Understanding Consistency Through Internal Representations in Large Vision-Language Models

Anonymous Authors[1]

Abstract

Large Vision-Language Models (LVLMs) have shown strong performance on a wide range of multimodal tasks, yet their reliability, especially consistency, remains imperfect across semantically equivalent inputs. Prior work has evaluated consistency by aggregating model responses from multiple paraphrased or restyled variants, such repeated sampling is computationally expensive, making them difficult to use in real time. In this paper, we consider a different, more efficient, and competitively effective alternative—asking whether consistency can be predicted directly from the model's internal states. Specifically, we introduce *single-pass consistency prediction*, which estimates LVLM's consistency from a single forward pass. Intuitively, consistent examples occupy coherent, in-distribution regions of the representation space, whereas unstable examples exhibit distinctive deviations that can be detected before inconsistency arises in generated tokens. Across several consistency and robustness evaluations, we find that features available from a single forward pass, including hidden-state representations and output logits, contain predictive signals of future model stability. Further systematic analysis—examining layers, components, and token positions—provides insights into where consistency-related information resides within the model. Together, our findings suggest that internal representations provide both a practical, effective mechanism for low-cost consistency monitoring and a useful lens for understanding the internal basis of reliable multimodal reasoning in LVLMs.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Over the past few years, Large Vision-Language Models (LVLMs) such as LLaVA (Liu et al., 2023) and Qwen-VL (Bai et al., 2023) have emerged as practical tools leveraging the advanced natural language capabilities of Large Language Models (LLMs) for visual perception and comprehension. LVLMs have also been deployed for real-world applications that require more nuanced reasoning skills, such as visual reasoning (Kamath et al., 2023; Chen et al., 2024; Pothiraj et al., 2025), medical diagnosis (Tu et al., 2024; Wang et al., 2024), or autonomous driving (Hwang et al., 2024; Tian et al., 2025).

Despite their strong performance across diverse tasks, LVLMs still face important reliability challenges, specifically *hallucinations and inconsistency*. LVLMs sometimes generate hallucinated (incorrect) content that is not well grounded in visual evidence (Liu et al., 2024; Zhou et al., 2024; Jiang et al., 2025a). However, even when outputs are correct, reliability is not guaranteed: LVLMs with relatively strong average task performance have been shown to exhibit substantial *inconsistency* across semantically equivalent inputs (Chou et al., 2025; Jia et al., 2025). For example, as shown in Figure 1, a model may produce the correct answer on the original input but return incorrect outputs when the question is rephrased or the image is transformed. Such *inconsistency* can be particularly concerning in high-stakes applications, where practical deployment also requires knowing when a model's prediction is likely to be unreliable. In a broader sense, inconsistency can also make benchmark evaluations less faithful because an unstable model may appear stronger or weaker depending on which input variant is tested.

Yet despite its importance, consistency remains much harder to measure than standard predictive performance. Current approaches to estimating consistency typically require constructing many semantically equivalent variants of the sample and then aggregating the responses for a consistency score (Rabinovich et al., 2023; Chou et al., 2025). For multimodal models, this process is particularly challenging because equivalence can arise from changes to the text, the image, or their interaction, making variant construction and evaluation more complex than in unimodal settings. Conse-

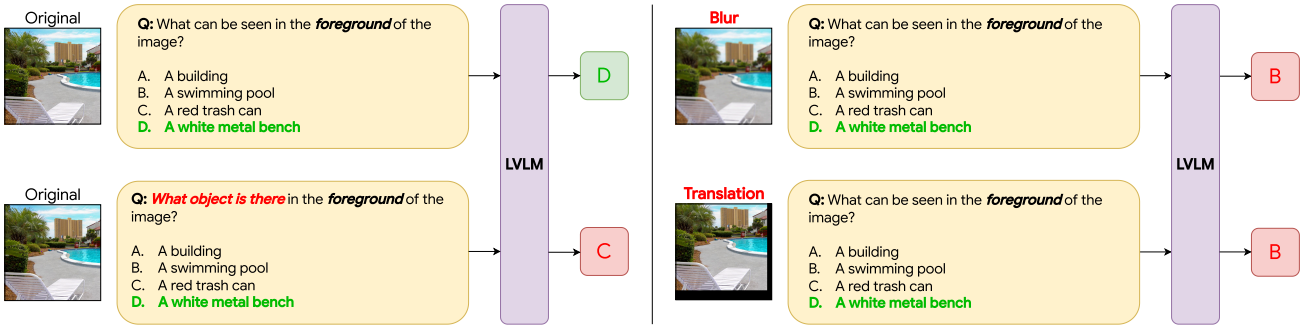


Figure 1. Examples of semantics-preserving perturbations used to test LVLM consistency. Although the correct answer remains unchanged, question rephrasing (bottom left) and image transformations (right) can cause the model to switch predictions.

quently, this substantially escalates the inference cost and may limit the practicality of real-time reliability monitoring.

In this work, we take a different view: instead of asking whether consistency can only be observed after multiple rollouts, we ask whether it is already reflected in the model's internal representations. Specifically, we study whether features extracted from a single forward pass can predict whether an LVLM will remain stable under semantics-preserving multimodal perturbations. This framing shifts consistency evaluation from an external output-sampling procedure to an internal reliability prediction problem. The resulting approach uses only one representative input at inference time, so its efficiency follows naturally from leveraging internal model information rather than repeatedly querying the model on many variants.

The intuition behind this approach is that instance-level consistency depends on where an input lies with respect to the model's learned data distribution. Inputs that fall in coherent, well-represented regions of the representation space are more likely to preserve the same prediction under semantics-preserving perturbations. In contrast, inputs that are far from the training distribution, underrepresented in the training data, or located in low-density regions of the learned representation space may induce less stable internal processing. For these unstable inputs, even minor meaning-preserving perturbations can shift the representation toward a different decision region. Such perturbations may also expose ambiguity in the model's visual-linguistic grounding, consistent with recent findings that multimodal models struggle under distribution shifts and domain-specific inputs (Zhang et al., 2025). Thus, just as hallucinations (Slobodkin et al., 2023; Snyder et al., 2024; Orgad et al., 2025; Kogilathota et al., 2026) or rich global information (Tao et al., 2024; Jiang et al., 2025b) can be decoded from internal states for both LLMs and LVLMs, these internal signals, including intermediate representations and output logits, may likely encode signals related to the ability to produce consistent responses across equivalent inputs.

Building on this idea, we formulate a *single-pass consistency prediction* using features extracted from a single original input. We first construct consistency labels from model responses over multiple text and image variants, and then train lightweight probes to predict these labels using the input's extracted features. We compare probes trained on hidden states with output-based baselines derived from logits, confidence, and token distributions. Our experiments show that hidden-state probes provide a stronger consistency signal than output-level confidence metrics, indicating that instability is often reflected in intermediate representations before it is exposed in the final answer distribution. We further analyze this signal across layers and token positions, showing where consistency-related information is most concentrated inside the model. These results suggest that internal representations offer a practical and interpretable basis for reliability monitoring in LVLMs.

In summary, our contributions are three-fold:

- We introduce *single-pass consistency prediction* for LVLMs, a new reliability prediction setting in which consistency across semantically equivalent multimodal inputs is estimated from only one forward pass.
- We demonstrate that lightweight probes over hidden states can accurately predict consistency and achieve comparative performance with multi-sample agreement estimators using five variants, substantially reducing the cost of consistency assessment.
- We conduct a systematic analysis of the representation space of LVLMs, identifying which layers and token positions carry the strongest consistency signal, thereby shedding light on the internal basis of reliable multimodal reasoning.

2. Related Work

Hallucinations in LVLMs. Despite strong progress in multimodal reasoning, LVLMs remain vulnerable to hal-

lucinations, namely responses that are weakly grounded in visual input (Zhou et al., 2024; Leng et al., 2024; Cao et al., 2024). Prior studies have analyzed object hallucination mechanistically and proposed mitigation strategies through attention intervention (Jiang et al., 2025b), contrastive decoding (Leng et al., 2024) or latent-space control (Duan et al., 2025). These works establish hallucination as a core reliability challenge, but they mainly focus on detecting or reducing ungrounded outputs rather than estimating whether a model will respond consistently across semantically equivalent multimodal inputs.

Consistency and robustness under equivalent inputs. A separate line of research studies consistency as a reliability criterion, asking whether a model gives aligned answers when the input is modified without changing its semantics. In language-only settings, semantic consistency across paraphrased questions has been used both as an evaluation target and as a signal for performance prediction (Rabinovich et al., 2023). In VQA, stable answers across rephrased question neighborhoods correlate with higher accuracy, whereas instability signals uncertainty and unreliable predictions (Khan & Fu, 2024). For LVLMs, this invariance principle extends beyond wording to visual presentation: reliable models should remain stable under meaning-preserving image and text variations (Schmalfuss et al., 2025). However, recent studies show that semantically equivalent multimodal variants can still elicit different LVLM responses (Chou et al., 2025; Jia et al., 2025). Such inconsistency may indicate unstable grounding, where predictions depend on surface textual cues, spurious visual correlations, or fragile multimodal associations.

Existing studies evaluate consistency in LVLMs by constructing multiple equivalent variants of the same instance and then aggregating the resulting responses into a consistency score (Cao et al., 2024; Chou et al., 2025; Jia et al., 2025). Although informative, these methods require repeated inference over many variants and are therefore costly to deploy for real-time monitoring. Our work differs in both objective and setting: instead of estimating consistency by explicitly sampling multiple variants, we ask whether consistency can be predicted from a *single* forward pass on only one input instance.

Internal representations as reliability signals. Recent work on LLM reliability suggests that hidden representations can encode useful information about uncertainty and factuality beyond what is available from output probabilities alone. Probing-based methods have shown that internal states can help distinguish truthful from false statements, improve confidence estimation, and even anticipate hallucination risk before generation (Azaria & Mitchell, 2023; Ji et al., 2024; Beigi et al., 2024; Kossen et al., 2024).

Positioning of our work. Taken together, recent work has either (i) measured consistency by aggregating predictions across multiple input variants, or (ii) used internal representations to predict hallucination, confidence, or truthfulness. Our work connects these two directions by introducing *single-pass consistency prediction*: estimating whether an LVLM will remain stable under semantics-preserving multimodal perturbations from a single forward pass, using lightweight probes over hidden representations.

3. Preliminaries

3.1. Preliminary on Large Vision-Language Models

Notation. In this study, we focus on investigating LVLMs such as the LLaVA series (Liu et al., 2023). These LVLMs typically contain three core components: a vision encoder, a multimodal projector and a pretrained LLM. The input to the LVLM is an image-text pair. The image (I) is divided into n image patches and passed through the vision encoder, yielding n image features. These image features are mapped into the LLM embedding space via the multimodal projector, resulting in d-dimensional vision embeddings $\mathbf{X}_v \in \mathbb{R}^{n\times d}$. The text prompt ($q$) is tokenized into m tokens, which are embedded as $\mathbf{X}_t \in \mathbb{R}^{m\times d}$. For clarity, let $T = n + m$ denote the total input sequence length. The input to the LLM component is obtained by concatenating vision and text embeddings along the sequence dimension: $\mathbf{X} = [\mathbf{X}_v; \mathbf{X}_t] \in \mathbb{R}^{T\times d}$.

A pretrained LLM $\mathcal{M}$ is a traditional Transformer (Vaswani et al., 2017) decoder-only network with L layers. Each layer contains a Multi-Head Attention (MHA) module and a Multi-Layer Perceptron (MLP) module. Let $h_0 = \mathbf{X}$ be the input to the LLM. For each layer $\ell \in \{1, \dots, L\}$, the model takes $h_{\ell-1}$ as input and updates the residual stream, resulting in hidden representations $\mathbf{h}_\ell$. At the last layer, the hidden states of the last token position, denoted by $\mathbf{h}_L^{(T)}$ are projected into the vocabulary space $\mathcal{V}$ with an unembedding matrix $\mathbf{W}_U$ to generate the next token. The model generates text autoregressively, where the probability of the next token is given by:

$$p(x_{T+1} \mid \mathbf{X}) = \mathrm{Softmax}(\mathbf{h}_L^{(T)}\mathbf{W}_U).$$

3.2. Problem Formulation: Consistency Prediction

Let $\mathcal{D} = \{\mathcal{E}_i\}_{i=1}^N$ denote a dataset of N samples with N corresponding *equivalence sets*, where each set $\mathcal{E}_i = \{z_i^{(1)}, z_i^{(2)}, \dots, z_i^{(K_i)}\}$ contains K_i semantically equivalent multimodal inputs. In this work, we define semantically equivalent multimodal prompts to be those that differ only in modality-specific surface form, while preserving the user's intended task and the task-relevant visual content. Equivalence is therefore task-specific: a transformation is

semantics-preserving only if it does not alter any information needed to determine the correct response. Each element $z_i^{(k)} = (I_i^{(k)}, q_i^{(k)})$ is an image-text pair that preserves the same underlying semantic content, but may differ through meaning-preserving transformations such as question paraphrases, or benign image augmentations. By construction, all members of $\mathcal{E}_i$ should induce the same ideal answer. For more details on the image augmentations, refer to Section A.

Given an LVLM $\mathcal{M}$, we denote its predicted answer on the k-th variant by $\hat{y}_i^{(k)} = \mathcal{M}\left(z_i^{(k)}\right)$. We define an instance to be *consistent* if the model produces the same answer for all variants in the equivalence set, and *inconsistent* otherwise. We also consider more relaxed alternatives in Section B, defining a sample as consistent when the model's agreement across all variants is above a specified threshold.

Formally, the binary consistency label is

$$c_i = \mathbf{1}\left[\forall a, b \in \{1, \ldots, K_i\},\ \hat{y}_i^{(a)} = \hat{y}_i^{(b)}\right]. \qquad (1)$$

Here, answer equality is evaluated after task-specific normalization, such as canonicalizing short free-form outputs into a standardized answer format.

Single-pass consistency prediction. In this work, we study whether the consistency label c_i can be inferred from only one forward pass on a single representative input. For each equivalence set E_i, we observe one representative instance $z_i^{(r)} \in E_i$ and run the LVLM once to obtain single-pass information

$$\phi_i = \Phi(\mathcal{M}, z_i^{(r)}),$$

where $\Phi(\cdot)$ denotes features available from this pass, such as hidden states, logits, or confidence-related statistics. The goal of single-pass consistency prediction is to determine, using only ϕ_i, whether the model would produce the same answer across all semantically equivalent variants in E_i. Importantly, the consistency label c_i is computed offline from the full equivalence set, but at inference time only the representative input and its single-pass representation are available.

Scope of this work. Our objective is not to improve the consistency of the LVLM itself, but to *predict* whether it will behave consistently across semantically equivalent inputs using internal signals from a single forward pass. This formulation enables a practical and efficient alternative to exhaustive multi-variant consistency estimation.

3.3. LVLMs Exhibit Substantial Inconsistency

As shown in Figure 1, current LVLMs may produce inconsistent outputs when the prompt or image is perturbed.

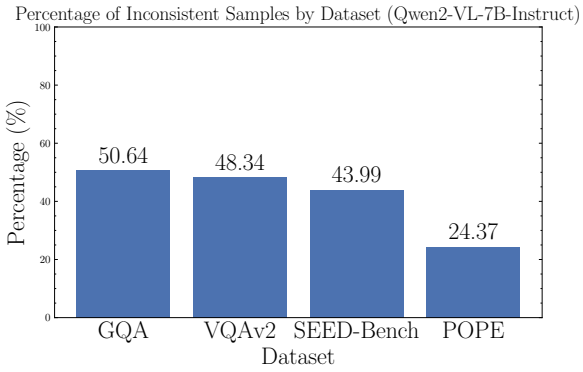


Figure 2. The proportion of *inconsistent* samples varies substantially by dataset (with GQA and VQAv2 having approximately 50% *inconsistent* samples), highlighting the need for efficient consistency monitoring. Refer to Section 3.2 for the definition of *inconsistent* samples.

We quantitatively examine the stability of standard LVLMs when they are given semantically equivalent inputs. Figure 2 depicts the ratio of inconsistent samples following the definition in Eq. 1, showing that inconsistency is widespread in current LVLMs. For Qwen2-VL-7B-Instruct, a substantial fraction of samples are inconsistent across benchmark datasets: 50.64% on GQA, 48.34% on VQAv2, 43.99% on SEED-Bench, and 24.37% on POPE. These results indicate that even strong modern LVLMs can be highly sensitive to benign variations of the same input. Therefore, measuring and predicting consistency is important for both practical deployment and faithful evaluation: a model that changes its answer under equivalent inputs may appear correct on one variant but fail on another. However, directly measuring consistency requires evaluating many variants per instance, which is expensive. This motivates our central question: can we predict, from a single forward pass on the original input, whether an LVLM will remain consistent across equivalent multimodal variants?

4. Methodology

4.1. Consistency Is Encoded in The Representation Space

In this section, we aim to predict the consistency of an LVLM under semantics-preserving perturbations using only a single forward pass on the original input. We hypothesize that the model's sensitivity to input variants is already partially reflected in its computation on the unperturbed image-question pair.

Recent probing work suggests that hidden states in LLMs and LVLMs contain reliability- and semantics-related information, including signals of truthfulness, hallucination risk, and visual grounding (Azaria & Mitchell, 2023; Ji et al., 2024; Orgad et al., 2025; Tao et al., 2024; Kogilathota et al.,

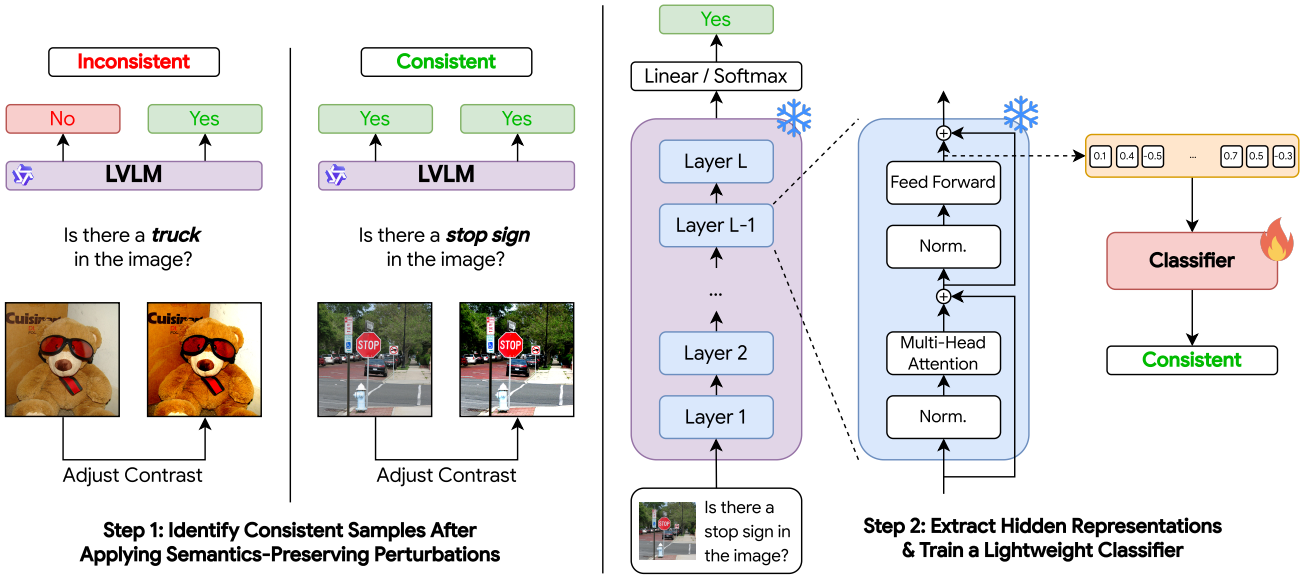


Figure 3. **Overview of single-pass consistency prediction.** We first label samples by applying semantics-preserving perturbations. A sample is labeled *consistent* if the model preserves the same prediction across variants, and *inconsistent* otherwise. At test time, we use hidden representations from one forward pass on the original input to train a lightweight classifier that predicts consistency.

2026). We extend this line of reasoning from hallucination to consistency. If the model changes its answer under benign changes in wording or image appearance, the original input may induce a less stable internal computation: the representation may encode competing visual-textual interpretations, rely on spurious cues that are perturbed by the transformation, or fail to form an invariant mapping from equivalent inputs to the same response. We do not assume that inconsistent examples are globally out-of-distribution. Instead, we hypothesize that they occupy regions of the model's internal space that are more sensitive to semantics-preserving perturbations.

To validate this hypothesis, we examine whether offline consistency labels exhibit observable structure in the representation space using Principal Component Analysis (PCA). Figure 4 provides an initial qualitative motivation for using internal representations as consistency features. When last-token hidden states are projected with PCA, consistent and inconsistent examples form partially separable regions across multiple datasets. This separation suggests that consistency is not only revealed after observing multiple perturbed outputs, but is also partially encoded in the representation produced by the original input itself.

4.2. Detailed Framework

Figure 3 presents our overall framework. Our method has three main stages: offline consistency labeling, single-pass feature extraction, and probe training. First, we evaluate the LVLM on all variants in each equivalence set to obtain consistency labels. Then, for each set, we run the model on only one representative input and extract internal or output-level features. Finally, we train a lightweight classifier to predict the offline consistency label from these single-pass features.

Offline consistency labeling. For each equivalence set $\mathcal{E}_i$, we run the LVLM on all K_i variants and collect the predictions $\{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(K_i)}\}$. We then compute the binary consistency label c_i using the agreement criterion defined in Eq. 1.

This step requires multiple forward passes, but only during dataset construction. The resulting labels are used as supervision for training the consistency predictor.

Single-pass feature extraction. Motivated by this qualitative evidence, we now define the single-pass features used for prediction. For each equivalence set, we select one representative input $z_i^{(r)}$ and run the LVLM once. From this forward pass, we extract features

$$\phi_i = \Phi(\mathcal{M}, z_i^{(r)}).$$

The feature function $\Phi(\cdot)$ may include hidden states from selected Transformer layers, the final-token representation, output logits, token probabilities, entropy, or confidence margins. These features are intended to capture signals that correlate with whether the model's prediction is likely to be stable under semantic-preserving perturbations.

Probe training and inference. We train a lightweight classifier f to predict the consistency label from the single-

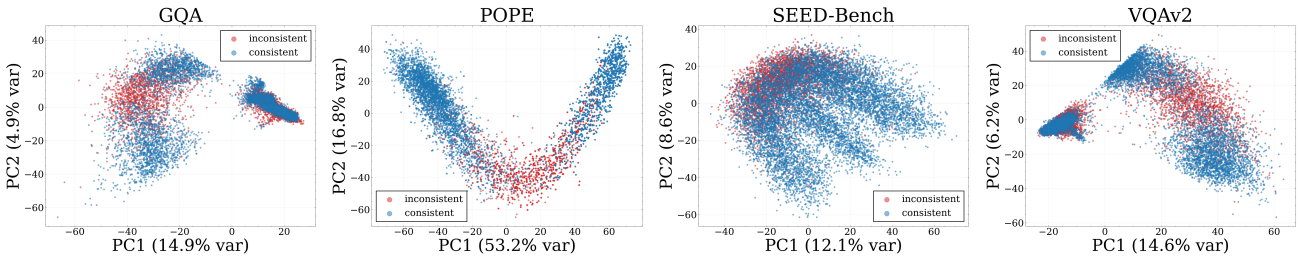


Figure 4. **PCA visualization of Qwen2-VL-7B-Instruct (Layer 23) for consistent and inconsistent samples across datasets**. Each point represents a single original image-question input, colored by its offline consistency label computed from responses to semantics-preserving variants. Across GQA, POPE, SEED-Bench, and VQAv2, consistent and inconsistent examples occupy partially distinct regions of the representation space, suggesting that LVLM hidden states contain signals predictive of future consistency.

pass feature vector:

$$p_i = f(\phi_i),$$

where p_i is the predicted probability that the LVLM will behave consistently across the equivalence set. The probe is optimized with binary cross-entropy loss.

At inference time, the method requires only one image-text input and one LVLM forward pass. The extracted feature vector is passed to the trained probe, and the resulting probability can be thresholded to obtain the predicted consistency label $\hat{c}_i$.

5. Experiments

This section empirically evaluates whether consistency under semantics-preserving multimodal perturbations can be predicted from information available in a single LVLM forward pass. We first describe the experimental setup, including the datasets, models, evaluation metrics, and probing baselines used throughout the study in Section 5.1. We then compare hidden-state probes against output-level metrics (Section 5.3), assess whether the learned consistency signal transfers across datasets (Section 5.2), and analyze the tradeoff between single-pass prediction and multi-sample consistency estimation (Section 5.4).

5.1. Experimental Setup

5.2. Cross-Dataset Generalization

Datasets and models. We evaluate single-pass consistency prediction on four standard multimodal benchmarks: POPE (Li et al., 2023), SEED-Bench (SEED) (Li et al., 2024), VQAv2 (VQA) (Goyal et al., 2017), and GQA (Hudson & Manning, 2019). These datasets cover a range of visual question answering and multimodal reasoning settings, allowing us to test whether consistency-related signals generalize across different task formats and dataset distributions.

We conduct experiments with two widely used open-source LVLMs: Qwen2-VL-7B-Instruct (Bai et al., 2023) and LLaVA-1.5-7B (Liu et al., 2023). For each model and dataset, consistency labels are constructed using the model's own responses to the corresponding equivalence sets. During prediction, however, the classifier only has access to features extracted from a single forward pass on the original input.

Evaluation metrics. We formulate consistency prediction as a binary classification problem, and the proportion of inconsistent samples varies substantially across datasets. Following Orgad et al. (2025), we report the area under the receiver operating characteristic curve (AUROC) as the main evaluation metric.

Classification methods. First, we consider output-based metrics computed from the LVLM's last-token prediction distribution which is available with one forward pass. These include the predicted-answer logit, predicted-answer probability, the logit margin between the top two candidate answers, the corresponding probability margin, and the entropy of the output distribution.

Second, we train lightweight probes on hidden representations extracted from a single forward pass. Unless otherwise stated, we use the hidden state at the last input token as the feature representation. We evaluate three classifiers with increasing nonlinearity: logistic regression as a linear probe, random forest, and histogram gradient boosting (HGB). All classifiers are trained to predict the binary consistency label constructed from the full equivalence set, while using only features from the original input at test time.

5.3. Intermediate Representations Capture Consistency Signals

Table 1 presents the AUROC comparison between output-level features and hidden-state features across two LVLMs and four datasets. Overall, the AUROC scores are high, showing that consistency can often be predicted from a single forward pass. For Qwen2-VL-7B, hidden-state probes

Table 1. AUROC comparison with Qwen2-VL-7B and LLaVA-1.5-7B. Best (blue) and second-best (red) results per dataset.

	Qwen2-VL-7B				LLaVA-1.5-7B			
Method	**POPE**	**SEED**	**VQA**	**GQA**	**POPE**	**SEED**	**VQA**	**GQA**
Logit	0.932	0.845	0.743	0.783	0.706	0.772	0.693	0.715
Probability	0.931	0.871	0.838	0.830	0.718	0.849	0.720	0.850
Logit Margin	0.928	0.872	0.824	0.813	0.718	0.862	0.706	0.846
Probability Margin	0.930	0.872	0.834	0.826	0.718	0.860	0.722	0.865
Entropy	0.933	0.859	0.829	0.816	0.718	0.822	0.701	0.788
Linear	0.950	0.866	0.837	0.802	0.855	0.840	0.789	0.757
Random Forest	0.942	0.860	0.829	0.793	0.853	0.865	0.805	0.791
HGB	0.952	0.872	0.849	0.813	0.853	0.871	0.807	0.789

reach 0.950 on POPE, 0.866 on SEED-Bench, 0.837 on VQAv2, and 0.802 on GQA with a linear probe. Across both LVLMs, probes trained on hidden states are generally competitive with or better than with output-level features. This improvement is especially clear for LLaVA-1.5-7B, where hidden-state probes achieve substantially higher AUROC on most datasets.

Increasing probe complexity provides only modest additional gains. HGB often achieves the highest AUROC, but the linear probe remains close and is sometimes the strongest method. This suggests that consistency-related information is already accessible in the hidden representation and can often be recovered with a simple decision boundary, rather than requiring a highly expressive classifier. Thus, the main performance gain comes from using internal representations, while more complex probes provide only incremental improvements.

Out-of-distribution setting. The previous experiments evaluate whether a probe trained and tested on the same dataset can predict consistency from a single forward pass. To test whether hidden-state probes learn a general reliability signal rather than dataset-specific artifacts, we evaluate *cross-dataset generalization*. For each LVLM, we train the probe on one source dataset and evaluate it directly on a different target dataset, without using target-domain examples for training, validation, normalization, hyperparameter tuning, or threshold selection.

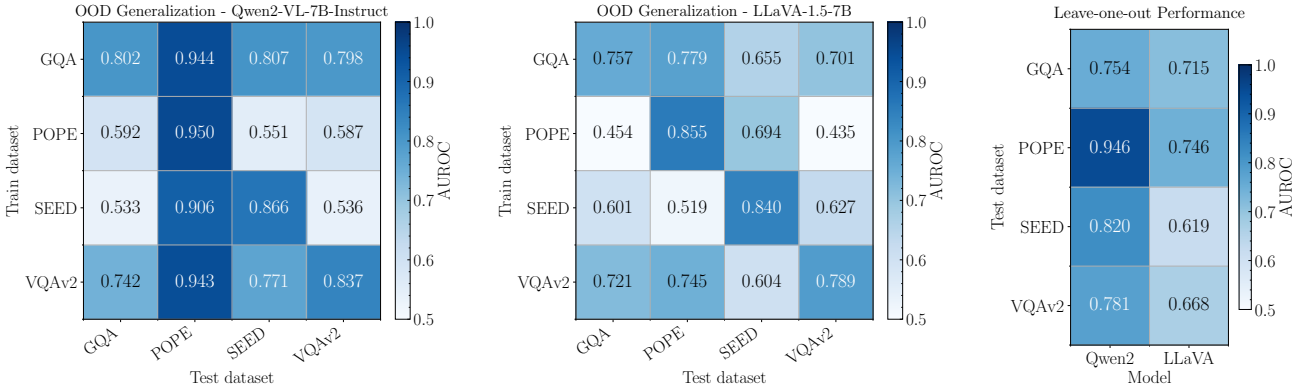
Figures 5a, 5b suggest that hidden-state probes capture a partially transferable consistency signal, but this signal is not fully dataset-invariant. Many off-diagonal AUROCs remain well above random, indicating that a probe trained on one benchmark can still identify unstable predictions on another without target-domain labels. This is strongest for Qwen2-VL-7B, where several source datasets transfer effectively to POPE and, to a lesser extent, to GQA, SEED, and VQA. However, transfer is asymmetric: a dataset can be an effective target without being an effective source, as seen with POPE, whose probes often generalize poorly despite high in-domain performance. LLaVA-1.5-7B shows weaker but still meaningful transfer, suggesting that the strength and universality of the consistency signal may depend on the underlying LVLM. These patterns indicate that hidden representations encode general information about model stability, while also retaining dataset-specific structure that limits zero-shot generalization.

Leave-one-dataset-out setting. We also evaluate a leave-one-dataset-out (LOO) setting, where the probe is trained on all datasets except the target dataset. The results are shown in Figure 5c. Across all held-out datasets, the probe achieves above-random AUROC for both LVLMs, suggesting that the learned consistency signal transfers beyond the datasets used for training.

Qwen2 generalizes strongly, with AUROC ranging from 0.754 on GQA to 0.946 on POPE, while LLaVA shows more moderate transfer, including a drop to 0.619 on SEED-Bench. Overall, these results support the hypothesis that single-pass hidden representations encode a general signal related to output consistency, while also showing that this signal remains sensitive to model architecture and task distribution shift.

5.4. Cost-Quality Trade-off against Multi-Sample Consistency Estimation.

Direct consistency estimation requires running the LVLM on multiple equivalent variants and measuring agreement among the outputs. This is reliable but expensive. Our method instead uses one forward pass, so we evaluate how much prediction quality is retained at much lower inference cost. We compare the single-pass hidden-state probe with partial multi-sample estimators. For each equivalence set $\mathcal{E}_i$, the full consistency label c_i is computed using all K_i variants. We then estimate this label using different inference budgets.



(a) Qwen2-VL-7B *(b)* LLaVA-1.5-7B *(c)* LOO

Figure 5. AUROC comparison for out-of-distribution generalization with a linear probe evaluated on Qwen2-VL-7B, LLaVA-1.5-7B, and the leave-one-dataset-out (LOO) setting.

Table 2. AUROC comparison between agreement of subset of size 2, 3 and 5 with linear probing. Best (blue) and second-best (red) results per dataset. **Linear probe achieves comparable results while being at least 5 times more efficient**.

	Qwen2-VL-7B				LLaVA-1.5-7B			
Subset Size	**POPE**	**SEED**	**VQAv2**	**GQA**	**POPE**	**SEED**	**VQAv2**	**GQA**
$B = 2$	0.654	0.649	0.693	0.685	0.691	0.666	0.675	0.675
$B = 3$	0.748	0.734	0.771	0.764	0.782	0.735	0.752	0.755
$B = 5$	0.836	0.826	0.851	0.839	0.877	0.828	0.845	0.837
Linear	0.950	0.866	0.837	0.802	0.855	0.840	0.789	0.757

The proposed method uses one representative input and one LVLM forward pass. The multi-sample baselines use $B \in \{2, 3, 5\}$ variants and compute agreement among their predicted answers. For a sampled subset $\mathcal{S}_i^B$, we define

$$s_i^B = \frac{1}{B(B-1)} \sum_{a \neq b} \mathbf{1}\left[\hat{y}_i^{(a)} = \hat{y}_i^{(b)}\right].$$

This agreement score is used to predict the full-set consistency label. During inference, we sample B variants without replacement for each sample.

We report AUROC against the number of LVLM forward passes, comparing the single-pass probe with 2-, 3-, 5-variant agreement with Linear classifier.

The results in Table 2 show a favorable cost-quality tradeoff for the single-pass probe. Although partial multi-sample agreement improves steadily as the budget increases from $B = 2$ to $B = 5$, the hidden-state probe already outperforms all $B = 2$ and $B = 3$ baselines using only one LVLM forward pass. It also remains competitive with the much more expensive $B = 5$ estimator, achieving the best AUROC on three of eight model-dataset pairs and the second-best result on the remaining five. This suggests that a single forward pass contains substantial information about whether the model would remain consistent across equivalent variants. In other words, consistency is not only observable after sampling multiple outputs; it is partially encoded in the model's internal representation, making it instance-level consistency monitoring feasible without generating a single token.

6. Conclusion

We introduced *single-pass consistency prediction* for LVLMs: predicting whether a model will remain stable under semantics-preserving multimodal perturbations from only one forward pass. Across two LVLMs and four benchmarks, we showed that hidden representations contain strong consistency signals, well above for both output-based metrics and hidden state probes. Our analyses further show that this signal is partially transferable across datasets, is recoverable with lightweight probes, and is concentrated in specific layers and token positions. Compared with multi-sample consistency estimation, our approach provides a substantially cheaper yet competitive alternative for reliability monitoring. Overall, our findings suggest that LVLM internal states can be used not only to predict consistency efficiently, but also to better understand the internal basis of reliable multimodal reasoning.

References

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, December 2023.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Beigi, M., Shen, Y., Yang, R., Lin, Z., Wang, Q., Mohan, A., He, J., Jin, M., Lu, C.-T., and Huang, L. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12847–12865. Association for Computational Linguistics, November 2024.

Cao, Q., Cheng, J., Liang, X., and Lin, L. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12161–12176. Association for Computational Linguistics, August 2024.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Chou, S.-H., Chandhok, S., Little, J., and Sigal, L. Mm-r3: On (in-) consistency of vision-language models (vlms). In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788, 2025.

Duan, J., Kong, F., Cheng, H., Diffenderfer, J., Kailkhura, B., Sun, L., Zhu, X., Shi, X., and Xu, K. Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7372–7382, October 2025.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Hwang, J.-J., Xu, R., Lin, H., Hung, W.-C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., Zhou, Y., Guo, J., Anguelov, D., and Tan, M. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

Ji, Z., Chen, D., Ishii, E., Cahyawijaya, S., Bang, Y., Wilie, B., and Fung, P. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104. Association for Computational Linguistics, November 2024.

Jia, B., Zhang, J., Zhang, H., and Wan, X. Exploring and evaluating multimodal knowledge reasoning consistency of multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11966–11981, November 2025.

Jiang, N., Kachinthaya, A., Petryk, S., and Gandelsman, Y. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=94kQgWXojH.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., and Yang, X. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025b.

Kamath, A., Hessel, J., and Chang, K.-W. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023.

Khan, Z. and Fu, Y. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 10854–10863, 2024.

Kogilathota, S. A., G, S. V. E., Sun, L., and Zhou, J. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, March 2026.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large

vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, June 2024.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, June 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., and Wen, J.-R. 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.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Pothiraj, A., Stengel-Eskin, E., Cho, J., and Bansal, M. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8001–8010, October 2025.

Rabinovich, E., Ackerman, S., Raz, O., Farchi, E., and Anaby Tavor, A. Predicting question-answering performance of large language models through semantic consistency. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 138–154, December 2023.

Schmalfuss, J., Chang, N., VS, V., Shen, M., Bruhn, A., and Alvarez, J. M. Parc: A quantitative framework uncovering the symmetries within vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25081–25091, 2025.

Slobodkin, A., Goldman, O., Caciularu, A., Dagan, I., and Ravfogel, S. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, December 2023.

Snyder, B., Moisescu, M., and Zafar, M. B. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 2721–2732, 2024.

Tao, M., Huang, Q., Xu, K., Chen, L., Feng, Y., and Zhao, D. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13050–13056, May 2024.

Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., and Zhao, H. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning*, pp. 4698–4726. PMLR, 2025.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. Towards generalist biomedical ai. *NEJM AI*, 1(3): AIoa2300138, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., and Shen, D. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133, 2024.

Zhang, X., Li, J., Chu, W., Hai, J., Xu, R., Yang, Y., Guan, S., Xu, J., Jing, L., and Cui, P. On the out-of-distribution generalization of large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10315–10326, 2025. doi: 10.1109/CVPR52734.2025.00965.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZDJKTlOUe.

A. Constructing Equivalence Sets

We construct equivalence sets to evaluate whether a multimodal model gives consistent responses to inputs that preserve the same underlying meaning. Each equivalence set contains multiple versions of the same example, where the image or question may vary in superficial ways while the intended task and correct answer remain unchanged.

In this work, two inputs are considered equivalent if they preserve the information needed to answer the question. Equivalence is therefore task-dependent: a change is acceptable only when it does not alter the user's intent, remove relevant visual evidence, or change the expected answer. For example, an image modification may be harmless for a general object-recognition question but unsuitable for a question that depends on small text, fine visual details, or precise colors.

The augmentation modes used to create equivalent image variants are summarized in Table 3. These transformations introduce variation in image appearance while preserving the semantic content of the original example.

Augmentation Mode	Description
Original	The unmodified input image.
Blur	A slightly blurred version of the image.
Lighting	A version with altered brightness.
Crop	A cropped version that retains the main visual content.
Translation	A shifted version of the image.
Noise	A version with added random visual noise.
Mask	A version with small randomly masked regions.
Sharpen	A version with enhanced sharpness.
Pixelate	A lower-resolution, pixelated version of the image.
Shot Noise	A version with noise resembling low-light image degradation.
JPEG Compression	A compressed version of the image.
Posterize	A version with reduced color detail.
Elastic Deformation	A mildly warped version of the image.
Perspective Shift	A version with a changed perspective or viewpoint.

Table 3. Augmentation modes used to construct image variants in each equivalence set. Each augmented image is treated as equivalent only when it preserves the task-relevant content and expected answer.

By grouping these variants into equivalence sets, we can measure whether model behavior is stable under benign changes in input form. A robust model should produce the same answer for all members of an equivalence set, since the user's intended task and the correct response remain the same. Figures 6, 7 present some examples of these image augmentation modes for the POPE dataset.

B. Adaptive Thresholding for Defining Consistency

In the main experiments, we use the strict all-agreement criterion in Eq. 1: an equivalence set is labeled consistent only if the LVLM produces exactly the same normalized answer for every variant in the set. This definition is simple and conservative, but it may be overly strict when a model produces a dominant answer for most variants while changing its prediction on only a small number of perturbed inputs. To test whether our conclusions depend on this strict binary definition, we also consider threshold-based alternatives that assign consistency labels from the dispersion of the model's answers within each equivalence set.

For each equivalence set $E_i = \{z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(K_i)}\}$, let $\hat{y}_i^{(k)} = M(z_i^{(k)})$ denote the normalized answer produced by the LVLM on the k-th variant. We define the pairwise consistency score as

$$s_i = \frac{1}{K_i(K_i-1)} \sum_{a \neq b} \mathbf{1}\left[\hat{y}_i^{(a)} = \hat{y}_i^{(b)}\right]. \tag{2}$$

The score $s_i \in [0, 1]$ measures the fraction of ordered variant pairs that receive the same answer. A value of $s_i = 1$ corresponds exactly to the strict all-agreement case, while lower values indicate increasing disagreement across semantically equivalent variants.

Given a threshold γ, we define the thresholded consistency label as

$$c_i^{(\gamma)} = \mathbf{1}[s_i \geq \gamma].$$

Figure 6. Examples of image augmentations for sample 1160 in the POPE dataset. The question for this sample is: *Is there a person in the image?*

Thus, equivalence sets with sufficiently high pairwise consistency are labeled consistent, while those below the threshold are labeled inconsistent. The strict definition used in the main experiments is recovered by setting $\gamma = 1$.

We consider two thresholding strategies. The first uses the median pairwise consistency score over the training equivalence sets:

$$\gamma_{\text{med}} = \text{median}\left(\{s_i\}_{i=1}^{N}\right).$$

The median is used to create balanced binary classes to facilitate the training of classifiers.

Following the procedure in Kossen et al. (2024), we consider an optimal one-dimensional threshold (Opt. Threshold) that separates the training equivalence sets into low-consistency and high-consistency groups. For a candidate threshold γ, define

$$\mathcal{I}_{\text{low}}(\gamma) = \{i : s_i < \gamma\}, \qquad \mathcal{I}_{\text{high}}(\gamma) = \{i : s_i \geq \gamma\}.$$

Let the mean pairwise consistency scores within the two groups be

$$\hat{s}_{\text{low}}(\gamma) = \frac{1}{|\mathcal{I}_{\text{low}}(\gamma)|} \sum_{i \in \mathcal{I}_{\text{low}}(\gamma)} s_i, \qquad \hat{s}_{\text{high}}(\gamma) = \frac{1}{|\mathcal{I}_{\text{high}}(\gamma)|} \sum_{i \in \mathcal{I}_{\text{high}}(\gamma)} s_i.$$

We then choose

$$\gamma^* = \arg\min_{\gamma} \left[\sum_{i \in \mathcal{I}_{\text{low}}(\gamma)} \left(s_i - \hat{s}_{\text{low}}(\gamma)\right)^2 + \sum_{i \in \mathcal{I}_{\text{high}}(\gamma)} \left(s_i - \hat{s}_{\text{high}}(\gamma)\right)^2 \right].$$

This objective selects the split that minimizes the within-group variance of the pairwise consistency scores, producing a data-adaptive separation between low-consistency and high-consistency equivalence sets. The threshold is selected using only the training split and is then applied to the corresponding evaluation split.

Table 4 reports the resulting AUROC scores under the strict all-agreement label, the median threshold, and the optimal threshold. Overall, the results are stable across labeling choices. The threshold-based definitions yield performance comparable to the strict criterion, and in some cases slightly improve AUROC. This indicates that the hidden-state signal studied in the main experiments is not tied to a single hard definition of consistency, but remains predictive under softer pairwise-consistency criteria.

Figure 7. Examples of image augmentations for sample 6533 in the POPE dataset. The question for this sample is: *Is there a hot dog in the image?*

Method	Qwen2-VL-7B				LLaVA-1.5-7B			
	POPE	SEED	VQA	GQA	POPE	SEED	VQA	GQA
All Consistent	0.950	0.866	0.837	0.802	0.855	0.840	0.789	0.757
Median Threshold	0.950	0.866	0.837	0.810	0.924	0.840	0.789	0.761
Opt. Threshold	0.956	0.871	0.840	0.781	0.872	0.839	0.791	0.751

Table 4. Results for Qwen2-VL-7B and LLaVA-1.5-7B across thresholding methods.

C. Implementation Details

C.1. Datasets

We evaluate consistency prediction on four standard multimodal benchmarks: POPE (Li et al., 2023), SEED-Bench (Li et al., 2024), VQAv2 (Goyal et al., 2017), and GQA (Hudson & Manning, 2019). These datasets cover complementary visual question answering and multimodal reasoning settings, allowing us to test whether consistency-related signals are present across different input distributions, answer formats, and reasoning requirements. Unless noted otherwise, the training split for these datasets is 70/15/15 for train, validation and test, respectively.

POPE. POPE is an object-hallucination evaluation benchmark for large vision-language models. Each example asks a yes/no question about whether a particular object is present in the image. Because the task has a constrained binary answer space and directly tests visual grounding, POPE provides a useful setting for studying whether model predictions remain stable under semantics-preserving visual perturbations. We use 9,000 POPE examples.

SEED-Bench. SEED-Bench is a multimodal benchmark designed to evaluate a broad range of image-understanding and reasoning abilities. It contains multiple-choice questions that require models to interpret visual content and select the correct answer from candidate options. Compared with POPE, SEED-Bench has a more diverse task structure and answer space, making it useful for testing whether consistency prediction extends beyond binary object-presence questions. We only use single-image samples from SEED-Bench, totaling 14,233 examples.

VQAv2. VQAv2 is a widely used visual question answering benchmark containing natural-language questions about images. The questions cover diverse visual concepts, including objects, attributes, counting, spatial relations, and scene understanding. We include VQAv2 to evaluate consistency prediction in an open-ended VQA setting with varied question types and short textual answers. We sample 20,000 VQAv2 examples from the *validation* set. The training split for VQAv2 is 10000/5000/5000 for train, validation and test, respectively.

GQA. GQA is a visual reasoning dataset built around compositional questions over real-world images. Its questions often require reasoning over objects, attributes, relations, and scene structure, making it a challenging benchmark for evaluating multimodal reasoning consistency. We use GQA to test whether internal representations can predict stability on examples that require more structured visual-linguistic reasoning. We use the *testdev_balanced* set, containing 12,578 GQA examples.

For each dataset, every original image-question pair is treated as one base sample. We construct an equivalence set for each base sample by applying the semantics-preserving image transformations described in Appendix A, including blur, lighting changes, cropping, translation, noise, masking, sharpening, pixelation, shot noise, JPEG compression, posterization, elastic deformation, and perspective shift. The consistency label for each base sample is computed from the LVLM's predictions over all variants in its equivalence set, while the consistency predictor only observes features from a single representative input at test time.

C.2. Classifiers

In this section, we provide technical details on the hyperparameters and setups used from the methods used in the paper as follows:

- Linear: We use LogisticRegression from the *scikit-learn* library. We use L1 regularization with $C = 0.01$ and $max\_iter = 1000$.
- Random Forest: We use RandomForestClassifier from the *scikit-learn* library. All settings are set to default.
- HGB: We use HistGradientBoostingClassifier from the *scikit-learn* library. All settings are set to default.

  All classifiers are set with $random\_state = 42$. We train the classifier on all layers, and select the layer with the highest AUROC score on the validation set. For intermediate features, we utilize outputs of the MLP module unless noted otherwise.

All experiments are conducted wit RTX A5000 GPUs. For all probing experiments, we utilize the Linear classifier unless noted otherwise.

D. Ablation Studies

D.1. Where is the Consistency Signal Encoded?

The main results show that hidden states are predictive of consistency, but they do not show where this signal appears inside the LVLM. We therefore analyze the effect of model component, layer depth, and token position.

Component analysis. We compare probes trained on three representation types: MLP outputs, MHA outputs, and hidden states. Figure 8 reports the results for Qwen2-VL-7B and LLaVA-1.5-7B across the four datasets. Overall, hidden states provide the strongest and most stable consistency signal. Probes trained on hidden states are consistently competitive across datasets and models, while probes trained only on MHA or MLP outputs are generally weaker or less stable.

This suggests that consistency-related information is most accessible in the integrated residual representation, rather than being isolated in a single Transformer submodule. MHA and MLP outputs may each contain partial information about stability, but the hidden state combines information from attention, feed-forward computation, and residual accumulation. This makes it a more reliable representation for predicting whether the model's answer will remain stable under equivalent perturbations.

We also observe that increasing probe complexity provides limited additional benefit. Random forest and HGB probes sometimes improve over the linear probe, but the gains are modest and not consistent across datasets. This supports the

conclusion that consistency information is often linearly accessible from hidden states, rather than requiring a highly expressive nonlinear classifier.

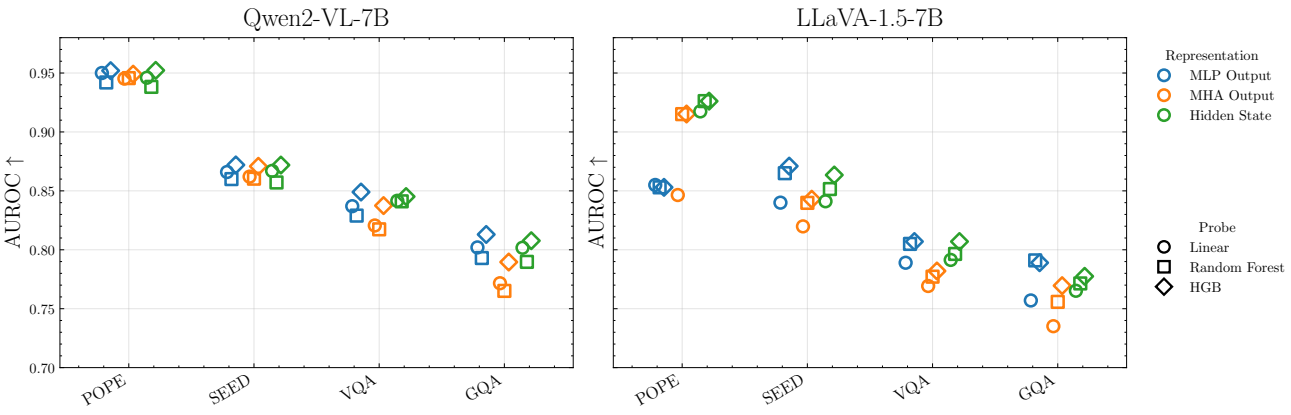


Figure 8. Component-level ablation of consistency prediction across representation types and probe classifiers. We compare probes trained on MLP outputs, MHA outputs, and hidden states for Qwen2-VL-7B and LLaVA-1.5-7B across four datasets. Hidden states generally provide the strongest or most competitive consistency signal, while more expressive classifiers yield only modest gains over the linear probe.

Layer analysis. We examine how consistency-prediction performance changes across layers. Figure 9 reports AUROC when the probe is trained on the hidden state from each individual layer. Across datasets, early layers generally provide weaker consistency signals, while performance increases substantially in the middle layers and remains strong through the later layers. This trend is especially clear for SEED-Bench, GQA, and VQAv2, where AUROC rises steadily from the early layers and peaks around the mid-to-late layers. POPE shows a relatively strong signal even at the embedding layer, but after an early drop it follows the same pattern of improving toward the middle and late layers.

These results suggest that consistency information becomes more accessible after the model has integrated visual and textual evidence through several Transformer blocks. The signal is not confined to the final layer; instead, it emerges in the middle layers and remains available across later layers. This supports the view that consistency is encoded as part of the model's internal multimodal reasoning process rather than appearing only in the final output distribution.

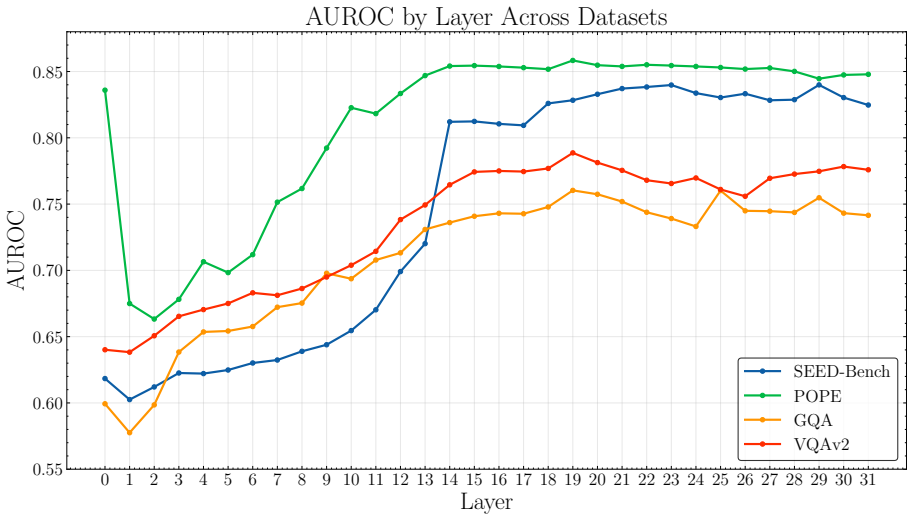


Figure 9. Layer-wise consistency prediction performance. We train separate probes on hidden states from each layer and report AUROC across datasets. Consistency signals are generally weak in early layers, increase substantially in the middle layers, and remain strong in later layers, suggesting that consistency-related signals emerge after multimodal information has been integrated through the Transformer stack.

Token-position analysis. We next compare different token positions for feature extraction. Table 5 shows that the last query token substantially outperforms the last image token across all datasets. The last image token performs close to chance, with AUROC scores between 0.513 and 0.571, indicating that image-token representations alone do not provide sufficient information for consistency prediction.

In contrast, the last query token achieves the best performance on every dataset, with AUROC scores of 0.950 on POPE, 0.866 on SEED-Bench, 0.837 on VQAv2, and 0.802 on GQA. The second-last answer token is also highly predictive, but is consistently slightly worse than the last query token. These results suggest that the consistency signal is strongest immediately before generation, after the model has processed both the image and the full question. At this point, the representation reflects the model's integrated multimodal interpretation and is therefore most informative about whether the subsequent answer is likely to be stable.

Overall, the ablation results indicate that consistency information is primarily encoded in the integrated hidden state at late text positions, rather than in isolated image tokens or individual Transformer submodule outputs. This supports our use of the last-query-token hidden state as the default feature for single-pass consistency prediction.

Table 5. Results using different token positions.

Token position	POPE	SEED	VQA	GQA
Last Image Token	0.513	0.571	0.558	0.514
Second Last Answer Token	0.945	0.846	0.830	0.784
Last Query Token	0.950	0.866	0.837	0.802

D.2. Sample Efficiency

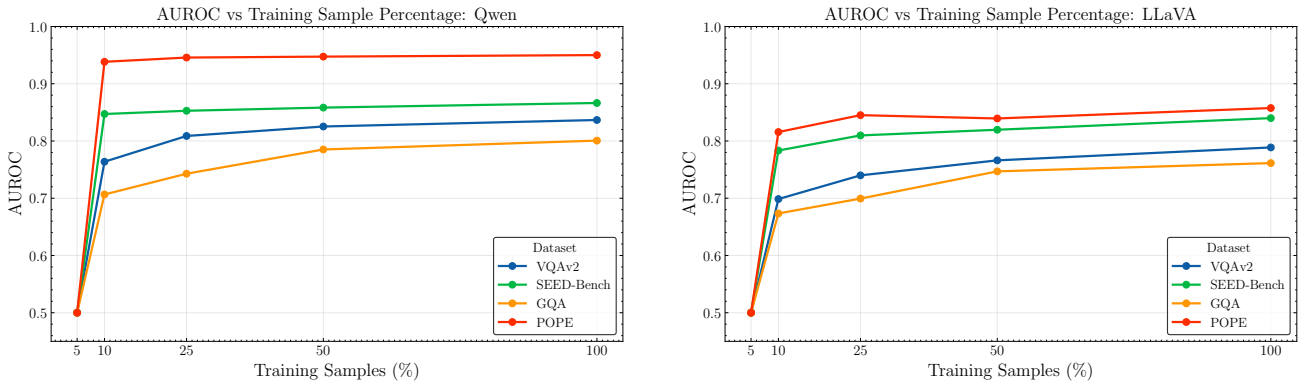


Figure 10. **Sample efficiency of consistency probing.** AUROC improves rapidly with small training fractions and then saturates, indicating that hidden-state probes require limited supervision to recover most consistency-prediction performance.

To evaluate whether the consistency signal requires a large amount of supervision, we train the probe using only a fraction of the available training equivalence sets and report test AUROC on each dataset. As shown in Fig. 10, performance improves sharply once the training fraction increases from 5% to 10%, and then grows only gradually as more data is added. This trend is consistent across both Qwen2-VL-7B and LLaVA-1.5-7B.

The results indicate that the probe is highly sample-efficient. With only 10% of the training data, the probe already recovers most of its final performance, especially on POPE and SEED-Bench. Increasing the training set to 50% or 100% gives only modest additional gains, suggesting that the learned consistency direction is relatively simple and does not require extensive supervision. GQA and VQAv2 benefit more from additional data, indicating that their consistency signals are weaker or more heterogeneous.

D.3. Representative-Choice Robustness

Single-pass prediction assumes that one representative input is enough to estimate set-level consistency. We test whether probe performance depends strongly on which equivalent variant is chosen.

For each equivalence set, we treat each variant as the representative input in turn. For variant $z_i^{(k)}$, the probe prediction is $p_i^{(k)} = f\left(\Phi(\mathcal{M}, z_i^{(k)})\right)$. The ground-truth consistency label c_i remains fixed for all variants in the set.

Table 6. AUROC under different representative inputs. Values for variants report the change in test AUROC relative to the original setting. The last column reports the mean and standard deviation over absolute AUROC values across the original and variant inputs.

Dataset	Original	Blur Δ	Lighting Δ	Noise Δ	Posterize Δ	Mean ± Std.
GQA	0.802	-0.024	-0.006	+0.000	-0.004	0.795 ± 0.009
POPE	0.950	-0.044	-0.008	-0.004	-0.006	0.938 ± 0.016
SEED-Bench	0.866	-0.004	-0.002	-0.011	-0.003	0.862 ± 0.004
VQAv2	0.837	-0.008	+0.004	+0.001	+0.000	0.836 ± 0.004
Average	**0.864**	**-0.020**	**-0.003**	**-0.004**	**-0.003**	$\mathbf{0.858 \pm 0.007}$

Table 6 reports the AUROC scores of using different variants as representative inputs for the linear probe. AUROC remains stable across lighting, noise, and posterization, with average drops of only 0.003–0.004 relative to the original input. Blur is the most damaging variant, reducing average AUROC by only 0.02, with the largest degradation on POPE. These results suggest that consistency-related information is not tied to a particular clean representative input and remains accessible under several semantics-preserving visual transformations.

D.4. Accuracy-Consistency Disentanglement

Accuracy and consistency are related but distinct aspects of reliability. A model may answer the representative input correctly while still changing its prediction under semantically equivalent perturbations. Conversely, a model may produce the same answer across variants while being consistently wrong. Therefore, strong consistency-prediction performance could in principle arise from predicting correctness on the original input rather than consistency across the equivalence set.

To test this possibility, we evaluate the probe on a correctness-filtered subset containing only examples where the LVLM answers the representative input correctly. In this setting, the probe must distinguish between examples that are both correct and consistent and examples that are correct but unstable under equivalent variants.

The correctness-filtered results (Figure 11) show that the probe is not merely predicting whether the LVLM answers the representative input correctly. After restricting evaluation to examples where the representative prediction is already correct, AUROC remains well above chance across all datasets and both models. For Qwen2-VL-7B, performance decreases only modestly, e.g., from 0.950 to 0.946 on POPE and from 0.837 to 0.814 on VQA. For LLaVA-1.5-7B, the filtered results are also close to the original results, and even slightly improve on SEED and GQA. This suggests that correctness and consistency are related but distinct: even among examples the model answers correctly, hidden representations still contain information about whether the answer is stable under equivalent perturbations. Therefore, the probe captures a consistency-specific reliability signal beyond standard accuracy on the original input.

E. Results of Equivalence Sets of Only Augmentations against Both Augmentations and Paraphrases

In this paper, we solely focus on image augmentations because VQA consistency failures are primarily about whether the LVLM's visual grounding remains stable under benign changes in visual presentation. Question paraphrases are useful, but they introduce an additional source of variation: paraphrase quality, syntactic ambiguity, and possible changes in linguistic priors. In contrast, image augmentations keep the question fixed and isolate sensitivity to the visual input. Since the goal is to predict whether the model's answer is stable under semantics-preserving perturbations, image-only equivalence sets are a clean and scalable proxy. In addition, our focus is on single-pass consistency prediction, not exhaustive consistency evaluation. The equivalence set is used to construct supervision offline; it does not need to enumerate every possible semantic perturbation. However, it needs to be informative enough that the resulting labels support the same probe conclusions.

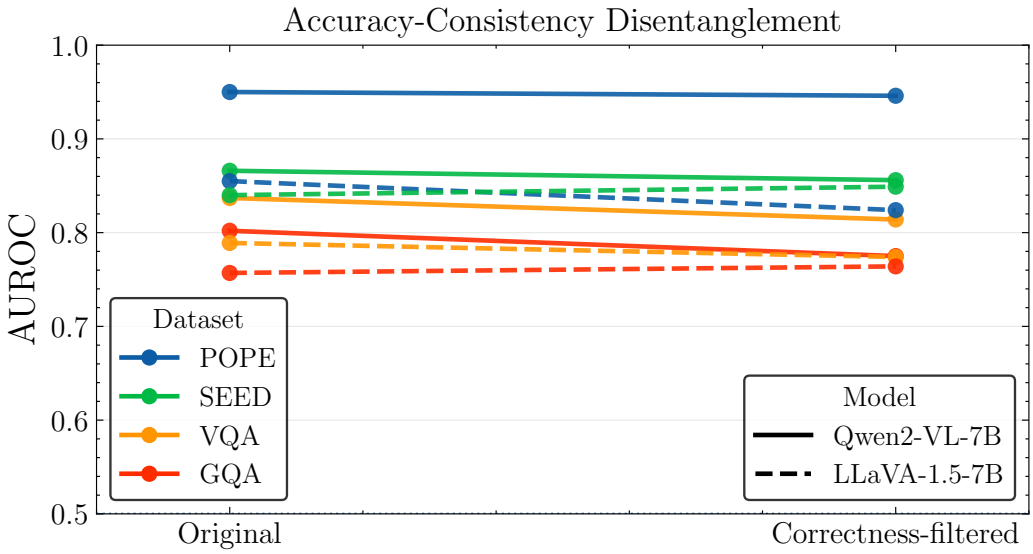


Figure 11. Accuracy-consistency disentanglement analysis. We compare probe performance on the full test set and on the subset of examples where the LVLM answers the representative input correctly. AUROC remains well above chance after correctness filtering, showing that the probe captures consistency-specific information beyond simply predicting whether the original answer is correct.

To test whether our conclusions extend to a broader equivalence setting, we additionally construct equivalence sets that combine image augmentations with question paraphrases. We generate three paraphrases for POPE and VQAv2: POPE uses a fixed set of manually written templates, while VQAv2 uses GPT-5-nano to generate paraphrases. We then sample five image augmentation modes and pair them with the paraphrases, producing 15 variants per example, and evaluate the resulting equivalence sets with Qwen2-VL-7B-Instruct.

Table 7 compares consistency probes trained and evaluated under image-only equivalence sets and equivalence sets that combine image augmentations with question paraphrases. Image-only supervision already provides a strong consistency signal, achieving AUROC scores of 0.8550 on POPE and 0.7836 on VQAv2. Moreover, when a probe trained with image-only labels is evaluated against the broader image-plus-text setting, performance remains competitive, with AUROC scores of 0.8367 on POPE and 0.7285 on VQAv2. This suggests that visual perturbation consistency captures a substantial part of the general stability signal.

Adding question paraphrases can improve performance when the train and test definitions are matched, especially on POPE, where the Both/Both setting achieves the best AUROC of 0.8994. However, the gains are not uniform: on VQAv2, Image/Image slightly outperforms Both/Both (0.7836 vs. 0.7768). In addition, cross-setting transfer is asymmetric. Probes trained on Both and tested on Image perform worse than probes trained directly on Image, particularly on POPE (0.7619 vs. 0.8550). This indicates that paraphrase-based supervision introduces additional variation that is not fully aligned with purely visual consistency. Overall, these results support our decision to focus on image augmentations: they keep the question fixed, isolate sensitivity to visual grounding, and avoid confounds from paraphrase quality, linguistic ambiguity, and shifts in textual priors.

Table 7. Effect of image augmentations and question paraphrases for constructing equivalence sets. "Image" denotes image-only augmentations, while "Both" denotes image augmentations plus question paraphrases. Results are AUROC with Qwen2-VL-7B-Instruct.

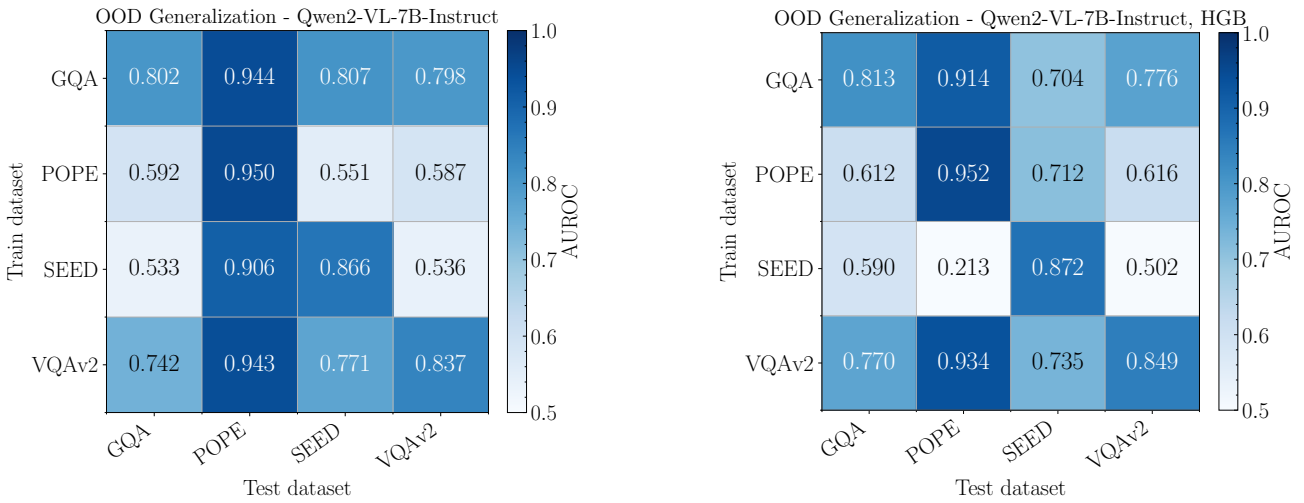
Train	Test	POPE	VQAv2
Image	Image	0.8550	**0.7836**
Image	Both	0.8367	0.7285
Both	Image	0.7619	0.7368
Both	Both	**0.8994**	0.7768

F. Additional Results

F.1. OOD Generalization - Histogram-based Gradient Boosting

The main paper evaluates cross-dataset generalization using a linear probe to test whether consistency-related signals in hidden representations transfer beyond the dataset used for training. In this section, we provide additional OOD results to

further examine how the choice of classifier affects transfer performance. In particular, we compare the linear probe with histogram-based gradient boosting (HGB), a more expressive nonlinear classifier.



(a) Linear Probe *(b)* Histogram-based Gradient Boosting

Figure 12. Additional out-of-distribution generalization results for Qwen2-VL-7B-Instruct. We compare cross-dataset AUROC for a linear probe and a histogram-based gradient boosting probe. While HGB can improve some in-domain and transfer results, it also exhibits less stable cross-dataset behavior, indicating greater sensitivity to dataset-specific structure.

Figure 12 provides additional out-of-distribution (OOD) generalization results for Qwen2-VL-7B-Instruct, comparing a linear probe with a histogram-based gradient boosting (HGB) probe. The linear probe corresponds to the main cross-dataset transfer setting, while the HGB probe tests whether a more expressive nonlinear classifier improves transfer under dataset shift.

Overall, the results show that increasing probe nonlinearity does not uniformly improve OOD generalization. HGB achieves slightly higher in-domain AUROC on several datasets, such as GQA and VQAv2, and improves some off-diagonal transfer results, including training on POPE and testing on SEED-Bench. However, its cross-dataset behavior is less stable than the linear probe. In particular, when trained on SEED-Bench and evaluated on POPE, HGB drops sharply, suggesting that the nonlinear probe can overfit to dataset-specific structure that does not transfer reliably.

These results reinforce the conclusion that hidden representations encode a transferable consistency signal, but also show that this signal is not fully invariant across datasets. The linear probe often provides a better balance between in-domain accuracy and OOD robustness, likely because its limited capacity encourages it to recover broad directions in representation space associated with model stability rather than dataset-specific artifacts. This supports our use of linear probing as the default setting for most analysis: it is simple, computationally efficient, and more reliable under cross-dataset transfer.

F.2. Relationship Between Pairwise Consistency and Accuracy

Setup. Extending the experiment of Khan and Fu (Khan & Fu, 2024), we further examine whether consistency under semantics-preserving perturbations is related to predictive accuracy. For each dataset, we evaluate Qwen2-VL-7B-Instruct on the equivalence sets constructed from the original image-question pairs and their augmented variants listed in App. A. For each equivalence set E_i, we compute the pairwise consistency score as in Eq. B, which measures the fraction of variant pairs that receive the same normalized answer. We also compute the average accuracy over the variants in the same equivalence set. We then analyze the relationship between pairwise consistency and average accuracy across GQA, POPE, SEED-Bench, and VQAv2.

Results. Figure 13 shows a strong positive relationship between pairwise consistency and average accuracy across all four datasets. The correlation is high for each benchmark: GQA achieves Pearson correlation $P = 0.938$ and Spearman correlation $S = 0.936$, POPE achieves $P = 0.974$ and $S = 1.000$, SEED-Bench achieves $P = 0.875$ and $S = 0.891$, and VQAv2 achieves $P = 0.964$ and $S = 0.978$. This demonstrates the fact that samples or settings with higher pairwise

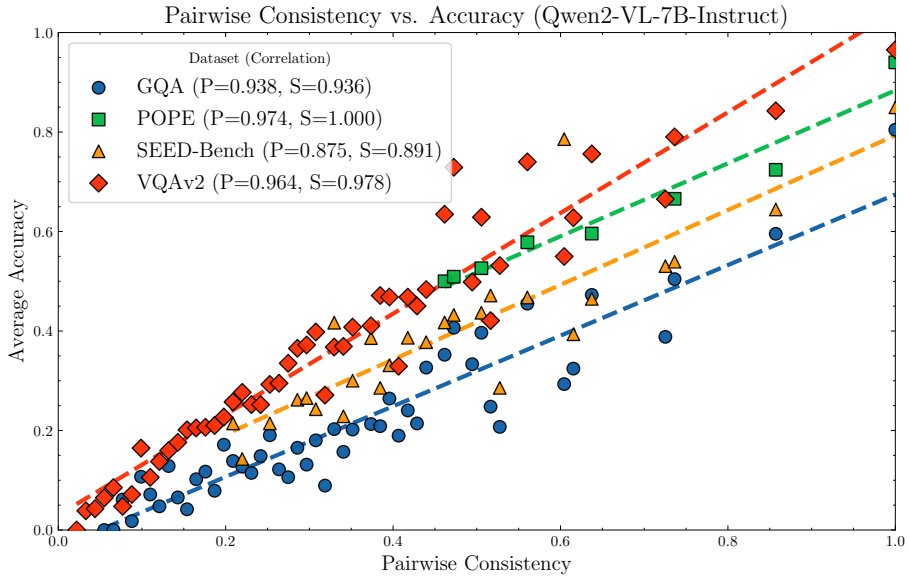


Figure 13. Pairwise consistency (Eq. B) positively correlates with accuracy across multiple VQA datasets, suggesting that more stable predictions tend to be more accurate. P denotes Pearson correlation and S denotes Spearman's rank correlation.

consistency tend to achieve higher accuracy, indicating that consistency is not merely a superficial robustness property and is closely tied to whether the model has formed a reliable multimodal prediction. In other words, when a model is unstable across equivalent variants, its prediction is also more likely to be unreliable.

G. Limitations

This work has several limitations. First, our experiments cover two LVLMs and four VQA-style benchmarks, so the conclusions may not fully generalize to larger models, closed-source systems, or specialized domains such as medicine, robotics, or autonomous driving. Second, our equivalence sets mainly use image augmentations. Although these transformations are designed to preserve task-relevant semantics, equivalence is task-dependent, and some perturbations may affect fine-grained questions involving text, color, counting, or spatial details. Third, our labels measure behavioral consistency rather than correctness: a model can be consistent but wrong, or correct on the original input but inconsistent across variants. Finally, the probe requires offline multi-variant evaluation to construct training labels and provides a probabilistic reliability signal rather than a formal guarantee.

H. Broader Impacts

This work aims to make LVLM reliability monitoring more efficient by predicting consistency from a single forward pass. This could help identify unstable predictions, reduce the cost of robustness evaluation, and support safer use of LVLMs in applications where stable multimodal reasoning is important.

However, consistency prediction should not be treated as a complete safety measure. Consistent outputs can still be incorrect, biased, or harmful, and probes trained on benchmark data may not generalize equally across domains, languages, or user populations. In deployment, such predictors should therefore be used as monitoring or triage tools, together with uncertainty reporting, additional verification, and human oversight in high-stakes settings.