Looking Beyond the One: Operationalizing and Eliciting Visual Ambiguity in VLLMs

Yuchong Chen[1], Bowei Zou[2], Yuhan Chen[1], Yifan Fan[1], Xinyu Li[1], Shujun Cao[1], Yu Hong[1]*
[1] School of Computer Science and Technology, Soochow University, Suzhou, China
[2] Institute for Infocomm Research, A*STAR, Singapore
{ycchen0421,cyhhh1121,yifanfannlp,lxy299385,tianxianer}@gmail.com
zou_bowei@a-star.edu.sg; 20245227110@stu.suda.edu.cn

Abstract

Visual questions are often ambiguous: the same image–question pair may admit multiple valid answers depending on which region is referenced. However, current Visual Question Answering (VQA) systems typically collapse this ambiguity, committing to a single interpretation during decoding and evaluation. In this work, we study visual question ambiguity from a grounded, region-centric perspective. We operationalize ambiguity as the existence of multiple distinct answer-supporting regions in an image, each independently yielding a valid answer. This formulation makes ambiguity observable without requiring exhaustive multi-answer annotations. Based on this definition, we conduct a systematic empirical study of state-of-the-art Visual Large Language Models (VLLMs). We find that, under default decoding, VLLMs consistently under-report ambiguity—even when multiple valid visual groundings are present. Importantly, probing model hidden states reveals that ambiguity-related signals are already encoded in their internal representations, despite not being reliably expressed in outputs. Finally, we show that selectively activating multi-focus answering based on these signals can recover additional valid answers while avoiding excessive hallucination. Together, our results suggest that ambiguity in VQA is not merely an annotation artifact or capability limitation, but a property that VLLMs internally recognize yet often fail to surface under standard decoding assumptions.

1 Introduction

Visual Large Language Models (VLLMs) are commonly evaluated under the assumption that a visual question admits a single, well-defined answer. Under this paradigm, models are expected to commit to one interpretation and produce one response, while alternative interpretations are often treated as errors or hallucinations. However, in real-world visual scenes, this assumption frequently fails. Many questions are inherently ambiguous: multiple objects or regions in an image may each plausibly answer the same question.

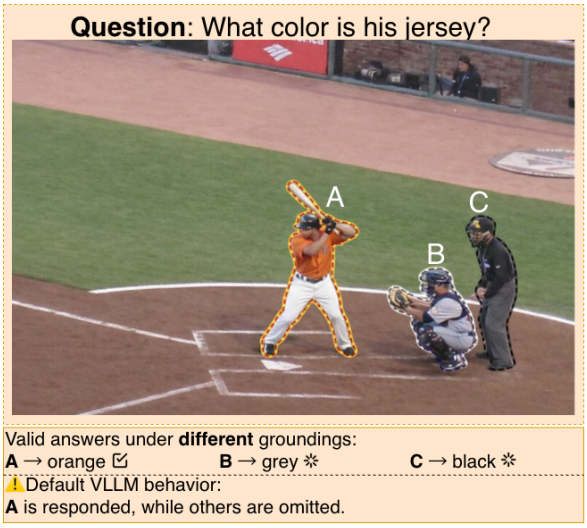


Figure 1: An example showcasing the typical behavior overlooking ambiguity among best VLLMs' generation.

For example in Figure 1, a question such as "*What color is his jersey?*" posed over an image containing multiple players does not uniquely specify a referent. Humans naturally recognize such ambiguity and may ask for clarification or enumerate multiple plausible answers. In contrast, current VLLMs are typically optimized to commit to a single interpretation during decoding.

Prior work has examined ambiguity from several perspectives, including uncertainty estimation, multi-answer annotation, and clarification question generation (Antorán et al., 2021; Chen et al., 2023; Lee et al., 2023; Jian et al., 2025). However, it remains unclear whether VLLMs fail to express ambiguity because they lack the underlying representations, or because standard decoding conven-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24306–24323

tions suppress these signals.

In this work, we study visual question ambiguity through grounded visual evidence. We consider a question to be ambiguous when multiple distinct regions in the image can each support a distinct plausible answer. This region-centric formulation treats ambiguity as an observable property of the image–question pair, enabling systematic analysis without relying on exhaustive multi-answer annotations but still aligning with current answer-centric resources.

Using this formulation, we conduct a series of controlled experiments to characterize how state-of-the-art VLLMs respond to ambiguity. We analyze their default single-pass behavior, their responses under iterative probing, and their behavior when explicitly informed that a question may admit multiple valid answers.

Complementing our behavioral analysis, we further probe the internal representations of VLLMs to assess whether ambiguity-related information is encoded in their hidden states. Using a simple classifier trained on final-layer representations, we show that ambiguity—as defined by the presence of multiple distinct answer-supporting regions—is predictable from the model's internal features. This suggests that VLLMs may internally "know" when a question is ambiguous, even if this knowledge is not consistently reflected in their outputs. To facilitate this analysis, we introduce a lightweight gated answering setup that separates ambiguity recognition from answer generation. Importantly, this mechanism is not intended as a standalone solution, but rather as a diagnostic tool that allows us to examine how selectively activating multi-focus decoding affects answer diversity, redundancy, and hallucination.

Together, these analyses provide a comprehensive picture of how ambiguity is processed or overlooked by current VLLMs. Our findings suggest that the under-reporting of ambiguity in default VLLM outputs is not solely due to a lack of capability, but is also shaped by decoding conventions and evaluation assumptions that favor single-answer responses.

Our contributions are summarized as follows:

(1) We introduce a region-based operationalization of visual question ambiguity that enables systematic analysis without exhaustive multi-answer ground truth.

(2) We provide representational evidence that ambiguity-related signals are encoded in VLLMs' hidden states, and demonstrate how this signal can be used to activate multi-focus answering.

(3) We present a comprehensive behavioral study of VLLMs under ambiguous and non-ambiguous conditions, examining default decoding, iterative probing, and explicit ambiguity-aware prompting.

2 Methodology

2.1 Problem Formulation and Notation

In the standard VQA paradigm, a visual(-language) model f_θ is tasked with generating a single answer A given an image I and a query Q. We extend this formulation to the *visual ambiguous* setting, where a query may admit multiple valid answers, each grounded in a distinct spatial region of the image.

To operationalize it, we prompt the model to generate K candidate answers, where each answer a_i is accompanied by a natural-language description l_i of its corresponding image location. To facilitate quantitative analysis, we employ Qwen3-VL as a spatial parser to map l_i into a normalized bounding box b_i. Formally, the process is denoted as

$$f_\theta(Q, I) \rightarrow \{(a_i, l_i)\}_{i=1}^{K} \rightarrow \{(a_i, b_i)\}_{i=1}^{K}, \quad (1)$$

where b_i is represented in $[0, 1000]$-normalized coordinates.

2.2 Multi-Turn Answer Probing

To explore the model's internal set of interpretations, we employ a multi-turn probing strategy. Let $S^{(0)} = \{(a_i^{(0)}, l_i^{(0)})\}_{i=1}^{K_0}$ denote the answer-location pair produced in the initial round. For subsequent iterations $t > 0$, the model is given the query Q, the image I, and the full history containing all prior discoveries $S^{(<t)} = \bigcup_{j=0}^{t-1} S^{(j)}$.

During each turn, the model is explicitly instructed to provide new answers that are grounded in different image regions than those previously identified.

The iterative process terminates when either (i) a pre-defined maximum number of rounds is reached, or (ii) the set of answers generated at step t is redundant (has appeared) with respect to $S^{(<t)}$. It is designed to exhaust the model's internal knowledge of plausible answers and uncover the full extent of the perceived visual ambiguity.

The exact prompt templates are provided in Appendix A.

2.3 Ambiguity-Aware Answer Probing

To differentiate between a model's inherent ability to uncover ambiguity and its performance under explicit instruction, we define an *ambiguity-aware* probing variant. Unlike the above setting where ambiguity is elicited implicitly over multiple turns, such variant employs a single-round prompt that characterizes the question as potentially ambiguous (refer to Appendix A for the specific prompt). The model aims to identify the full set of interpretations simultaneously

$$f_\theta^{\text{amb}}(Q, I) \rightarrow \{(a_i, l_i)\}_{i=1}^{K}. \tag{2}$$

All downstream components remain identical. The comparison sheds light on whether VLLMs require iterative probing to see beyond a single answer or if they can perceive the full scope of visual ambiguity in a single forward pass.

2.4 Region Grounding and Answer Categorization

To determine whether a candidate answer $a^{(i)}$ introduces a distinct interpretation of an ambiguous question, we evaluate its spatial novelty and semantic validity.

Spatial Novelty. We compute a region match score $R^{(i)}$ for each bounding box $b^{(i)}$ based on its maximum Intersection over Union (IoU) with all previously identified regions b' for the same question:

$$R^{(i)} = \max_{b' \in \mathcal{B}^{(<i)}} \text{IoU}(b^{(i)}, b'), \tag{3}$$

where $\mathcal{B}^{(<i)}$ denotes the set of bounding boxes associated with answers generated in previous rounds. A lower $R^{(i)}$ indicates that the answer is grounded in a spatially distinct region of the image, representing a unique visual focus.

Semantic Validation. We assess the correctness of the answer relative to its proposed region given a tuple $(I, Q, a^{(i)}, b^{(i)})$. We visually superimpose $b^{(i)}$ onto the original image I and employ Gemini-2.5-Pro (Comanici et al., 2025) as an oracle judge, to determine whether $a^{(i)}$ is a valid response to Q when conditioned specifically on the highlighted region.

Based on the oracle's verdict and the match score $R^{(i)}$, we classify each answer into one of three categories:

- **Invalid**: The answer is judged as incorrect to the specified region.
- **New-Valid**: The answer is judged as correct and satisfies the novelty constraint $R^{(i)} < \tau$.
- **Redundant**: The answer is judged as correct but overlaps significantly with a prior interpretation $R^{(i)} \geq \tau$.

We set the IoU threshold to $\tau = 0.3$ by default, as it provides a robust balance between spatial diversity and grouping near-identical objects. The sensitivity to different values of τ is illustrated in Appendix E. Smaller thresholds (e.g., $\tau = 0.1$) are overly sensitive to minor localization noise introduced by automatic grounding, while larger thresholds risk collapsing visually distinct regions.

2.5 Probing Internal Representations for Latent Ambiguity

To investigate whether a model's internal representations encode visual ambiguity prior to explicit answer generation, we perform a probing analysis on the hidden states. We use the initial prompt as described in the multi-turn probing setting (Section 2.2) to ensure the model is not explicitly alerted to the potential for multiple answers. We also experiment on smaller VLLMs.

For each input pair (Q, I), we extract the hidden vector of the last non-padding token from the final transformer layer and denote the representation as $F(Q, I) \in \mathbb{R}^d$.

On top of this feature, we train a lightweight multilayer perceptron (MLP) F_{mlp} to predict whether the question Q is ambiguous:

$$\hat{y} = F_{\text{mlp}}(F(Q, I)), \quad \hat{y} \in \{0, 1\}. \tag{4}$$

Training data is derived from VQA-Therapy (Chen et al., 2023).

Following the logic in Section 2.4, we label a question as ambiguous ($\hat{y} = 1$) if the ground-truth annotations contain multiple distinct answer regions with a low region match score R. Conversely, questions with a single, consistent focal region are labeled as non-ambiguous ($\hat{y} = 0$).

Implementation details are shown in Appendix F.

2.6 Gated Ambiguity-Aware Answering

To leverage the model's internal awareness of ambiguity during inference, we introduce a gated answering framework. The trained MLP probe (Section 2.5) serves as an inference-time router to determine the optimal prompting strategy based on

the latent features of the input (Q, I). For each query, we first compute the ambiguity probability $\hat{y} = F_{\text{mlp}}(F(Q, I))$. Based on this prediction, the framework then branches into two distinct execution paths:

- If $\hat{y} = 1$ (predicted ambiguous), we apply the ambiguity-aware probing prompt and obtain multiple candidate answers.
- If $\hat{y} = 0$ (predicted non-ambiguous), we fall back to the model's default behavior.

We then evaluate the resulting outputs using the answer categories in Section 2.4.

This gated strategy allows the system to selectively activate multi-answer decoding only when the model's internal states indicate a high likelihood of visual ambiguity.

3 Experiments

Our experiments are mainly conducted on the RAcQUEt-GENERAL dataset, which contains 500 image–question pairs annotated as inherently ambiguous. Following Testoni et al. (2025), the dataset contains no ground truth answers because the questions can be interpreted in multiple ways depending on different groundings. We further use subsets from VQ-FocusAmbiguity and VQA-v2 as supplemental and contrastive testbeds. We do not claim that the automatic judge provides perfect correctness labels; instead, it serves as a consistent reference for comparing relative behaviors across settings. Across these datasets we study the following questions:

- **RQ1**: What are VLLMs' default behaviors tackling ambiguity?
- **RQ2**: Do VLLMs generate more valid answers through iterative questioning?
- **RQ3**: Are VLLMs capable of generation with multiple focuses?

Finally, we ask whether ambiguity is encoded in the model's hidden states, and whether this signal can be used to route between either hinting ambiguity before decoding inside a QA pipeline.

3.1 RQ1: What's VLLMs' default behaviors tackling ambiguity

Table 1 reports the behavior of three strong VLLMs (GPT-5 (OpenAI, 2025), Qwen3-VL-235B-A22B, Claude-4.5-Sonnet (Anthropic, 2025)) on RAcQUEt-GENERAL under a single-pass multi-region (SP-MR), non-ambiguous-style prompt. For each question, we ask the model to answer once and evaluate the output using Gemini as an automatic judge. We also conduct experiments with single-answer + single supporting region (SA+Loc) prompt in Appendix B.

We report three metrics: (i) **agreement**, the proportion of model answers judged correct by Gemini; (ii) **avg.# valid answers**, the average number of distinct answers per question judged as correct; and (iii) **single answer**, the proportion of questions for which the model outputs only one answer.

Although every question in RAcQUEt is constructed to have multiple valid answers, both GPT and Qwen model behave conservatively: they keep high agreement ($\geq$ 0.88) while returning only about 2 valid answers on average and answering with a *single* answer for 24–29% of questions. While the Claude model generates more answers, it slightly decreases agreement and still leaves 6% of ambiguous questions undiscovered. This suggests that, under default decoding, VLLMs systematically under-report the ambiguity present in the data, implicitly committing to one interpretation instead of exposing multiple focuses.

We next ask whether models could have produced more answers if prompted differently, in particular when we explicitly require additional answers over multiple turns.

3.2 RQ2: Do the VLLMs generate more valid answers through iterative questioning

Using the multi-turn answer probing setup in Section 2.2, we first apply the SP-MR prompt (without mentioning ambiguity) to obtain the initial set of answers $S^{(0)}$. At each subsequent step t, the model is given the question, the image, and the full history $S^{(<t)}$, and is instructed not to repeat previous answers or locations. Bounding boxes are obtained via Qwen3-VL-235B-A22B grounding and then evaluated with Gemini.

Figure 2 plots, for each step t, (i) the Gemini agreement rate over all answers at that step and (ii) the average region match R with respect to previous steps. As the step index increases, agreement gradually decreases, while the region match R increases. In other words, later answers tend to be more likely to be incorrect and more likely to overlap with earlier focus regions.

To provide a finer-grained view, we use the answer categories from Section 2.4 and compute, at

model	prompt type	agreement↑	avg. #valid answers↑	single answer↓
GPT-5	SP-MR	0.90	2.1	0.29
Qwen3-VL-235B-A22B	SP-MR	0.88	2.0	0.24
Claude-4-5-sonnet	SP-MR	0.82	2.5	0.06

Table 1: Default single-pass multi-region (SP-MR) behavior of three VLLMs on RAcQUEt-GENERAL. **agreement** is the acceptance rate judged by Gemini; **avg.# valid answers** is the average count of distinct Gemini-accepted answers per question; **single answer** is the proportion of questions where the model outputs only one answer.

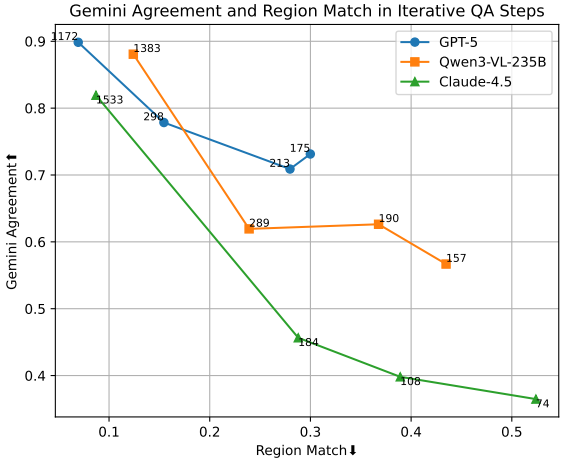


Figure 2: Gemini agreement and region match R across probing steps on RAcQUEt-GENERAL. Points from left to right correspond to increasing probing steps. Higher agreement indicates better answer quality, while higher R indicates that newly generated answers focus on regions already covered by previous steps. Curves closer to the top-left represent a better performance.

each step t, the proportion of *new-valid*, *redundant* and *invalid* answers. The results with are shown in Figure 3 (*new-valid* trends) and Figure 4 (in detail): the fraction of new-valid answers drops quickly after the first step, while redundant and invalid answers become more frequent. This confirms a diminishing-return pattern: iterative probing can indeed elicit additional valid answers beyond the default single-pass decoding, but the marginal gains quickly shrink and are accompanied by more redundancy and hallucination.

3.3 RQ3: Are the VLLMs capable of generation with multiple focuses

The results above show that VLLMs can indeed produce additional valid answers when repeatedly probed, but this behavior only emerges under iterative questioning. We now investigate whether VLLMs are able to generate multi-focus answers *in a single shot* when explicitly instructed that the question may be ambiguous.

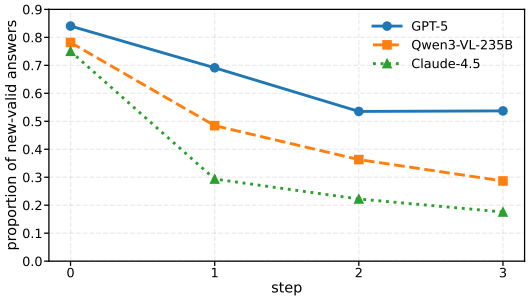


Figure 3: Proportion of *new-valid* answers at each probing step for three VLLMs on RAcQUEt-GENERAL.

To this end, we prepend an ambiguity-aware prefix to the prompt, explicitly stating that the question may admit multiple valid answers corresponding to different image regions, and ask the model to list all such answers in one pass. Table 2 compares this ambiguity-aware multi-focus prompt (m) with the SP-MR prompt (sr) on RAcQUEt-GENERAL. We report three metrics: **agreement**, measured by Gemini acceptance; **Avg. #VA**, the average number of distinct valid answers per question; and **SA**, the proportion of questions answered with a single answer.

Across all three models, ambiguity-aware prompting substantially increases the average number of valid answers and sharply reduces the rate of single-answer outputs. For example, GPT-5 reduces its single-answer rate from 29% to 3%, while increasing Avg. #VA from 2.1 to 2.8. This indicates that once ambiguity is made explicit, VLLMs are capable of exposing more of their latent answer space in a single decoding pass, albeit with a modest decrease in overall agreement.

To better characterize the quality of the generated answers, we further analyze their composition using the region-based categories introduced in Section 2.4. Specifically, we classify answers as *new-valid*, *redundant*, or *invalid* based on Gemini judgment and region overlap. The results are summarized in Table 3.

model	agreement↑	Avg. #VA↑	SA↓
GPT-5	0.90 / 0.87	2.1 / 2.8	0.29 / 0.03
Qwen	0.88 / 0.87	2.0 / 2.6	0.24 / 0.06
C4-5	0.82 / 0.75	2.5 / 2.5	0.06 / 0.01

Table 2: Comparison between SP-MR prompts (sr) and ambiguity-aware multi-focus prompts (underlined m) on RAcQUEt-GENERAL.

method	New.↑	Re.↓	Inv.↓
Total samples: 1172/1858*/1574			
GPT-5-sr	0.84/0.75*	0.06/0.09*	0.10/0.16*
GPT-5-m	0.82	0.05	0.13
Total samples: 1383/2019*/1480			
Qwen-sr	0.78/0.66*	0.10/0.13*	0.12/0.20*
Qwen-m	0.84	0.04	0.12
Total samples: 1533/1899*/1592			
C4-5-sr	0.75/0.65*	0.07/0.09*	0.18/0.26*
C4-5-m	0.73	0.02	0.25

Table 3: Answer composition under different prompting strategies on ambiguous questions from RAcQUEt-GENERAL. **New.**, **Re.**, and **Inv.** denote the proportions of *new-valid*, *redundant*, and *invalid* answers, respectively, defined in Section 2.4. The suffix sr indicates SP-MR prompting, where values before/after the slash correspond to the first step and the final step of iterative probing (*), while m denotes single-shot ambiguity-aware prompting. *Total samples* reports the total number of generated answers for each setting. Across models, ambiguity-aware prompting consistently reduces redundancy and increases the density of new-valid answers compared to iterative SP-MR prompting, at the cost of a moderate increase in invalid answers for some models.

Compared with iterative probing under the SP-MR prompt, ambiguity-aware single-shot generation produces fewer total answers, but a higher proportion of them are new-valid. In particular, ambiguity-aware prompting consistently reduces redundancy across models, suggesting that explicitly signalling ambiguity encourages more diverse focus selection rather than repeatedly exploiting the same salient regions. At the same time, the proportion of invalid answers slightly increases for some models (e.g., GPT-5 and Claude-4.5), reflecting a trade-off between coverage and precision.

Finally, we evaluate the same ambiguity-aware prompt on non-ambiguous questions to assess its potential risks. Table 9 reports results on 100 ambiguous questions from VQA-v2. In this setting, ambiguity-aware prompting brings little benefit and instead increases the invalid rate, indicating a tendency to over-generate speculative answers when ambiguity is absent. This highlights that ambiguity-aware prompting should not be applied indiscriminately, and motivates the need for selectively activating multi-focus generation only when ambiguity is detected.

3.4 Probing ambiguity from VLLMs' hidden state

We now turn to RQ4: *Can a model's internal representations predict whether a question is ambiguous?* Following the procedure described in Section 2.5, we probe the hidden states of VLLMs using a lightweight MLP classifier. Specifically, we extract the last-layer hidden state of the final non-padding token and train an MLP on top of this representation to classify questions as ambiguous or non-ambiguous, using Binary Cross-Entropy (BCE) loss.

Training data is derived from the VQA-Therapy dataset (Chen et al., 2023). While we focus on ambiguity, spatial novelty may serve as a general signal for discovering unexhausted visual evidence in multimodal reasoning. We label a question as *ambiguous* if it contains more than one distinct group of answer-bearing regions, where each group is defined by bounding boxes with mutual region match $R < 0.3$ (Section 2.4). Questions not satisfying this criterion are labeled as non-ambiguous. We use samples from both the training and validation splits, maintaining an ambiguous-to-non-ambiguous ratio of 1:1.5, resulting in 185 training questions in total.

Table 4 compares our probes with Gemini-2.5-Pro used directly as an ambiguity detector, detailed prompt is in Appendix A. On the train split and valid split of VQ-FocusAmbiguity dataset, our Qwen3-VL-30B-A3B classifier achieves an overall accuracy of 0.76, with an ambiguity recall of 0.77 and precision of 0.70, which is comparable to Gemini (0.69 / 0.88 / 0.59). Probes trained on smaller models, including Gemma3-4B-it (Team et al., 2025) and InternVL2-2B (Chen et al., 2024a,b), exhibit similar trends with moderately lower performance.

In addition, our Qwen classifier achieves an accuracy of 0.91 on ambiguous questions (RAcQUEt dataset) and an accuracy of 0.81 on non-ambiguous questions (100 questions from VQA-v2 dataset). These results suggest that ambiguity-related signals are indeed encoded in the internal representations of the VLLM and can be recovered by a simple linear probe, without modifying the base model.

We use Qwen-based probe as a routing mod-

method	Acc.	Amb. R	Amb. P	Amb. F1
Gemini	0.69	0.88	0.59	0.71
Ours-Q	**0.76**	0.77	0.70	**0.73**
Ours-G	0.64	0.88	0.55	0.68
Ours-I	0.65	0.87	0.56	0.68

Table 4: Ambiguity detection performance on VQ-FocusAmbiguity. **Acc.** denotes overall classification accuracy, while **Amb. R**, **Amb. P**, and **Amb. F1** denote recall, precision, and F1 score on the ambiguous class, respectively. Gemini refers to Gemini-2.5-Pro used directly as an ambiguity detector, while **Ours** denotes MLP probes trained on top of **Q**wen3-VL, **G**emma3 and **I**nternVL2 hidden states.

ule in the next subsection to selectively activate ambiguity-aware decoding.

3.5 Gated Answering

We evaluate a gated ambiguity-aware answering setup as a **controlled intervention** to study how selectively activating multi-answer decoding affects model behavior. Importantly, this mechanism is **not intended as a standalone solution**, but rather as a diagnostic tool that allows us to separate ambiguity recognition from answer generation.

We evaluate the gated ambiguity-aware answering strategy on Qwen3-VL-235B-A22B. To ensure a fair and unbiased assessment, we conduct experiments on both ambiguous questions (RAcQUEt) and non-ambiguous questions (100 questions from VQA-v2).

On ambiguous questions (Table 5), the gated strategy effectively identifies samples that benefit from ambiguity-aware decoding. Compared with the SP-MR prompt, our method substantially reduces the **single-answer** rate, indicating improved recognition of ambiguity, while increasing the proportion of **new-valid** answers. At the same time, it avoids the over-generation behavior observed when ambiguity-aware prompting is applied indiscriminately, maintaining a balanced trade-off between coverage and answer quality.

On non-ambiguous questions (Table 6), the gated strategy behaves similarly to the SP-MR prompt. In particular, it prevents unnecessary exploration of multiple answers and keeps the invalid rate close to the baseline. This contrasts with always-on ambiguity-aware prompting, which tends to introduce additional invalid answers on questions with a single correct focus.

Overall, these results show that the gating mechanism mitigates the under-reporting of ambiguity observed in default decoding, while preserving answer quality by selectively combining the strengths of plain and ambiguity-aware prompts rather than relying exclusively on either. Further, we evaluate the gating mechanism in Appendix H.

3.6 Reliability

Our conclusions rely on relative trends, not absolute scores. As our evaluation relies on Gemini-2.5-Pro as an automatic judge of model generations, we conduct a manual audit to assess its reliability. We randomly sample 10% of the questions and corresponding model responses from the RAcQUEt dataset and compare Gemini's verdicts with human annotations. On this random subset, Gemini agrees with human judgment in 90% of the cases.

We further examine the more challenging subset where Gemini labels the model output as *disagree*. On these cases, the agreement rate with human annotations is 82%. This indicates that Gemini may make mistakes on borderline or ambiguous cases; however, its judgments remain sufficiently reliable for analyzing overall trends and comparative behaviors across different methods.

In addition, bounding boxes in our evaluation pipeline are generated by Qwen3-VL based on the location descriptions produced alongside model answers. We randomly sample and manually verify these grounded bounding boxes, observing a 95% agreement rate with human judgment, suggesting that the grounding procedure is reasonably robust.

Notably, samples labeled as *disagree* exhibit a lower agreement rate than the overall average, motivating a closer inspection of failure modes. In the following section, we provide a qualitative error analysis to investigate whether such failures arise from incorrect answers, inaccurate grounding, or misjudgments by the automatic evaluator.

Also, we conduct an additional multi-judge analysis on a sampled subset of ambiguous questions in Appendix I to further assess the robustness of our evaluation pipeline.

3.7 Error Analysis

In our evaluation pipeline, Gemini assesses each model output based on both the generated answer and its associated rationale, i.e., the highlighted bounding box on the image. Accordingly, erroneous cases can be grouped into four categories: *correct answer with wrong rationale* (the answer itself is correct but the highlighted region is insufficient to support the answer), *wrong answer with*

method	Avg. #valid answers ↑	single answer↓	new-valid↑	redundant↓	invalid↓
Qwen-sr	2.2	0.24	0.78	0.10	0.12
Qwen-m	2.5	0.06	0.84	0.04	0.12
Ours	2.4	0.08	0.83	0.04	0.13

Table 5: Gated ambiguity-aware answering on the RAcQUEt dataset. **Qwen-sr** denotes SP-MR prompting, **Qwen-m** denotes always-on ambiguity-aware prompting, and **Ours** denotes the gated strategy. The gated strategy recovers most of the gains in new-valid answers achieved by ambiguity-aware prompting, while avoiding excessive redundancy and maintaining a comparable invalid rate.

method	Avg. #valid answers	single answer↑	new-valid↑	redundant↓	invalid↓
Qwen-sr	0.95	0.92	0.85	0.00	0.15
Qwen-m	0.96	0.84	0.79	0.01	0.20
Ours	0.94	0.90	0.82	0.00	0.18

Table 6: Results on 100 non-ambiguous questions from the VQA-v2 dataset. Metrics follow the same definitions as in Table 5. While always-on ambiguity-aware prompting increases the number of generated answers, it also introduces more invalid outputs. In contrast, the gated strategy behaves similarly to SP-MR prompting, preserving answer quality and preventing unnecessary multi-answer exploration when ambiguity is absent.

correct rationale, *wrong answer with wrong rationale*, and *misjudgment*, where both the answer and the rationale are correct but the automatic judge rejects the output.

In practice, approximately 13% of the inspected cases fall into a borderline category that cannot be reliably classified. These include cases where the judged rationale overlaps other objects (which is inevitable when the judged object is not rectangle), answers involving subjective or vague attributes (e.g., "grey" vs. "dark color"), or images that are blurred, making verification ambiguous. We label these cases as *uncertain* and exclude them from the quantitative breakdown below, while still counting them as errors in previous statistics.

Table 7 summarizes the remaining error cases. Excluding misjudgments, the dominant failure mode is *wrong answer with correct rationale*. This indicates that, in many cases, the model successfully localizes a plausible answer-bearing region, but fails to produce a correct answer grounded in that region. In contrast, errors involving incorrect rationales are relatively rare. We provide error and uncertain annotated examples in Appendix J.

This pattern suggests that current VLLMs often possess reasonably accurate spatial grounding, but underperform at translating localized visual evidence into precise answers—especially under ambiguous or fine-grained conditions.

	a-correct	**a**-incorrect
r-correct	0.23	0.71
r-incorrect	0.03	0.02

Table 7: Error breakdown on 100 manually inspected error cases from the RAcQUEt dataset. Prefix **a** denotes answer correctness and prefix **r** denotes rationale correctness (bounding box grounding). Approximately 13% of samples are labeled as *uncertain* and excluded from this breakdown due to ambiguity or unverifiable visual evidence.

4 Related Work

There has been abundant work on Visual Question Answering. Classical Visual Question Answering (VQA) benchmarks such as VQA (Goyal et al., 2017; Zhang et al., 2016; Antol et al., 2015) collects multiple human crafted answers per question. Several datasets have been collected focusing on disentangling questions themselves as the source of ambiguity (Stengel-Eskin et al., 2023; Chen et al., 2025a; Testoni et al., 2025). In this work, we focus on referential ambiguity, following Testoni et al. (2025)'s work, but our evaluation directly measures the diversity and novelty of answer-supporting visual evidence.

Ambiguity is not only researched in the field of VQA. In text-only setting, Min et al. (2020) introduced a dataset to study ambiguity in open-domain question answering. To tackle ambiguities, Stelmakh et al. (2023) released ASQA dataset. Furthermore, supporting the generation with citations (Gao et al., 2023; Chen et al., 2025c). There has been studies through asking clarification questions (Lee et al., 2023; Chen et al., 2025b), to eliminate the ambiguity through interactions.

To provide better grounding and interpretability, prior VQA studies use segmentation masks and bounding boxes (Chen et al., 2025a, 2023; Das et al., 2016). There also have been works leveraging attention mechanism (Das et al., 2016; Zhang et al., 2018). Current VLLMs such as Qwen3-VL (Bai et al., 2025) have shown remarkable performance on bounding box groundings. A large body of work has been learning internal representation for latent knowledge (Mallen et al., 2024; Maiya et al., 2025).

5 Conclusion

We study visual question ambiguity from a grounded, region-centric perspective. Our analysis shows that VLLMs systematically under-report ambiguity under default decoding, even when multiple valid answer-supporting regions are present. We further find that ambiguity-related signals are encoded in model hidden states and can be recovered with lightweight probes. These findings suggest that ambiguity in VQA is not merely an annotation issue, but a property that models internally recognize yet often fail to express under standard decoding assumptions.

Limitations

Reliance on automatic grounding and evaluation. Our analysis depends on automatically generated bounding boxes and the use of Gemini as an external judge for answer correctness. Although our manual audits suggest reasonably high agreement with human judgment, such automatic pipelines may still introduce noise, especially in borderline cases involving subtle visual distinctions or subjective attributes. Incorporating more extensive human evaluation or alternative grounding models could further strengthen the robustness of the analysis.

Diagnostic nature of gated answering. The proposed gated ambiguity-aware answering mechanism is intended as an analytical tool rather than a fully optimized deployment strategy. Our goal is to study how selectively activating multi-focus decoding affects answer diversity and quality, not to propose a production-ready ambiguity resolution system. Designing end-to-end systems that jointly detect, clarify, and resolve ambiguity remains an open direction for future work.

Lack of user intent and interaction modeling. Our study focuses on ambiguity as a property of the image–question pair and the model's internal behavior, without modeling user intent or interactive clarification. In real-world applications, users may prefer disambiguation through follow-up questions or contextual cues. Integrating grounded ambiguity detection with interactive or user-adaptive frameworks is an important direction for future research.

Acknowledgments

The research is supported by National Science Foundation of China (62376182).

References

Anthropic. 2025. Claude 4.5. https://www.anthropic.com. Accessed: 2026-01-05.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2021. Getting a clue: A method for explaining uncertainty estimates. *Preprint*, arXiv:2006.06848.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023. Vqa therapy: Exploring answer differences by visually grounding answers. *Preprint*, arXiv:2308.11662.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. 2025a. Acknowledging focus ambiguity in visual questions. *Preprint*, arXiv:2501.02201.

Yuchong Chen, Yifan Fan, Yanling Li, Chuyao Ding, and Yu Hong. 2025b. ADP: answer-oriented distinction perception for end-to-end clarification question generation. In *International Joint Conference on Neural Networks, IJCNN 2025, Rome, Italy, June 30 - July 5, 2025*, pages 1–8. IEEE.

Yuhan Chen, Bowei Zou, Yifan Fan, Yuchong Chen, Shujun Cao, and Yu Hong. 2025c. Enhancing attributed question answering using tailored progressive curriculum learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7947–7956, Suzhou, China. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Preprint*, arXiv:1606.03556.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *Preprint*, arXiv:2305.14627.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, Vienna, Austria. Association for Computational Linguistics.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain qa. *Preprint*, arXiv:2305.13808.

Sharan Maiya, Yinhong Liu, Ramit Debnath, and Anna Korhonen. 2025. Improving preference extraction in llms by identifying latent knowledge through classifying probes. *Preprint*, arXiv:2503.17755.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. 2024. Eliciting latent knowledge from quirky language models. *Preprint*, arXiv:2312.01037.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5. `https://openai.com`. Accessed: 2026-01-05.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. Asqa: Factoid questions meet long-form answers. *Preprint*, arXiv:2204.06092.

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2023. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10220–10237, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Alberto Testoni, Barbara Plank, and Raquel Fernández. 2025. RAcQUEt: Unveiling the dangers of overlooked referential ambiguity in visual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23638–23658, Suzhou, China. Association for Computational Linguistics.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2018. Interpretable visual question answering by visual grounding from attention supervision mining. *Preprint*, arXiv:1808.00265.

A Prompt Templates

We list the prompt templates we used in our experiments. All prompts are shown verbatim unless otherwise stated.

A.1 SP-MR Answering Prompt

> You are a vision model that answers a question by locating all relevant regions.

Task: 1. Based on the image and the question, find ALL distinct regions (people / objects) that can answer the question. 2. For each region, output: - answer: the answer to the question - location: describe the location of the region in the image in detail, using natural language - confidence: a float between 0 and 1

Answer ONLY in JSON list format. Example: ["answer": "batter", "location": "detailed description of the location", "confidence": 0.95 , "answer": "catcher", "location": "detailed description of the location", "confidence": 0.88]

Question: question

This prompt is used to measure the model's default single-pass behavior without explicitly signaling ambiguity.

A.2 SA+Loc Answering Prompt

Answer the question based on the image.

Then, identify the single image region that best supports your answer.

Output : - answer: a short answer - location: detailed natural language description of the supporting region - confidence: a float between 0 and 1

Answer ONLY in JSON list format. Example:

["answer": "batter", "location": "detailed description of the location", "confidence": 0.95 ,]

Question: question

This prompt is used to measure the model's default behavior without explicitly signaling multiple supporting regions.

A.3 Multi-turn Answer Probing Prompt

You are a vision model that answers a question by locating all relevant regions. You are refining your previous predictions about relevant regions.

Based on the image and the question, find ALL distinct regions (people / objects) that can answer the question. For each region, output: - answer: the answer to the question - location: describe the location of the region in the image in detail, using natural language - confidence: a float between 0 and 1

The question is: question

You have ALREADY proposed these regions (answer, location): history

Now your task: 1. Propose ONE NEW region that can also answer the question and is semantically different from all previous ones. 2. Or, if there is NO NEW DISTINCT region, answer with an empty list [].

Output format (JSON list, either empty [] or with ONE object): ["answer": "...", "location": "...", "confidence": 0.xx (from 0 to 1)]

At each step, the model is instructed not to repeat previously generated answers or regions.

A.4 Ambiguity-aware Prompt

You are a vision model that answers a question by locating all relevant regions.

The question may be AMBIGUOUS and can have MULTIPLE valid answers, depending on which region of the image you focus on.

Task: 1. Based on the image and the question, find ALL distinct regions (people / objects) that can answer the question. 2. For each region, output: - answer: the answer to the question - location: describe the location of the region in the image in detail, using natural language - confidence: a float between 0 and 1

Answer ONLY in JSON list format. Example: ["answer": "batter", "location": "detailed description of the location", "confidence": 0.95 , "answer": "catcher", "location": "detailed description of the location", "confidence": 0.88]

Question: question

This prompt explicitly signals ambiguity and is used to test whether models can expose multiple focuses in a single pass.

A.5 Prompt used for Qwen Grounding

You are a precise visual grounding module.

Given: - an image - a question about the image - a candidate textual answer - a natural language description of the region that supports this answer

Your task: 1. Locate the SINGLE most appropriate rectangular region in the image that matches the region description and supports the answer to the question. 2. The box should tightly cover the described person/object/region, including the full body or object if possible, but avoid unnecessary background. 3. Always ensure x1 < x2 and y1 < y2. Do NOT change or reinterpret the answer text. Your job is ONLY to find a region that matches the given answer and region description.

Output format (JSON only, no explanation, no extra fields):

```
[
{
"bbox_2d": [x1, y1, x2, y2],
"confidence":
float-between-0-and-1
}
]
```

All coordinates must be normalized to the range [0, 1000].

If the description is very vague or you are uncertain, still choose the best single region and reflect the uncertainty in the confidence score.

Now process the following:

Question: question Answer: answer Region description: region_text

This prompt is used for instructing Qwen3-VL model before converting a lexical location to a normalized bounding box.

A.6 Prompt for Gemini Judging Answer and Rationale Correctness

You are given a question, an answer, an image where a RED rectangle marks a specific region. You MUST ONLY look INSIDE the red box. Ignore everything outside the box.

Question: "question" Answer: "answer"

Task: 1. Check whether question can be answered based on red box inside the image. 2. Answer in JSON: "is_correct": true/false, "reason": "short explanation"

This prompt is used for instructing Gemini-2.5-pro model before verifying the generated answer with highlighted region on the image as grounding.

A.7 Prompt for Gemini as Ambiguity Dectector

You are given an image and a question about the image.

We say a question is AMBIGUOUS if, based on the image alone, there are multiple different, reasonable answers depending on which person/object/region in the image the question refers to. Examples: - "What color is the player's shirt?" when there are several players wearing different colored shirts. - "What is the man holding?" when there are multiple men holding different objects.

We say a question is NOT ambiguous if: - There is only one reasonable interpretation based on the image, OR - The question has a unique correct answer given the image.

Your task: 1. Decide whether the question is AMBIGUOUS or NOT AMBIGUOUS with respect to the image. 2. Briefly explain your reasoning.

Answer ONLY in JSON format: "ambiguous": true/false, "reason": "<short explanation>" Question: question

This prompt is for Gemini-2.5-pro model to detect if a question is ambiguous.

B Analysis on ambiguous questions with SA-Loc prompting

In Table 8, we compare SP-MR prompting and SA-Loc prompting on ambiguous questions from RAcQUEt dataset.

C Analysis on non-ambiguous questions with ambiguity-aware prompting

In Table 9, we compare SP-MR prompting and ambiguity-aware prompting on non-ambiguous questions from VQA-v2 dataset.

model	prompt type	agreement↑	avg. #valid answers↑	single answer↓
GPT-5	SP-MR	0.90	2.1	0.29
GPT-5	SA-Loc	0.93	0.93	1.00
Qwen3-VL-235B-A22B	SP-MR	0.88	2.0	0.24
Qwen3-VL-235B-A22B	SA-Loc	0.94	1.0	0.92

Table 8: Default single-pass multi-region (SP-MR) and single-answer + single supporting region (SA+Loc) behavior of two VLLMs on RAcQUEt-GENERAL.

method	New.↑	Re.↓	Inv.↓	samples↓
Qwen-sr	0.85	0.00	0.15	112
Qwen-m	0.79	0.01	0.20	122

Table 9: Answer composition on non-ambiguous questions from VQA-v2. We compare SP-MR prompting (sr) and ambiguity-aware prompting (m) using Qwen3-VL-235B-A22B. Unlike the ambiguous setting, ambiguity-aware prompting does not improve answer quality on non-ambiguous questions and instead increases the invalid rate, indicating a tendency to overgenerate speculative answers when ambiguity is absent.

D Additional Analysis on Iterative Probing

In Figure 4, we show detailed models' behavior when being probed iteratively.

E Effects of different thresholds τ

We investigate different threshold τ related with dividing *new-valid* and *redundant* answers, which is shown in Figure 5. We select $\tau = 0.3$ as a balanced choice.

F MLP Probe Architecture

The ambiguity probe is implemented as a lightweight MLP with two linear layers. The input is the final-layer hidden state of the last non-padding token. We use a hidden dimension of 256 with ReLU activation, followed by a linear output layer with sigmoid activation. The model is trained using binary cross-entropy loss. We also experimented with near-linear and deeper MLPs, but observed no consistent improvement.

G MLP Probe Robustness: Multi-Seed and Threshold-Free Evaluation

We further evaluate the robustness of the ambiguity probe, addressing concerns regarding limited training data (approximately 185 training questions) and potential sensitivity to threshold selection.

Backbone	ROC-AUC ↑	AP ↑
Gemma3-4B	0.6807 ± 0.0116	0.6383 ± 0.0077
InternVL2-2B	0.7287 ± 0.0034	0.6798 ± 0.0070
Qwen3-VL-30B (Thinking)	0.7888 ± 0.0198	0.7691 ± 0.0102

Table 10: Probe performance across different backbones, averaged over 5 random seeds.

G.1 Threshold-Free Evaluation

Instead of relying on a fixed decision threshold, we report threshold-independent metrics, including ROC-AUC and Average Precision (AP). These metrics assess the separability between ambiguous and non-ambiguous instances without requiring any threshold tuning.

G.2 Multi-Seed and Multi-Backbone Results

We train and evaluate the probe across 5 random seeds (0–4) on multiple backbone models. Results are reported as mean $\pm$ standard deviation.

G.3 Discussion

According to Table 10, across all backbones, the probe achieves ROC-AUC in the range of 0.68–0.79 and AP in the range of 0.64–0.77, indicating a consistent separation between ambiguous and non-ambiguous instances at the representation level.

Importantly, the use of threshold-free metrics ensures that these results are not an artifact of decision threshold tuning. The relatively low variance across seeds further suggests that the probe is stable despite the limited training set.

Together, these findings support our claim that ambiguity-related signals are reliably encoded in the model's final-layer representations. Therefore, the observed failure to express ambiguity in model outputs is unlikely to stem from missing internal information, but rather from decoding and evaluation conventions.

H Gating Validation: Routing Errors and End-to-End Impact

We evaluate the gating mechanism from two complementary perspectives: (i) routing quality under

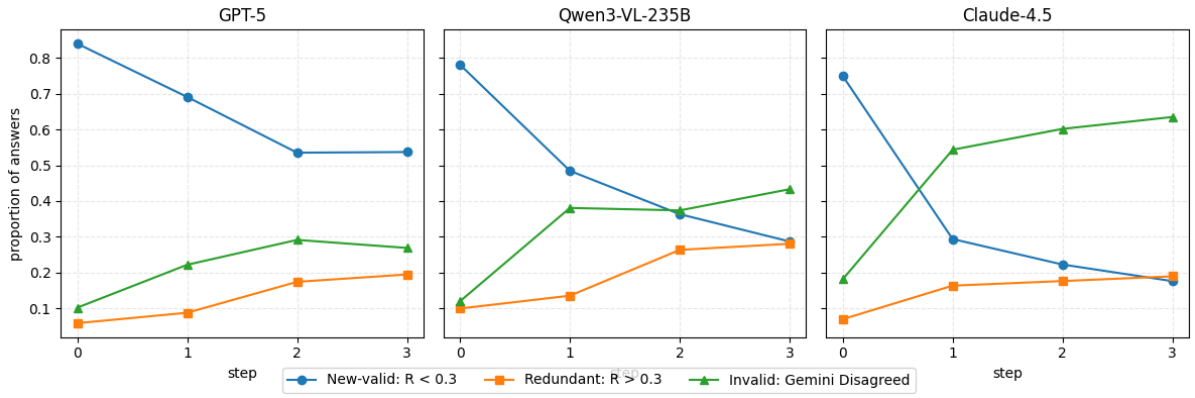


Figure 4: Proportion of *new-valid*, *redundant* and *invalid* answers at each probing step for three VLLMs on RAcQUEt-GENERAL. New-valid answers are Gemini-accepted and have low region match R; redundant answers are accepted but overlap heavily with previous regions; invalid answers are rejected by Gemini.

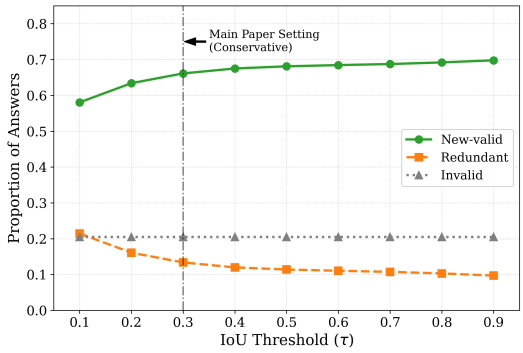


Figure 5: The proportion of new-valid answers remains consistently high (>58%) even under strict overlap penalties ($\tau = 0.1$), confirming that the generated multi-focus answers correspond to distinct visual regions. The threshold $\tau = 0.3$ (marked) used in our main analysis represents a conservative choice to balance distinctiveness and localization noise.

controlled false positive rates, and (ii) end-to-end impact on ambiguous question answering.

H.1 Routing Operating Points

We first analyze routing performance on a mixed set of ambiguous and non-ambiguous questions. Results are averaged over 5 random seeds (0–4).

- OP @ FPR = 5%: TPR = 0.293 ± 0.063, Precision = 0.810 ± 0.031, Trigger rate = 0.154 ± 0.027
- OP @ FPR = 10%: TPR = 0.523 ± 0.076, Precision = 0.795 ± 0.023, Trigger rate = 0.281 ± 0.033

These operating points quantify routing errors under explicit false-positive constraints, capturing the trade-off between detecting ambiguous cases (TPR) and avoiding over-triggering on non-ambiguous inputs (FPR).

Method	Avg. #New-Valid ↑	Single-Answer ↓	Agreement ↑
SP-MR (plain)	2.162	0.236	0.881
AA (always)	2.478	0.060	0.876
Gated ($t = 0.8398$)	2.338	0.166	0.874
Gated ($t = 0.7972$)	2.396	0.122	0.873

Table 11: End-to-end comparison of decoding strategies on RACQUET.

H.2 End-to-End Gated Decoding on RACQUET

We further evaluate the end-to-end impact of gating on the RACQUET dataset (500 ambiguous questions, $\tau = 0.3$, seed = 0).

Since RACQUET contains only ambiguous instances (i.e., no negative samples), metrics such as FPR and precision are not meaningful in this setting. Instead, we focus on coverage and diversity effects under different decoding strategies.

The corresponding trigger rates on RACQUET are 0.664 for $t = 0.8398$ and 0.806 for $t = 0.7972$.

H.3 Discussion

Gating provides a controllable trade-off between conservative and ambiguity-aware decoding. Lowering the threshold increases the trigger rate ($0.664 \rightarrow 0.806$ on RACQUET), routing more cases to ambiguity-aware decoding. As a result, performance smoothly transitions from SP-MR toward always-on ambiguity-aware prompting, yielding higher answer diversity (more new-valid answers) and fewer single-answer outputs.

Importantly, unlike always-on ambiguity-aware prompting, gating avoids indiscriminate over-generation. The FPR-controlled operating

points reported above provide a complementary, deployment-relevant view of routing error risk on non-ambiguous inputs.

I Multi-Judge Evaluation Analysis

To further assess the robustness of our evaluation pipeline, we conduct an additional multi-judge analysis on a sampled subset of ambiguous questions. In addition to the original Gemini-2.5-Pro judge, we re-evaluate the same model outputs using GPT-5.2 and Gemini-3-Flash as independent judges.

I.1 Cross-Judge Agreement

We first measure pairwise agreement between different judges:

- Gemini-2.5-Pro vs. Gemini-3-Flash: $\sim$92% agreement ($\kappa \approx 0.70$)
- Gemini-2.5-Pro vs. GPT-5.2: $\sim$74–80% agreement ($\kappa \approx 0.30$–0.45)

Gemini-3-Flash exhibits strong consistency with Gemini-2.5-Pro. In contrast, GPT-5.2 is systematically stricter, leading to a lower overall acceptance rate. Nevertheless, cross-judge agreement remains substantial across all pairs, indicating that the evaluation signal is largely stable.

I.2 Robustness of Qualitative Trends

We further verify whether our main behavioral findings hold under a stricter judge (GPT-5.2). The key trends remain unchanged:

Average number of valid answers per question

- Plain: 2.32
- Ambiguity-aware: 2.44

First-step single-answer rate

- Plain: 0.18
- Ambiguity-aware: 0.06

Answer composition (Plain $\rightarrow$ Ambiguity-aware)

- New-valid: 0.673 $\rightarrow$ 0.661
- Redundant: 0.071 $\rightarrow$ 0.053
- Invalid: 0.256 $\rightarrow$ 0.287

Question:Which direction is the airplane heading? response
Answer: right

judge
is_correct: False,
reason: The airplane inside the red box is clearly heading towards the left. The nose of the plane is on the left side and the tail is on the right.

Figure 6: An incorrect answer with correct rationale example: the plane is heading left.

These results show that ambiguity-aware prompting consistently increases answer diversity and significantly reduces single-answer commitment. At the same time, it introduces a moderate increase in invalid answers, reflecting the coverage–precision trade-off discussed in the main paper.

I.3 Discussion

While absolute acceptance rates vary across judges—particularly due to the stricter behavior of GPT-5.2—the qualitative trends remain consistent. This suggests that our conclusions are not tied to a specific automatic evaluator, but instead reflect robust behavioral patterns of the underlying models.

J Error and Uncertain Examples

From Figure 6 to Figure 13, we provide three types of errors made by tested models, as well as some samples labeled as uncertain and misjudged by our annotators.

K Annotation Recruitment, Payment and Instructions

We recruit 3 PhD students with fluent level of English, and pay $3 for each valid annotation.

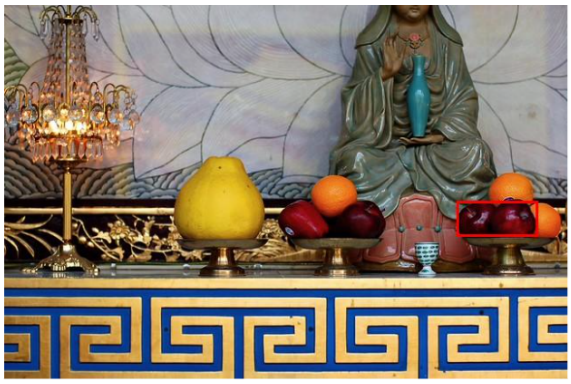
K.1 Evaluating Gemini's Answer–Region Judgment

Task Name Answer Correctness Given a Highlighted Region 1 Task Overview In this task, you will be shown: 1) an image

Question: What's the number on the train? response
Answer: 395 026

is_correct: False, judge
reason: The red box is highlighting an empty area on the side of the train. The number \'395 026\' is located directly below the red box, not inside it.

Figure 7: A correct answer with incorrect rationale example: the rationale–bounding box mismatches the answer.



Question: What color is the fruit? response
Answer: purple

is_correct: False, judge
reason: The fruit inside the red box is dark red, not purple.

Figure 9: An uncertain example: both dark red and purple can be right.

Question: What is the person wearing? response
Answer: white collared shirt and dark pants

is_correct: False, judge
reason: While there is a person wearing a white collared shirt inside the red box, their pants are not visible. Therefore, it is not possible to confirm they are wearing dark pants.

Figure 8: An incorrect answer with incorrect rationale example: the answer can be incorrect, and the rationale is inaccurate.

Question: What color is the couch? response
Answer: red

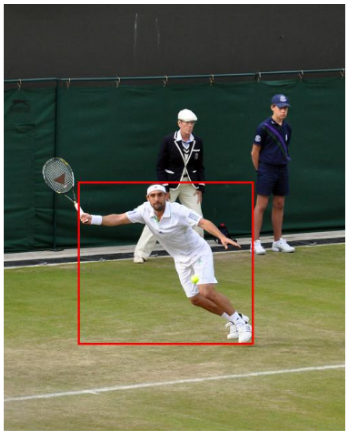
is_correct: False, judge
reason: The red box encloses a single-seater armchair, not a couch. Therefore, the question about the color of the couch cannot be answered by only looking inside the red box

Figure 10: An uncertain example: the boxed object can be defined as both couch and armchair.

Question: What does the sign say? response
Answer: KRA

is_correct: False, judge
reason: The text inside the red box reads "Kro", not "KRA". The third letter is clearly an \'o\'.

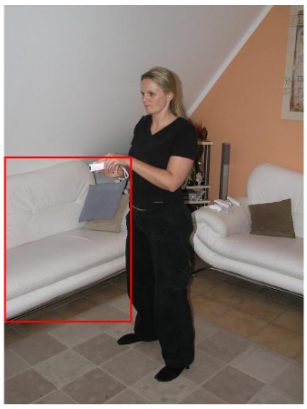
Figure 11: An uncertain example: hard to verify whether it's Kro or Kra.



Question: Does he wear a hat? response
Answer: yes

is_correct: True, judge
reason: The man inside the red box is clearly wearing a white cap, which is a type of hat.

Figure 12: A misjudged example: the player is not wearing a hat, but a headband.



Question: Where is the couch? response
Answer: couch

is_correct: True, judge
reason: The red box clearly contains a white couch, which is the object the question is asking about.

Figure 13: A misjudged example: the answer is not sufficient to represent the location of the couch.

2) a question about the image 3) a candidate answer 4) a highlighted rectangular region (bounding box)

The highlighted region corresponds to the location used by the model when producing the answer. Your task is to judge whether the answer is correct with respect to the highlighted region.

2 Core Principle You must evaluate the answer using ONLY the information inside the highlighted region. Ignore everything outside the red box. Do not use global image context. Do not assume model intent.

3 Labels You must assign one of the following three labels:

1) Correct

Choose Correct if: The highlighted region contains sufficient visual evidence to support the answer A reasonable human observer would agree that the answer follows from this region Minor wording differences are acceptable (e.g., grey vs dark grey)

Examples Region shows a player wearing a red jersey → Answer: "red" Region shows a man holding a tennis racket → Answer: "racket"

2) Incorrect

Choose Incorrect if: The highlighted re-

gion does not support the answer The answer refers to something outside the region The region is misplaced or too incomplete to justify the answer

Examples Region highlights Player A, answer describes Player B Region shows shoes, answer refers to a hat Region is background only

3) Uncertain

Choose Uncertain if: Visual evidence is genuinely ambiguous or unverifiable The attribute is subjective or borderline (e.g., purple vs dark red) Image quality (blur, occlusion) prevents confident judgment Use this label sparingly. If you can reasonably decide correct vs incorrect, do not choose uncertain.

4 What You Are NOT Judging

Do NOT judge: Whether the question itself is ambiguous Whether other regions could also support valid answers Whether the answer is the best or only answer This task evaluates local grounding correctness only.

K.2 Verifying Qwen's Bounding Box Grounding from Location Descriptions

Task Name Bounding Box Accuracy for Location Descriptions

1 Task Overview In this task, you will be shown: 1) an image 2) a question 3) a candidate answer 4) a natural-language location description 5) a bounding box produced by Qwen The bounding box is automatically generated from the location description. Your task is to judge whether the bounding box correctly matches the location description.

2 What You Are Judging You are evaluating: Does the bounding box accurately cover the region described in the location text? This task focuses on spatial grounding, not answer correctness.

3 Labels You must assign one of the following labels:

Correct Choose Correct if: The box tightly covers the described object or person The described entity is clearly identifiable inside the box Minor extra background is acceptable if unavoidable

Examples "the player wearing a blue jersey on the left" → box covers that player "the couch near the window" → box covers the couch

Incorrect Choose Incorrect if: The box highlights the wrong object or person The described entity is missing or mostly outside the box The box is clearly misaligned or meaningless Examples Description refers to a player, box highlights the referee Description refers to a couch, box highlights the floor

Uncertain Choose Uncertain if: The description itself is vague or underspecified Multiple objects equally fit the description The object is partially visible and hard to localize precisely

4 What You Are NOT Judging Do NOT judge:

Whether the answer is correct Whether the question is ambiguous Whether a better box could be drawn Only judge whether this box reasonably matches the given description.

5 Output Format "label": "correct / incorrect / uncertain", "reason": "Short explanation (1–2 sentences)"