FROM KNOWING TO ABSTAINING: BRIDGING THE REPRESENTATION–ACTION GAP IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability of VLMs to correctly abstain from answering unanswerable questions is as important as their ability to generate accurate answers to answerable ones. Recently, several benchmarks have emerged to evaluate and improve VLM abstention. However, they suffer from substantial limitations. First, their samples often contain unintended shortcut cues in either the images or questions that inadvertently reveal answerability; moreover, they typically provide an explicit "unanswerable" option, preventing an accurate assessment of whether VLMs can abstain spontaneously. Second, when used as training data, they generally provide only binary labels without fine-grained explanations to support deeper supervision. To address these limitations, we introduce Visual Answerability Diagnosis with Rationales (VAD-R), a new benchmark constructed through a two-stage pipeline of shortcut filtering and quality verification to prevent answerability leakage. Each example is annotated with step-by-step rationales and causal evidence-gap labels. Evaluation of state-of-the-art open- and closed-source VLMs on VAD-R reveals strikingly limited spontaneous abstention capabilities, with the two model groups achieving average recall rates of only 11.4% and 16.3%, respectively. We then conduct probing analyses, revealing that hidden-state representations in certain layers can effectively distinguish answerability, yet this internal distinction fails to manifest in VLMs' final responses. Motivated by this observation, we introduce Rep2Act, a representation-to-action alignment method that translates latent answerability awareness into explicit abstention decisions. Rep2Act substantially improves action accuracy on VAD-R from 56.67% to 86.33% for Qwen2.5-VL-3B and from 59.33% to 88.67% for Qwen2.5-VL-7B. On the out-of-distribution TUBench, Rep2Act achieves an average F1 score of 53.3% with only a 3B model, surpassing the closed-source GPT-4 Turbo and GPT-4o by 16.2% and 1.1%, respectively.

1 INTRODUCTION

Vision-language models (VLMs) (Liu et al., 2023a; Bai et al., 2025; Dai et al., 2023; Hurst et al., 2024; Alayrac et al., 2022; Zhu et al., 2025; Ye et al., 2024) have achieved remarkable progress in multimodal question answering. Beyond generating correct responses to answerable questions, recent work has investigated whether and how VLMs can recognize unanswerable questions, where the requisite visual evidence is absent or otherwise insufficient, and abstain rather than fabricate answers (Favero et al., 2024; Zhou et al., 2024; Varshney & Baral, 2023). Accordingly, several abstention benchmarks have been developed (He et al., 2026; Guo et al., 2024; Miyai et al., 2025; Chandu et al., 2025). However, they exhibit limitations as summarized in Table 1. *First*, they often provide an explicit "unanswerable" option through binary-judgment or multiple-choice prompts. Thus, they evaluate prompted answerability judgments rather than whether models abstain spontaneously when responding to questions directly. *Second*, they typically provide only binary answerability labels without explaining why the available evidence is insufficient, thereby offering limited supervision for training models to identify the missing evidence and articulate their reasoning step by step. *Third*, heuristic dataset construction can introduce unintended shortcuts. For example, TUBench (He et al., 2026) generates unanswerable tabular questions by blanking out cells, enabling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of VAD-R with existing VLM unanswerability benchmarks. Text-only and image-only shortcut probes are trained on frozen Qwen2.5-VL-7B-Instruct features, with accuracy evaluated over three runs and reported as (overall average accuracy / highest subclass accuracy).

Benchmark	Reasoning Rationale	Causal Label	No Hint for Abstention	Shortcut Probe Diagnosis	
				Text-only	Image-only
UNK-VQA (Guo et al., 2024)	✗	✗	✗	54.04% / 61.46%	52.59% / 58.09%
MM-UPD (Miyai et al., 2025)	✗	✗	✗	90.17% / 99.84%	54.37% / 74.71%
TUBench (He et al., 2026)	✗	✗	✗	68.71% / 77.93%	60.88% / 99.67%
VAD-R (ours)	✓	✓ (6 Categories)	✓	**48.22% / 50%**	**47.56% / 48.81%**

Table 2: Summary evaluation of diverse VLMs on VAD-R test set under Direct Answering (P_D), where models receive only the image and question and Qwen3.6-27B serves as Action Judge, and Explicit Judgment (P_J), where models are explicitly prompted to judge whether the question is answerable with a binary "yes" or "no". Metrics include Action Accuracy (A-Acc), Answerable Recall (Ans-R), Abstain Recall on unanswerable queries (Abs-R), Abstain F1 (Abs-F1), and Abstain Recall Gap ΔAbs-R $=$ Abs-R$_{P_J}$ $-$ Abs-R$_{P_D}$. Action Accuracy (A-Acc) measures answer/abstain correctness from generated responses under P_D and explicit yes/no judgments under P_J. Complete per-model evaluation results are provided in Table 11 in Appendix C.1.

	Direct Answering Mode (P_D)				Explicit Judgment Mode (P_J)				Gap
Model Group	**A-Acc**	**Ans-R**	**Abs-R**	**Abs-F1**	**A-Acc**	**Ans-R**	**Abs-R**	**Abs-F1**	**Δ Abs-R**
Open-Source Model (8)	55.7	100.0	11.4	19.4	77.6	77.2	78.0	77.2	**+66.6**
Closed-Source Model (3)	57.2	98.0	16.3	27.6	81.7	68.4	94.9	84.5	**+78.6**

an image-only probe to achieve over 99% accuracy without access to the question. To systematically assess answerability-label leakage across these benchmarks, we conduct a comprehensive probing analysis. We feed each input into a frozen Qwen2.5-VL-7B-Instruct backbone and extract the hidden state of the final prompt token immediately before generation. We consider two unimodal probing conditions: *Question-Only*, which includes the question text and candidate options but excludes the image, and *Image-Only*, which includes the image but excludes the question text and task instructions. We train a two-layer MLP probe on the extracted features and evaluate it using five-fold stratified group cross-validation across three independent runs. As shown in Table 1, MM-UPD and TUBench exhibit strong unimodal predictability, indicating substantial answerability-label leakage.

To address these limitations, we construct **Visual Answerability Diagnosis with Rationales (VAD-R)**, a benchmark based on unedited, natural photographs in which unanswerability arises from genuine gaps in visual evidence. To mitigate answerability-label leakage, we employ a rigorous two-stage quality-control pipeline rarely adopted by prior benchmarks: it first removes linguistically predictable shortcuts using a probe trained on frozen representations and then independently audits each question's validity against the corresponding visual evidence. Each example is accompanied by a step-by-step rationale, while every unanswerable question is additionally assigned one or more of the six causal labels characterizing its underlying evidence gap. A stratified manual audit of 100 test examples confirms an annotation accuracy of 93%.

Using VAD-R, we reveal a pronounced capability gap in VLMs between prompted answerability judgment and spontaneous abstention during direct answering. As shown in Table 2, explicit judgment yields an average abstention recall of 78.0% for open-source models and 94.9% for closed-source models, whereas direct answering achieves only 11.4% and 16.3%, respectively. Further layerwise probing reveals that mid-to-late hidden-state representations clearly separate answerable from unanswerable queries, indicating that VLMs already encode answerability internally. Although this latent capability is elicited under explicit judgment prompting, the same signal fails to steer generation toward abstention during direct answering, i.e., **representation-to-action disconnect**. Existing approaches (Kuhn et al., 2023; Zhang et al., 2024; Leng et al., 2024; Vardi et al., 2026; Zhao et al., 2024) do not adequately resolve this issue. As in Figure 3, they may still fabricate answers, produce generic refusals, or generate explanations unsupported by the available visual evidence.

To bridge this gap, we propose **Representation-to-Action (Rep2Act)**, which externalizes the model's internal answerability signal as a discrete action $b \in \{\mathrm{A}, \mathrm{U}\}$ before response generation, where A and U denote answerable and unanswerable, respectively. Rep2Act combines a prior derived from deep hidden representations with vocabulary-level logits to select either an answer or abstain token. We further introduce a signed margin loss to sharpen this decision and a route-

contrastive loss that assigns a higher likelihood to the ground-truth continuation under the correct route than under its inversion, thereby preventing the generated response from contradicting the selected action. Our main contributions are as follows:

- **Reasoning-Annotated, Leakage-Resistant Benchmark**: VAD-R provides a step-by-step rationale for every question and a six-category causal taxonomy for evidence gaps in unanswerable cases, while employing a two-stage quality-control pipeline to eliminate superficial shortcuts.
- **Identification of Representation-to-Action Disconnect**: We systematically identify and characterize a fundamental disconnect in VLM abstention: answerability is encoded in deep hidden representations but fails to translate into spontaneous abstention during direct answering.
- **Rep2Act Framework**: We introduce Rep2Act, which externalizes internal answerability representations as an explicit discrete routing variable that directly conditions subsequent response generation. Rep2Act substantially improves action accuracy on VAD-R from 56.67% to 86.33% for Qwen2.5-VL-3B and from 59.33% to 88.67% for Qwen2.5-VL-7B. On the out-of-distribution TUBench benchmark, Rep2Act achieves an average F1 score of 53.3% with only a 3B-parameter model, outperforming GPT-4 Turbo and GPT-4o by 16.2% and 1.1%, respectively.

2 VAD-R: VISUAL ANSWERABILITY DIAGNOSIS WITH RATIONALES

To investigate whether VLMs genuinely understand multimodal evidence rather than exploit superficial shortcuts, we construct VAD-R as a controlled benchmark.

Benchmark Overview. VAD-R comprises 1,852 visual question-answering instances, partitioned into 1,226 training, 326 development, and 300 test examples. As shown in Figure 1(a), VAD-R covers four visual dimensions. Every question includes a visual reasoning rationale. For unanswerable questions, the six evidence-gap categories in Figure 1(b) explain why the image cannot support an answer. Multiple labels may apply to a single question. Detailed definitions and breakdowns are provided in Appendix B.3.

Data Construction Pipeline and Quality Control. Our protocol is specifically designed to eliminate superficial shortcuts and ensure a faithful evaluation of VLMs' genuine abstention capabilities. (1) We source images and descriptions from DOCCI (Onoe et al., 2024), whose detailed annotations capture fine-grained visual information. Qwen3.5-122B-A10B generates an initial pool of candidate questions across four visual dimensions, together with step-by-step visual reasoning rationales and causal evidence-gap labels. An initial quality-filtering stage yields 30,386 questions. (2) To prevent answerability from being inferred solely from questions, we train a two-layer question-only multilayer perceptron (MLP) probe on frozen text representations from Qwen2.5-VL-7B-Instruct. We rank questions by the probe's cross-entropy loss and remove those whose answerability labels are predictable from linguistic cues, retaining 8,357 of the most challenging questions, including 4,066 unanswerable ones. (3) Finally, Qwen3.6-27B independently audits 4,066 candidate unanswerable questions against the corresponding visual evidence, of which 926 pass quality verification. Pairing them with an equal number of answerable examples produces the balanced benchmark of 1,852 VQA questions. Appendix B.3 provides details on construction protocol and filtering statistics.

3 IN-DEPTH INVESTIGATION OF VLM ABSTENTION CAPABILITIES

Using the genuinely challenging VAD-R benchmark, we investigate whether VLM abstention failures arise from an inability to capture visual evidence or to translate such evidence into decisions.

Examining the Separability of Answerability Representations. For an image–question pair $x = (\boldsymbol{i}, \boldsymbol{q})$, answerability indicates whether the image provides sufficient evidence to answer the question, with label $y(x) \in \{\mathrm{A}, \mathrm{U}\}$. Following prior representation analyses (Alain & Bengio, 2017; Burns et al., 2023; Azaria & Mitchell, 2023; Liu et al., 2023b), we examine this signal under Direct Answering (D) and Explicit Judgment (J). At layer l, we define the answerability direction $\hat{\boldsymbol{d}}_{M,l}$ as the normalized difference between the mean hidden representations of answerable and unanswerable training samples under prompting mode $M \in \{D, J\}$. The answerability score is

$$z_{M,l}(x) = \frac{(\boldsymbol{h}_l^M(x) - \boldsymbol{c}_{M,l})^\top \hat{\boldsymbol{d}}_{M,l}}{\sigma_{M,l}} \tag{1}$$

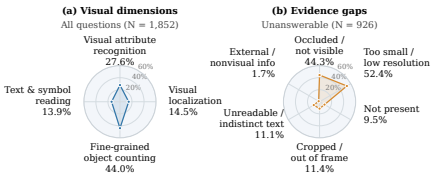

Figure 1: VAD-R composition across all splits. (a) Four visual dimensions among 1,852 questions. (b) Six evidence-gap categories among 926 unanswerable questions.

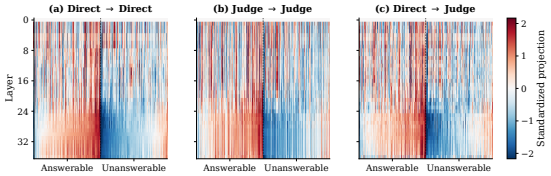

Figure 2: Answerability projections of Qwen2.5-VL-3B on the VAD-R development set: **(a)** Direct features on the Direct axis; **(b)** Judge features on the Judge axis; **(c)** Direct features on the Judge axis.

where $\boldsymbol{h}_l^M(x)$ is the final prompt token's hidden state, $\boldsymbol{c}_{M,l}$ is the midpoint of the two class means, and $\sigma_{M,l}$ is the training projection scale. Positive and negative answerability scores indicate the answerable and unanswerable sides of this axis. Appendix B.2 details how the answerability direction, center, and scale are computed. As shown in Figure 2(a)(b), projecting the intermediate representations from each prompting mode onto their corresponding answerability directions reveals a clear separation between answerable and unanswerable samples in the later layers. This demonstrates that, under both prompting modes, the VLM has already encoded information relevant to determining answerability. As shown in Figure 2(c), the representations remain clearly separable across prompting modes. Specifically, at layer 31 on the development split, direct-answer representations achieve an AUROC of 90.01% when projected onto the judgment-mode answerability direction, compared with 91.95% for judgment-mode representations projected onto their own direction, with a gap of only 1.94%. These results indicate that two prompting modes share a common answerability direction.

Representation-Action Gap and Initial Attempt. However, Table 2 shows that the direct-answer mode achieves an average abstention recall of only 11.4%. The gap between internal answerability representations and external abstention behavior naturally raises two questions:

Q1: *Why does VLM's internal recognition of insufficient evidence fail to manifest in its answer?* We hypothesize that standard VLM training overwhelmingly rewards answer generation. Instruction-tuning and preference-optimization pipelines reinforce helpful and content-producing responses (Liu et al., 2023a; Bai et al., 2025), biasing the autoregressive decoding distribution toward generating an answer regardless of whether the hidden states encode insufficient evidence.

Q2: *How can the answerability representation be effectively channeled into the final response?* A natural approach is to align the direct-answer representations with the judgment-mode projections by penalizing projection discrepancies with an auxiliary anchoring loss (Appendix C.4). However, this approach yields only negligible improvements (§5.3). The analysis in Appendix C.4 shows that the anchoring loss optimizes an internal projection score and affects the final A/U token decision only indirectly. These results suggest that regularizing continuous latent representations alone is insufficient to overcome the model's generative bias toward answering. We therefore introduce Rep2Act to directly optimize the A/U decision boundary and train subsequent generation to follow the selected route.

4 Rep2Act: **Rep**resentation-to-**Act**ion Alignment Framework

We propose **Representation-to-Action (Rep2Act)**, a method that translates internal answerability representations into explicit answer-or-abstain decisions that condition subsequent response.

As illustrated in Figure 3(d), rather than relying on unguided direct-answer generation, Rep2Act structures the VLM output as a triple (b, r, a) and autoregressively factorizes the corresponding generative probability as follows:

$$\pi_\theta(b, r, a \mid \boldsymbol{i}, \boldsymbol{q}) = \underbrace{\pi_\theta(b \mid \boldsymbol{i}, \boldsymbol{q})}_{\text{Discrete Routing Token}} \cdot \underbrace{\pi_\theta(r \mid b, \boldsymbol{i}, \boldsymbol{q})}_{\text{Route-conditioned Rationale}} \cdot \underbrace{\pi_\theta(a \mid b, r, \boldsymbol{i}, \boldsymbol{q})}_{\text{Grounded Answer Response}} \tag{2}$$

Here, π_θ denotes the VLM parameterized by θ, and $b \in \{\mathrm{A}, \mathrm{U}\}$ is the routing token that specifies whether the model should answer or abstain, whereas $y(x)$ denotes the ground-truth answerability label of sample x. The routing token is label-consistent when $b = y(x)$ and label-contradictory when $b \neq y(x)$. Furthermore, $r = (w_1^r, \ldots, w_{|r|}^r)$ denotes the reasoning

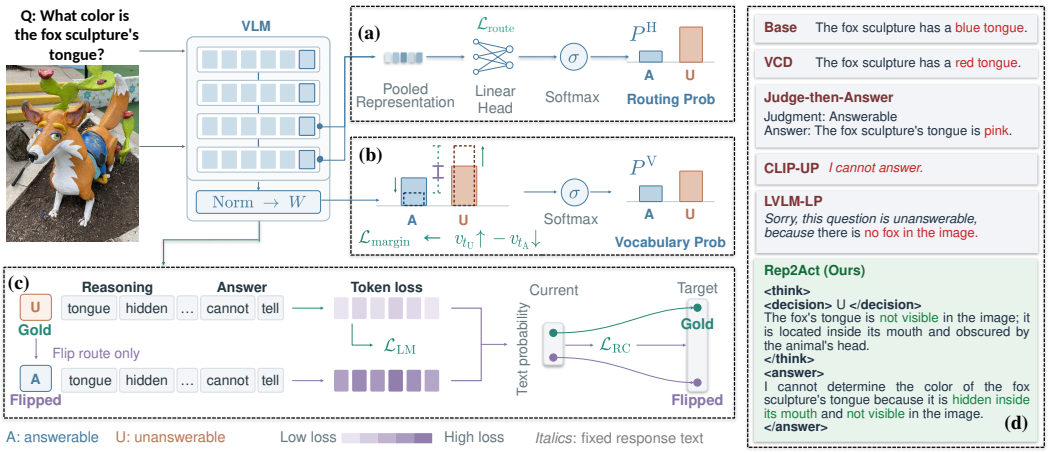

Figure 3: Rep2Act overview. (a) Representation-based routing with $\mathcal{L}_{\text{route}}$. (b) A/U vocabulary logits trained with Signed Margin Loss. (c) Language modeling and Route-Contrastive Loss train the response to follow the route. (d) A VAD-R example comparing six 3B methods.

trace, and a denotes the final response. Together, these components form the structured output: $\langle\text{think}\rangle\langle\text{decision}\rangle b\langle/\text{decision}\rangle r\langle/\text{think}\rangle\langle\text{answer}\rangle a\langle/\text{answer}\rangle$.

Rep2Act Training Scheme. An overview of the Rep2Act training pipeline is presented in Figure 3. Specifically, we introduce a set of complementary supervision signals that systematically guide the VLM in translating internal answerability representations into explicit answer-or-abstain decisions.

• *Routing-token Classification Loss*. Guided by the layerwise probing analysis in Appendix B.2, we select the layers $\mathcal{K}_{\text{H}}$ at which answerable and unanswerable representations are most clearly separable and aggregate their hidden states at the final prompt token as follows:

$$\bar{\boldsymbol{h}}_{\text{route}} = \frac{1}{|\mathcal{K}_{\text{H}}|} \sum_{l \in \mathcal{K}_{\text{H}}} \frac{\boldsymbol{h}_l^D(x)}{\|\boldsymbol{h}_l^D(x)\|_2} \in \mathbb{R}^d \tag{3}$$

We apply layer normalization (L-N) to the pooled representation $\bar{\boldsymbol{h}}_{\text{route}}$, followed by a linear classification head parameterized by the weight matrix $\boldsymbol{W}_h \in \mathbb{R}^{2\times d}$ and bias vector $\boldsymbol{\beta}_h \in \mathbb{R}^2$, to obtain the routing logits $(\eta_{\text{A}}, \eta_{\text{U}})$ and their corresponding probabilities (see Figure 3 (a)):

$$(\eta_{\text{A}}, \eta_{\text{U}}) = \boldsymbol{W}_h \,\text{L-N}(\bar{\boldsymbol{h}}_{\text{route}}) + \boldsymbol{\beta}_h, P^{\text{H}}(\text{A} \mid \boldsymbol{i}, \boldsymbol{q}) = \frac{e^{\eta_{\text{A}}}}{e^{\eta_{\text{A}}} + e^{\eta_{\text{U}}}}, P^{\text{H}}(\text{U} \mid \boldsymbol{i}, \boldsymbol{q}) = 1 - P^{\text{H}}(\text{A} \mid \boldsymbol{i}, \boldsymbol{q})$$

We then define the routing-token classification loss for each sample as follows:

$$\mathcal{L}_{\text{route}}(x) = -\mathbb{I}(y = \text{A}) \log P^{\text{H}}(\text{A} \mid \boldsymbol{i}, \boldsymbol{q}) - \mathbb{I}(y = \text{U}) \log P^{\text{H}}(\text{U} \mid \boldsymbol{i}, \boldsymbol{q}) \tag{4}$$

where $\mathbb{I}$ denotes the indicator function, which equals 1 when its condition holds and 0 otherwise.

• *Vocabulary Route Logits and Signed Margin Loss*. As shown in Figure 3(b), we further make the routing variable explicit. When predicting the A/U token following $\langle\text{think}\rangle\langle\text{decision}\rangle$, the VLM produces a vocabulary-logit vector $\mathbf{v} \in \mathbb{R}^{|\mathcal{V}|}$, where each entry is an unnormalized score for a token in the vocabulary $\mathcal{V}$. We extract $v_{t_{\text{A}}} = \mathbf{v}[t_{\text{A}}]$ and $v_{t_{\text{U}}} = \mathbf{v}[t_{\text{U}}]$, where t_{A} and t_{U} denote the vocabulary indices of the A and U tokens, respectively. We then apply a softmax over these two logits to obtain the answerability probabilities.

$$P^{\text{V}}(\text{A} \mid \boldsymbol{i}, \boldsymbol{q}) = \frac{\exp(v_{t_{\text{A}}})}{\exp(v_{t_{\text{A}}}) + \exp(v_{t_{\text{U}}})}, \quad P^{\text{V}}(\text{U} \mid \boldsymbol{i}, \boldsymbol{q}) = \frac{\exp(v_{t_{\text{U}}})}{\exp(v_{t_{\text{A}}}) + \exp(v_{t_{\text{U}}})} \tag{5}$$

To promote robust routing decisions and penalize ambiguous predictions, we introduce a signed margin loss as follows:

$$\mathcal{L}_{\text{margin}}(\theta) = \max\left(0,\ \gamma - s \cdot (v_{t_{\text{A}}} - v_{t_{\text{U}}})\right) \tag{6}$$

where $s \in \{+1, -1\}$ denotes the ground-truth direction, with $s = +1$ if $y = \text{A}$ and $s = -1$ if $y = \text{U}$, and $\gamma > 0$ is the target margin. As shown in Equation (6), achieving zero loss requires

$v_{t_A} - v_{t_U} \geq \gamma$ for answerable inputs and $v_{t_U} - v_{t_A} \geq \gamma$ for unanswerable inputs, thereby encouraging a wider decision margin.

• *Route-Contrastive Loss*. Even when the generated routing token matches the ground-truth label, the subsequent generation may contradict the routing decision, a phenomenon we term *route drift*. To address this issue, Rep2Act introduces a route-contrastive loss that encourages the continuation to remain consistent with the selected route, as illustrated in Figure 3(c). To compute the route-contrastive loss, we perform two teacher-forced forward passes, with the routing token set to $b = y(x)$ and $b = \bar{y}(x)$, respectively. Only the routing token changes; the target reasoning and answer remain unchanged. We compute $\mathcal{L}_{\mathrm{LM}}$ from the gold-route pass and use the continuation likelihoods from both passes to compute $\mathcal{L}_{\mathrm{RC}}$. To mitigate the effect of sequence length, for each component $k \in \{r, a\}$, corresponding to the reasoning process and the final answer, respectively, we compute the mean negative log-likelihood (NLL) separately:

$$N_k(b) = -\frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \pi_\theta(o_t \mid \mathbf{o}^b_{<t}, \boldsymbol{i}, \boldsymbol{q}) \tag{7}$$

where $\mathcal{T}_k$ denotes the set of token positions corresponding to component k in the gold output sequence $\mathbf{o}$. The context $\mathbf{o}^b_{<t}$ contains the selected routing token $b \in \{y, \bar{y}\}$ and all preceding gold tokens. Then, for each component, we subtract the gold-route NLL, $N_k(y)$, from the flipped-route NLL, $N_k(\bar{y})$, and then average the resulting differences:

$$g = \frac{1}{2} \underbrace{[N_r(\bar{y}) - N_r(y)]}_{\text{Reasoning NLL difference}} + \frac{1}{2} \underbrace{[N_a(\bar{y}) - N_a(y)]}_{\text{Answer NLL difference}} \tag{8}$$

A positive g indicates that, on average, the gold reasoning process and answer are easier to generate under the gold route than under the flipped route. We then define the route-contrastive loss as

$$\mathcal{L}_{\mathrm{RC}}(\theta) = \max(0, \delta - g) \tag{9}$$

where $\delta > 0$ is the target margin. This objective encourages the difference g to reach δ.

Putting It All Together. Overall, Rep2Act combines the language-modeling, route-classification, signed-margin, and route-contrastive objectives as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda_r \mathcal{L}_{\text{route}} + \lambda_m \mathcal{L}_{\text{margin}} + \lambda_c \mathcal{L}_{\mathrm{RC}} \tag{10}$$

The training of Rep2Act proceeds in two stages. *Stage 1: Routing-head training.* We freeze the VLM backbone and optimize only the representation-based routing head on VAD-R using $\mathcal{L}_{\mathrm{route}}$. *Stage 2: Joint training.* We initialize the routing head with the parameters learned in Stage 1 and jointly optimize it with the backbone's LoRA attention adapters using $\mathcal{L}_{\mathrm{total}}$.

Rep2Act Test-Time Inference Scheme. At inference time, Rep2Act combines the continuous prior derived from the deep representation, P^{H} with the vocabulary-level probability, P^{V}:

$$P^{\mathrm{E}}(\mathrm{A} \mid \boldsymbol{i}, \boldsymbol{q}) = \alpha P^{\mathrm{H}}(\mathrm{A} \mid \boldsymbol{i}, \boldsymbol{q}) + (1 - \alpha) P^{\mathrm{V}}(\mathrm{A} \mid \boldsymbol{i}, \boldsymbol{q}) \tag{11}$$

where $\alpha \in [0, 1]$ balances the representation signal against the output-layer confidence. We obtain the discrete routing token $\hat{b}$ by thresholding the fused probability: $\hat{b} = \mathrm{A}$ if $P^{\mathrm{E}}(\mathrm{A} \mid \boldsymbol{i}, \boldsymbol{q}) \geq 0.5$, and $\hat{b} = \mathrm{U}$ otherwise. We then insert $\hat{b}$ into the decision tag and generate the reasoning $\hat{r}$ and final answer $\hat{a}$ conditioned on the routing token, image, and question: $(\hat{r}, \hat{a}) \sim \pi_\theta(\cdot \mid \hat{b}, \boldsymbol{i}, \boldsymbol{q})$.

5 Evaluation

We evaluate Rep2Act by addressing the following research questions: **RQ1: In-Domain Effectiveness and Preservation of Standard VQA Capabilities.** How effectively does Rep2Act bridge the representation-to-action disconnect compared with existing baselines? How do its improvements vary across different visual skills? Can Rep2Act improve answerability discrimination without degrading performance on standard VQA tasks? **RQ2: Out-of-Distribution Generalization.** How robustly does Rep2Act generalize to previously unseen domains and types of unanswerability? **RQ3: Ablation Study.** What are the respective contributions of the signed margin loss, route-contrastive loss, and supervised target formulations?

Table 3: Results on VAD-R test set. Rep2Act, Rep2Act†, and Rep2Act‡ denote the fused-decision variant ($\alpha = 0.50$), routing-head-only variant ($\alpha = 1.0$), and vocabulary-only variant ($\alpha = 0.0$).

(a) Answerability classification.

Method	Acc	M-P	M-R	M-F1	AUROC
Qwen2.5-VL-3B					
SE	69.00	69.25	69.00	68.90	69.44
VCD	50.33	50.34	50.33	50.13	53.39
VL-U	58.67	62.11	58.67	55.50	61.01
LVLM-LP	80.67	82.13	80.67	80.44	90.71
Rep2Act†	82.33	83.59	82.33	82.17	92.60
Rep2Act‡	86.33	86.61	86.33	86.31	92.68
Rep2Act	**87.33**	**87.33**	**87.33**	**87.33**	**92.98**
Qwen2.5-VL-7B					
SE	67.67	68.26	67.67	67.40	69.30
VCD	46.33	46.19	46.33	45.83	46.17
VL-U	58.67	59.96	58.67	57.28	58.87
LVLM-LP	87.33	88.01	87.33	87.28	**95.82**
Rep2Act†	63.33	78.85	63.33	57.64	91.84
Rep2Act‡	**88.33**	88.42	**88.33**	**88.33**	93.75
Rep2Act	**88.33**	**88.54**	**88.33**	88.32	93.82

(b) Response generation and content.

Method	A-Acc	M-P	M-R	M-F1	C-Acc
Qwen2.5-VL-3B					
P_D	56.67	76.79	56.67	46.65	30.00
P_{JA}	72.00	76.53	72.00	70.75	32.33
VCD	58.33	75.44	58.33	49.91	30.33
CLIP-UP	63.33	76.63	63.33	58.10	30.00
LVLM-LP	81.33	82.64	81.33	81.14	26.00
Rep2Act†	81.67	81.91	81.67	81.63	53.67
Rep2Act‡	84.00	84.62	84.00	83.93	55.67
Rep2Act	**86.33**	**86.33**	**86.33**	**86.33**	**57.00**
Qwen2.5-VL-7B					
P_D	59.33	77.57	59.33	51.28	43.67
P_{JA}	85.00	85.04	85.00	85.00	47.00
VCD	60.33	76.39	60.33	53.22	42.00
CLIP-UP	61.67	71.34	61.67	56.77	29.00
LVLM-LP	87.00	87.74	87.00	86.94	47.33
Rep2Act†	68.67	77.20	68.67	66.00	48.33
Rep2Act‡	**88.67**	**88.69**	**88.67**	88.66	**60.67**
Rep2Act	**88.67**	88.67	**88.67**	**88.67**	**60.67**

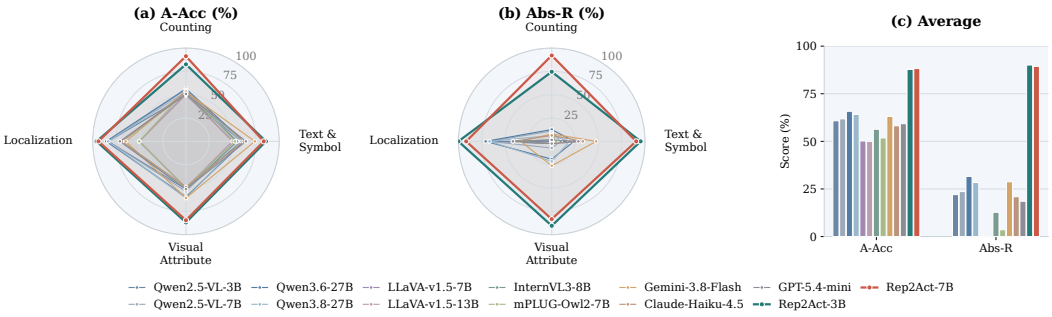

Figure 4: Direct Answering (P_D): four-skill radar plots of (a) action accuracy (A-Acc) and (b) abstention recall (Abs-R), with (c) their averages across 4 visual skills.

Evaluation Benchmarks. We evaluate Rep2Act on *VAD-R* for in-domain answerability diagnosis; *TUBench* (He et al., 2026) and *UNK-VQA* (Guo et al., 2024) for out-of-distribution generalization; and *POPE* (Li et al., 2023), *GQA* (Hudson & Manning, 2019), and *TextVQA* (Singh et al., 2019) for general multimodal perception. Complete benchmark descriptions are provided in Appendix B.4.

Baselines. We compare Rep2Act against the following baselines. *Semantic Entropy (SE)* (Kuhn et al., 2023) and *VL-Uncertainty (VL-U)* (Zhang et al., 2024) produce answerability classifications only. *Visual Contrastive Decoding (VCD)* (Leng et al., 2024), *LVLM-LP* (Zhao et al., 2024) support both answerability classification and complete response generation. *CLIP-UP* (Vardi et al., 2026), *Direct Answering* (P_D), and *Judge-then-Answer* (P_{JA}) are evaluated based on their complete responses, which either provide an answer or abstain. Implementation details are in Appendix B.5.

Evaluation Metrics. We evaluate answerability classification against the ground-truth labels using accuracy (Acc), macro-precision (M-P), macro-recall (M-R), macro-F1 (M-F1), and, for continuous uncertainty scores, the area under the receiver operating characteristic curve (AUROC). For complete generated responses, we measure action accuracy (A-Acc), using Qwen3.6-27B as an action judge to determine whether each response answers or abstains. On VAD-R, we additionally evaluate content accuracy (C-Acc) using Qwen3.8-27B, which requires factually correct responses to answerable questions and valid causal explanations of the missing visual evidence for unanswerable questions. All metrics are reported as percentages. The evaluation protocols and judge prompts are detailed in Appendix B.6 and Appendix D.2, respectively.

5.1 In-Domain Effectiveness and Baseline Comparison (RQ1)

In-Domain Results on VAD-R. As shown in Table 3(a), for 3B model, the standalone representation-based routing head, Rep2Act† with $\alpha = 1.0$, achieves an accuracy of 82.33%. Fus-

Table 4: OOD evaluation on TUBench and UNK-VQA.

(a) Answerability classification.

Method	Acc	M-P	M-R	M-F1	AUROC
TUBench					
SE	52.72	53.63	53.13	51.23	53.54
VCD	47.54	46.90	48.27	41.48	52.11
VL-U	51.10	53.67	51.88	44.77	53.12
LVLM-LP	50.30	50.27	50.27	50.27	57.07
Rep2Act	**54.21**	**54.17**	**54.16**	**54.16**	**57.19**
UNK-VQA					
SE	52.26	50.68	50.54	**48.73**	51.34
VCD	47.12	48.48	48.74	45.47	47.97
VL-U	53.83	52.11	50.70	42.29	50.96
LVLM-LP	54.49	56.49	50.95	39.26	55.55
Rep2Act	**55.37**	**57.23**	**52.17**	43.58	**57.71**

(b) Response-level actions.

Method	A-Acc	M-P	M-R	M-F1
TUBench				
Base-P_D	53.27	59.29	52.31	41.77
VCD	55.65	61.31	54.82	47.70
CLIP-UP	55.35	59.73	54.55	48.00
LVLM-LP	53.31	53.22	53.12	**52.86**
Rep2Act	**56.84**	**61.41**	**56.09**	50.71
UNK-VQA				
Base-P_D	50.64	55.89	53.24	45.83
VCD	52.43	52.18	52.18	**52.18**
CLIP-UP	51.17	52.84	52.51	50.26
LVLM-LP	54.69	**56.65**	51.24	40.38
Rep2Act	**55.38**	54.62	**53.47**	51.26

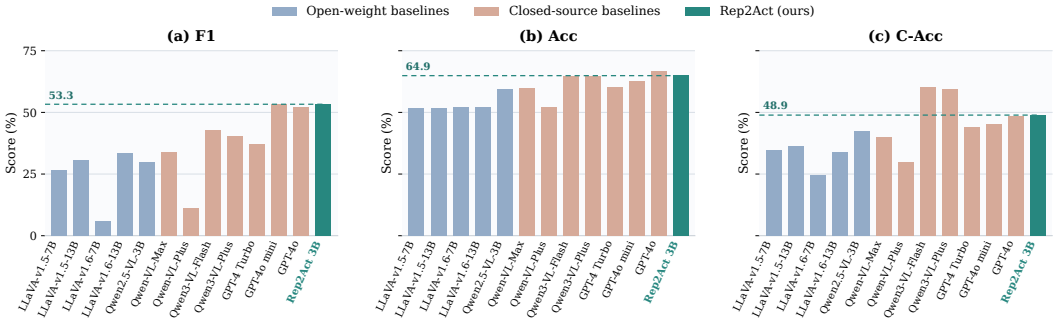

Figure 5: Results under the official TUBench evaluation protocol, averaged across four subsets: (a) F1, (b) accuracy (Acc), and (c) content accuracy (C-Acc).

ing its predictions with the vocabulary logits using $\alpha = 0.50$ increases accuracy to 87.33%, outperforming the vocabulary-only routing variant, Rep2Act‡, which achieves 86.33%. For 7B model, vocabulary-based routing alone already achieves 88.33% accuracy, matching the performance of the fused variant. Table 3(b) shows that LVLM-LP achieves 81.33% A-Acc but only 26.00% C-Acc with 3B model, falling below the base model's 30.00%. For inputs predicted to be unanswerable, LVLM-LP inserts the fixed prefix "Sorry, this question is unanswerable, because" and relies on the frozen 3B base model to generate the continuation. Consequently, the model may provide an incorrect justification for abstention, as illustrated in Figure 3(d). More examples are in Appendix E.

Fine-Grained Visual Skill Breakdown. Figure 4 shows that baseline performance varies substantially across visual skills, particularly in abstention recall. In contrast, both Rep2Act models achieve consistently high A-Acc and Abs-R across all four skills, demonstrating robust improvements across diverse dimensions of visual understanding. Detailed comparisons across visual skills and prompting modes are provided in Appendix C.1, particularly Table 12.

Preservation of Standard VQA Capabilities. Table 6 evaluates capability preservation on POPE, GQA, and TextVQA. Rep2Act preserves the general multimodal perception capabilities of base models, with performance deviations of no more than 0.38% for 3B model and 0.64% for 7B model.

5.2 Out-of-Distribution Generalization (RQ2)

To assess whether representation-to-action routing generalizes beyond the in-domain distribution, we evaluate Rep2Act on TUBench (He et al., 2026) and UNK-VQA (Guo et al., 2024).

Evaluation on TUBench. TUBench presents a formidable out-of-distribution challenge because its questions and images are collected using distinct protocols across multiple specialized subdomains. Table 4 shows that Rep2Act achieves superior performance among all evaluated 3B-scale methods, with 54.21% answerability-classification accuracy and 56.84% response-level A-Acc. We further evaluate Rep2Act using the official subset-specific prompts and parsers provided by TUBench, as shown in Figure 5, with complete results in Appendix C.2. Following He et al. (2026), F1 is com-

Table 5: Zero-shot evaluation on UNK-VQA under judgment and multiple-choice settings.

Model	BY Acc	MC Acc
Otter-MPT-7B	40.87	20.33
Open-Flamingo-MPT-7B	31.49	26.52
Instruct-BLIP-Vicuna-7B	43.98	22.38
Instruct-BLIP-Vicuna-13B	**54.82**	18.41
Qwen2.5-VL-3B	52.44	18.68
Rep2Act 3B	54.78	**28.22**

Table 6: VQA capabilities preservation on POPE, GQA and TextVQA.

Model	POPE	GQA	TextVQA
Qwen2.5-VL-3B	**86.04**	**60.09**	**79.56**
+ Margin Loss	85.95	59.64	79.40
+ Route-Contrastive	85.74	59.56	79.08
Rep2Act	85.85	59.71	79.25
Qwen2.5-VL-7B	**86.60**	**60.98**	**83.75**
Rep2Act	86.11	60.34	83.62

Table 7: Response-level component ablation on VAD-R. Cons. denotes route-action consistency.

Configuration	λ_m	λ_c	A-Acc	M-P	M-R	M-F1	Cons.
w/o Margin & RC	0.0	0.0	73.00	73.23	73.00	72.93	95.67
+ Margin Loss	0.2	0.0	78.33	78.55	78.33	78.29	89.00
+ Route-Contrastive	0.0	1.0	78.00	78.51	78.00	77.90	**99.67**
Rep2Act Combined	0.2	1.0	**86.33**	**86.33**	**86.33**	**86.33**	99.33

Table 8: Response-level comparison between FA-Route-SFT and Rep2Act on VAD-R.

Method	A-Acc	M-P	M-R	M-F1	C-Acc
Route-SFT	76.67	78.50	76.67	76.29	49.67
FA-Route-SFT	77.33	78.47	77.33	77.10	52.67
Rep2Act	**86.33**	**86.33**	**86.33**	**86.33**	**57.00**

puted by treating the unanswerable class as the positive class. Rep2Act-3B substantially outperforms all open-source baselines and remains competitive with frontier proprietary models.

Evaluation on UNK-VQA. As shown in Table 4, Rep2Act achieves superior performance among all evaluated 3B-scale methods with answerability-classification accuracy of 55.37% and A-Acc of 55.38%. We further evaluate Rep2Act under the official UNK-VQA protocol in Binary (BY) and Multiple-Choice (MC) settings, where the prompt templates provide four candidate answer options for each question. Table 5 shows that Rep2Act performs on par with Instruct-BLIP-Vicuna-13B in BY setting. In MC setting, Rep2Act improves upon 3B base model by 9.54% and outperforms all evaluated baselines despite their considerably larger parameters. These results demonstrate that Rep2Act generalizes effectively to out-of-distribution data.

5.3 Ablation Study (RQ3)

Effectiveness of Loss Design. We vary λ_m and λ_c in Equation 10, while retaining the language-modeling and representation-based routing-head supervision in every trained variant and fixing the inference-time fusion weight at $\alpha = 0.50$. As shown in Table 7, the signed margin loss improves A-Acc by 5.33% over Rep2Act w/o both loss. The route-contrastive loss increases route–action consistency to 99.67% and improves A-Acc by 5% . Combining both objectives yields A-Acc of 86.33%, representing a 13.33% improvement over Rep2Act w/o either loss. Complete sensitivity analyses for λ_m and λ_c are provided in Appendix C.5.

Discrete Routing vs. Continuous Feature Anchoring. We evaluate Feature-Anchor Route-SFT (FA-Route-SFT), which combines $\mathcal{L}_{\mathrm{LM}}$ with $\mathcal{L}_{\mathrm{anchor}}$ (Appendix C.4). As shown in Table 8, feature anchoring improves A-Acc by only 0.66% over Route-SFT. Because prompt switching induces a substantial representation shift orthogonal to the answerability direction, optimizing internal projections influences the final A/U token boundary only indirectly. In contrast, Rep2Act directly optimizes this decision boundary and trains subsequent generation to remain consistent with the selected route. Appendix C.4 provides the complete geometric decomposition and gradient derivations. We additionally evaluate alternative SFT response formats on VAD-R, including concise and detailed targets both w/ and w/o chain-of-thought reasoning. The results are provided in Appendix C.3.

6 Conclusion

In this work, we investigate the abstention capabilities of VLMs. We construct VAD-R, a genuinely challenging benchmark, and reveal that deep hidden representations clearly distinguish answerable from unanswerable queries, yet this internal signal fails to guide direct generation, exposing a representation-to-action disconnect. To bridge this gap, we introduce Rep2Act, which converts latent answerability into an explicit routing decision and aligns subsequent generation with the selected answer-or-abstain action. Extensive evaluations demonstrate that Rep2Act substantially improves both action and content accuracy, generalizes effectively to out-of-distribution benchmarks, and preserves standard VQA capabilities without inducing excessive refusal.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic benchmark data, support automated response evaluation, and improve the readability of the manuscript. We did not use generative AI to develop models or formulate hypotheses, and the remaining disclosure categories are not applicable to this work. Specifically, candidate benchmark examples and rationales generated by Qwen3.5-122B-A10B and audited by Qwen3.6-27B were subjected to automated shortcut probing and author-conducted manual auditing. The authors have reviewed all AI-assisted outputs and assume full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The VAD-R benchmark is constructed from the publicly available DOCCI dataset, which is distributed under standard open-research licensing terms and contains no human-subject data, personally identifiable information, or harmful content. Our research seeks to enhance the safety and reliability of multimodal foundation models by mitigating hallucinations through evidence-grounded abstention.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the mathematical formulation and training objectives of Rep2Act are detailed in Section 4 and Appendix B.2. The benchmark construction protocol, hyperparameter configurations, and evaluation prompt templates are documented in Section 2, Section 5, and Appendix B.6. The complete source code, evaluation scripts, and benchmark data splits will be made publicly available upon publication.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE, 2018.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017,Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

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.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Khyathi Chandu, Linjie Li, Anas Awadalla, Ximing Lu, Jae Sung Park, Jack Hessel, Lijuan Wang, and Yejin Choi. Certainlyuncertain: A benchmark and metric for multimodal epistemic and aleatoric awareness. In *International Conference on Learning Representations*, volume 2025, pp. 12676–12705, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14303–14312, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10284–10296, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Xingwei He, Cheng Huang, Qianru Zhang, A-Long Jin, Linlin Yu, Yuan Yuan, and Siu-Ming Yiu. Tubench: Benchmark large vision-language models on trustworthiness with unanswerable questions. *IEEE Transactions on Dependable and Secure Computing*, 2026.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1855–1888, Albuquerque, New Mexico, April 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

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

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, 2023b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6497–6540, 2025.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Ben Vardi, Oron Nir, and Ariel Shamir. Clip-up: Clip-based unanswerable problem detection for visual question answering. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5898–5908. IEEE, 2026.

Neeraj Varshney and Chitta Baral. Post-abstention: Towards reliably re-attempting the abstained instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 967–982, 2023.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1112–1122, 2018.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. Mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13040–13051. IEEE, 2024.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pp. 127–142. Springer, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Related Work

VLM Abstention Benchmarks. Several benchmarks have been developed to evaluate the abstention capabilities of VLMs. VizWiz (Gurari et al., 2018) collects photographs captured by visually impaired users, which often exhibit authentic image-acquisition challenges. In contrast, UNK-VQA, MM-UPD and TUBench construct unanswerable inputs through synthetic interventions such as masking, blanking, answer-option deletion, or cross-source mismatching (Guo et al., 2024; He et al., 2026; Miyai et al., 2025). Heuristic dataset construction can introduce visual or linguistic shortcuts that allow answerability labels to be inferred without genuine image–question reasoning (Goyal et al., 2017; Agrawal et al., 2018; Geirhos et al., 2020). Moreover, these benchmarks provide only binary answerability labels and lack step-by-step explanations identifying the missing evidence in unanswerable examples. Their evaluation protocols explicitly ask models to identify unanswerable inputs and therefore measure prompted judgment rather than spontaneous abstention during direct answering. In this work, we introduce VAD-R to address the limitations of existing benchmarks. Each VAD-R sample includes an answerability label and a step-by-step rationale, while unanswerable examples are additionally annotated with causal evidence-gap labels. We employ a two-stage pipeline combining shortcut filtering and quality verification to prevent answerability-label leakage. Furthermore, our evaluation protocol requires models to respond directly without receiving an explicit unanswerability hint, enabling a faithful assessment of spontaneous abstention.

Methods for Enhancing VLM Abstention Capabilities. Recent studies have investigated how to identify unanswerable questions and improve the abstention capabilities of VLMs. Some methods estimate unreliability from response variability, including consistency-based detection (Khan & Fu, 2024), semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), and multimodal perturbation consistency (Zhang et al., 2024). Other approaches exploit internal representations or decoding-time signals: LVLM-LP probes first-token logits (Zhao et al., 2024), CLIP-UP conditions generation on a learned visual prefix (Vardi et al., 2026), and VCD contrasts predictions obtained from clean and corrupted visual inputs (Leng et al., 2024). Although these methods estimate answerability or steer answer-versus-abstain behavior, they do not train models to generate complete responses that explain which visual evidence is missing, occluded, or insufficiently resolved and Rep2Act aims to fill this gap.

B Supplementary Design and Evaluation Details

B.1 Detailed Analysis of Unimodal Shortcuts

Diagnostic Findings and Comparative Analysis. As detailed in Table 9:

- **Text Shortcuts in Existing Benchmarks**: In MM-UPD, MMAAD achieves an AUROC of 100.00% and MMIASD achieves 99.77% purely from question texts, as negative samples were constructed by removing options or using standardized textual templates. In TUBench, UGeoQA and UVQA reach text-only AUROCs of 84.29% and 84.91%. Models can thus identify unanswerability from superficial phrasing without verifying visual evidence.
- **Image Shortcuts in Existing Benchmarks**: In MM-UPD's MMIVQD subset, randomly pairing images from different source distributions produces an Image AUROC of 83.99%. In TUBench's UTabMWP, visual masking artifacts yield a perfect Image AUROC of 100.00%.
- **VAD-R Shortcut Diagnosis**: On the final VAD-R test split, question-only accuracy is 48.22% with AUROC 48.05%, and image-only accuracy is 47.56% with AUROC 47.15%. These results indicate weak label predictability under the evaluated unimodal probes.

Table 9: Unimodal shortcut probing under frozen Qwen2.5-VL-7B representations across UNK-VQA, TUBench, MM-UPD, and VAD-R. Probes are retrained on all 300 questions of VAD-R using five-fold image-grouped cross-validation and three independent runs. Red cells indicate accuracy above 60%.

Benchmark / Sub-dataset	Input Modality	Samples	U / A Ratio	Accuracy	Bal-Acc	Macro-F1	AUROC
1. UNK-VQA Benchmark							
Question-Only Probing							
UNK-VQA, Overall	Question-Only	11,036	5,991 / 5,045	54.04	50.02	37.67	50.93 ± 0.07
I-1, Image Context Mislead	Question-Only	3,522	1,998 / 1,524	56.73	50.46	39.93	51.78 ± 0.40
I-2, Object Masking	Question-Only	807	496 / 311	61.46	49.42	39.68	49.30 ± 1.31
I-3, Object Copy-Move	Question-Only	812	449 / 363	55.30	49.27	37.21	47.12 ± 2.24
T-1, Text Over-Specification	Question-Only	4,936	2,519 / 2,417	51.03	50.10	35.90	51.85 ± 0.42
T-2, Text Missing Details	Question-Only	959	529 / 430	55.16	49.53	37.02	49.41 ± 0.83
Image-Only Probing							
UNK-VQA, Overall	Image-Only	2,984	1,592 / 1,392	52.59	51.27	49.87	53.15 ± 0.09
I-1, Image Context Mislead	Image-Only	946	530 / 416	56.03	49.81	48.03	49.82 ± 0.35
I-2, Object Masking	Image-Only	207	138 / 69	51.09	46.45	58.92	64.89 ± 1.31
I-3, Object Copy-Move	Image-Only	208	109 / 99	50.91	45.05	60.12	56.68 ± 2.24
T-1, Text Over-Specification	Image-Only	1,351	657 / 694	51.37	50.61	49.67	51.39 ± 0.31
T-2, Text Missing Details	Image-Only	272	158 / 114	58.09	51.66	51.25	51.30 ± 0.45
2. TUBench Benchmark							
Question-Only Probing							
TUBench, Overall	Question-Only	2,354	1,151 / 1,203	68.71	68.77	68.70	77.23 ± 0.04
UCR, Code Rendering	Question-Only	480	214 / 266	57.29	55.44	54.56	58.67 ± 0.15
UGeoQA, Geometry Reasoning	Question-Only	974	487 / 487	77.24	77.24	77.23	84.29 ± 0.08
UTabMWP, Table Math	Question-Only	400	200 / 200	50.00	50.00	44.12	50.00 ± 0.00
UVQA, Visual QA	Question-Only	500	250 / 250	77.93	77.93	77.93	84.91 ± 0.11
Image-Only Probing							
TUBench, Overall	Image-Only	2,354	1,151 / 1,203	60.88	60.77	60.65	69.11 ± 0.12
UCR, Code Rendering	Image-Only	480	214 / 266	61.88	59.66	58.96	64.32 ± 0.21
UGeoQA, Geometry Reasoning	Image-Only	974	487 / 487	50.00	50.00	49.79	50.00 ± 0.00
UTabMWP, Table Math	Image-Only	400	200 / 200	99.67	99.67	99.67	100.00 ± 0.00
UVQA, Visual QA	Image-Only	500	250 / 250	50.13	50.13	47.45	50.02 ± 0.06
3. MM-UPD Benchmark							
Question-Only Probing							
MM-UPD, ALL Combined	Question-Only	15,245	7,213 / 8,032	90.17	89.66	89.96	94.79 ± 0.03
MMAAD, Missing Option	Question-Only	5,537	2,359 / 3,178	99.84	99.85	99.84	100.00 ± 0.00
MMIASD, Inconsistent Annotation	Question-Only	7,052	3,526 / 3,526	98.10	98.10	98.10	99.77 ± 0.01
MMIVQD, Incompatible Question	Question-Only	2,656	1,328 / 1,328	50.01	50.01	47.36	50.01 ± 0.01
Image-Only Probing							
MM-UPD, ALL Combined	Image-Only	15,245	7,213 / 8,032	54.37	53.11	50.97	55.11 ± 0.11
MMAAD, Missing Option	Image-Only	5,537	2,359 / 3,178	57.40	50.00	36.47	49.81 ± 0.11
MMIASD, Inconsistent Annotation	Image-Only	7,052	3,526 / 3,526	50.00	50.00	48.79	50.00 ± 0.00
MMIVQD, Incompatible Question	Image-Only	2,656	1,328 / 1,328	74.71	74.71	74.52	83.99 ± 0.57
4. Ours: VAD-R Benchmark							
Question-Only Probing							
VAD-R	Question-Only	300	150 / 150	48.22	48.22	47.02	48.05 ± 0.68
Counting	Question-Only	126	63 / 63	48.68	48.68	47.59	48.37 ± 1.40
Text & Symbol	Question-Only	42	21 / 21	50.00	50.00	49.16	46.18 ± 1.82
Visual Attribute	Question-Only	84	42 / 42	46.43	46.43	45.04	45.31 ± 0.60
Visual Localization	Question-Only	48	24 / 24	48.61	48.61	45.23	53.07 ± 0.78
Image-Only Probing							
VAD-R	Image-Only	300	150 / 150	47.56	47.56	47.51	47.15 ± 0.25
Counting	Image-Only	126	63 / 63	48.15	48.15	48.07	45.54 ± 0.98
Text & Symbol	Image-Only	42	21 / 21	43.65	43.65	43.17	43.12 ± 2.33
Visual Attribute	Image-Only	84	42 / 42	48.81	48.81	48.72	49.39 ± 0.83
Visual Localization	Image-Only	48	24 / 24	47.22	47.22	47.03	50.35 ± 2.66

B.2 Layer-wise Answerability Diagnostics

Answerability Directions and Projection Scores. We first examine whether hidden-state representations can distinguish between answerable and unanswerable questions. Specifically, we construct two types of prompting modes:

- *Direct-Answer Prompt Mode (P_D)*, which consists solely of the original question from VAD-R;
- *Judgment Prompt Mode (P_J)*, which prepends the instruction "*Based only on the image, can a viewer reliably answer the question? Answer only yes or no.*" to the original question.

Inspired by prior work (Alain & Bengio, 2017; Burns et al., 2023; Azaria & Mitchell, 2023; Liu et al., 2023b), we adopt linear probing to examine the information encoded in VLM representations. Formally, a VQA instance is denoted by $x = (\boldsymbol{i}, \boldsymbol{q}) \in \mathcal{I} \times \mathcal{Q}$, where $\mathcal{I}$ denotes the image space and $\mathcal{Q}$ denotes the space of tokenized questions over vocabulary $\mathcal{V}$. The image $\boldsymbol{i} \in \mathbb{R}^{H \times W \times C}$ has height H, width W, and C channels, while the question $\boldsymbol{q} = (w_1^q, \ldots, w_{|\boldsymbol{q}|}^q)$ comprises $|\boldsymbol{q}|$ tokens, with each $w_j^q \in \mathcal{V}$. Let $y(x) \in \mathcal{Y} = \{\mathrm{A}, \mathrm{U}\}$ denote the ground-truth answerability label, where A and U correspond to answerable and unanswerable, respectively. On training data $\mathcal{D}_{\text{train}}$, let $\mathcal{S}_A = \{x \in \mathcal{D}_{\text{train}} \mid y(x) = \mathrm{A}\}$ and $\mathcal{S}_U = \{x \in \mathcal{D}_{\text{train}} \mid y(x) = \mathrm{U}\}$.

For each prompting mode $M \in \{D, J\}$ and layer $l \in \{0, \ldots, L\}$, where layer 0 corresponds to the embedding output and layer L to the final Transformer layer, we extract the hidden state of the final prompt token immediately before generation, denoted by $\boldsymbol{h}_l^M(x) \in \mathbb{R}^d$, with d representing the hidden-state dimensionality. We then average the representations within each class to obtain the corresponding layerwise centroid:

$$\boldsymbol{\mu}_{A,l}^M = \frac{1}{|\mathcal{S}_A|} \sum_{x \in \mathcal{S}_A} \boldsymbol{h}_l^M(x), \qquad \boldsymbol{\mu}_{U,l}^M = \frac{1}{|\mathcal{S}_U|} \sum_{x \in \mathcal{S}_U} \boldsymbol{h}_l^M(x) \tag{12}$$

For each prompting mode $M \in \{D, J\}$ at layer l, we then compute the following terms:

- *Unit Answerability Direction* is defined as the normalized difference between the answerable and unanswerable mode centroids:$\boldsymbol{d}_{M,l} = \boldsymbol{\mu}_{A,l}^M - \boldsymbol{\mu}_{U,l}^M, \hat{\boldsymbol{d}}_{M,l} = \frac{\boldsymbol{d}_{M,l}}{\|\boldsymbol{d}_{M,l}\|_2}$
- *Mode-Midpoint Centroid*: $\boldsymbol{c}_{M,l} = \frac{1}{2}\left(\boldsymbol{\mu}_{A,l}^M + \boldsymbol{\mu}_{U,l}^M\right)$

Then, we conduct the following analysis:

• *Intra-Mode Analysis.* We derive standardized projection score as follows:

$$z_{M,l}(x) = \frac{(\boldsymbol{h}_l^M(x) - \boldsymbol{c}_{M,l})^\top \hat{\boldsymbol{d}}_{M,l}}{\sigma_{M,l}}, \quad \sigma_{M,l} = \sqrt{\frac{1}{|\mathcal{D}_{\text{train}}| - 1} \sum_{x \in \mathcal{D}_{\text{train}}} \left((\boldsymbol{h}_l^M(x) - \boldsymbol{c}_{M,l})^\top \hat{\boldsymbol{d}}_{M,l}\right)^2} \tag{13}$$

The standardized score $z_{M,l}(x)$ quantifies the position of a sample's hidden-state representation along the answerability direction, centered at the midpoint between the two mode centroids and normalized by the sample standard deviation $\sigma_{M,l}$ of the centered training projections. Positive values indicate alignment with the answerable class, whereas negative values indicate alignment with the unanswerable class, and larger absolute values reflect greater separation from the decision boundary.

• *Inter-Mode Analysis.* We then evaluate whether the answerability direction estimated under the judgment prompting mode can also separate the intermediate representations elicited under the direct-answer mode. Successful transfer would indicate that both prompting modes share a common representation of answerability that can be accessed without an explicit judgment instruction.

Formally, we derive the following cross-mode score:

$$z_{D \to J,l}(x) = \frac{(\boldsymbol{h}_l^D(x) - \boldsymbol{c}_{D,l})^\top \hat{\boldsymbol{d}}_{J,l}}{\sigma_{D \to J,l}}, \; \sigma_{D \to J,l} = \sqrt{\frac{1}{|\mathcal{D}_{\text{train}}| - 1} \sum_{x \in \mathcal{D}_{\text{train}}} \left[(\boldsymbol{h}_l^D(x) - \boldsymbol{c}_{D,l})^\top \hat{\boldsymbol{d}}_{J,l}\right]^2} \tag{14}$$

Development Diagnostics. We analyze the last prompt token's hidden state at each layer of the frozen Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct models using 1,226 training and 326 development questions. Projection directions, centroids, and scales are estimated on the training split. We also standardize features using training statistics and fit a class-balanced logistic regression probe per layer and prompt. We select its inverse regularization strength $C_{\text{probe}} \in \{0.01, 0.1, 1, 10\}$ by development Macro-F1 and plot the resulting development scores.

Figure 6 reveals stronger answerability separation in the later layers under both prompting modes. For the 3B model, the judgment-mode and direct-answer probes achieve peak development-set macro-F1 scores of 91.72% at layer 27 and 90.79% at layer 31, respectively. For the 7B model, the corresponding maxima are 91.40% at layer 19 and 91.71% at layer 20. These development-set diagnostics support the use of later-layer representations in the routing head.

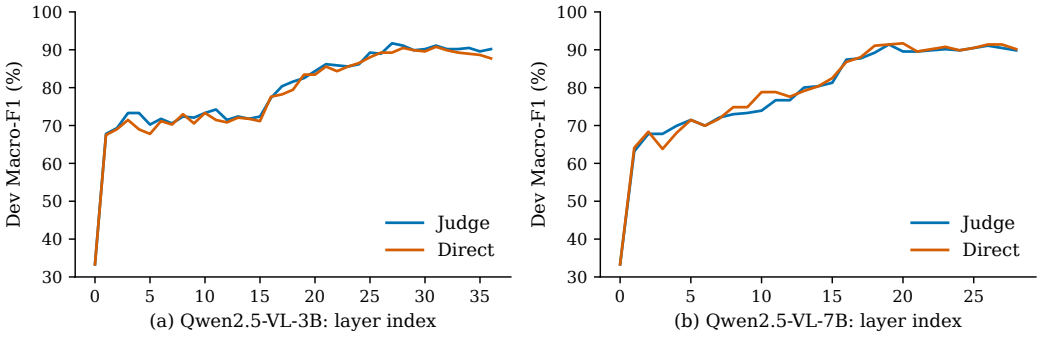

Figure 6: Layerwise linear-probe macro-F1 on the development set for (a) Qwen2.5-VL-3B-Instruct and (b) Qwen2.5-VL-7B-Instruct. The probes operate on frozen hidden states extracted from the final prompt token, with probe parameters fitted on the training set and regularization hyperparameters selected on the development set. The horizontal axes indicate hidden-state indices, where index 0 denotes the embedding output. Blue and orange represent the judgment and direct-answer prompting modes, respectively. Higher values indicate stronger discrimination between answerable and unanswerable questions..

Figure 7 presents the 7B projection heatmaps for the same 326 development examples used in Figure 2. At layer 22, which achieves the highest development-set AUROC for the direct-answer projection, the direct-answer, judgment-mode, and cross-prompt projections attain AUROCs of 94.72%, 94.51%, and 93.98%, respectively. These results demonstrate that the cross-prompt projection retains strong answerability separation at the 7B scale.

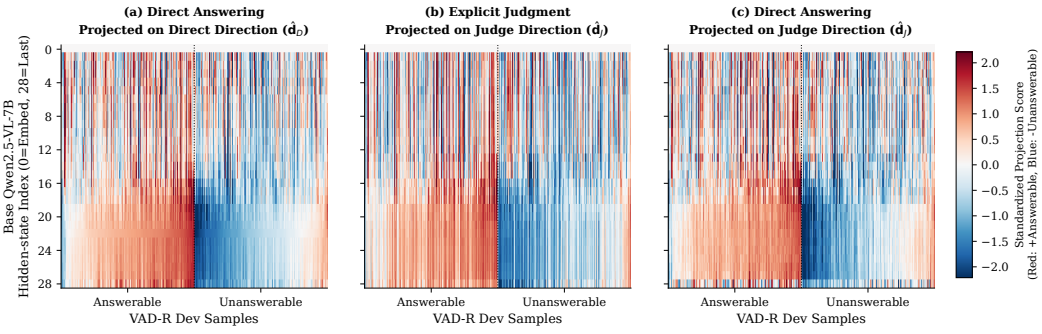

Figure 7: Layerwise per-sample answerability-projection heatmaps for Qwen2.5-VL-7B on the VAD-R development split, comprising 163 answerable instances on the left and 163 unanswerable instances on the right. The answerability directions, class centroids, and normalization scales are estimated from the training split. **(a)** Direct-answer representations projected onto their intrinsic answerability direction $\hat{\boldsymbol{d}}_{D,l}$. **(b)** Explicit-judgment representations projected onto the judgment-mode answerability direction $\hat{\boldsymbol{d}}_{J,l}$. **(c)** Direct-answer representations projected onto the judgment-mode direction $\hat{\boldsymbol{d}}_{J,l}$ after centering by $\boldsymbol{c}_{D,l}$ and normalizing by $\sigma_{D\to J,l}$. Positive values, shown in red, indicate alignment with the answerable direction, whereas negative values, shown in blue, indicate alignment with the unanswerable direction.

B.3 VAD-R Benchmark Construction, Taxonomies, and Distributions

Benchmark Motivation and Design Principles. Vision-language models frequently hallucinate answers to unanswerable VQAs due to strong language priors and training objectives that reward answering regardless of visual evidence sufficiency. Existing unanswerability datasets predominantly rely on coarse binary labels or fixed refusal phrases, lacking fine-grained evidence grounding or verifiable causal explanations. We design VAD-R as a diagnostic evaluation benchmark. Every question is annotated with a step-by-step visual reasoning rationale, and every unanswerable question includes an explicit explanation detailing precisely why the visual evidence is insufficient, mapped to a causal taxonomy.

Three-Stage Construction Pipeline. We first generate candidate questions, then apply shortcut filtering and independent quality verification:

1. **Stage 1: Qwen3.5-122B-A10B Generation.** Over a diverse collection of real-world images from DOCCI (Onoe et al., 2024) spanning four core visual domains, namely Fine-Grained Counting, Text & Symbol Reading, Visual Attribute Recognition, and Visual Localization, Qwen3.5-122B-A10B (Qwen Team, 2026a) generates candidate questions across both answerable and unanswerable queries referencing the visual scenes. Every candidate question receives a step-by-step reasoning rationale, and unanswerable candidates additionally receive causal evidence-gap labels. Initial quality filtering yields 30,386 questions.
2. **Stage 2: Question-Only Difficulty Filtering with Qwen2.5-VL-7B-Instruct (Top 1/3).** A common vulnerability in VLM unanswerability benchmarks is language shortcut leakage, where models predict unanswerability purely from text phrasing without inspecting visual evidence. To eliminate textual shortcuts, we train a two-layer Question-Only MLP probe on frozen Qwen2.5-VL-7B-Instruct hidden features extracted at the final prompt token. The difficulty of each question is quantified by its cross-entropy loss under the probe, where a higher loss signifies that the answerability label cannot be easily predicted from text alone. Specifically, the Question-Only MLP probe is trained exclusively on the candidate pool of the training split ($\mathcal{D}_{\text{train}}$). To establish an unbiased filtering criterion, we evaluate probe cross-entropy losses on the development split ($\mathcal{D}_{\text{dev}}$) to calibrate category-specific score thresholds that retain approximately the hardest top one-third ($1/3$) of candidate questions within each visual dimension and answerability category. Finally, we apply these fixed thresholds uniformly across candidate questions in the training, development, and test splits ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{dev}}$, and $\mathcal{D}_{\text{test}}$). This procedure retains 8,357 questions in total, including 4,066 unanswerable candidates, purging textual shortcuts while preventing distribution leakage across splits.
3. **Stage 3: Qwen3.6-27B Quality Verification and Class Balancing.** To ensure that unanswerable queries genuinely lack sufficient visual evidence, Qwen3.6-27B (Qwen Team, 2026b) independently evaluates each unanswerable candidate using the image, question, and candidate reference explanation, assigning a continuous verification confidence score in $[0, 1]$. We strictly retain candidates with verification scores of at least 0.8. After selecting the unanswerable questions, we retain an equal number of answerable questions, enforcing a balanced 1:1 ratio between answerable and unanswerable questions. Among the 4,066 audited unanswerable candidates, 926 pass verification, and pairing them with an equal number of answerable questions yields a balanced benchmark of 1,852 questions.

This pipeline produces the final benchmark of 1,852 questions, partitioned into 1,226 training, 326 development, and 300 testing questions.

Human Quality Audit. To validate annotation quality, an independent human expert conducted a stratified audit on 100 randomly sampled test questions across all four visual skills and answerability classes. The annotator independently verified whether the image genuinely supports the answer or substantiates the stated reason for unanswerability. Across answerable and unanswerable questions combined, 93 annotations were confirmed completely correct, 6 were incorrect, and 1 was borderline, yielding a confirmed accuracy of 93%.

Visual Dimensions and Causal Taxonomies. Table 10 provides the detailed quantitative distributions of VAD-R across visual dimensions and causal evidence-gap categories. Subtable (a) details

Table 10: Distributional breakdown of the VAD-R benchmark across visual dimensions and causal evidence-gap categories. Percentages in Subtable (b) are calculated relative to the 926 unanswerable questions. Multiple labels may apply to a single question.

(a) Distribution across visual skills ($N = 1{,}852$).

Visual Dimension	Count	Ratio (%)	Scope / Description
Fine-Grained Counting	814	43.95	Number of objects, including small, overlapping, or densely arranged instances
Visual Attribute	512	27.65	Queries about color, texture, material, shape, and fine-grained visual states
Visual Localization	268	14.47	Relative positions, spatial orientation, and coordinates of objects within a scene
Text & Symbol Reading	258	13.93	Scene text, logos, icons, alphanumeric markings, and typography
Total	**1,852**	**100.00**	Complete set of benchmark questions across training, dev, and test splits

(b) Distribution across causal evidence-gap categories for unanswerable queries ($N_{\mathrm{U}} = 926$).

Evidence-Gap Category	Count	Ratio (%)	Scope / Description
Too Small / Low Resolution	485	52.38	The relevant target region lacks sufficient resolution or detail for a reliable answer
Occluded / Not Visible	410	44.28	The queried object, feature, or attribute is physically blocked or not visible
Cropped / Out of Frame	106	11.45	The required visual entity or region extends beyond the image boundaries
Unreadable / Indistinct Text	103	11.12	Blurred, corrupted, low-contrast, or incomplete text prevents reliable transcription
Not Present	88	9.50	The queried entity is completely absent from the visual scene
External / Nonvisual Info	16	1.73	The query requires domain facts or external knowledge not depicted in the image

the distribution across 4 visual dimensions: Fine-Grained Object Counting (814 questions, 43.95%), Visual Attribute Recognition (512 questions, 27.65%), Visual Localization (268 questions, 14.47%), and Text & Symbol Reading (258 questions, 13.93%). Subtable (b) presents the distribution across 6 causal evidence-gap categories for the 926 unanswerable questions: *Too Small or Low Resolution* (52.38%), *Occluded or Not Visible* (44.28%), *Cropped or Out of Frame* (11.45%), *Unreadable or Indistinct Text* (11.12%), *Not Present* (9.50%), and *Requires External or Nonvisual Information* (1.73%). Multiple causal categories may apply to a single question. The balanced 300-question test split contains 150 answerable and 150 unanswerable queries.

External Training Datasets. Joint instruction tuning uses 1,226 VAD-R training examples and 3,406 external samples, comprising 1,703 answerable examples from ScienceQA (Lu et al., 2022) and 1,703 unanswerable examples from VizWiz (Gurari et al., 2018). Qwen3.6-27B generates the supervision targets for both external sources as follows:

• *ScienceQA*: we convert image-containing multiple-choice examples into open-ended visual questions with short text answers and reasoning rationales. The conversion model receives the image, original question, choices, correct answer, and available hints, lectures, and solutions. We retain examples verified to be answerable strictly from the visual scene and rewritten query, filtering out instances that depend on external knowledge or omitted options.

• *VizWiz*: we start from the official training split, requiring at least 7 of 10 annotators to vote unanswerable. Qwen3.6-27B assigns evidence-gap labels, writes an image-grounded rationale, and generates a natural refusal. Invalid annotations are excluded. To balance the external training data, we retain all 1,310 eligible examples with 9–10 votes and select another 393 from the 7–8-vote pool without replacement using a deterministic seed-42 ranking, yielding 1,703 examples.

For external examples, the routing tokens are masked during loss computation, allowing the models to preserve general instruction-following capabilities while learning routed abstention from VAD-R.

B.4 Evaluation Benchmarks

In-Domain Benchmark (VAD-R). VAD-R serves as our primary in-domain benchmark for evaluating whether models can spontaneously recognize evidence insufficiency and abstain without prompt hints. As detailed in Appendix B.3, the evaluation uses the balanced test split of 300 questions (150 answerable and 150 unanswerable), systematically assessing models across answerability classification, A-Acc, and C-Acc.

OOD Generalization Suite. To evaluate whether abstention capabilities generalize across diverse domains and synthetic perturbations without retraining, we evaluate on two prominent OOD unanswerability benchmarks:

- **TUBench** (He et al., 2026): Evaluates unanswerable question identification across four specialized multimodal reasoning domains: code-based reasoning (UCR, 480 questions), planar geometry problems (UGeoQA, 974 questions), tabular mathematical word problems (UTabMWP, 400 questions), and multimodal visual question answering (UVQA, 500 questions). The complete test split contains 2,354 samples, comprising 1,203 answerable and 1,151 unanswerable questions.
- **UNK-VQA** (Guo et al., 2024): Evaluates abstention capabilities across heuristic textual and visual perturbations applied to MS-COCO images and questions, with five subsets: Image Context Mislead (I-1), Object Masking (I-2), Object Copy-Move (I-3), Text Over-Specification (T-1), and Text Missing Details (T-2). The evaluated subset contains 4,539 answerable and 5,293 unanswerable questions.

  The evaluated UNK-VQA benchmark comprises 9,832 questions with available image assets, supplemented with original MS-COCO images.

General Multimodal Perception Benchmarks. To verify that our abstention training preserves core visual-understanding capabilities, we evaluate the trained models on three standard vision-language benchmarks:

- **POPE** (Li et al., 2023): Probes object hallucination on MS-COCO across three splits of increasing difficulty: Random (sampling non-existent objects uniformly), Popular (sampling frequently occurring co-present objects), and Adversarial (sampling objects that frequently co-occur according to segmentation statistics). It evaluates $9{,}000$ binary yes/no questions, measuring whether models falsely affirm non-existent entities under hallucination pressure.
- **GQA** (Hudson & Manning, 2019): Evaluates multi-step compositional spatial reasoning, relational scene graph navigation, and structural logic over real-world imagery using $12{,}578$ balanced validation questions, assessing whether abstention mechanisms interfere with complex multi-hop visual reasoning.
- **TextVQA** (Singh et al., 2019): Assesses optical character recognition (OCR) and open-ended multimodal question answering on $5{,}000$ validation questions across $3{,}166$ images containing scene text from the Open Images dataset, scored via standard soft VQA accuracy across 10 human annotations.

B.5 Baseline Implementation Details

- **Direct Answering (P_D).** Generates a concise response directly from the image and question without explicitly assessing answerability.
- **Judge-then-Answer (P_{JA}).** First assesses whether the available visual evidence is sufficient and then either answers the question or explains why a reliable answer cannot be determined, all within a single response. The complete prompts for both baselines are provided in Appendix D.1.
- **Semantic Entropy (SE) (Kuhn et al., 2023).** SE samples $K = 10$ candidate responses and groups semantically equivalent answers using DeBERTa-v2-xlarge (He et al., 2021), fine-tuned on MultiNLI for natural language inference (NLI) (Williams et al., 2018). Two responses are assigned to the same semantic group only if the NLI model predicts entailment in both directions. The semantic entropy computed over these groups is then used for answerability classification.
- **VL-Uncertainty (VL-U) (Zhang et al., 2024).** Estimates uncertainty from responses sampled under clean inputs, textual paraphrases, and joint text–image perturbations. We evaluate VL-U only on answerability classification.
- **Visual Contrastive Decoding (VCD) (Leng et al., 2024).** For classification, we first generate an answer from the original image. We then measure how much its mean token log-probability drops when noise is added to the image, averaging over five independently sampled noise patterns. A threshold selected on the development set converts this score into an answerability prediction. For response generation, VCD contrasts original-image and noisy-image logits at each decoding step.

- **LVLM-LP (Zhao et al., 2024).** Fits a logistic-regression classifier to the first-token vocabulary logits while keeping the VLM frozen. We evaluate both answerability classification and complete response generation. For inputs predicted to be unanswerable, the fixed prefix *"Sorry, this question is unanswerable, because"* is inserted before the model generates the continuation.
- **CLIP-UP (Vardi et al., 2026).** Computes the elementwise product of the CLIP image and text embeddings, $\mathbf{u}_{\text{CLIP}} = \mathbf{f}_{\text{img}} \odot \mathbf{f}_{\text{text}}$, and projects the resulting representation into a learned soft prefix that guides answer-or-abstain generation. We evaluate its complete generated responses.
- **Rep2Act (Ours).** Produces both an answerability classification and a complete response conditioned on the selected A/U routing token. Rep2Act, Rep2Act†, and Rep2Act‡ employ fused routing ($\alpha = 0.50$ for the 3B model and $\alpha = 0.15$ for the 7B model), routing-head-only inference ($\alpha = 1.0$), and vocabulary-only inference ($\alpha = 0.0$), respectively. All three variants are evaluated at both the answerability-classification and response-generation levels.

B.6 Rep2Act Implementation Details

Model Training and Hyperparameters. The training set contains 1,226 VAD-R samples and 3,406 external samples from ScienceQA and VizWiz. For Rep2Act, the representation route head is first trained for three epochs on VAD-R with the backbone frozen. The 3B and 7B joint stages then use two epochs, an effective batch size of 16, and a learning rate of 10^{-5} with AdamW and cosine decay. Joint training updates the route head and LoRA attention adapters; the original backbone weights remain frozen. The thresholds are $\gamma = 1.0$ for Signed Margin Loss and $\delta = 0.1$ for Route-Contrastive Loss, with weights $\lambda_r = 1.0$, $\lambda_m = 0.2$, and $\lambda_c = 1.0$. The head applies per-layer L2 normalization, mean pooling, LayerNorm, and a two-output linear classifier with softmax. Hidden-state indices are 25–35 for Qwen2.5-VL-3B-Instruct and 20–28 for Qwen2.5-VL-7B-Instruct.

Inference Fusion Parameter α Selection and Freezing. Inference uses Equation 11. The fusion weight α is selected on the 326-question development set via grid search over $\{0, 0.05, \ldots, 1\}$, yielding $\alpha = 0.50$ for Rep2Act-3B and $\alpha = 0.15$ for Rep2Act-7B. These parameters are fixed and evaluated across all in-domain and out-of-distribution benchmarks. For the controlled 3B loss ablation on VAD-R, $\alpha = 0.50$ is likewise fixed for all variants. General VQA uses vocabulary-only routing with $\alpha = 0$; the selected A/U token still conditions the continuation.

Evaluation Metrics and Evaluation Protocols. We compare answerability predictions with gold A/U labels and report Accuracy, Macro-Precision, Macro-Recall, Macro-F1, and AUROC, using continuous scores for AUROC. Qwen3.6-27B labels each generated response as answering or abstaining; A-Acc measures agreement with the gold label. For C-Acc on VAD-R, Qwen3.8-27B (Qwen Team, 2026c) compares each response with the reference answer and explanation: answerable questions require a correct answer, while unanswerable questions require both abstention and a correct explanation of the missing visual evidence. Official TUBench evaluation uses its subset-specific prompts and parsers and reports F1 for the unanswerable class. For UNK-VQA, we use the published Binary Judgment (BY) and Multiple-Choice (MC) prompts and implement output parsing to compute accuracy. BY predicts answerable or unanswerable; MC selects from three candidate answers and an unanswerable option. Parsing failures count as incorrect. All metrics are reported as percentages.

Hardware and Computing Environment. All model fine-tuning, probe training, and generative inference were conducted on a dedicated server equipped with $8\times$ NVIDIA RTX A6000 GPUs, each with 48 GB of VRAM. The software environment comprised Ubuntu Linux, NVIDIA driver 550.67, CUDA 12.4, PyTorch 2.4, and the vLLM inference engine, which was used to accelerate batched response generation and automated evaluation.

C Additional Experimental Results

C.1 Full Evaluation Results Across Individual VLMs on VAD-R

Table 11 presents the complete evaluation results for 8 open-source VLMs and 3 closed-source models on the VAD-R test split. Under direct answering (P_D), models across parameter scales predominantly attempt to answer every question, achieving near-perfect recall on answerable queries but minimal abstention recall on unanswerable ones. Specifically, abstention recall ranges from 0.0% to 23.3% for open-source models and from 11.9% to 22.5% for closed-source models. In contrast, when prompted to make an explicit judgment (P_J), the same models exhibit substantially higher abstention recall, reaching 96.7% for Qwen3.6-27B and 98.7% for Claude-Haiku-4.5. This striking contrast demonstrates that the representation-to-action disconnect is a pervasive phenomenon across both open-source models and commercial frontier models.

Table 11: Complete empirical evaluation across 8 open-source VLMs and 3 closed-source models on the VAD-R test set under Direct Answering (P_D) and Explicit Judgment (P_J) modes. Direct responses are classified by Qwen3.6-27B as Action Judge. Metrics include Action Accuracy (A-Acc), Answerable Recall (Ans-R), Abstain Recall on unanswerable queries (Abs-R), Abstain F1 (Abs-F1), and the Abstain Recall Gap ΔAbs-R $=$ Abs-R$_{P_J}$ $-$ Abs-R$_{P_D}$. A-Acc measures answer/abstain correctness from generated responses under P_D and explicit yes/no judgments under P_J.

	Direct Answering Mode (P_D)				**Explicit Judgment Mode (P_J)**				**Gap**
Model Architecture	**A-Acc**	**Ans-R**	**Abs-R**	**Abs-F1**	**A-Acc**	**Ans-R**	**Abs-R**	**Abs-F1**	**Δ Abs-R**
Open-Source Models									
Qwen2.5-VL-3B	56.7	**100.0**	13.3	23.5	74.7	81.3	68.0	73.4	+54.7
Qwen2.5-VL-7B	59.3	**100.0**	18.5	31.3	83.4	76.8	90.1	84.5	+71.6
Qwen3.6-27B	**61.7**	**100.0**	**23.3**	**37.8**	**91.3**	86.0	96.7	**91.8**	+73.4
Qwen3.8-27B	61.0	**100.0**	22.0	36.1	89.7	**93.3**	86.0	89.3	+64.0
LLaVA-v1.5-7B	50.3	**100.0**	0.7	1.3	62.9	76.2	49.7	57.3	+49.0
LLaVA-v1.5-13B	50.0	**100.0**	0.0	0.0	67.2	74.2	60.3	64.8	+60.3
InternVL3-8B	54.3	**100.0**	8.6	15.9	78.1	74.8	81.5	78.9	+72.9
mPLUG-Owl2-7B	52.3	**100.0**	4.6	8.9	73.2	55.0	91.4	77.3	+86.8
Open-Source Average	**55.7**	**100.0**	**11.4**	**19.4**	**77.6**	**77.2**	**78.0**	**77.2**	+66.6
Closed-Source Models									
Gemini-3.8-Flash	59.9	97.4	22.5	36.6	86.8	76.8	96.7	88.0	+74.2
Claude-Haiku-4.5	55.6	96.7	14.6	24.9	68.5	38.4	**98.7**	75.8	+84.1
GPT-5.4-mini	56.0	**100.0**	11.9	21.3	89.7	90.1	89.4	89.7	+77.5
Closed-Source Average	**57.2**	**98.0**	**16.3**	**27.6**	**81.7**	**68.4**	**94.9**	**84.5**	+78.6

Fine-Grained Visual Skill Breakdown. Table 12 reports results under both prompting modes across 4 visual skills. For most open-source and closed-source models, explicit judgment yields high abstention recall for text recognition, attribute identification, and localization, whereas direct answering substantially reduces it. Counting is particularly challenging during direct generation: every base model abstains on fewer than 13% of unanswerable counting questions. Localization is comparatively less affected, particularly for Qwen models. These results demonstrate that the representation-to-action disconnect persists across model families and is most pronounced for counting and fine-grained visual details. Rep2Act improves direct-answering accuracy across all skills and achieves the strongest average performance at both model scales. Rep2Act-7B performs best on counting, while Rep2Act-3B leads on the other three skills.

C.2 Complete TUBench Results under Official Evaluation Protocol

Table 13 presents the complete numerical results corresponding to Figure 5. Rep2Act-3B achieves an average F1 score of 53.3%, outperforming GPT-4 Turbo and GPT-4o by 16.2% and 1.1%, respectively, while delivering leading performance on specialized subsets such as UTabMWP. These results demonstrate that explicit representation-to-action routing transfers effectively to diverse out-of-distribution structures without task-specific fine-tuning.

Table 12: Fine-grained performance breakdown across four visual skills on VAD-R. A-Acc measures answer/abstain correctness from generated responses under P_D and explicit yes/no judgments under P_J.

	Counting		Text & Symbol		Visual Attribute		Localization		Average	
Model / Prompting Setup	**A-Acc**	**Abs-R**	**A-Acc**	**Abs-R**	**A-Acc**	**Abs-R**	**A-Acc**	**Abs-R**	**A-Acc**	**Abs-R**
Explicit Judgment Mode (P_J)										
Qwen2.5-VL-3B	69.05	42.86	88.10	**100.00**	78.57	69.05	**95.83**	**100.00**	82.89	77.98
Qwen2.5-VL-7B	82.54	77.78	85.71	**100.00**	80.95	**100.00**	89.58	**100.00**	84.69	94.44
Qwen3.6-27B	92.06	96.83	**92.86**	95.24	**91.67**	97.62	89.58	**100.00**	**91.54**	97.42
Qwen3.8-27B	93.65	98.41	85.71	76.19	83.33	69.05	**95.83**	95.83	89.63	84.87
LLaVA-v1.5-7B	54.76	28.57	66.67	**100.00**	61.90	45.24	83.33	70.83	66.67	61.16
LLaVA-v1.5-13B	69.05	65.08	61.90	90.48	63.10	45.24	75.00	50.00	67.26	62.70
InternVL3-8B	74.60	68.25	80.95	**100.00**	77.38	85.71	87.50	95.83	80.11	87.45
mPLUG-Owl2-7B	76.98	82.54	59.52	**100.00**	63.10	**100.00**	93.75	95.83	73.34	94.59
Gemini-3.8-Flash	84.13	98.41	85.71	95.24	89.29	92.86	89.58	**100.00**	87.18	96.63
Claude-Haiku-4.5	65.08	**100.00**	64.29	**100.00**	72.62	95.24	75.00	**100.00**	69.25	**98.81**
GPT-5.4-mini	**94.44**	**100.00**	**92.86**	**100.00**	79.76	66.67	93.75	91.67	90.20	89.59
Direct Answering Mode (P_D)										
Qwen2.5-VL-3B	49.21	0.00	59.52	19.05	51.19	2.38	83.33	66.67	60.81	22.02
Qwen2.5-VL-7B	55.56	11.11	54.76	9.52	53.57	7.14	83.33	66.67	61.80	23.61
Qwen3.6-27B	56.35	12.70	61.90	23.81	59.52	19.05	85.42	70.83	65.80	31.60
Qwen3.8-27B	53.17	6.35	59.52	19.05	60.71	21.43	83.33	66.67	64.18	28.37
LLaVA-v1.5-7B	50.79	1.59	50.00	0.00	50.00	0.00	50.00	0.00	50.20	0.40
LLaVA-v1.5-13B	50.00	0.00	50.00	0.00	50.00	0.00	50.00	0.00	50.00	0.00
InternVL3-8B	50.79	1.59	54.76	9.52	51.19	2.38	68.75	37.50	56.37	12.75
mPLUG-Owl2-7B	54.76	9.52	52.38	4.76	50.00	0.00	50.00	0.00	51.79	3.57
Gemini-3.8-Flash	53.17	7.94	73.81	47.62	60.71	26.19	64.58	33.33	63.07	28.77
Claude-Haiku-4.5	53.17	6.35	61.90	33.33	48.81	2.38	68.75	41.67	58.16	20.93
GPT-5.4-mini	50.79	1.59	64.29	28.57	51.19	2.38	70.83	41.67	59.28	18.55
Rep2Act-3B	82.54	74.60	**85.71**	**95.24**	**86.90**	**90.48**	**95.83**	**100.00**	87.74	**90.08**
Rep2Act-7B	**91.27**	**92.06**	83.33	90.48	84.52	83.33	93.75	91.67	**88.22**	89.39

Table 13: TUBench OOD evaluation under official subset-specific prompts and parsers. F1 is evaluated on the unanswerable class.

	UCR			UVQA			UGeoQA			UTabMWP			Average		
Model	**F1**	**Acc**	**C-Acc**	**F1**	**Acc**	**C-Acc**	**F1**	**Acc**	**C-Acc**	**F1**	**Acc**	**C-Acc**	**F1**	**Acc**	**C-Acc**
Open-Source Models															
LLaVA-v1.5-7B	0.0	55.4	28.1	0.0	50.0	31.6	63.9	51.3	**45.0**	41.8	50.5	34.2	26.4	51.8	34.7
LLaVA-v1.5-13B	0.0	55.4	25.4	0.0	50.0	32.6	**64.2**	51.8	**45.0**	59.2	49.2	42.2	30.8	51.6	36.3
LLaVA-v1.6-7B	0.0	55.4	27.7	0.0	50.0	30.4	5.1	50.0	14.5	17.8	53.8	26.0	5.7	52.3	24.6
LLaVA-v1.6-13B	16.0	54.2	29.2	7.7	51.8	30.2	48.3	48.4	30.6	61.5	53.8	45.8	33.4	52.0	33.9
Qwen2.5-VL-3B	7.0	55.4	27.9	69.1	71.4	62.0	23.2	54.5	29.1	19.0	55.3	50.5	29.6	59.2	42.4
Closed-Source Models															
Qwen-VL-Max	16.1	56.5	30.4	67.7	74.4	61.8	43.6	57.3	33.3	7.7	52.0	34.8	33.8	60.0	40.1
Qwen-VL-Plus	23.8	56.0	31.5	7.6	51.6	41.0	8.8	50.8	15.4	3.9	50.7	31.8	11.0	52.3	29.9
Qwen3-VL-Flash	49.2	**67.7**	**57.9**	48.2	65.6	58.2	9.8	52.6	**51.2**	63.5	73.2	73.0	42.7	64.8	**60.1**
Qwen3-VL-Plus	28.3	61.0	53.5	50.0	66.4	55.6	10.1	52.7	50.0	**72.6**	**78.5**	**78.5**	40.3	64.7	59.4
GPT-4 Turbo	57.7	57.5	45.0	77.6	**80.6**	68.4	5.6	51.4	23.2	7.7	52.0	38.8	37.1	60.4	43.8
GPT-4o mini	**57.8**	51.9	41.0	**79.3**	77.6	66.4	32.6	57.9	27.0	44.2	64.0	47.0	**53.5**	62.8	45.4
GPT-4o	53.0	60.8	39.8	76.8	80.2	**68.6**	19.5	55.1	26.6	59.6	70.5	59.5	52.2	**66.7**	48.6
Rep2Act 3B	42.1	59.4	34.4	68.6	68.2	58.0	40.1	**60.2**	37.8	62.2	71.8	65.5	53.3	64.9	48.9

C.3 SFT Response Target Construction

Table 14 examines the effects of response-target formatting on VAD-R using concise and detailed outputs for both answerable and unanswerable queries. *Short* targets pair concise answers with brief abstention responses such as "I cannot answer", whereas *Detailed* targets pair complete answers with

explicit causal rationales identifying the missing visual evidence. We evaluate both formats with and without chain-of-thought reasoning. Without an explicit decision token, standard SFT baselines struggle to balance answering and abstaining. *Short w/o CoT* answers nearly every query, achieving 98.00% answer recall but only 28.00% abstention recall. Introducing chain-of-thought reasoning in *Short w/ CoT* increases abstention recall to 73.33% but reduces answer recall to 81.33%. Models trained with detailed targets similarly plateau at A-Acc values between 76.67% and 77.67%, with abstention recall remaining near 65%. In contrast, Rep2Act achieves more balanced performance, with 86.67% answer recall, 86.00% abstention recall, and 86.33% macro-F1.

Table 14: Ablation of response-target formatting on VAD-R. Ans-R and Abs-R denote answer recall and abstention recall, respectively.

Response Target	Loss	A-Acc	Ans-R	Abs-R	M-F1
Short w/o CoT	LM Only	63.00	**98.00**	28.00	57.83
Short w/ CoT	LM Only	77.33	81.33	<u>73.33</u>	<u>77.30</u>
Detailed w/ CoT	LM Only	76.67	88.00	65.33	76.36
Detailed w/o CoT	LM Only	<u>77.67</u>	<u>92.00</u>	63.33	77.20
Rep2Act	Full objective	**86.33**	86.67	**86.00**	**86.33**

C.4 Feature Anchoring versus Direct Routing Supervision

Feature Anchor Method. We implement this approach using an anchoring loss that minimizes the discrepancy between the student model's direct-answer projection score and the frozen base model's judgment-mode projection score through the Huber penalty H:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{x \in \mathcal{D}_{\text{train}}} H\big(z^{\theta}_{D \to J,l}(x) - z^{\text{base}}_{J,l}(x)\big), \quad H(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le 1 \\ |e| - \frac{1}{2}, & |e| > 1 \end{cases} \tag{15}$$

Here, θ denotes the trainable parameters of the student model.

The feature-anchor baseline uses layer $l = 31$ of Qwen2.5-VL-3B-Instruct for both the student Direct features and the frozen Base Judge features. FA-Route-SFT trains the model to predict the correct A/U token and generate the reference response, as in Route-SFT. It also uses a Huber loss, weighted by 0.1, to bring the student's answerability projection under Direct prompting closer to the frozen base model's projection under Judge prompting.

Results and Analysis Table 8 shows that feature anchoring improves A-Acc by only 0.66% over standard Route-SFT. To understand this limited gain, we analyze the cross-prompt centroid shift and the objective optimized by the anchoring loss. The centroids and judgment-mode answerability direction are estimated from the training representations. For the base Qwen2.5-VL-3B model at the final layer (hidden-state index 36), the parallel and orthogonal components have ℓ_2 norms of $\|\Delta \boldsymbol{c}_{\parallel}\|_2 = 10.33$ and $\|\Delta \boldsymbol{c}_{\perp}\|_2 = 91.09$, respectively. The dominant orthogonal component may partly reflect the change in task induced by prompt switching: the model transitions from answering a question to determining whether that question is answerable. Prior work similarly demonstrates that prompts can alter the task encoded in hidden representations (Hendel et al., 2023; Todd et al., 2024; Kirsanov et al., 2025). Plausible contributors to $\Delta \boldsymbol{c}_{\perp}$ include:

- the current task mode: directly answering a question or judging whether it is answerable;
- the expected output space: an open-ended answer or a binary yes/no judgment;
- the content and length of the instruction;
- the identity and position of the final prompt token from which the representation is extracted;
- prompt-dependent attention pathways;
- the model's internal preparation for the subsequent decoding task.

Together, these factors may account for the substantial orthogonal shift. Because the anchoring loss aligns only the models' projection scores along the answerability direction, its influence on the final A/U token decision remains indirect.

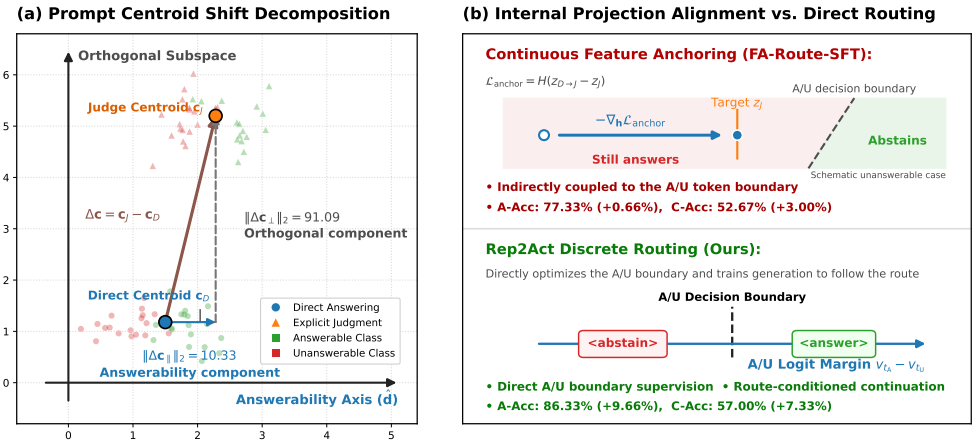

Figure 8: Geometric decomposition of prompt centroid shift and comparison between internal projection alignment and direct routing supervision. **(a)**: The total displacement $\Delta \boldsymbol{c} = \boldsymbol{c}_J - \boldsymbol{c}_D$ contains a component parallel to the answerability direction, $\Delta \boldsymbol{c}_\parallel$, and a larger orthogonal prompt-mode component, $\Delta \boldsymbol{c}_\perp$. **(b)**: Feature anchoring aligns the Direct cross-prompt score $z_{D\to J}$ with the frozen Judge score z_J, whereas Rep2Act directly optimizes the A/U decision boundary and conditions the continuation on the selected route.

The anchor loss uses the student's Direct feature $\boldsymbol{h}_l^{D,\theta}(x)$ and the frozen Base Judge feature $\boldsymbol{h}_l^{J,\text{base}}(x)$ at layer $l = 31$. Both are projected onto the frozen unit Judge direction $\hat{\boldsymbol{d}}_{J,l}^{\text{base}}$. For N training samples and prompt mode $M \in \{D, J\}$, the implemented centering and normalization are

$$\begin{aligned} \boldsymbol{c}_{M,l}^{\text{base}} &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}_l^{M,\text{base}}(x_i) \\ \sigma_{M\to J,l}^{\text{base}} &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left[(\boldsymbol{h}_l^{M,\text{base}}(x_i) - \boldsymbol{c}_{M,l}^{\text{base}})^\top \hat{\boldsymbol{d}}_{J,l}^{\text{base}}\right]^2} \end{aligned} \tag{16}$$

Thus, both scales measure variation along the Judge direction, with $\sigma_{J\to J,l}^{\text{base}} = \sigma_{J,l}^{\text{base}}$. This is the same sample-standard-deviation normalization used in the heatmaps. The centers, scales, direction, and Judge target are fixed during training. The two scores used by the anchor loss are

$$\begin{aligned} z_{D\to J,l}^{\theta}(x) &= \frac{(\boldsymbol{h}_l^{D,\theta}(x) - \boldsymbol{c}_{D,l}^{\text{base}})^\top \hat{\boldsymbol{d}}_{J,l}^{\text{base}}}{\sigma_{D\to J,l}^{\text{base}}} \\ z_{J,l}^{\text{base}}(x) &= \frac{(\boldsymbol{h}_l^{J,\text{base}}(x) - \boldsymbol{c}_{J,l}^{\text{base}})^\top \hat{\boldsymbol{d}}_{J,l}^{\text{base}}}{\sigma_{J,l}^{\text{base}}} \end{aligned} \tag{17}$$

To differentiate the loss for one sample, write $\boldsymbol{h} = \boldsymbol{h}_l^{D,\theta}(x)$ and $e = z_{D\to J,l}^{\theta}(x) - z_{J,l}^{\text{base}}(x)$. Since Judge target is frozen, its derivative with respect to $\boldsymbol{h}$ is zero. Differentiating the linear projection gives

$$\nabla_{\boldsymbol{h}} e = \nabla_{\boldsymbol{h}} z_{D\to J,l}^{\theta}(x) = \frac{\hat{\boldsymbol{d}}_{J,l}^{\text{base}}}{\sigma_{D\to J,l}^{\text{base}}} \tag{18}$$

The implemented Huber loss has threshold 1: $H(e) = \frac{1}{2}e^2$ for $|e| \le 1$ and $H(e) = |e| - \frac{1}{2}$ otherwise. Applying the chain rule yields

$$\begin{aligned} \nabla_{\boldsymbol{h}} H(e) &= \frac{\mathrm{d}H(e)}{\mathrm{d}e}\nabla_{\boldsymbol{h}} e \\ &= \begin{cases} e\,\dfrac{\hat{\boldsymbol{d}}_{J,l}^{\text{base}}}{\sigma_{D\to J,l}^{\text{base}}}, & |e| \le 1 \\ \operatorname{sign}(e)\,\dfrac{\hat{\boldsymbol{d}}_{J,l}^{\text{base}}}{\sigma_{D\to J,l}^{\text{base}}}, & |e| > 1 \end{cases} \end{aligned} \tag{19}$$

For the weighted mean anchor loss over n eligible VAD-R samples, the contribution of sample i to its hidden-state gradient is $\frac{0.1}{n}\nabla_{\boldsymbol{h}_i} H(e_i)$, using the implemented anchor weight 0.1. The parameter gradient follows by backpropagating through the student feature:

$$\nabla_\theta\big(0.1\mathcal{L}_{\text{anchor}}\big) = \frac{0.1}{n}\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{h}_i}{\partial \theta}\right)^{\top}\nabla_{\boldsymbol{h}_i} H(e_i) \tag{20}$$

Here $\text{sign}(e)$ is $+1$ for positive errors and -1 for negative errors.

Equation 19 shows that the gradient with respect to the hidden representation is a scalar multiple of the fixed judgment-mode direction. When the student's projection is too high ($e > 0$), gradient descent decreases it; when the projection is too low ($e < 0$), gradient descent increases it. The anchoring loss therefore trains the student to match the judgment-mode projection. However, because this gradient depends on the projection error rather than the correctness of the A/U token decision, it influences that decision only indirectly. In contrast, Rep2Act's signed margin loss derives its gradient directly from the ground-truth A/U label and the competing token logits, while the route-contrastive loss further encourages the generated response to remain consistent with the correct route.

C.5 Sensitivity Analysis of Training-Loss Hyperparameters

Table 15 presents detailed sensitivity analyses of the signed margin loss weight λ_m and route-contrastive loss weight λ_c on VAD-R.

Signed Margin Loss Weight λ_m. Signed Margin Loss directly widens the signed vocabulary-logit difference $s(v_{t_\text{A}} - v_{t_\text{U}})$ at the A/U decision step. At $\lambda_m = 0.0$, the model reaches 74.33% A-Acc. Increasing λ_m to 0.2 raises accuracy to 82.33% and yields balanced recall between answer and abstain categories, at 81.33% and 83.33% respectively. Further increasing λ_m to 0.5 reduces accuracy to 80.67% as the model begins favoring the answer route. Raising λ_m to 1.0 and 2.0 decreases abstain recall to 66.67% and 61.33%, bringing final accuracy down to 75.67%.

Route-Contrastive Loss Weight λ_c. Route-Contrastive Loss penalizes counterfactual generations, ensuring that the gold continuation remains significantly more probable under the chosen route than under the inverted route. At $\lambda_c = 0.0$, route-action consistency is 89.00% and A-Acc is 78.33%. Increasing λ_c through 0.1, 0.5, and 1.0 progressively elevates A-Acc to 82.33%, 85.00%, and 86.33%, while driving consistency up to 93.67%, 99.00%, and 99.33%. Setting λ_c to 2.0 slightly degrades A-Acc to 85.33% at this larger loss weight.

Table 15: Hyperparameter sensitivity analysis on VAD-R. Cons. denotes the percentage of route–action-consistent responses.

(a) Margin Loss weight λ_m, with fixed $\lambda_c = 0.1$.

λ_m	A-Acc	M-P	M-R	M-F1	Cons.
0.0	74.33	74.39	74.33	74.32	**98.67**
0.2	**82.33**	**82.35**	**82.33**	**82.33**	93.67
0.5	80.67	81.47	80.67	80.54	89.33
1.0	77.67	79.07	77.67	77.39	84.33
2.0	75.67	77.96	75.67	75.16	80.67

(b) RC Loss weight λ_c, with fixed $\lambda_m = 0.2$.

λ_c	A-Acc	M-P	M-R	M-F1	Cons.
0.0	78.33	78.55	78.33	78.29	89.00
0.1	82.33	82.35	82.33	82.33	93.67
0.5	85.00	85.26	85.00	84.97	99.00
1.0	**86.33**	**86.33**	**86.33**	**86.33**	**99.33**
2.0	85.33	85.34	85.33	85.33	98.33

D PROMPT TEMPLATES

This section documents the prompt templates used throughout our experiments, including those for baseline methods, diagnostic probing, and automated evaluation by the action and content judges.

D.1 BASELINES AND PROBING PROMPTS

Direct Answering (P_D) requests a response without explicitly eliciting intermediate reasoning:

Prompt Template: Direct Answering (P_D)

```
{question}
```

Explicit Judgment (P_J) probes whether the base model can explicitly assess answerability through a binary query:

Prompt Template: Explicit Judgment (P_J)

```
Based only on the image, can a viewer reliably answer the question?
Answer only yes or no.
Question: {question}
```

Judge-then-Answer (P_{JA}) explicitly separates the assessment of visual-evidence sufficiency from subsequent answer generation or the articulation of an abstention rationale:

Prompt Template: Judge-then-Answer (P_{JA})

```
You are given an image and a question.
Please follow these two steps:
Step 1 [Judgment]: First determine whether the visual evidence in
the image is sufficient to answer the question.
- If the question cannot be answered from the image (e.g., the
required object or information is missing, occluded, unanswerable,
or ambiguous), output: ''Judgment: Unanswerable'' and explain why.
- If the question can be answered from the image, output:
''Judgment: Answerable'', followed by the concise answer in the
format ''Answer: <your answer>''.
Question: {question}
```

D.2 AUTOMATIC JUDGE PROMPTS

Action Judge: Qwen3.6-27B classifies each response as either an attempted answer or evidence-grounded abstention, using greedy decoding with reasoning disabled:

Prompt Template: Action Judge: Strict Binary Action Classifier

```
[System Prompt]
You are a strict binary classifier. Do not reveal reasoning.
[User Prompt]
Question:
{question}
Model response:
{response}
Does the model response refuse to answer because the image does not
provide enough evidence?
Do not explain. End with exactly one word: yes or no.
```

Content Judge: Qwen3.8-27B evaluates C-Acc on VAD-R using greedy decoding with reasoning disabled. For answerable queries ($y = \mathrm{A}$), it assesses the factual correctness of each response against the visual evidence and reference answer:

Prompt Template: Content Judge: Answerable Verification ($y = \mathrm{A}$)

```
[User Prompt]
Question:
{question}
Reference Answer:
{reference_answer}
Candidate Response:
{candidate_response}
Based on the image, is the candidate response correct? Reply with
exactly yes or no.
```

For unanswerable queries ($y = \mathrm{U}$), Qwen3.8-27B determines whether the candidate response provides a valid explanation that correctly identifies the missing visual evidence:

Prompt Template: Content Judge: Evidence-Gap Explanation Verification ($y = \mathrm{U}$)

```
[User Prompt]
Question:
{question}
Reference Missing-Evidence Explanation:
{reference_reason}
Candidate Response:
{candidate_response}
Based on the image, does the candidate response explicitly state the
correct missing visual evidence that makes the question impossible
to answer reliably? A refusal without a reason is no. Reply with
exactly yes or no.
```

E ADDITIONAL RESPONSE EXAMPLES

Figures 9 and 10 present two additional response comparisons on VAD-R across 6 methods shown in Figure 3(d), all built on the same 3B base model.



Question: How many distinct items are reflected on the side of the car above the wheel?

Base Direct

There are two distinct items reflected on the side of the car above the wheel: a person and some trees.

Judge-then-Answer

Judgment: Answerable
Answer: Two

VCD

There are two distinct items reflected on the side of the car above the wheel.

CLIP-UP

Two

LVLM-LP

Sorry, this question is unanswerable, because there is no question in the image.

Rep2Act (Ours)

<think>
<decision>
U
</decision>
The reflection is too small and blurry to discern individual objects or people clearly enough to count them accurately.
</think>
<answer>
I cannot determine this because the reflection is too small and blurry to see any distinct objects or people.
</answer>

Figure 9: VAD-R counting example. The baselines either provide a definitive count or abstain for an unrelated reason when asked about objects reflected on the car. Rep2Act correctly identifies that the reflection is too indistinct to support a reliable count. Green and gray denote responses that pass and fail Qwen3.8-27B content judge evaluation, respectively.



Question: What number is displayed on the distant blue billboard sign?

Base Direct

The number displayed on the distant blue billboard sign is 5.

Judge-then-Answer

Judgment: Unanswerable
Explanation: The image does not provide any information about a distant blue billboard sign or its number.

VCD

The blue billboard sign in the background has the number "6" displayed on it.

CLIP-UP

I cannot answer.

LVLM-LP

Sorry, this question is unanswerable, because there is no blue billboard sign in the image.

Rep2Act (Ours)

<think>
<decision>
U
</decision>
The blue billboard in the background is too far away and out of focus to discern any text or numbers.
</think>
<answer>
I cannot determine the number because the distance and focus make it impossible to read the text on the distant blue billboard sign.
</answer>

Figure 10: VAD-R text-reading example. *Base Direct* and *VCD* produce inconsistent numbers for a distant billboard, while *Judge-then-Answer* and *LVLM-LP* incorrectly claim that the billboard is absent. Rep2Act correctly explains that the text is unreadable because of the viewing distance and image focus. Green and gray denote responses that pass and fail Qwen3.8-27B content judge evaluation, respectively.