CAN'T FIND WALDO: EVALUATING VLMS' SENSITIVITY TO IMAGE RESOLUTION AND DETAIL LEVEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Language Models (VLMs) have achieved remarkable success across diverse tasks, yet they struggle with high-resolution inputs where critical information resides in small regions or detailed, cluttered scenes. While several approaches address this limitation, a systematic understanding of *why* models fail at high resolutions is lacking. We introduce a controlled evaluation framework that disentangles resolution-related performance degradation from task difficulty through semantics-preserving transformations. We propose two simple metrics: Area Under the Scaling Curve (AUSC), which quantifies scaling robustness independent of baseline accuracy, and Prediction Variance Score (PVS), which measures resolution-induced prediction instability. Through comprehensive experiments across four model families and five benchmarks, we investigate architecture-dependent resolution sensitivity in relation to three pipeline mechanisms: (1) information loss from downsampling at vision token limits, (2) tokenization and positional sensitivity under changes in patch layout and aspect ratio, and (3) attention dilution as token counts increase. Our analysis reveals that widely used VLM architectures suffer from performance drops when processing high-resolution images, with degradation patterns varying systematically by architectural family. We provide actionable insights for model architecture design and data augmentation strategies to mitigate these limitations.

1 INTRODUCTION

Deploying Vision–Language Models (VLMs) in domains such as medical imaging (Tran et al., 2025; Dai et al., 2025), satellite imagery (Luo et al., 2025), and document understanding (Liu et al., 2025) requires reliable extraction of fine-grained visual detail across varying scales and resolutions. Yet the visual information a model can effectively access is shaped not only by image semantics, but also by how that information is encoded and processed: the input resolution, and the size, number, and position of task-relevant visual cues. As a result, a model may fail not because of an inability to interpret and reason on the input, but because critical visual evidence was lost during preprocessing, distorted by patch tokenization, or diluted across an overlong token sequence.

Current VLM benchmarks span heterogeneous image collections with varying resolutions, aspect ratios, and information densities (Fu et al., 2024; Zhang et al., 2024a; Yu et al., 2024). Although intended to reflect real-world conditions, this diversity introduces a confound: performance differences may reflect how well each architecture handles incidental image properties rather than differences in perceptual or reasoning ability, complicating fair comparison and failure attribution. To isolate these effects, we introduce a diagnostic framework that applies controlled, semantics-preserving transformations to existing VQA benchmarks, systematically varying image resolution and aspect ratio, level of detail, and the spatial position of task-relevant content. Comparing responses across these transformations reveals distinct patterns of performance degradation and helps identify which failure mechanisms warrant investigation for each model architecture.

On top of this framework, we propose two complementary metrics. *Area Under the Scaling Curve* (AUSC) quantifies how model performance degrades as resolution increases, and *Prediction Variance Score* (PVS) measures instability in model predictions across resolution variants of the same image, revealing that the model possesses the underlying reasoning capacity but that its answer is contingent on tokenization artifacts.

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

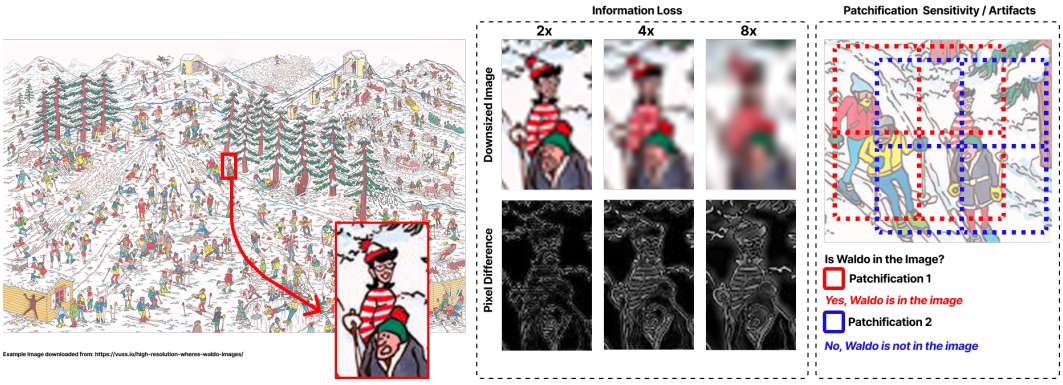

Figure 1: Information loss may be attributable to diverse causes, including downsizing, patch artifacts, and attention dilution.

Through extensive experiments across four architecturally distinct model families and multiple model sizes, we characterize resolution sensitivity using three mechanisms associated with successive stages of visual processing:

1. **Downsampling-induced information loss**, where fine-grained detail is irreversibly discarded when image resolution exceeds the model's vision token budget.
2. **Tokenization artifacts** arising from patch boundary misalignment and information density mismatch, causing non-monotonic, resolution-dependent prediction instability.
3. **Attention dilution**, where increasing token counts cause the LLM decoder to lose effective focus on spatially localized task-relevant cues.

The onset and severity of each failure mode varies by architecture, reflecting differences in dynamic resolution strategies, token budget management, and attention mechanisms—differences to which standard single-resolution benchmarks are blind.

Our contributions are threefold:

- A controlled evaluation framework isolating visual encoding sensitivity from reasoning ability via semantics-preserving variations in resolution, aspect ratio, and spatial position of task-relevant cues.
- Two architecture-agnostic metrics (AUSC and PVS) quantifying resolution robustness independently of baseline accuracy.
- A systematic analysis framework leveraging cross-scenario prediction trajectories to identify patterns of instability and provide insights into architectural and training factors potentially contributing to resolution sensitivity, together with possible mitigation strategies.

2 Related Work

2.1 Scaling Vision Encoders to High Resolution

Processing high-resolution images in Vision–Language Models involves a fundamental tension between preserving fine-grained visual detail and managing computational cost (Bakhtiarnia et al., 2024). Early models, including LLaVA 1.5 and BLIP-2, resample inputs to a fixed resolution, discarding information before reasoning (Liu et al., 2023; Li et al., 2023). Dynamic-tiling methods instead process local crops alongside a global thumbnail (Liu et al., 2024; Xu et al., 2024; Li et al., 2024; Grattafiori et al., 2024), improving flexibility but increasing computation and retaining implicit resolution and aspect-ratio constraints. Native-resolution models remove fixed input dimensions at the cost of variable token counts (Agrawal et al., 2024; Wang et al., 2024a; Bai et al., 2025). Content-adaptive tokenization and selective-encoding approaches further reduce computation by allocating tokens according to visual complexity or task relevance (Shen et al., 2025; Zhou et al., 2025; Shi et al., 2025), but may introduce new sensitivities to image layout and content position.

2.2 Evaluating High-Resolution Visual Processing

Recent benchmarks expose persistent limitations in fine-grained visual processing. MME-RealWorld reports substantial degradation on high-resolution images (Zhang et al., 2024b), while HRScene identifies spatially systematic failures that intensify with image size (Zhang et al., 2025). VisualOverload evaluates dense scenes and reveals severe failures in counting, OCR, and logical consistency (Gavrikov et al., 2025). VER-Bench tests evidence extraction from small visual regions and distinguishes evidence coverage from reasoning quality (Qiang et al., 2025). V*Bench and Visual Haystacks similarly emphasize precise spatial grounding and retrieval over raw model scale (Wu & Xie, 2024; Wu et al., 2025). However, these benchmarks generally evaluate each image at a single resolution, conflating resolution-induced failures with semantic complexity, task difficulty, and image content.

2.3 Evaluating Robustness in Vision–Language Models

Robustness studies further connect VLM failures to architectural design. VLM-RobustBench finds that geometric and resampling perturbations are more damaging than perceptually severe photometric corruptions (Saxena et al., 2026). V^2R-Bench attributes scale and spatial variation sensitivity to error accumulation in the vision encoder and insufficient multimodal alignment (Fan et al., 2025), while HRScene links resolution-dependent spatial failures to uneven token compression under dynamic patching (Zhang et al., 2025). These studies establish important resolution and spatial sensitivities. We extend them by introducing a controlled diagnostic framework that reveals model-specific sensitivities that remain concealed when considering only aggregate accuracy. Our framework connects them to potential bottlenecks in the VLM pipeline, guiding targeted architectural and preprocessing improvements.

3 Background

3.1 Problem Formulation

Let $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a VQA benchmark, where $x_i \in \mathbb{R}^{H_i \times W_i \times 3}$ is an image and y_i is the corresponding question. A VLM f maps each pair to a predicted answer $\hat{z}_i = f(x_i, y_i)$, evaluated against ground truth z_i. Standard benchmarks measure aggregate accuracy $\frac{1}{N}\sum_i \mathbf{1}[\hat{z}_i = z_i]$ at a single fixed resolution per image, implicitly assuming performance is invariant to image scale and tokenization structure. We challenge this assumption and study model behavior under a family of semantics-preserving transformations $\mathcal{T} = \{T_1, \ldots, T_M\}$, where each T_j modifies an image along a controlled axis (pixel density or token allocation) without altering the visual content relevant to y_i.

Processing a visual query involves a sequential causal chain:

$$x \xrightarrow{\text{preproc.}} \text{pixels} \xrightarrow{\text{tokenization}} \underset{\text{tokens}}{\text{vision}} \xrightarrow{\text{encoder}} \text{visual features} \xrightarrow{\text{projection}} \underset{\text{tokens}}{\text{aligned}} \xrightarrow{\text{LLM}} \hat{z} \tag{1}$$

Evaluating performance across controlled transformations and scales provides a structured view of model behavior beyond aggregate accuracy. It enables us to decompose performance into more informative components, identify likely failure modes and attribute them to specific model components.

3.2 Failure Modes

We identify three prominent sources of performance degradation across VLM components: (1) information loss induced by image downsizing during preprocessing, (2) tokenization artifacts introduced by the visual encoder, and (3) attention dilution in the LLM.

3.2.1 Downsizing-Induced Information Loss

Downsizing-induced information loss arises when the preprocessing pipeline reduces image resolution before visual encoding. This is a major source of performance degradation in models that apply

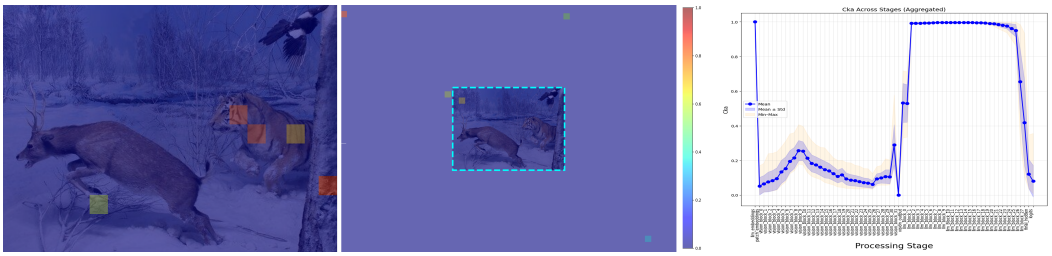

Figure 2: **On the left:** Illustration of attention dilution in Qwen2-VL-2B. Despite identical visual content within the cyan bounding box, attention weights diffuse across uninformative blank regions, reducing concentration on task-relevant content and degrading performance; **On the right:** Centered Kernel Alignment (CKA) calculated for all layers of the Qwen2-VL-2B VLM comparing internal representations for two images: (1) original image padded with white field, (2) upscaled image

aggressive resizing (e.g., LLaVA-1.5) and also in models that otherwise support high-resolution inputs once their vision token budget is exceeded (e.g., Pixtral, Qwen-family). In the latter case, the effective input is still constrained through automatic downsampling or cropping, which limits the spatial detail available to downstream components.

3.3 Tokenization Artifacts

Vision Transformers partition images into fixed-size patches (typically 14×14 or 16×16 pixels) that serve as input tokens. This discrete tessellation introduces a fundamental dependency: Model performance can vary based on where patch boundaries fall relative to semantic content, independent of the information itself. We identify two distinct sources of tokenization artifacts: (a) **spatial shift artifacts**, where translating the patch grid by a few pixels changes which visual features are split across patch boundaries, and (b) **compression artifacts**, where downsampling increases the semantic density per patch, forcing individual tokens to encode different level of detail.

To examine how these artifacts propagate through the model, we compared layer-wise representations of the same visual content placed at different background locations in Qwen2-VL-2B-Instruct. The resulting representations become more similar in later layers of the vision encoder, but positional differences remain clearly encoded. In the LLM, visual information has little influence in the early and middle layers, where representations are largely dominated by text; location-dependent differences emerge only in the later decoder layers (Figure 2).

3.4 Attention Dilution

Native resolution models such as the Qwen-VL series process images without resizing, converting each image into a variable number of visual tokens proportional to its dimensions. As padding increases the image size, the resulting token sequence grows quadratically, forcing the model's attention mechanism to distribute its capacity across increasingly sparse representations of the same visual content. This attention dilution occurs when the attention distribution becomes diffuse across many redundant tokens, reducing the relative weight assigned to semantically meaningful regions. To illustrate this effect, we analyzed attention maps for images where the correct answer flipped to an incorrect one after introducing image transformations (Figure 2).

4 Evaluation Framework

We design two complementary transformation protocols, each targeting a distinct stage of the processing pipeline. Together, they form a *decomposition along the causal chain*: Pixel-Resolution probes the full pipeline response to increased pixel density; Token Resolution isolates tokenization-level effects by holding inter-pixel content fixed. The difference in model response between the scenarios provides diagnostic evidence about the processing pipeline.

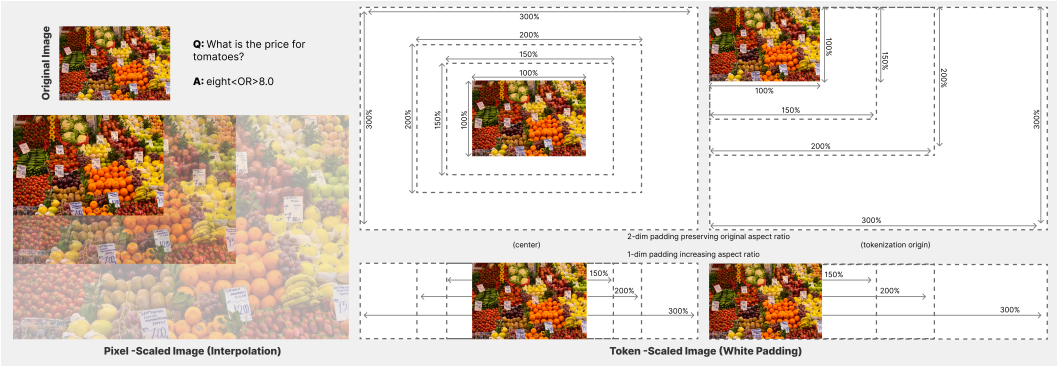

Figure 3: Pixel- and Token- Resolution Scaling Example

4.1 Pixel-Level Resolution Scaling

We generate higher-resolution variants of each benchmark image via bicubic interpolation upsampling. This transformation scales spatial dimensions uniformly by factor s, preserving global semantic content and the relative scale of all objects within the frame. No additional task-relevant information is introduced: objects subtend the same proportion of the image regardless of s. From the model's perspective, interpolation reduces *information density per token*: the same semantic content is distributed across more pixels and therefore more vision tokens, decreasing the task-relevant signal per token. Pixel resolution, therefore, probes how the model perceives changes in information density, while detail level stays constant.

4.2 Token-Level Resolution Scaling via Canvas Padding

Alternatively, we can increase the effective image resolution by embedding the original image within a larger white canvas, leaving pixel content entirely unchanged. This forces the model's preprocessing pipeline to handle a larger image, typically triggering more vision tokens, shifted patch boundaries, and, in dynamic tiling models, a different tile arrangement, without any change to the inter-pixel distances in the relevant image region. This transformation isolates tokenization dynamics—patch boundary misalignment, increased token sequences, and attention dilution—from any pixel-level change, enabling attribution of performance changes to the tokenization stage and beyond. It is important to note that, while canvas padding may in part target out-of-distribution regimes, we empirically confirm that across all models it is not adversarial (Appendix C.1).

Spatial position variants. To further probe the effect of patch boundary alignment and spatial attention bias, we place the original image at two positions on the canvas: *top-left*, and *center*. These positions were selected to expose distinct mechanistic effects with minimal computational overhead. Top-left preserves the original patch boundary alignment for models that do not resize, serving as a controlled baseline. Center placement tests spatial attention bias and tile-boundary splitting.

4.3 Metrics

We propose two complementary metrics to quantify resolution scaling behavior, designed to be architecture-agnostic and independent of baseline performance level. Both metrics condition on subsets of $\mathcal{B}$ stratified by correctness at original resolution, and we recommend $|\mathcal{B}_{\text{correct}}| \geq 75$ and $|\mathcal{B}_{\text{wrong}}| \geq 75$ for statistical reliability.

4.3.1 Area Under the Scaling Curve (AUSC)

Motivation. Raw accuracy conflates baseline capability with scaling robustness. AUSC isolates scaling behavior from baseline performance by conditioning exclusively on examples the model has proven to have sufficient reasoning ability to answer.

Definition. Let $\mathcal{B}_{\text{correct}} \subseteq \mathcal{B}$ denote the subset of examples answered correctly at the original resolution:

$$\mathcal{B}_{\text{correct}} = \{(x_i, y_i) \in \mathcal{B} \mid f(x_i, y_i) = z_i\} \tag{2}$$

For transformation sequence $\mathcal{T} = \{T_1, \ldots, T_M\}$, let $\mathrm{Acc}(T_j)$ denote accuracy on $\{(T_j(x_i), y_i) \mid (x_i, y_i) \in \mathcal{B}_{\text{correct}}\}$. AUSC is computed via the trapezoidal rule:

$$\mathrm{AUSC} = \frac{1}{M-1} \sum_{j=1}^{M-1} \frac{1}{2} \Big[\mathrm{Acc}(T_j) + \mathrm{Acc}(T_{j+1}) \Big] \tag{3}$$

AUSC $\in [0, 1]$, where 1 indicates perfect robustness across all transformations and 0 indicates complete brittleness.

4.3.2 Prediction Variance Score (PVS)

Motivation. A complementary signal is whether resolution changes can *rescue* initially incorrect predictions, indicating that the model possesses the underlying reasoning capacity, but that its answer is sensitive to tokenization or encoding artifacts.

Definition. Let $\mathcal{B}_{\text{wrong}} = \mathcal{B} \setminus \mathcal{B}_{\text{correct}}$. For each transformation T_j:

$$\mathrm{PVS}(T_j) = \frac{1}{|\mathcal{B}_{\text{wrong}}|} \sum_{(x_i, y_i) \in \mathcal{B}_{\text{wrong}}} \mathbb{1}\big[f(T_j(x_i),\, y_i) = z_i \big] \tag{4}$$

PVS $\in [0, 1]$ measures the fraction of initially incorrect predictions corrected by T_j. Values near 0 indicate that errors are stable and likely reflect genuine reasoning limitations. Higher values reveal resolution-dependent instability.

4.4 Attribution Analysis

The two transformation protocols jointly support diagnostic attribution by relating observed sensitivity to candidate mechanisms in the processing pipeline. We track each example's answer correctness as a binary sequence across scaling levels. Based on this, we identify trajectory patterns, summarized in Table 1, with cross-scenario fingerprints to help prioritize candidate explanations for further investigation.

Table 1: Answer trajectory patterns, their diagnostic interpretation, and the cross-scenario fingerprint enabling attribution.

Pattern	Trajectory	Attribution	Fingerprint
A	$1 \cdots 1\, 0 \cdots 0$ (threshold)	Downsampling loss	Pixel Resolution scaling only, at token budget
B	Alternating	Boundary / density	Token Resolution Scaling $\rightarrow$ boundary; Pixel Resolution Scaling $\rightarrow$ density
C	$0 \cdots 0\, 1 \cdots$	Tokenization masking	PVS signal; both scenarios
D	$1\, 1\, 0\, 0 \cdots$	Attention dilution	Both scenarios, same token count
E	$0 \cdots 0$	Reasoning failure	Excluded from metrics
F	$1 \cdots 1$	Robust correct	AUSC numerator
G	Position-varying	Spatial bias / tiling	Token-Resolution Scaling, position-stratified

A more detailed version of the analysis complemented with VLM architecture attribution and recommended measures is provided in Appendix B.

5 Experiments

5.1 Experimental Set-up

Scaling Range. We empirically select an image-scale range that provides enough granularity to expose encoder instabilities while keeping the computational cost manageable. We evaluate models

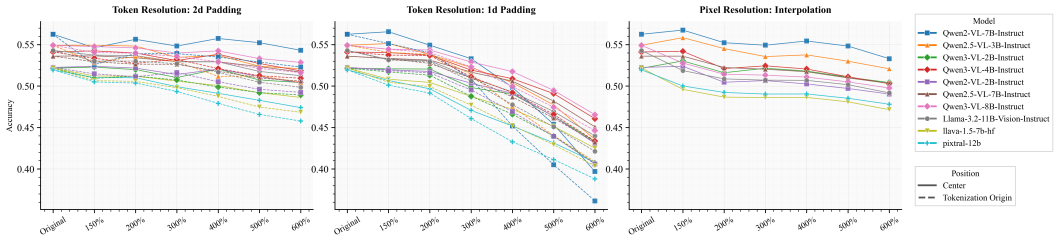

Figure 4: Accuracy degradation under different scaling factors.

over a fixed range of image scales (1.5×–6×) to measure how performance changes with increasing visual detail. The lower end is sufficient to reveal sensitivity to small changes in resolution, while the upper end approaches the practical limits of current VLMs, where larger inputs become increasingly constrained by context length and memory. This range provides enough coverage to capture meaningful scaling behavior without making evaluation prohibitively expensive.

Models. We chose several commonly used open and closed source models, specifically LLaVa-1.5 (Liu et al., 2024), Llama-3.2 (Grattafiori et al., 2024), Qwen-2-XB (Wang et al., 2024a), Qwen-2.5-XB (Xu et al., 2025), Qwen-3-XB (Yang et al., 2025), and Pixtral (Agrawal et al., 2024).

Benchmarks. To assess differences in scaling across different types of VQA, we chose the MMVet (Yu et al., 2024), CharXiv (Wang et al., 2024b), MMEval (Fu et al., 2024), MMStar (Chen et al., 2024), and MathVerse (Zhang et al., 2024a) benchmarks for further analysis. Our choice is motivated by the desire to create a diverse set of image–question pairs across different aspects of model performance and image sizes. However, the proposed method is benchmark-agnostic and could be applied to any set of benchmarks that best matches the evaluation purpose. Further per-benchmark analysis is provided in Appendix C.4.

Sampling Strategy. The overall size of the evaluation sample used in the experiments is 1,000 image–question pairs. For a diverse stratified sample, we randomly sample an equal number of instances, $\frac{1}{5}$-th of the overall size, from each of the source benchmarks.

Evaluation Protocol To evaluate correctness of VQA answers, we use a hybrid approach combining answer parsing for cases where it is possible and LLM-as-a-judge (gpt-5-nano) for more complex cases. Detailed evaluation details are provided in Appendix F.

5.2 Experimental Results

5.2.1 Accuracy Analysis

We report accuracy across five transformation conditions designed to isolate distinct failure modes (Figure 4). Under 2D padding, degradation is modest ($\leq$10%), with native-resolution models barely affected, confirming that preprocessing downsampling, not the added padding itself, drives information loss. Placing the image at the top-left corner (the tokenization origin) consistently underperforms centering, revealing a learned spatial bias toward centered content rather than position-invariant perception, in contrast to the findings by Zhang et al. (2025).

The most revealing contrast is between 1D and 2D padding. Despite producing fewer total tokens, 1D padding causes more severe degradation (5–18%). This directly disentangles two candidate failure modes: if attention dilution over uninformative tokens were dominant, 2D padding should perform worse. Instead, the extreme aspect ratios push positional encodings out of distribution, identifying positional encoding fragility as an independent failure mode.

Bicubic interpolation reveals a non-monotonic accuracy profile. Several models improve at 1.5× before declining, while others degrade immediately, suggesting model-specific optimal information density ranges shaped by pre-training distributions. We further investigate this by analyzing AUSC, PVS, and answer trajectories.

Table 2: Original accuracy, AUSC, and PVS across augmentation variants for different VLMs.

		AUSC					PVS				
Model	**Orig. Acc.**	**1D-C**	**1D-T**	**2D-C**	**2D-T**	**Intra**	**1D-C**	**1D-T**	**2D-C**	**2D-T**	**Intra**
Qwen2-VL-2B	0.448	0.752	0.716	0.845	0.826	0.827	0.368	0.368	0.290	0.307	0.300
Qwen2-VL-7B	0.562	0.786	0.733	0.850	0.830	0.851	0.385	0.397	0.353	0.348	0.334
Qwen2.5-VL-3B	0.536	0.830	0.814	0.832	0.823	0.846	0.368	0.352	0.361	0.319	0.352
Qwen2.5-VL-7B	0.604	0.858	0.853	0.870	0.864	0.871	0.382	0.397	0.359	0.351	0.349
Qwen3-VL-2B	0.466	0.803	0.799	0.827	0.817	0.836	0.414	0.376	0.416	0.394	0.445
Qwen3-VL-4B	0.600	0.836	0.820	0.855	0.851	0.775	0.454	0.475	0.475	0.454	0.429
Qwen3-VL-8B	0.627	0.842	0.832	0.866	0.858	0.683	0.428	0.414	0.414	0.384	0.398
Llama-3.2-11B-V	0.499	0.569	0.578	0.719	0.704	0.777	0.493	0.483	0.477	0.456	0.475
Llava-1.5-7b-hf	0.352	0.723	0.589	0.673	0.593	0.751	0.307	0.317	0.365	0.334	0.306
Pixtral-12b	0.499	0.521	0.507	0.707	0.694	0.890	0.432	0.483	0.422	0.450	0.349

5.2.2 AUSC AND PVS

While AUSC measures how well models maintain correct predictions across resolutions, PVS reveals a complementary phenomenon: cases where resolution changes rescue initially incorrect predictions. High PVS values indicate that a model's failures are partially attributable to resolution-dependent artifacts rather than fundamental reasoning limitations. Table 2 presents AUSC and PVS metrics across different scaling factors for representative models from each architectural family. We observe three distinct patterns.

Positional fragility dominates token dilution. The AUSC gap between 1D and 2D padding is consistent and substantial—up to 20 points for Pixtral (0.521 vs. 0.707) and Llama (0.569 vs. 0.719). Since 1D padding produces fewer total tokens yet causes greater degradation, aspect ratio distortion of positional encodings is demonstrably more damaging than attention dilution over uninformative padding regions. Generational improvements within the Qwen family (Qwen2: 0.752 $\rightarrow$ Qwen2.5: 0.830) suggest this fragility is partially addressable through diverse aspect ratio training.

Scale robustness and positional robustness are independent axes. Pixtral achieves the highest interpolation AUSC (0.890) yet the lowest 1D padding AUSC (0.507)—highly resilient to density changes but fragile to positional shift. Qwen3-VL-8B exhibits the inverse: strong padding tolerance but degraded interpolation stability (0.683), indicating possible overfitting to a narrow information density range. This dissociation confirms that these represent fundamentally different failure mechanisms requiring distinct solutions.

PVS and AUSC reveal prediction volatility. Models with high PVS and low AUSC (e.g., Llama-3.2 and Pixtral) under 1D padding—exhibit volatile representations where predictions flip in both directions, suggesting near-decision-boundary uncertainty. In contrast, LLaVA-1.5 shows low PVS with moderate AUSC, indicating rigid confidence in both correct and incorrect predictions. Notably, PVS values of 0.30–0.49 across all conditions imply that a substantial fraction of baseline errors stem from suboptimal default tokenization rather than fundamental capability gaps.

6 CONCLUSION

VLM evaluation on VQA tasks is typically based solely on accuracy, implicitly treating all errors as failures of multimodal reasoning. Our analysis shows that this view is incomplete: only a small fraction of incorrect predictions stem from fundamental reasoning limitations. Many errors arise from information loss introduced during image preprocessing, instabilities in the vision encoder, and attention dilution caused by longer visual token sequences. These findings suggest that improving VLM performance requires not only stronger reasoning modules, but also more robust visual encoders, adapters, and evaluation protocols that disentangle perception, tokenization, and reasoning effects.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Efficient high-resolution deep learning: A survey. *ACM Computing Surveys*, 56(7):1–35, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Dawei Dai, Yuanhui Zhang, Qianlan Yang, Long Xu, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pathologyvlm: a large vision-language model for pathology image understanding. *Artificial Intelligence Review*, 58(6):1–19, 2025.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. Unveiling the lack of LVLM robustness to fundamental visual variations: Why and path forward. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20222–20242, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1037. URL https://aclanthology.org/2025.findings-acl.1037/.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Paul Gavrikov, Wei Lin, M Jehanzeb Mirza, Soumya Jahagirdar, Muhammad Huzaifa, Sivan Doveh, Serena Yeung-Levy, James Glass, and Hilde Kuehne. Visualoverload: Probing visual understanding of vlms in really dense scenes. *arXiv preprint arXiv:2509.25339*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri ..., and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Chenxu Li, Zhicai Wang, Yuan Sheng, Xingyu Zhu, Yanbin Hao, and Xiang Wang. Res-bench: Benchmarking the robustness of multimodal large language models to dynamic resolution input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31545–31553, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models, 2024. URL https://arxiv.org/abs/2311.06607.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning, 2025. URL https://arxiv.org/abs/2503.07588.

Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12698–12705, 2025.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint arXiv:2603.06148*, 2026.

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. Cat: Content-adaptive image tokenization, 2025. URL https://arxiv.org/abs/2501.03120.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and Hongxu Yin. Scaling vision pre-training to 4k resolution, 2025. URL https://arxiv.org/abs/2503.19903.

Manuel Tran, Paul Schmidle, Ruifeng Ray Guo, Sophia J Wagner, Valentin Koch, Valerio Lupperger, Brenna Novotny, Dennis H Murphree, Heather D Hardway, Marina D'Amato, et al. Generating dermatopathology reports from gigapixel whole slide images with histogpt. *Nature Communications*, 16(1):4886, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024b. URL https://arxiv.org/abs/2406.18521.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Tsung-Han Wu, Giscard Biamby, , Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.13766.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images, 2024. URL https://arxiv.org/abs/2403.11703.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024a. URL https://arxiv.org/abs/2403.14624.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Yusen Zhang, Wenliang Zheng, Aashrith Madasu, Peng Shi, Ryo Kamoi, Hao Zhou, Zhuoyang Zou, Shu Zhao, Sarkar Snigdha Sarathi Das, Vipul Gupta, et al. Hrscene: How far are vlms from effective high-resolution image understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22922–22933, 2025.

Xirui Zhou, Lianlei Shan, and Xiaolin Gui. Dynrsl-vlm: Enhancing autonomous driving perception with dynamic resolution vision-language models, 2025. URL https://arxiv.org/abs/2503.11265.

A LIMITATIONS

We caution that the newly introduced metrics are dependent on the chosen image–question pairs and hence can solely be used for comparing between models evaluated on the same sample and for determining reasons for potential model performance instabilities, not as an absolute numeric measure.

The identified failure modes provide a structured basis for diagnosing resolution sensitivity across the VLM pipeline, highlighting how interacting effects shape model behavior under different transformations. These mechanisms are not meant to be mutually exclusive and only serve as a diagnostic tool allowing to identify problematic model elements.

B FUTURE WORK DISCUSSION

Table 3 synthesizes the failure modes identified in our analysis into a component-level attribution and intervention framework. For each failure pattern, we indicate the relative involvement of the image encoder, vision–language alignment module, and language model, with checkmark multiplicity denoting the estimated strength of attribution. We then outline corresponding directions for future work across three complementary dimensions: architectural design, training objectives, and data curation. These attributions are not intended to imply isolated causes; rather, they highlight the pipeline components most likely to constrain performance and provide a structured basis for developing targeted mitigation strategies.

C ADDITIONAL ANALYSIS

C.1 IS PADDING AN OUT-OF-DISTRIBUTION ARTIFACT?

Since padding augmentation introduces images with large uniform borders, we asked whether the observed degradation is partly driven by the out-of-distribution nature of the augmented images than by reduced effective information density.

To test this, we compare model performance on original high-resolution images (Wu & Xie, 2024) against versions where all irrelevant content is masked out and replaced with a uniform white background, while preserving the question-relevant region. If white padding were inherently harmful, performance would drop in the masked setting. Instead, all models perform better when irrelevant content is replaced with white space.

This shows that white regions are not intrinsically out-of-distribution for VLMs and are not the primary cause of the degradation observed under white-padding augmentation. The performance drop is instead driven by reduced effective information density.

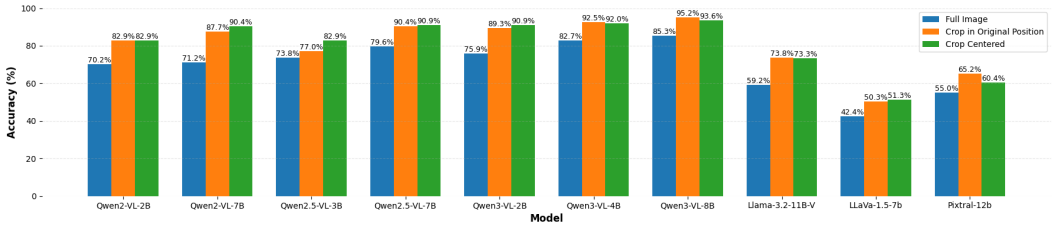

Figure 5: All models perform better when irrelevant content is replaced with white background, showing that white padding itself is not the primary cause of performance degradation.

C.2 LIMITATIONS OF SINGLE-RESOLUTION VQA ACCURACY

Standard VQA accuracy evaluates whether a model answers each question correctly for a single image representation. To examine what this aggregate score obscures, we conduct an additional analysis of prediction outcomes across semantically preserving image transformations. The results

Table 3: Failure mode attribution to pipeline components and recommended interventions. Checkmark weight indicates strength of component involvement.

Failure Mode (Pattern)	Img. Enc.	VL Align.	LLM	Architecture Changes	Training Tasks Changes	Training Data Changes
1. Down- sampling Loss	✓✓✓			Adaptive resolution; hierarchical/pyramid encoder; multi-scale feature extraction; increase native resolution (1024×1024+)	Pre-train on high-res tasks; multi-scale prediction tasks	High-resolution datasets; preserve original resolutions
2a. Patch Boundary Sensitivity	✓✓✓			Overlapping patches; continuous tokenization (Perceiver); sliding window approaches	Tasks requiring cross-patch reasoning; random patch offset augmentation	Annotations spanning patch boundaries
2b. Information Density Mismatch	✓✓	✓	✓✓	Adaptive token allocation; variable patch sizes; learned token pruning/expansion	Detail-focused pretraining; dense region description; small object detection	Variable-density images; sparse/dense scene mixtures; fine-grained datasets
3. Attention Dilution			✓✓✓	Hierarchical attention; sparse/local attention; region-of-interest modules; cross-attention gating	Spatial attention supervision; localization-aware tasks; progressive detail focusing	Ground-truth attention maps; saliency-annotated data
4. Vision–Language Alignment Gap	✓	✓✓✓	✓	Richer projection layers; multi-layer alignment; feature-preserving projectors	Fine-grained VQA; contrastive attribute learning; pixel-word grounding; referring expression tasks	Detailed attribute annotations; fine-grained VQA; grounding datasets (RefCOCO); descriptive region captions
5. Spatial Attention Bias	✓	✓	✓✓	Position-aware attention; learned spatial priors; tile-aware cross-attention	Position-randomized training; spatial localization tasks	Spatially balanced datasets; annotations at varied image positions

reveal that accuracy alone cannot distinguish persistent model failures from predictions that change because of incidental image statistics. This distinction is particularly important for high-resolution inputs, for which detail size, resizing, and patch alignment can substantially affect the visual evidence available to the model.

Figure 6 demonstrates that benchmark errors are not necessarily persistent across image transformations. For Llama-3.2-11B-Vision, for example, approximately half of the original predictions are incorrect. However, only 11.4% of all samples remain incorrect under every tested condition, whereas 38.7% are answered correctly under at least one alternative transformation. Thus, most of this model's original errors are recoverable under the tested variations and cannot be attributed solely to an invariant inability to solve the corresponding questions. Conversely, many originally correct predictions become incorrect after a change in image statistics, revealing a substantial source of instability that single-resolution accuracy does not capture.

This analysis demonstrates that two models with similar VQA accuracy may differ substantially in both the persistence of their errors and the stability of their correct predictions. Accuracy alone therefore provides an incomplete account of model behavior under resolution changes. These observations motivate AUSC and PVS as complementary metrics: AUSC measures the preservation of initially correct predictions, whereas PVS characterizes variation among initially incorrect pre-

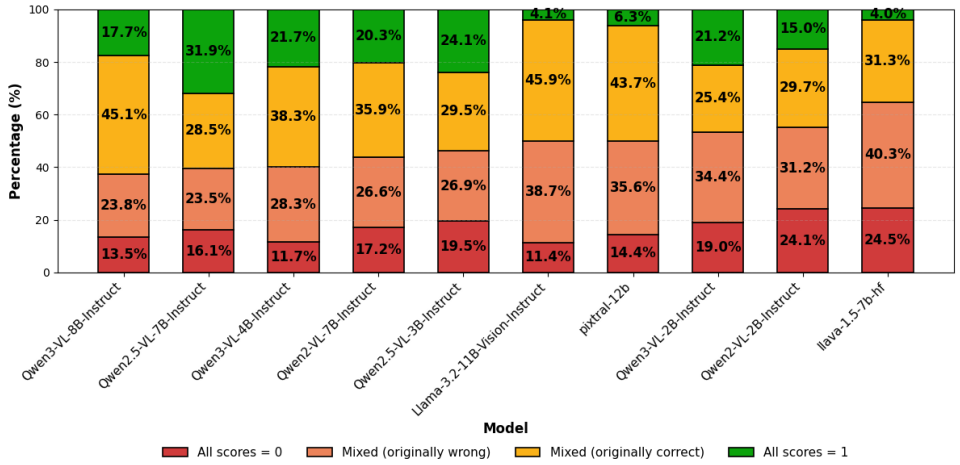

Figure 6: Distribution of prediction outcomes across the evaluated image transformations. Samples are divided into those answered incorrectly under all tested conditions (*all scores* $= 0$), those whose predictions vary and are incorrect for the original image (*mixed, originally wrong*), those whose predictions vary and are correct for the original image (*mixed, originally correct*), and those answered correctly under all tested conditions (*all scores* $= 1$). A large proportion of benchmark errors are therefore recoverable under an alternative, semantically equivalent image representation.

Table 4: Comparison of AUSC and PVS with existing robustness metrics. Higher values indicate better robustness for AUSC, whereas lower values indicate greater robustness for ACE, RCE, Flip$^+$, and Flip$^-$. For PVS, higher values indicate greater prediction variability among samples that are initially answered incorrectly.

Model	Acc_{orig}	AUSC	PVS	ACE	RCE	Flip$^+$	Flip$^-$
Qwen2.5-VL-7B-Instruct	0.604	0.863	0.368	0.054	0.091	0.086	0.067
Qwen2.5-VL-3B-Instruct	0.536	0.829	0.351	0.070	0.137	0.096	0.073
Qwen3-VL-4B-Instruct	0.600	0.827	0.457	0.086	0.151	0.109	0.079
Qwen3-VL-8B-Instruct	0.627	0.816	0.408	0.109	0.195	0.121	0.063
Qwen3-VL-2B-Instruct	0.466	0.816	0.409	0.080	0.174	0.090	0.085
Qwen2-VL-7B-Instruct	0.562	0.810	0.363	0.109	0.214	0.112	0.070
Qwen2-VL-2B-Instruct	0.448	0.793	0.326	0.078	0.191	0.097	0.072
Llama-3.2-11B-Vision-Instruct	0.499	0.669	0.477	0.173	0.429	0.173	0.083
LLaVA-1.5-7B-HF	0.352	0.666	0.326	0.102	0.348	0.123	0.060
Pixtral-12B	0.499	0.664	0.427	0.170	0.435	0.171	0.083

dictions. In the following section, we compare these metrics with existing robustness measures and show that they capture distinct properties of model behavior.

C.3 Comparison with Related Robustness Metrics

We compare our metrics with the most closely related recent proposals: ACE and RCE (Li et al., 2026) for AUSC, and Flip$^+$ and Flip$^-$ (Saxena et al., 2026) for PVS. The differences are not merely definitional: as shown in Table 4, the metrics can produce different model rankings because they capture distinct aspects of robustness to resolution changes.

AUSC versus ACE and RCE. ACE and RCE are computed over the complete evaluation set and quantify accuracy changes between adjacent resolution levels. Consequently, losses on samples that are initially answered correctly can be offset by gains on initially incorrect samples. AUSC instead conditions on initially correct samples and measures the retention of these predictions over

the complete resolution range. Thus, ACE and RCE characterize local changes in the accuracy curve, whereas AUSC captures overall preservation of initially correct predictions.

This distinction can change the resulting model ranking. For example, ACE ranks Qwen2-VL-2B as more robust than Qwen3-VL-8B (0.078 versus 0.109), whereas AUSC reverses this ordering (0.793 versus 0.816). Qwen3-VL-8B exhibits larger changes between adjacent resolutions but retains a greater proportion of its initially correct predictions over the full resolution range. Metrics that emphasize local variation therefore do not necessarily capture the global performance retention measured by AUSC.

PVS versus Flip rates. The principal difference between PVS and Flip^+/Flip^- is the normalization population. Flip rates are computed relative to the complete dataset and therefore depend on baseline accuracy: a model with lower initial accuracy has a larger pool of initially incorrect samples that may change their predictions. PVS instead normalizes by the number of initially incorrect samples, isolating their conditional prediction variability and making comparisons across models with different baseline accuracies more meaningful.

Table 4 illustrates the effect of this normalization. Flip^- identifies Qwen3-VL-2B as slightly more unstable than Llama-3.2-11B (0.085 versus 0.083), whereas PVS reverses the ordering (0.409 versus 0.477). Normalizing Flip^- by the fraction of initially incorrect samples,

$$\frac{\text{Flip}^-}{1 - \text{Acc}_{\text{orig}}},$$

produces the same reversal: the normalized values are 0.159 for Qwen3-VL-2B and 0.166 for Llama-3.2-11B. Hence, the higher unnormalized Flip^- value of Qwen3-VL-2B primarily reflects its lower baseline accuracy rather than greater conditional variability among initially incorrect samples.

Overall, these comparisons show that AUSC and PVS are not interchangeable with existing metrics. AUSC separates the preservation of initially correct predictions from compensating accuracy gains, while PVS separates conditional prediction variability from the prevalence of baseline errors. These properties are central to attributing resolution-induced failures to the behavior of a model rather than to differences in its initial accuracy.

C.4 Per-Benchmark Analysis of Resolution Sensitivity

Aggregate AUSC and PVS scores summarize model behavior across benchmarks but can conceal substantial variation between visual tasks. We therefore disaggregate both metrics by benchmark and transformation type. This analysis examines whether prediction stability depends on benchmark characteristics and whether different image transformations expose distinct failure modes.

Stability depends on benchmark characteristics. Figure 7 reveals pronounced benchmark-specific differences in the preservation of initially correct predictions. Across many of the Qwen models, AUSC is highest on MMEval, MMStar, and MMVet, while greater degradation is frequently observed on CharXiv and MathVerse. However, the magnitude of these differences remains model dependent; no benchmark is uniformly more or less stable across all architectures.

These results are consistent with the expected role of task-relevant detail size. Predictions should remain stable for questions determined by large visual structures or the overall scene because the relevant evidence remains accessible across a wider range of transformations. In contrast, questions whose answers depend on a small localized region are more sensitive to changes in resolution, interpolation, and patch alignment. Benchmark composition can therefore affect measured robustness even when the transformations preserve the semantic content of the image. Because the benchmarks are not explicitly annotated by the spatial extent of their task-relevant evidence, this interpretation should be viewed as a plausible explanation of the observed trend rather than a direct causal attribution.

Transformation discrepancies reveal model-specific behavior. The distance between transformation-specific curves provides further information about potential failure mechanisms. For Qwen2.5-VL-3B and Qwen2.5-VL-7B, the AUSC curves largely overlap, suggesting

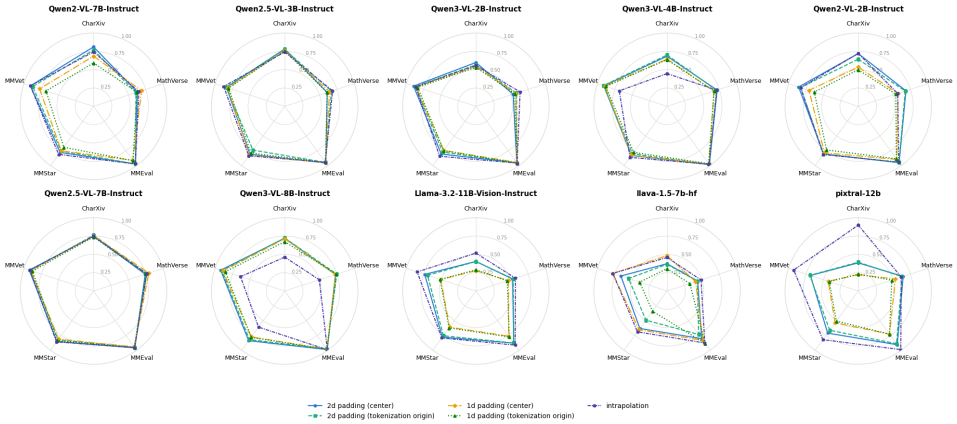

Figure 7: AUSC disaggregated by benchmark and transformation type. Each panel corresponds to one model, and each axis to one benchmark. Higher values indicate stronger retention of predictions that are correct on the original images. The five curves represent centered and tokenization-aligned 1D and 2D padding, as well as interpolation.

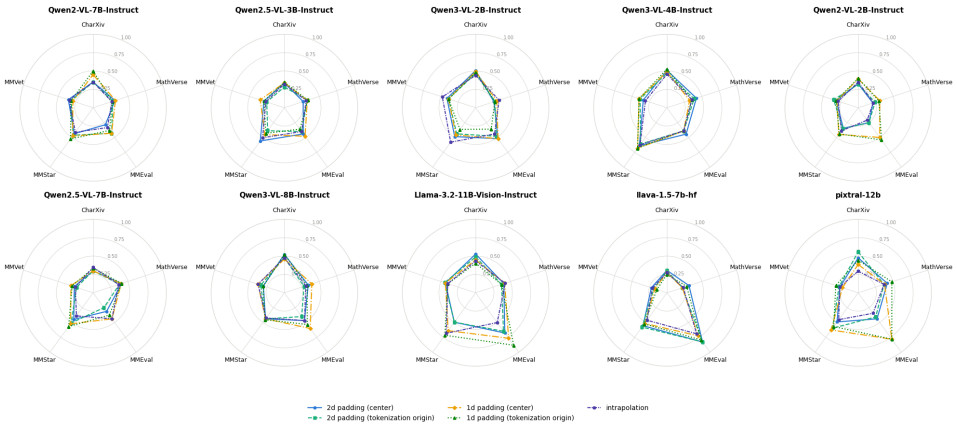

Figure 8: PVS disaggregated by benchmark and transformation type. Higher values indicate greater prediction variability among samples that are answered incorrectly on the original images. Differences between transformation types reveal that the recoverability of these errors depends on how the image representation is modified.

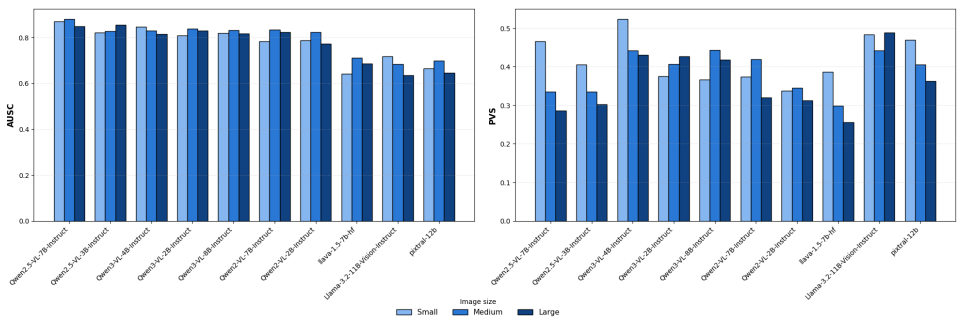

Figure 9: AUSC and PVS stratified by original image size. AUSC varies only modestly across size strata, whereas PVS generally decreases from small to large images, indicating stronger transformation sensitivity among initially incorrect predictions on smaller inputs. This trend is model dependent and is not observed uniformly across architectures.

that degradation is relatively insensitive to the particular transformation used. Other models exhibit substantially larger discrepancies. Most notably, interpolation produces considerably lower AUSC than padding for Qwen3-VL-8B on several benchmarks, whereas the opposite pattern appears for Pixtral-12B, for which interpolation is often substantially less damaging. LLaVA-1.5 also exhibits large differences between centered and tokenization-aligned padding on several benchmarks.

The absence of a universally most damaging transformation indicates that these effects cannot be explained by input size alone. Instead, they likely reflect interactions among the transformation, the model's image preprocessing pipeline, its tokenization or patchification strategy, and the visual encoder architecture. Consequently, evaluating only one type of resolution-preserving transformation may overlook model-specific vulnerabilities.

PVS exposes asymmetric behavior on initially incorrect samples. The per-benchmark PVS results in Figure 8 do not simply mirror the AUSC patterns. A model may preserve its initially correct predictions while simultaneously exhibiting considerable variability among its initially incorrect predictions. This behavior is especially visible on MMEval: several models achieve high AUSC on this benchmark but also show high PVS under particular transformations. Thus, robustness of correct predictions and recoverability of incorrect predictions represent distinct properties and should be analyzed separately.

PVS also exhibits pronounced transformation-specific asymmetries. For example, the variability of Llama-3.2-11B-Vision on MMEval is substantially higher under tokenization-aligned 1D padding than under interpolation. Pixtral-12B shows a similarly large separation on MMEval, with 1D padding entailing markedly higher PVS than interpolation. In contrast, several Qwen models display comparatively aligned PVS curves, indicating less dependence on the precise transformation. These asymmetries suggest that an image modification can recover initially incorrect predictions for one architecture while having little effect for another.

Overall, the per-benchmark results show that resolution sensitivity is jointly determined by the benchmark, model, and image transformation. Aggregate metrics remain useful for ranking models, but disaggregated AUSC and PVS expose where instability occurs and whether it primarily affects initially correct or initially incorrect predictions.

C.5 Per Image Size Analysis

Figure 9 shows that AUSC is comparatively stable across image-size strata, with no consistent monotonic relationship between original image size and the retention of initially correct predictions. PVS exhibits a stronger size dependence: for seven of the ten models, PVS is lower for large than for small images, with particularly pronounced decreases for Qwen2.5-VL-7B, LLaVA-1.5, and Pixtral-12B. This suggests that initially incorrect predictions on smaller images are more likely to change under the evaluated transformations. The exceptions, including Qwen3-VL-2B and Llama-3.2-11B-Vision, show that this effect remains architecture dependent. Overall, original image size explains

more variation in the behavior of initially incorrect predictions than in the preservation of initially correct ones.

D Human Review of Augmented Images

To verify that the evaluated transformations preserve semantic content, we conducted a human validation review of a subset of transformed images. We randomly sampled images from all benchmarks and applied the same padding and bicubic interpolation procedures used in the main experiments. Human annotators were shown the original and transformed images side-by-side and asked whether the transformation altered the semantic interpretation required to answer the task correctly.

Across all evaluated samples, annotators judged the vast majority of padded and interpolated images as semantically equivalent to the originals. Padding transformations were consistently rated as preserving semantic content, while bicubic interpolation occasionally introduced mild blur but rarely changed task-relevant information. In contrast, preliminary experiments with super-resolution methods produced noticeable hallucinations and texture artifacts, particularly in fine-grained regions, and were therefore excluded from the final evaluation.

These results support our assumption that the transformations used in the paper primarily modify resolution characteristics and information density while preserving the underlying semantic content of the image.

E Detailed Metric Specification

E.1 Area Under the Scaling Curve (AUSC)

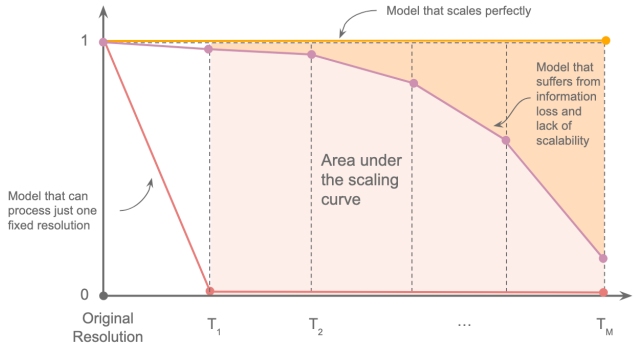

Figure 10: Illustration of the Area Under the Scaling Curve (AUSC) metric. AUSC is computed on $\mathcal{B}_{\text{correct}}$ and measures the area under the accuracy-vs-scale curve, normalized to $[0, 1]$.

Let $\mathcal{B}_{\text{correct}} \subseteq \mathcal{B}$ denote the subset of image-question pairs for which the model provides the correct answer at the original resolution:

$$\mathcal{B}_{\text{correct}} = \{(x_i, y_i) \in \mathcal{B} \mid f(x_i, y_i) = z_i\} \tag{5}$$

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ be a set of image transformations as described in Section 3.2, where M represents the number of transformation levels (e.g., $M = 8$ for padding with margins from $0.25\times$ to $2.0\times$ image size in $0.25\times$ increments). For each transformation T_j, we construct a transformed benchmark subset:

$$\mathcal{B}^{(j)}_{\text{correct}} = \{(T_j(x_i), y_i) \mid (x_i, y_i) \in \mathcal{B}_{\text{correct}}\} \tag{6}$$

For each size of the margin $j = 1, 2, ..., M$, we evaluate the model $f(x)$ based on the chosen metric and evaluation protocol (for example, VQA accuracy).

Given the sequence of accuracies $[\text{Acc}(T_1), \text{Acc}(T_2), \ldots, \text{Acc}(T_M)]$, we calculate the area under the scaling curve using the trapezoidal rule:

$$\text{AUSC} = \frac{1}{M-1} \sum_{j=1}^{M-1} \frac{1}{2} \Big[\text{Acc}(T_j) + \text{Acc}(T_{j+1}) \Big] \tag{7}$$

Interpretation. AUSC values lie in $\in [0, 1]$ and have an intuitive interpretation:

- **AUSC** $= 1$: The model scales perfectly under the set of transformations $\mathcal{T}$. This indicates *maximum robustness*.
- **AUSC** $\in (0, 1)$: Model architecture and image preprocessing steps incur visual information loss. Higher AUSC values reflect smoother degradation and more stable performance across transformations. Lower AUSC values indicate that the model is sensitive to certain transformations and quickly degrades.
- **AUSC** $= 0$: The model can process just one fixed resolution and fails to recognize any information in images of higher resolutions. This indicates *complete brittleness*.

Statistical Significance. It is important to consider that the AUSC has a dependence on the size of $\mathcal{B}_{\text{correct}}$. For the purpose of statistical significance, we recommend $|\mathcal{B}_{\text{correct}}| \geq 75$.

E.1.1 Prediction Variance Score (PVS)

For additional insight, we assess how variable the predictions are under the set of transformations $\mathcal{T}$ when considering instances that the model originally did not answer correctly. This reveals whether resolution changes can, in effect, repair incorrect predictions, indicating sensitivity to tokenization or visual encoding artifacts.

Let $\mathcal{B}_{\text{wrong}} = \mathcal{B} \setminus \mathcal{B}_{\text{correct}}$ denote examples, on which the model initially yields incorrect predictions. For each transformation T_j, we calculate the Prediction Variance Score as

$$\text{PVS}(T_j) = \frac{1}{|\mathcal{B}_{\text{wrong}}|} \sum_{(x_i, y_i) \in \mathcal{B}_{\text{wrong}}} \mathbb{1}\big[f(T_j(x_i), y_i) = z_i \big], \tag{8}$$

where $\mathbb{1}$ denotes an indicator function, in this case assessing the correctness of predictions.

Interpretation. This metric captures how often the transformation T_j causes the model to produce the correct answer on an image-question pair for which the model produced an incorrect prediction at the original resolution. The PVS lies in $[0, 1]$:

- **PVS** $= 0$: The model is invariant toward shifts and size changes of visual information. Incorrect predictions remain incorrect regardless of resolution.
- **PVS** $\in (0, 1)$: Resolution changes flip predictions from incorrect to correct. For example, $\text{PVS}(T_j) = 0.3$ means that 30% of originally incorrect answers became correct after applying transformation T_j.
- **PVS** $= 1$: All incorrect predictions become correct after transformation (extremely rare and would indicate severe resolution bias).

Diagnostic Value. High PVS values may signal:

- **Potential data augmentation benefits:** The vision encoder could benefit from training with similar transformations
- **Lack of robustness:** Shifts in model attention or reasoning pathways, indicating brittleness toward image transformations
- **Tokenization artifacts:** Token boundary alignment critically affects model predictions

Similarly to AUSC, we recommend to choose sufficiently large and challenging benchmarks with $|\mathcal{B}_{\text{wrong}}| \geq 75$.

F LLM-AS-A-JUDGE EVALUATION

To evaluate semantic equivalence between model predictions and ground-truth answers across heterogeneous benchmarks, we employ an LLM-as-a-judge protocol using `GPT-5-nano`. The judge model receives the question, the ground-truth answer and the model prediction and outputs a binary correctness score.

The following prompt was used for all evaluations:

```
Compare two answers to the question and output 1 if the
predicted answer contains the same information as the
ground truth. Output 0 if some of the information is
missing or the answer is completely wrong.

ex1:
Ground truth: "a"
Prediction: "the style depicted in this image is a:
impressionism."
Score: 1

ex2:
Ground truth: "$\frac(1)(2000)$"
Prediction: "the average rate of change in the graph
is 5."
Score: 0

Is the predicted answer correct?

Question: {question}
Ground truth: {ground_truth}
Prediction: {prediction}

Output only the numeric score (0 or 1)
```

The binary outputs produced by the judge model were used as the final correctness labels in all experiments requiring open-form answer matching.

To validate the reliability of the automatic evaluation procedure, we manually reviewed 500 randomly sampled answer pairs spanning all benchmarks and transformation settings. Human annotations were compared against the `GPT-5-nano` judgments, yielding an agreement accuracy of over **97%**. These results indicate that the automated evaluation protocol provides a reliable approximation of human judgment.

We must note that this type of evaluation is sensitive to the used prompt and the choice of the judge model.

G COMPUTE RESOURCES

All experiments and evaluations were conducted using a compute server equipped with 4×NVIDIA RTX A6000 GPUs. The evaluation pipeline is also compatible with a single-GPU setup, although runtimes increase proportionally due to the large number of evaluated transformations and model inferences.