illustrated in the knowledge quadrants of **Fig. 1 (left)**, vanilla VLMs often fall into the *Unknown Unknowns* (IDK-IDK) segment, where they generate confident but fabricated claims for entities they do not actually master. Current alignment paradigms emphasize "helpfulness" at the expense of "truthfulness" [17], inadvertently forcing models to guess in the absence of evidence.

To overcome this challenge, In this paper, we propose the **Visual-Idk alignment framework**, a systematic pipeline to calibrate Visual Assistants with their internal knowledge boundaries. Our methodology consists of three core stages: (i) Visual-Semantic Sample Selection, starting with an existing dataset, we utilize a powerful VLM to filter low-quality samples; (ii) Visual Knowledge Probing via multi-sample consistency to quantify parametric mastery; and (iii) Preference Pair Generation, which constructs contrasting samples to supervise the model's decision-making. We then employ preference-aware optimization (e.g., DPO, ORPO) to reshape the model's epistemic boundaries, cultivating a more truthful and prudent assistant.

We conduct extensive experiments, and the results on the V-Idk test set show that the framework significantly enhances reliability, improving the TRUTHFUL

rate of LLaVA-1.5-7B from **57.9% to 67.3%**. Notably, ORPO achieves a superior balance and preserves mastered knowledge (IK-IK) more effectively than Supervised Fine-Tuning (SFT). Besides, the proposed framework is domain-agnostic. It could generalize robustly to ScienceQA [21] dataset as well as OOD scenarios such as blurred images (VizWiz [9]) and specialized medical imagery (PMC-VQA [33]). The generalization results suggest that the alignment internalizes a fundamental awareness of knowledge mastery rather than overfitting to specific entities. Furthermore, internal probing of model Logprob [10] and PPL confirms that the model truly learns to recognize its own limits instead of just memorizing a response pattern.

To sum up, our contributions are summarized as follows:

(1) We introduce a framework that effectively improves epistemic uncertainty by teaching models to express unknowns for the question that goes beyond parametric knowledge.

(2) We propose a novel pipeline for constructing knowledge-intensive preference datasets, specifically designed to identify the epistemic boundaries of VLMs regarding their parametric knowledge.

(3) We conduct extensive experiments on multiple datasets, and the results show that the proposed method could improve truthfulness effectively.

2 Related Work

2.1 Knowledge Boundary

Despite the rapid advancement of Large Vision-Language Models (VLMs) such as LLaVA [20], InstructBLIP [7], and Qwen-VL [26], factual hallucination remains a persistent bottleneck [14, 15]. Prior work commonly categorizes VLM hallucinations into object-level errors (e.g., mentioning nonexistent objects) and attribute-level errors (e.g., incorrect colors, counts, or spatial relations). To mitigate these issues, existing approaches have explored (i) scaling up high-quality multimodal instruction data [18], (ii) post-hoc verification with external tools or retrieval modules [29], and (iii) architectural or training refinements that improve cross-modal alignment [30]. While effective to varying degrees, these methods largely share a common goal: making models answer more accurately. Yet, an equally important capability is often overlooked—knowing when not to answer. In practice, VLMs are frequently incentivized to guess plausible responses even when confronted with long-tail or specialized entities that lie beyond their parametric knowledge, resulting in confident but unfounded claims. This motivates a complementary perspective: aligning VLMs to be explicitly aware of their knowledge boundaries and to appropriately abstain when visual evidence is clear but the required world knowledge is missing or uncertain.

The pursuit of honest AI has been more systematically investigated in the text-only domain, where researchers have shown that LLMs can exhibit an intrinsic ability to assess the likelihood of their own claims. Building on this observation, Cheng et al. [6] proposed the "Idk" (I don't know) framework, which teaches LLMs to recognize and verbalize their knowledge limits by constructing model-specific datasets and applying supervised fine-tuning (SFT) followed

by preference optimization. Related efforts, including R-Tuning [31] and Hindsight Instruction Relabeling [32], further explored learning-to-refuse behaviors for unanswerable or underspecified queries. Our work follows this technical lineage but extends it to the multimodal setting, where refusal decisions depend on a more entangled interplay between visual perception (what is present in the image) and internal knowledge mastery (what the model actually knows about what it sees). Different from InBoL [27], which explores generic knowledge boundaries by focusing on questions that are unanswerable by nature, our framework targets epistemic gaps in more knowledge-intensive scenarios involving long-tail and professional domains. Consequently, we focus on cultivating VLMs' self-awareness of knowledge boundaries—not merely improving answer quality, but enabling prudent abstention when the query exceeds the model's reliable knowledge.

2.2 Preference-Aware Optimization

Direct Preference Optimization (DPO) [24] and its variants (e.g., ORPO [11], CPO [25]) have emerged as practical alternatives to reinforcement learning from human feedback (RLHF) by directly optimizing models from pairwise preference data without the need for explicit reward modeling. Compared with traditional RLHF pipelines that require reward model training and reinforcement learning stages, preference-based optimization provides a more stable and computationally efficient framework for aligning large models.

In the text-only domain, the preference optimization has been widely adopted to align helpfulness and reduce unsafe or untruthful model behaviors [6] [3]. In the multimodal setting, recent studies have also explored preference-based alignment for VLMs, improving instruction-following and factuality and thereby reducing hallucinations [28] [23]. Unlike existing VLM preference-optimization that focuses on answer quality, we target the answer–refusal boundary to mitigate over-refusal. We build upon these preference-aware techniques to explicitly calibrate the answer–refusal boundary, encouraging truthful abstention under epistemic uncertainty while preserving helpfulness on answerable queries, thereby alleviating the over-refusal (alignment tax) often observed in SFT-only refusal tuning.

3 The Visual-Idk Alignment Framework

In this section, we detail our framework for aligning Visual Assistants to recognize their knowledge boundaries. Our approach follows a transition from a non-training baseline to sophisticated training-based alignment, aimed at calibrating the model's epistemic uncertainty.

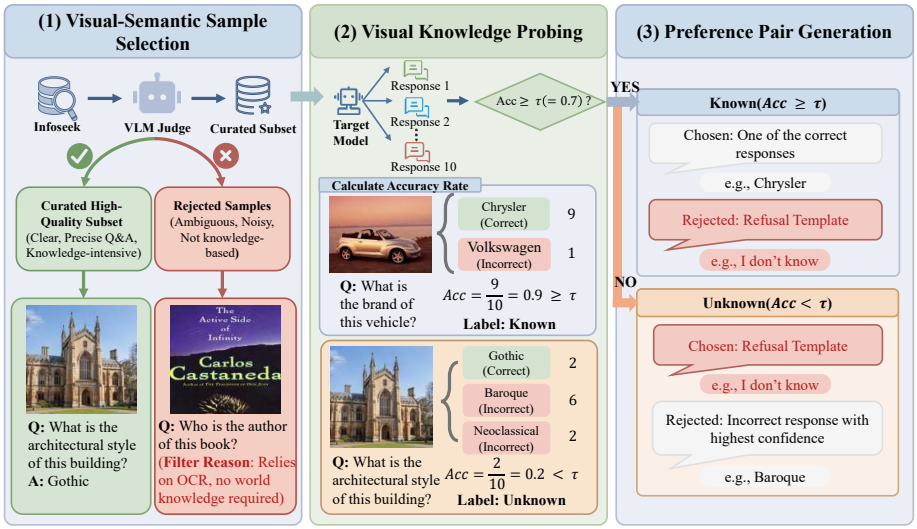

Fig. 2: The pipeline of Visual-Idk dataset construction. The process comprises three major stages: (1) Visual-Semantic Sample Selection for quality control; (2) Visual Knowledge Probing to identify model-specific knowledge boundaries; and (3) Preference Pair Generation.

3.1 Construction of the Visual-Idk Dataset

The foundation of our approach is the creation of a model-specific Visual-Idk dataset. As illustrated in Fig. 2, this process identifies the specific parametric knowledge gaps of the target model while ensuring data quality.

Visual-Semantic Sample Selection. To isolate *epistemic uncertainty* from *perceptual errors*, we first perform a high-quality sample selection process from an existing dataset (Stage 1 in Fig. 2). We leverage a high-capability VLM model (e.g., Qwen2.5-VL) as a semantic judge to curate image-question pairs from the initial InfoSeek dataset. We retain only those samples meeting three specific criteria: (1) visual quality is clear and unambiguous, (2) the ground-truth answer is precise, and (3) the query is knowledge-intensive rather than relying on basic optical character recognition (OCR) or simple perception. This rigorous selection ensures that model failures on the resulting curated subset can be reliably attributed to parametric knowledge gaps.

Visual Knowledge Probing. On the filtered dataset, we assess the target model π_θ via an empirical probing process (Stage 2 in Fig. 2). We sample $N = 10$ independent responses for each image-question pair (x, q) using a high decoding temperature. We define a mastery threshold τ (the *Ik threshold*). A sample is categorized as known if its average accuracy rate of 10 responses $A(x, q) \geq \tau$, indicating the model can consistently retrieve the correct fact. Otherwise, it is labeled as Unknown.

Preference Pair Generation. Based on the probing results, we construct the V-Idk training set $\mathcal{D}_{V-Idk}$ by generating preference pairs (y_w, y_l) to contrast desirable behaviors (Stage 3 in Fig. 2):

(1) For Known Facts ($A \geq \tau$): To prevent over-conservatism (the "alignment tax"), we set the correct response as the chosen y_w and a refusal template (e.g., "I don't know") as the rejected y_l.

(2) For Unknown Facts ($A < \tau$): To suppress hallucinations, we set the refusal template as the chosen y_w and the model's previously generated incorrect response with the highest confidence as the rejected y_l.

Dataset Statistics and Source. We construct the Visual-Idk dataset based on the InfoSeek benchmark [5]. We prioritize InfoSeek over traditional VQA datasets (e.g., VQA v2 [8] or OK-VQA [22]) due to its extreme knowledge-intensity and long-tail entity distribution. While the original InfoSeek benchmark contains a certain degree of perceptual noise and ambiguous queries, our visual-semantic sample selection ensures that the resulting curated subset maintains high visual and textual quality. This refinement renders the data highly conducive to the task of knowledge boundary identification: by minimizing visual ambiguity and reasoning noise, model failures on this subset can be more reliably attributed to gaps in internal parametric knowledge rather than external perceptual errors.

The finalized V-Idk dataset comprises 20,000 training samples and 3,000 test samples. Regarding its linguistic characteristics, the questions are designed to be concise and information-dense, with a mean length of approximately 10 words, while the corresponding ground-truth answers are highly succinct, averaging 3 to 4 words. Based on our multi-sample consistency probing (with $\tau = 0.7$), the dataset is curated to maintain a specific distribution: approximately 60% of the samples are categorized as "Known" (IK), while the remaining 40% are "Unknown" (IDK). This balanced ratio ensures that the alignment process provides sufficient supervision for both preserving helpfulness on mastered facts and suppressing hallucinations on unknown entities.

3.2 Experimental Methods

We include several popular methods. The first is simple prompting. Besides, we also investigate four training strategies to explicitly align the model with its knowledge boundaries using the V-Idk dataset.

Idk-Prompting. As a zero-training baseline, we evaluate the model's latent self-awareness via Idk-Prompting. We prepend an explicit instruction to the input: *"Answer the following question based on the image. If you do not know the answer, please respond with 'I'm sorry, this question is beyond my knowledge. I don't know the answer.' "* This strategy relies purely on the model's pre-trained instruction-following capability without any parameter updates.

Supervised Fine-tuning (SFT). We first perform Idk-SFT by maximizing the log-likelihood of the chosen responses y_w in $\mathcal{D}_{V-Idk}$:

$$\mathcal{L}_{SFT} = -\mathbb{E}_{(x,q,y_w)\sim\mathcal{D}_{V-Idk}} \left[\log \pi_\theta(y_w|x,q)\right] \tag{1}$$

SFT serves as the first stage of training to teach the model the standardized refusal format.

Direct Preference Optimization (DPO). To further calibrate the decision boundary, we use DPO to contrast honest refusals with hallucinations. The loss is defined as:

$$\mathcal{L}_{DPO} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,q)}{\pi_{ref}(y_w|x,q)} - \beta\log\frac{\pi_\theta(y_l|x,q)}{\pi_{ref}(y_l|x,q)}\right)\right] \tag{2}$$

where π_{ref} is the model after SFT.

Contrastive Preference Optimization (CPO). We also evaluate CPO, which avoids the need for a reference model by directly optimizing a contrastive objective between y_w and y_l. It encourages the model to stay close to the chosen distribution while moving away from the rejected one.

Odds Ratio Preference Optimization (ORPO). Finally, we utilize ORPO to combine supervised learning and preference alignment into a single objective:

$$\mathcal{L}_{ORPO} = \mathcal{L}_{SFT} + \lambda\cdot\mathcal{L}_{OR}, \quad \mathcal{L}_{OR} = -\log\sigma\left(\log\frac{odd_{\pi_\theta}(y_w|x,q)}{odd_{\pi_\theta}(y_l|x,q)}\right) \tag{3}$$

where $odd_\pi(y) = \frac{\pi(y)}{1-\pi(y)}$. ORPO regularizes the model to increase the relative odds of being honest for unknown facts while remaining helpful for known facts.

3.3 Evaluation Metrics

To formalize the assessment of a model's knowledge boundaries, we categorize the aligned model's inference behaviors into four quadrants, as illustrated in **Fig. 1 (left)**. These quadrants, *Known Knowns*, *Known Unknowns*, *Unknown Knowns*, and *Unknown Unknowns*, provide a granular mapping of the model's epistemic state by intersecting its initial mastery level $A(x,q)$ (from Sec. 3.1) with the semantic correctness of its output during testing. We define the evaluation dataset $\mathcal{D}$ and denote the subset of initially unmastered samples (the "Unknowns" column in **Fig. 1**) as the *Unknown* subset $\mathcal{D}_{unk} = \{(x,q)\in\mathcal{D} \mid A(x,q) < \tau\}$.

Given that model responses are in natural language, we utilize a high-capability LLM as a semantic judge to evaluate model behavior. We define a logical verification predicate $\mathcal{V}(y_{pred}, \text{target})$ that returns *True* if the semantic content of y_{pred} aligns with the intended *target* (either the ground-truth answer y_{gt} or a refusal intent). Based on this, we define three key metrics:

IK-IK Rate ($\rho_{\mathbf{IK}}$) (Knowledge Accuracy): Quantifies the model's effective factual knowledge base during inference. As visually represented by the transition in **Fig. 1 (right)**, a sample is classified as IK-IK if the model provides a correct factual response, regardless of whether the sample was initially categorized as Known or Unknown. It is defined as:

$$\rho_{\text{IK}} = \frac{\sum_{(x,q)\in\mathcal{D}}\mathbb{I}\left(\mathcal{V}(y_{pred}, y_{gt}) = \text{True}\right)}{|\mathcal{D}|} \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This metric reflects the model's ability to retrieve and articulate correct facts across the entire domain.

IK-IDK Rate ($\rho_{\mathbf{IDK}}$) (Admitting Ignorance): Measures the model's self-awareness by quantifying its ability to correctly refuse queries for which it lacks sufficient parametric knowledge ($A(x, q) < \tau$). Conceptually, this represents the model successfully landing in the *Known Unknown* quadrant (the top-right cell of **Fig. 1**). It is defined as:

$$\rho_{\text{IDK}} = \frac{\sum_{(x,q)\in\mathcal{D}_{unk}} \mathbb{I}\left(\mathcal{V}(y_{pred}, \text{refusal}) = \text{True}\right)}{|\mathcal{D}|} \tag{5}$$

TRUTHFUL Rate ($\mathcal{T}$) (Overall Reliability): Our primary evaluation metric, representing the total frequency of "honest" behaviors, either providing a correct factual answer or correctly admitting ignorance when mastery is low:

$$\mathcal{T} = \rho_{\text{IK}} + \rho_{\text{IDK}} \tag{6}$$

Maximizing the TRUTHFUL rate is the central objective of our alignment framework, as it captures the optimal balance between being helpful and being honest.

Consistent with this framework, the remaining behaviors are classified as IDK-IK (the "alignment tax," where the model refuses a known fact) and IDK-IDK (hallucinations, where the model provides an incorrect answer to an unknown fact, illustrated by the Base Model in **Fig. 1 (right)**). All metrics are normalized by the total dataset size $|\mathcal{D}|$.

4 Experiments

In this section, we evaluate the efficacy of our framework in aligning Visual Assistants with their internal knowledge boundaries. We present the results across five experimental dimensions: (1) main performance on V-Idk, (2) cross-dataset generalization on ScienceQA, (3) Generalization to out-of-distribution (OOD) scenarios, (4) internal probabilistic probing, and (5) case study.

4.1 Implementation Details

Training Configurations. We implement our framework using PyTorch and the DeepSpeed library. All models are trained on a server equipped with 2 NVIDIA H800 (80G) GPUs. We adopt Low-Rank Adaptation (LoRA) [12] for parameter-efficient fine-tuning, targeting all linear modules in the LLM backbone with a rank $r = 64$ and $\alpha = 128$.

The alignment process consists of two stages: (1) Supervised Fine-tuning (SFT): The model is trained on the curated 20,000 V-Idk training samples for 1 epoch. We employ the AdamW optimizer with a learning rate of 1×10^{-5} and a global batch size of 32. (2) Preference Optimization: For DPO, CPO, and ORPO stages, the model is further aligned for 3 epochs using 20,000 preference

Table 1: Main Results on V-Idk. Comparison of alignment methods on LLaVA-1.5 models across two inference settings. Subscripts denote improvement (↑) or decrement (↓) over the baseline.

Model	Method	Zero-shot			Few-shot		
		IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL
Data Proportion		*60.0*	*40.0*	*100.0*	*60.0*	*40.0*	*100.0*
LLaVA-1.5 7B	Base	45.8	1.8	47.6	**45.5**	12.4	57.9
	SFT	34.4 ↓11.4	**32.7** ↑30.9	**67.1** ↑19.5	27.6 ↓17.9	**36.8** ↑24.4	64.4 ↑6.5
	DPO	**51.9** ↑6.1	5.5 ↑3.7	57.4 ↑9.8	45.3 ↓0.2	17.8 ↑5.4	63.1 ↑5.2
	CPO	44.4 ↓1.4	13.9 ↑12.1	58.3 ↑10.7	42.8 ↓2.7	24.5 ↑12.1	**67.3** ↑9.4
	ORPO	45.9 ↑0.1	14.1 ↑12.3	60.0 ↑12.4	40.0 ↓5.5	26.5 ↑14.1	66.5 ↑8.6
LLaVA-1.5 13B	Base	49.9	4.3	54.2	49.6	19.8	69.4
	SFT	34.4 ↓15.5	**39.1** ↑34.8	**73.5** ↑19.3	27.5 ↓22.1	**39.8** ↑20.0	67.3 ↓2.1
	DPO	**50.2** ↑0.3	6.7 ↑2.4	56.9 ↑2.7	**49.7** ↑0.1	21.4 ↑1.6	71.1 ↑1.7
	CPO	44.9 ↓5.0	19.0 ↑14.7	63.9 ↑9.7	43.7 ↓5.9	30.3 ↑10.5	74.0 ↑4.6
	ORPO	45.8 ↓4.1	20.4 ↑16.1	66.2 ↑12.0	42.7 ↓6.9	33.9 ↑14.1	**76.6** ↑7.2

pairs. The learning rate is reduced to 1×10^{-6} with a cosine decay scheduler, maintaining the same LoRA configuration ($r = 64$).

Inference and Evaluation. During inference, we evaluate the aligned models on the 3,000 V-Idk test samples with a distribution of 60% Known (IK) and 40% Unknown (IDK) samples. We utilize greedy decoding to ensure deterministic results. As established in Sec. 3.1, each test sample is categorized into knowledge quadrants by comparing the model's output against the ground-truth and refusal templates via a high-capability LLM judge. To assess robustness, we report results under both *zero-shot* and *few-shot* settings as described in Sec. 4.2.

4.2 Main Results on V-Idk

We evaluate the performance of various alignment methods on the V-Idk test dataset under two distinct inference protocols to assess the model's robustness in discerning knowledge boundaries: **(1) Zero-shot**: The model is provided only with the image and the query. No explicit instructions to abstain are given, nor are any refusal templates provided in the prompt. This setting evaluates the *intrinsic* truthfulness and self-awareness internalized by the model during the alignment process. **(2) Few-shot**: The model is explicitly instructed: *"If you do not know the answer, please respond with 'I don't know'."* Furthermore, two in-context demonstrations are provided: one showing a correct factual response to a known entity and another illustrating a standard refusal for an unknown entity. This setting measures the model's ability to perform *guided* boundary calibration via in-context learning.

Table 1 compares the baseline *Prompting* with four training-based strategies across LLaVA-1.5 7B and 13B models. Based on the results, we have the following observations:

(1) Vanilla VLMs exhibit inherent overconfidence. The *Prompting* baseline in the zero-shot setting (7B) achieves a high IK-IK rate (45.8) but a

negligible IK-IDK rate (1.8). This confirms that pre-trained VLMs, while knowledgeable, default to generating fabricated responses when encountering parametric gaps. Even under few-shot guidance, overconfidence remains a bottleneck, as the model continues to guess for the majority of unknown queries.

(2) SFT enhances refusal capability but induces substantial suppression of mastered knowledge. While SFT on V-Idk yields substantial gains in IK-IDK, it incurs a significant "alignment tax." Specifically, for the 7B model, IK-IK accuracy decreases by 11.4% in zero-shot and even more substantially by 17.9% in the few-shot setting. This suggests that while SFT teaches a refusal format, it tends to over-generalize this behavior into a conservative prior, causing the model to erroneously reject queries it has actually mastered.

(3) Preference optimization methods achieve a more calibrated knowledge–refusal balance. Preference-aware methods (DPO, CPO, and ORPO) provide a superior trade-off compared to SFT. Instead of a blanket refusal prior, these methods learn to differentiate between mastery and ignorance. Among them, **ORPO** demonstrates the most prominent overall performance, achieving the leading TRUTHFUL scores in the 13B few-shot setting (76.6) and 7B zero-shot setting (60.0). This indicates that jointly optimizing for supervised likelihood and preference odds allows the model to curb hallucinations effectively while maintaining high factual utility.

(4) Model scaling facilitates more precise boundary identification. Larger models (13B) consistently outperform their 7B counterparts across both truthfulness and knowledge retention. This suggests that increased parametric capacity provides a more stable representation of internal knowledge, enabling the model to respond more robustly to alignment objectives.

(5) Few-shot demonstrations facilitate superior truthfulness calibration. Across all methods, Few-shot inference consistently outperforms Zero-shot in terms of TRUTHFUL rate. This improvement is attributed to the contextual anchors provided by the demonstrations, which help the model better calibrate its response threshold. Notably, the *intrinsic honesty* of aligned models is evident: the ORPO-aligned 7B model in Zero-shot (60.0) even slightly exceeds the performance of the base model in the Few-shot setting (57.9), suggesting that the knowledge boundary has been successfully internalized.

4.3 Insightful Analysis

Cross-Dataset Evaluation on ScienceQA. To evaluate the cross-dataset generalization of the learned honest behavior, we directly evaluate the Visual-Idk-trained models on the ScienceQA benchmark [21]. This experiment assesses whether the alignment internalized from encyclopedic knowledge can successfully transfer to the distinct domain of scientific reasoning.

The evaluation protocol remains strictly consistent with the main experiments: we first perform the same multi-sample consistency probe on the ScienceQA test set to identify the model's mastered and unmastered facts. This unified measurement allows for a direct assessment of how the model's awareness of its knowledge boundaries generalizes across different task distributions

Table 2: Cross-Dataset Evaluation on ScienceQA (Few-shot). Comparison of alignment methods trained solely on Visual-Idk. The evaluation set is balanced with a 50%:50% ratio between Known and Unknown samples. Subscripts (↑) denote the improvement over the baseline.

Method	IK-IK	IK-IDK	TRUTHFUL
Data Proportion	*50.0%*	*50.0%*	*100.0%*
Prompting	30.3	25.8	56.1
SFT	$24.0_{\downarrow 6.3}$	$35.5_{\uparrow 9.7}$	$59.5_{\uparrow 3.4}$
DPO	$\mathbf{33.9}_{\uparrow \mathbf{3.6}}$	$26.2_{\uparrow 0.4}$	$60.1_{\uparrow 4.0}$
CPO	$26.8_{\downarrow 3.5}$	$\mathbf{35.8}_{\uparrow \mathbf{10.0}}$	$62.6_{\uparrow 6.5}$
ORPO	$28.1_{\downarrow 2.2}$	$35.7_{\uparrow 9.9}$	$\mathbf{63.8}_{\uparrow \mathbf{7.7}}$

and data sources. Table 2 summarizes the results for LLaVA-1.5 7B. We observe the following:

(1) Alignment capability successfully transfers to scientific domains. Despite being trained solely on V-Idk, all aligned methods demonstrate a superior ability to discern scientific knowledge boundaries compared to the baseline. ORPO achieves a leading TRUTHFUL score of 63.8%, representing a 7.7% improvement. This suggests that our framework calibrates a fundamental internal confidence that is not overfitted to specific entities but represents a domain-agnostic honesty that generalizes to complex scientific facts.

(2) Consistent behavioral patterns in cross-dataset knowledge retention. The behavioral trade-offs observed on ScienceQA echo our primary findings: SFT remains highly conservative, resulting in a high IK-IDK rate but the lowest IK-IK accuracy (24.0%). Conversely, DPO stands out as the only method that effectively preserves and enhances scientific knowledge mastery, achieving the highest IK-IK rate of 33.9%. ORPO and CPO strike the most effective balance, maintaining superior overall truthfulness while successfully curbing the "alignment tax" on previously unseen scientific knowledge.

Truthfulness Generalization on OOD Scenarios. To evaluate the generalization of the learned honesty, we extend our analysis to VizWiz-Unans [9] and PMC-VQA [33], representing visual and knowledge domain shifts, respectively. The VizWiz-Unans dataset consists of low-quality, blurry, or obscured real-world images that introduce significant perceptual uncertainty. For VizWiz-Unans, the TRUTHFUL rate is therefore defined strictly as the successful refusal rate, since the correct behavior is to abstain from answering when visual evidence is insufficient. On PMC-VQA, we follow the same evaluation protocol as in the main experiments, and compute TRUTHFUL as $\rho_{\text{IK}} + \rho_{\text{IDK}}$.

As illustrated in Fig. 3, we observe the following:

(1) SFT captures generic refusal patterns under visual noise. On the VizWiz-Unans dataset, the SFT-aligned model achieves a peak TRUTHFUL rate of 95.6%, a gain of 11.3% over the baseline. This suggests that supervised fine-

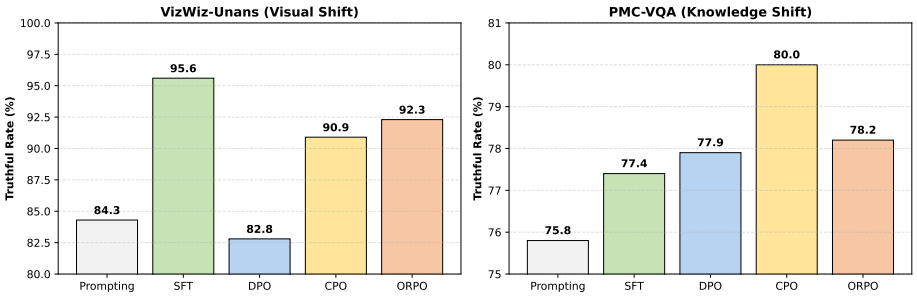

Fig. 3: Generalization on OOD scenarios. Numeric labels denote the TRUTHFUL rate (%).

tuning is effective at establishing a conservative prior that prioritizes abstention when visual signals are severely degraded.

(2) Preference-aware methods excel in specialized domain shifts. While SFT performs well under visual noise, CPO and ORPO demonstrate superior calibration when encountering professional knowledge gaps. Specifically, CPO achieves the leading score of 80.0% on PMC-VQA. This indicates that preference learning better enables the model to identify its lack of expertise in specialized fields like medicine, rather than merely relying on a fixed refusal template.

Internal Uncertainty Analysis. To delve into the model's internal state during the decision-making process, we perform a probabilistic probing analysis focused on the refusal response. Specifically, we force the model to generate the refusal template ("I don't know.") across both "Known" (IK) and "Unknown" (IDK) subsets, then analyze the resulting average log-probabilities (**Logprob** [10]) and **Perplexity (PPL)**. This allows us to discern whether the model's refusal is based on genuine epistemic calibration or mere template mimicry. Table 3 summarizes the transition from *Pre-Align* to *Post-Align* behavior.

(1) Internal Recalibration of Refusal Behavior. After alignment, the model demonstrates a significant shift toward a more confident and stable refusal mode, especially for unmastered facts. For "Unknown" (IDK) questions, the model's confidence in admitting ignorance increases (Logprob: $-0.80 \rightarrow -0.63$), while its uncertainty regarding the refusal template reaches the lowest level (PPL: 1.89). Crucially, the model exhibits higher confidence when refusing truly unknown facts (-0.63) compared to erroneously refusing known facts (-0.76). This probabilistic gap proves that the model has internalized its knowledge boundaries, enabling it to distinguish between justified abstention and unnecessary refusal at a foundational level.

(2) Quantifying the Alignment Tax via Probabilistic Pressure. In the "Known" (IK) category, the Logprob of the refusal template also rises from -0.88 to -0.76. This rise provides probabilistic evidence of the "alignment tax": as the model is incentivized to be truthful, its internal probability mass for refusal

Table 3: Internal Uncertainty Analysis on Visual-Idk. We analyze the average log-probability (Logprob) of generated answers and the Perplexity (PPL) of the refusal template. Pre-Align shows the intrinsic state of the model, while Post-Align reflects the ORPO aligned behavior.

Category	Answer Logprob (↑)		Refusal PPL (↓)	
	Pre-Align	Post-Align	Pre-Align	Post-Align
Known Questions	-0.88	-0.76	2.42	2.14
Unknown Questions	-0.80	**-0.63**	2.22	**1.89**

increases even for mastered facts. However, the refusal PPL for IK remains higher than that for IDK (2.14 vs. 1.89), suggesting the model maintains a healthy internal skepticism toward abstaining from questions it actually knows. This nuanced calibration explains why preference-aware optimization preserves higher IK-IK rates compared to the blanket refusal prior typically learned by SFT.

In conclusion, these results provide robust evidence that our Visual-Idk alignment genuinely recalibrates the model's epistemic uncertainty. The assistant does not just learn to repeat a template; it learns to allocate its probability mass strategically based on its internal knowledge mastery.

4.4 Case Study

To provide a more intuitive understanding of how different alignment strategies reshape the model's knowledge boundaries, we present qualitative examples from the Visual-Idk test set in Fig. 4. These cases highlight the transition from blind overconfidence to prudent abstention. We present qualitative examples in Fig. 4 to illustrate the behavioral shifts after alignment.

Mitigating the Alignment Tax. As shown in the left example, the base model possesses the knowledge of the location (Norway). While SFT induces excessive conservatism and erroneously refuses to answer, preference optimization methods successfully maintain helpfulness by preserving this mastered knowledge. This validates our framework's ability to refine the decision boundary without sacrificing existing utility.

Suppressing Hallucinations. In the right example, the model encounters a specialized entity beyond its mastery. The Base and DPO models generate confident but incorrect responses ("Thomas Telford"). In contrast, CPO and ORPO demonstrate superior epistemic awareness by recognizing the knowledge gap and electing to abstain. Notably, ORPO provides a more context-grounded refusal, suggesting a more calibrated internal understanding of its knowledge limits.

These cases confirm that V-Idk alignment effectively transforms VLMs from overconfident guessers into prudent visual assistants, promoting the development of more reliable and trustworthy multimodal AI systems.

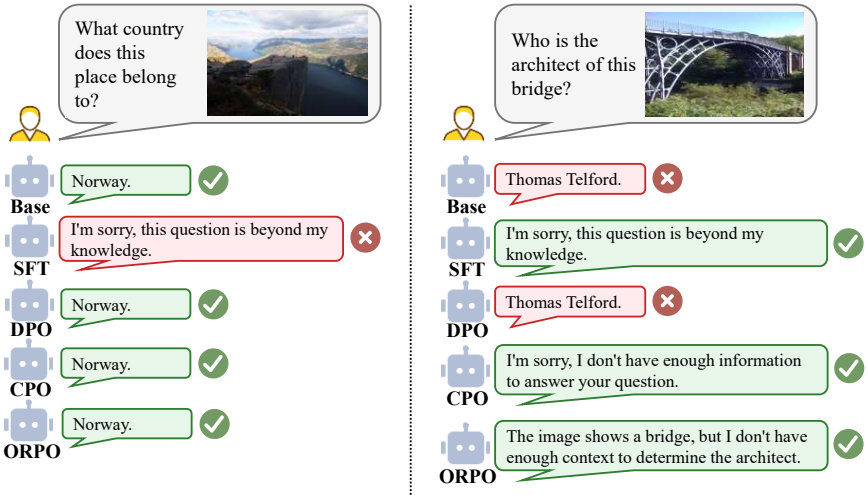

Fig. 4: Examples of model responses before and after alignment. The examples illustrate model behaviors on a mastered fact (**Left**) and an unmastered fact (**Right**) from the Visual-Idk test set, comparing the base LLaVA-1.5 with various alignment methods.

5 Conclusion

This paper presents the Visual-Idk alignment framework to delineate the knowledge boundaries of VLMs. Specifically, we curate a model-specific preference dataset through a systematic three-stage construction pipeline, followed by alignment training using diverse strategies. We conduct extensive Experiments on multiple datasets. The results on the Visual-Idk dataset demonstrate that this framework effectively transforms factual hallucinations into honest abstentions while preserving mastered knowledge and mitigating the "alignment tax." This learned honesty generalizes to out-of-distribution visual and knowledge shifts datasets, as further validated by internal probabilistic probing. Overall, our work offers a robust methodology for developing prudent visual assistants, contributing to the broader pursuit of more reliable and trustworthy multimodal systems. As for limitations, the number of training data is kind of small, and further work could enrich the training data for better exploration or efficient tuning [1, 2].

Acknowledgement

The project is supported by the Shandong Provincial Natural Science Foundation (ZR2025QC1570), National Natural Science Foundation of China (62402293), and CAAI-Lenovo Blue Sky Research Fund (2025CAAI-LENOVO-11).

References

1. Bi, J., Wang, Y., Yan, D., Huang, W., Jin, Z., Ma, X., Yan, S., Hecker, A., Ye, M., Xiao, X., et al.: Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. arXiv preprint arXiv:2502.12119 (2025)
2. Bi, J., Wang, Y., Chen, H., Xiao, X., Hecker, A., Tresp, V., Ma, Y.: Llava steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15230–15250 (2025)
3. Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al.: The art of saying no: Contextual noncompliance in language models. Advances in Neural Information Processing Systems **37**, 49706–49748 (2024)
4. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? ArXiv **abs/2302.11713** (2023)
5. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 14948–14968 (2023)
6. Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., He, Z., Chen, K., Qiu, X.: Can ai assistants know what they don't know? In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
7. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
9. Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3608–3617 (2018)
10. Hills, J., Anadkat, S.: Using logprobs. OpenAI Cookbook (2023), `https://cookbook.openai.com/examples/using_logprobs`
11. Hong, J., Lee, N., Thorne, J.: Orpo: Monolithic preference optimization without reference model. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 11170–11189 (2024)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
13. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**, 1 – 55 (2023)
14. Jia, Y., Du, Y., Jiang, K., Liang, Y., Ren, Q., Xin, Y., Yang, R., Feng, F., Chen, M., Lu, H., et al.: Benchmarking multimodal knowledge conflict for large multimodal models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 22283–22291 (2026)

15. Jiang, K., Du, Y., Ding, Y., Ren, Y., Jiang, N., Gao, Z., Zheng, Z., Liu, L., Li, B., Li, Q.: When large multimodal models confront evolving knowledge: Challenges and explorations. In: The Fourteenth International Conference on Learning Representations (2026)
16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., rong Wen, J.: Evaluating object hallucination in large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (2023)
17. Lin, S.C., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Annual Meeting of the Association for Computational Linguistics (2021)
18. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
21. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: Advances in Neural Information Processing Systems. pp. 2507–2521 (2022)
22. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
23. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: European Conference on Computer Vision. pp. 395–413. Springer (2024)
24. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
25. Tan, M., Xu, J., Liu, S., Feng, J., Zhang, H., Yao, C., Chen, S., Guo, H., Han, G., Wen, Z., et al.: Co-packaged optics (cpo): status, challenges, and solutions. Frontiers of optoelectronics **16**(1), 1 (2023)
26. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
27. Wang, Y., Zhu, Z., Liu, H., Liao, Y., Liu, H., Wang, Y., Wang, Y.: Drawing the line: Enhancing trustworthiness of MLLMs through the power of refusal. arXiv preprint arXiv:2412.11196 (2024)
28. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 13258–13273 (2024)
29. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
30. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)

31. Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H., Zhang, T.: R-tuning: Instructing large language models to say 'i don't know'. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7113–7139 (2024)
32. Zhang, T., Liu, F., Wong, J., Abbeel, P., Gonzalez, J.E.: The wisdom of hindsight makes language models better instruction followers. In: International Conference on Machine Learning. pp. 41414–41428. PMLR (2023)
33. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 2507–2517 (2023)