Does Acceleration Cause Hidden Instability in Vision Language Models? Uncovering Instance-Level Divergence Through a Large-Scale Empirical Study

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) are powerful yet computationally intensive for widespread practical deployments. To address such challenge without costly re-training, post-training acceleration techniques like quantization and token reduction are extensively explored. However, current acceleration evaluations primarily target minimal overall performance degradation, overlooking a crucial question: does the accelerated model still give the same answers to the same questions as it did before acceleration? This is vital for stability-centered industrial applications where consistently correct answers for specific, known situations are paramount, such as in AI-based disease diagnosis. We systematically investigate this for accelerated VLMs, testing four leading models (LLaVA-1.5, LLaVA-Next, Qwen2-VL, Qwen2.5-VL) with eight acceleration methods on ten multimodal benchmarks. Our findings are stark: despite minimal aggregate performance drops, accelerated models changed original answers up to 20% of the time. Critically, up to 6.5% of these changes converted correct answers to incorrect. Input perturbations magnified these inconsistencies, and the trend is confirmed by case studies with the medical VLM LLaVA-Med. This research reveals a significant oversight in VLM acceleration, stressing an urgent need for instance-level stability checks to ensure trustworthy real-world deployment.

1 Introduction

Large Vision-Language Models (VLMs) are demonstrating remarkable capabilities in understanding and generating content across visual and textual modalities (Liu et al., 2024a,b; Bai et al., 2025). Despite their impressive performance, the substantial computational demands of state-of-the-art VLMs critically limit their practical deployment, particularly in resource-constrained environments (Chen et al., 2024; Zhang et al., 2025; Tang et al., 2024). To mitigate these challenges without the necessity of costly re-training, post-training acceleration techniques—such as quantization (Lin et al., 2024; Frantar et al., 2022; Dettmers et al., 2022) and token reduction (Chen et al., 2024; Yang et al., 2024c; Xing et al., 2024)—are widely adopted. The primary objectives of these techniques have been two-fold: achieving substantial computational efficiency gains while ensuring minimal degradation in aggregate performance metrics. Yet, this prevailing focus obscures other vital impacts of acceleration, posing the question: Are these two criteria truly sufficient to guarantee the reliable deployment of accelerated VLMs in practice?

We contend that for many practical applications, particularly in critical domains like medicine (Zhang et al., 2023; Li et al., 2023a), the answer is highly risky to be "No". In such fields, system development and validation often adhere to a "case-

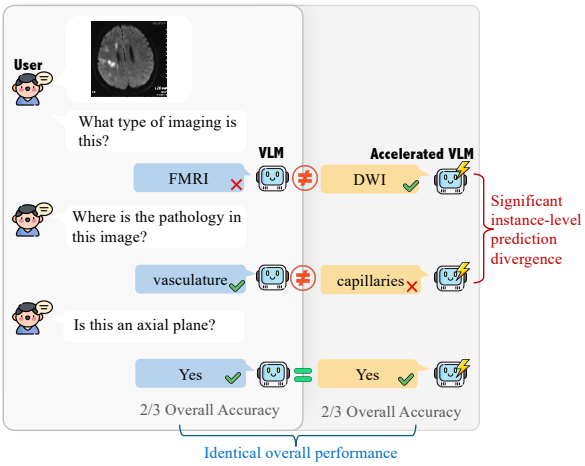

Figure 1: Current VLM acceleration methods focus on improving efficiency while minimizing overall performance drop relative to the base model. However, this focus may obscure a critical risk: accelerated models can exhibit significant changes in instance-level predictions compared to their original counterparts. Such instability poses serious concerns in sensitive domains such as healthcare, where producing stable and reliable outputs is essential.

driven" paradigm (Bodendorf, 2025; Liao and Xiao, 2023; Weidinger et al., 2025), where a fundamental requirement is the AI system's ability to consistently and correctly resolve specific, known crucial instances, even post-optimization or updates. Consider a medical VLM adept at identifying a rare disease from patient scans; it is paramount that this specific diagnostic capability remains invariant after an acceleration process aimed at enhancing efficiency. However, as illustrated in Figure 1, this crucial aspect of instance-level stability is largely unaddressed within the evaluation of current acceleration methodologies (Lin et al., 2024; Frantar et al., 2022; Chen et al., 2024; Yang et al., 2024c), potentially masking significant operational risks.

This paper confronts this oversight by systematically investigating the instance-level stability of accelerated VLMs. Our central aim is to evaluate whether and to what extent existing post-training acceleration techniques, despite ostensibly preserving overall performance, can induce substantial and often detrimental inconsistencies in models' response to individual inputs. To rigorously quantify this instability, we introduce two intuitive yet powerful metrics: Divergence Ratio (DR) and Negative Divergence Ratio (NDR). DR measures the frequency with which an accelerated model yields a different prediction for the same input compared to its original, unaccelerated counterpart. NDR quantifies a more critical failure mode: the proportion of instances where a correct prediction from the original model becomes incorrect after acceleration. Low DR and NDR values signify that an accelerated VLM maintains behavioral fidelity and reliability. Conversely, high values—even when accompanied by negligible shifts in aggregate performance—would indicate that the accelerated model's behavior has become alarmingly unpredictable relative to its original state.

To empirically validate our hypothesis, we undertook an extensive study. We assessed eight distinct acceleration methods applied to four leading open-source VLMs (LLaVA-1.5 (Liu et al., 2024a), LLaVA-Next (Liu et al., 2024b), Qwen2-VL (Wang et al., 2024), and Qwen2.5-VL (Bai et al., 2025)) across ten diverse multi-modal benchmarks. To probe the resilience of instance-level stability under practical conditions, we further evaluated model performance on perturbed inputs (spanning both visual and textual modalities) designed to mimic real-world data variations. Underscoring the high stakes involved, we conducted targeted case studies on LLaVA-Med (Li et al., 2023a), a VLM tailored for medical applications where predictive consistency is non-negotiable. Our experiments reveal several striking findings:

1. Despite acceleration methods inducing only a negligible drop in overall performance (average of 0.8%), they precipitated surprisingly high Divergence Ratios (DR) of up to 20% and, more critically, Negative Divergence Ratios (NDR) reaching up to 7%.

2. Input data perturbations, characteristic of real-world scenarios, further exacerbated this divergence.

3. Application of acceleration to the medical VLM (LLaVA-Med) corroborated these high DR and NDR values, highlighting the acute potential risks in safety-critical domains.

To the best of our knowledge, this work represents the first large-scale empirical investigation dedicated to the instance-level stability of VLM acceleration techniques. Our research uncovers a significant, potentially hazardous, oversight in current VLM acceleration practices, emphasizing an urgent imperative for incorporating rigorous instance-level stability checks to ensure these models are genuinely faithful and trustworthy for real-world deployment.

2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (VLMs) have advanced rapidly in integrating visual and textual understanding. Early models like CLIP (Radford et al., 2021) employed contrastive learning to align these modalities. Subsequent architectures, such as BLIP-2 and Instruct-BLIP (Li et al., 2023b; Dai et al., 2023), introduced Q-Former to bridge pre-trained vision encoders with Large Language Model (LLM) backbones. More recent state-of-the-art models, including LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), and the Qwen-VL series (Wang et al., 2024; Bai et al., 2025), leverage powerful LLMs (e.g., Vicuna, LLaMA, Qwen2 (Peng et al., 2023; Dubey et al., 2024; Yang et al., 2024a)) and lightweight vision-text connectors (typically linear layers) for advanced multimodal reasoning. However, VLM vision encoders often generate a high volume of visual tokens (hundreds or thousands (Radford et al.,

2021)). The LLM backbone processing of these numerous tokens incurs significant computational costs, hindering the practical deployment of VLMs.

2.2 Post-Training Acceleration Techniques for Vision-Language Models

Post-training acceleration techniques are widely applied to reduce computational demands of VLMs without costly retraining. Token Reduction methods aims to substantially remove the redundant visual tokens for VLMs, thereby reducing the input sequence length and lowering inference costs. Recent methods implementing this approach during inference include VisionZip (Yang et al., 2024c), PyramidDrop (Xing et al., 2024), FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024), and HiRed (Arif et al., 2025). Quantization techniques reduces model size and computational overhead by utilizing lower-precision numerical formats (e.g., 8-bit, 4-bit) for model weights and/or activations. Post-Training Quantization (PTQ), which applies this technique after model training, has become a common practice, such as LLM.int8() (Dettmers et al., 2022), GPTQ (Frantar et al., 2022), and AWQ (Lin et al., 2024). Although these methods often report minimal degradation on standard benchmarks, their impact on instance-level stability remains largely unexplored. This work systematically investigates the instance-level prediction stability of VLMs under both token reduction and quantization, moving beyond standard benchmark evaluations.

2.3 Evaluation for LM Acceleration

The typical approach to evaluating model acceleration techniques tends to emphasize negligible loss in aggregate performance and improved computational efficiency. However, there's a growing recognition that such criteria, while important, may overlook other critical impacts. Recent investigations, for example, have shown that quantization can diminish the reasoning capabilities of LLMs (Li et al., 2025), and that prompt compression can affect their ability to retain information (Lajewska et al., 2025). Similarly, Dutta et al. (2024) demonstrates that accuracy alone is not enough for assessing LLM quantization, leading to proposals like the "flip" metric for instance-level changes. Wen et al. (2025) argues that the fundamental designs of token reduction methods for VLMs can cause biased performance on different task types. Moreover, a specialized benchmark, LLMCBench, has been introduced targeting the practical efficiency of model compression techniques for real-world deployment (Yang et al., 2024b). Distinct from these explorations, our work concentrates on a crucial aspect: the instance-level stability and reliability of accelerated VLMs, ensuring they consistently solve the problems they were initially capable of solving.

3 Experimental Settings

3.1 Tasks and Datasets

We utilize a diverse suite of ten benchmark datasets covering various Visual-Language understanding capabilities. These include AI2D (Kembhavi et al., 2016) for diagram understanding, GQA (Hudson and Manning, 2019) for real-world compositional reasoning, MMBench (Liu et al., 2024c) for diverse multi-modal abilities, MMMU (Yue et al., 2024) for expert-level multi-discipline reasoning, OK-VQA (Marino et al., 2019) requiring external knowledge, POPE (Li et al., 2023c) for evaluating object hallucination, ScienceVQA (Lu et al., 2022) focusing on science diagrams, TextVQA (Singh et al., 2019) requiring reading text within images, VizWiz (Gurari et al., 2018) using images from visually impaired users, and the widely-used large-scale VQA benchmark VQAv2 (Goyal et al., 2017). Finally, we use VQA-RAD (Lau et al., 2018) to extend to medical domain tasks. Details of the benchmarks are presented in Appendix A.

3.2 Base Models and Acceleration Techniques

We select four state-of-the-art open-source VLMs as base models for our acceleration experiments. LLaVA-1.5 (Liu et al., 2024a) is a widely recognized VLM demonstrating strong general vision-language capabilities. LLaVA-Next (Liu et al., 2024b) extends LLaVA-1.5, improving performance particularly for high-resolution inputs. Qwen2-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025) are recent released VLMs, which are adept at handling various image resolutions and video inputs. Additionally, we also use LLaVA-Med (Li et al., 2023a), which is a specialised medical domain VLM. We adopt the 7B model size for all VLMs throughout our study, unless stated otherwise.

We investigate two main categories of post-training acceleration: token reduction and quantization. For token reduction, we evaluate five of the latest and widely applied methods, including

VisionZip (Yang et al., 2024c), which selects informative tokens and merges others; PyramidDrop (Xing et al., 2024), which progressively drops tokens in deeper layers; SparseVLMs (Zhang et al., 2024), which prunes tokens based on relevance scores; FastV (Chen et al., 2024), dynamically pruning based on attention scores during inference; and HiRed (Arif et al., 2025), designed for high-resolution inputs, allocating token budgets based on attention. For all the token reduction methods, we choose the signature or best-performing hyper-parameter settings as reported in the corresponding papers, which are listed in Appendix B. For Quantization, which reduces numerical precision, we apply: llm.int8() (Dettmers et al., 2022) (W8A16), a mixed-precision quantisation scheme; AWQ (Lin et al., 2024) (W4A16), an activation-aware 4-bit weight quantization; and GPTQ (Frantar et al., 2022) (W4A16), a layer-wise 4-bit weight quantization method.

3.3 Evaluation Metrics

We report standard top-1 accuracy for all tasks except for POPE, where F1 score is the standard metric. We also calculate the Accuracy or F1 Drop for all the acceleration methods compared with the corresponding baseline VLMs. To assess the instance-level instability of accelerated models compared to their original counterparts, we introduce two additional metrics: **1) Divergence Ratio (DR)**, defined as the proportion of test samples where the accelerated model's prediction differs from the original model's prediction, irrespective of correctness. **2) Negative Divergence Ratio (NDR)**, which quantifies harmful instability by measuring the proportion of samples that were correctly predicted by the original model but incorrectly predicted by the accelerated model.

3.4 Input Perturbation

To better understand the instance-level stability of accelerated VLMs under practical settings, we adopt a comprehensive set of input perturbation methods proposed by Chen et al. (2023) to simulate the real-world user scenarios. Specifically, we use 96 visual perturbation methods (e.g. noise, blur, weather effects) and 87 textual perturbation methods (e.g., typos, paraphrasing, character substitutions), whose details are shown in Appendix C. We apply these visual and textual perturbations separately to the inputs of the accelerated models and assess their impact on performance and prediction stability.

4 Experimental Results

This section presents our empirical findings on the instance-level stability of accelerated Vision-Language Models (VLMs). Our experiments are structured in three stages:

1. We first evaluate Divergence Ratios (DR) and Negative Divergence Ratios (NDR) for widely used post-training acceleration methods (Token Reduction and Quantization) on standard benchmarks in section 4.1. This establishes their fundamental impact on instance-level stability under laboratory conditions.

2. Next, we further assess the instance-level stability under more realistic conditions by applying input perturbations to large-scale Visual Question Answering (VQA) benchmarks (VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019)), simulating typical input noise encountered in practice as discussed in section 4.2.

3. Finally, we analyze an accelerated medical VLM to demonstrate the potential downstream consequences and critical risks of instance-level instability in a high-stakes domain in section 4.3.

4.1 Instance-Level Instability on Standard Benchmarks

This section presents our quantitative findings on the instance-level stability of various post-training acceleration techniques applied to leading Vision-Language Models (VLMs). The detailed results for token reduction techniques are shown in Table 1 and those for quantization methods are summarized in Table 2. Qualifying examples are demonstrated in appendix G.

The Illusion of Stability: High Divergence Despite Low Aggregate Performance Drops. The most striking revelation from our experiments is the significant instance-level instability introduced by many common acceleration methods, even when these methods exhibit only minimal degradation in overall aggregate performance. This creates an illusion of stability if one only considers coarse-grained metrics. Across multiple VLMs and benchmarks, we consistently observed that **accelerated**

Method	Metric	VQAv2	AI2D	GQA	ScienceQA Img	TextVQA	OKVQA	VizWiz	MMBench	MMMU	POPE	Average
LLaVA-1.5 *(Baseline)*	Acc/F1 (%) ↑	76.64	55.25	61.92	69.46	46.09	53.44	54.05	64.09	36.22	85.85	60.30
PyramidDrop *(CVPR 2025)*	Acc/F1 (%) ↑	75.93	54.60	60.05	69.21	45.32	52.67	53.19	64.60	37.22	84.81	59.76
	Acc/F1 Drop (%) ↓	0.71	0.65	1.87	0.25	0.76	0.77	0.86	-0.52	-1.00	1.03	**0.54**
	DR (%) ↓	8.97	12.82	11.34	5.75	18.68	10.64	11.79	4.87	10.89	3.31	**9.91**
	NDR (%) ↓	2.29	4.21	4.72	2.33	2.56	3.23	3.17	1.82	2.33	2.04	**2.87**
SparseVLM *(ICML 2025)*	Acc/F1 (%) ↑	75.26	54.83	59.37	68.77	44.88	51.93	54.82	64.00	35.78	85.59	59.52
	Acc/F1 Drop (%) ↓	1.38	0.42	2.54	0.69	1.21	1.51	-0.77	0.09	0.44	0.26	**0.78**
	DR (%) ↓	12.10	14.86	13.97	6.64	23.74	12.33	15.72	6.56	11.78	4.49	**12.22**
	NDR (%) ↓	3.33	4.83	5.81	2.93	3.48	4.12	3.52	2.66	2.89	2.33	**3.59**
VisionZip *(CVPR 2025)*	Acc/F1 (%) ↑	74.93	55.73	59.13	68.57	44.66	52.72	54.06	63.66	36.67	85.26	59.54
	Acc/F1 Drop (%) ↓	1.71	-0.49	2.78	0.89	1.43	0.72	-0.01	0.43	-0.45	0.59	**0.76**
	DR (%) ↓	13.10	17.71	14.46	7.98	24.34	13.91	19.17	7.55	14.89	4.89	**13.80**
	NDR (%) ↓	3.64	5.54	6.15	3.42	3.76	4.02	4.79	3.16	3.44	2.74	**4.07**
FastV *(ECCV 2024)*	Acc/F1 (%) ↑	75.86	55.38	60.09	68.86	45.91	52.95	54.51	64.18	35.89	82.47	59.61
	Acc/F1 Drop (%) ↓	0.77	-0.13	1.83	0.59	0.18	0.49	-0.47	-0.09	0.33	3.38	**0.69**
	DR (%) ↓	7.88	6.80	10.85	2.88	13.40	5.87	8.71	3.47	4.89	5.22	**7.00**
	NDR (%) ↓	2.17	1.75	4.41	1.39	1.58	1.70	1.90	1.27	1.44	3.82	**2.14**
HiRED *(AAAI 2025)*	Acc/F1 (%) ↑	76.51	53.50	61.23	67.97	48.58	53.39	53.54	62.97	35.60	84.06	59.74
	Acc/F1 Drop (%) ↓	0.12	1.75	0.69	1.49	-2.50	0.05	0.51	1.12	0.62	1.79	**0.56**
	DR (%) ↓	14.72	17.65	13.26	11.40	42.58	12.62	20.38	6.91	35.33	7.42	**18.23**
	NDR (%) ↓	3.50	6.35	4.47	5.06	7.40	3.23	5.16	3.28	10.00	4.43	**5.29**
LLaVA-Next *(Baseline)*	Acc/F1 (%) ↑	80.06	65.32	64.26	70.25	64.82	44.23	60.74	67.10	36.67	86.41	63.98
PyramidDrop *(CVPR 2025)*	Acc/F1 (%) ↑	79.46	64.31	63.38	69.16	64.18	46.07	61.16	66.84	36.56	86.60	63.77
	Acc/F1 Drop (%) ↓	0.60	1.00	0.88	1.09	0.64	-1.85	-0.42	0.26	0.11	-0.18	**0.21**
	DR (%) ↓	8.18	10.04	10.44	5.95	17.22	10.19	9.86	4.30	11.33	3.06	**9.06**
	NDR (%) ↓	1.97	3.79	3.96	2.78	3.40	1.78	2.36	1.71	2.00	1.41	**2.52**
SparseVLM *(ICML 2025)*	Acc/F1 (%) ↑	78.42	64.83	62.89	68.27	62.29	43.99	59.75	66.67	37.11	87.00	63.12
	Acc/F1 Drop (%) ↓	1.63	0.49	1.38	1.98	2.53	0.24	0.99	0.43	-0.44	-0.58	**0.86**
	DR (%) ↓	11.83	12.21	13.48	7.54	21.84	11.40	10.14	5.43	13.11	3.80	**11.08**
	NDR (%) ↓	3.28	4.34	5.20	3.57	5.00	3.37	3.36	1.89	2.11	1.63	**3.38**
VisionZip *(CVPR 2025)*	Acc/F1 (%) ↑	78.34	64.80	61.89	68.22	62.94	46.06	60.49	65.46	36.56	87.24	63.20
	Acc/F1 Drop (%) ↓	1.72	0.52	2.37	2.03	1.88	-1.84	0.25	1.63	0.11	-0.82	**0.78**
	DR (%) ↓	12.28	14.80	14.03	10.36	20.76	14.11	13.36	7.88	18.67	3.71	**13.00**
	NDR (%) ↓	3.44	5.02	6.01	4.96	4.56	2.95	3.43	3.47	3.78	1.53	**3.91**
FastV *(ECCV 2024)*	Acc/F1 (%) ↑	79.63	64.51	63.87	69.11	63.88	43.69	60.25	66.49	35.22	86.19	63.28
	Acc/F1 Drop (%) ↓	0.43	0.81	0.39	1.14	0.94	0.54	0.49	0.60	1.45	0.23	**0.70**
	DR (%) ↓	5.77	7.16	6.34	3.12	11.80	4.40	4.17	2.22	5.56	2.01	**5.25**
	NDR (%) ↓	1.44	2.78	2.31	1.69	2.48	1.55	1.30	0.97	1.44	1.06	**1.70**
HiRED *(AAAI 2025)*	Acc/F1 (%) ↑	77.57	62.05	61.33	67.97	61.54	46.70	58.53	65.38	36.22	85.10	62.24
	Acc/F1 Drop (%) ↓	2.49	3.27	2.93	2.28	3.28	-2.47	2.20	1.72	0.45	1.31	**1.75**
	DR (%) ↓	15.40	21.96	17.47	11.85	25.78	23.07	20.44	10.90	25.00	5.20	**17.71**
	NDR (%) ↓	4.43	9.46	7.60	5.80	6.36	4.80	6.58	4.46	5.67	3.04	**5.82**

Table 1: **Instance-Level Instability in Token Reduction Methods.** For each acceleration method, we report: Accuracy (Acc) for most benchmarks (F1 score for POPE (Li et al., 2023c)), Acc/F1 drop (performance degradation vs. baseline), Divergence Ratio (DR), and Negative Divergence Ratio (NDR) to evaluate instance-level prediction changes. Red values indicate the largest NDR per baseline model within each benchmark column. Across all benchmarks and token reduction methods, results reveal high DR and NDR values despite negligible Acc/F1 drops, signifying considerable instance-level prediction instability.

models altered their original predictions on identical inputs up to 20% of the time (DR), a concerning level of divergence. More critically, our findings indicate that **up to 6.5% of these changes converted previously correct answers into incorrect ones (NDR)**, directly undermining the model's reliability on specific, previously solved cases.

Instance-Level Instability in Token Reduction Methods. Our investigation into token reduction techniques reveals substantial instance-level instability (Table 1). The HIRED method, for example, when applied to LLaVA-1.5 and LLaVA-Next, caused minimal average aggregate performance drops (∼0.2-0.6%) but still led to high average DRs of ∼18% and average NDRs approaching 6%. Specific benchmarks under this method saw NDRs reach up to 9-10% and DRs over 25%. Other token reduction techniques like VisionZip and SparseVLM similarly produced notable DR and NDR values (e.g., average DRs often exceeding 12-13%) despite their modest impact on overall accuracy scores. Since the Qwen-VL model series (Wang et al., 2024; Bai et al., 2025) already features integrated token compression modules, we do not separately evaluate the impact of external token reduction methods.

Instance-level Instability in Quantization Methods. The phenomenon of high instance-level in-

Method	Metric	VQAv2	AI2D	GQA	ScienceQA Img	TextVQA	OKVQA	VizWiz	MMBench	MMMU	POPE	Average
LLaVA-1.5 *(Baseline)*	Acc/F1 (%) ↑	76.64	55.25	61.92	69.46	46.09	53.44	54.05	64.09	36.22	85.85	60.30
AWQ (W4A16)	Acc/F1 (%) ↑	76.23	53.24	60.92	67.87	48.40	53.71	50.87	62.71	35.89	83.93	59.38
	Acc/F1 Drop (%) ↓	0.41	2.01	1.00	1.59	-2.31	-0.27	3.18	1.37	0.33	1.92	**0.92**
	DR (%) ↓	16.32	20.76	14.84	14.48	44.92	17.72	25.28	9.45	20.44	7.44	**19.17**
	NDR (%) ↓	3.93	7.42	5.17	6.49	7.84	4.10	8.15	4.02	5.33	4.49	**5.69**
GPTQ (W4A16)	Acc/F1 (%) ↑	75.77	51.68	60.86	66.73	48.21	48.63	54.53	62.46	34.33	85.31	58.85
	Acc/F1 Drop (%) ↓	0.87	3.56	1.06	2.73	-2.12	4.81	-0.48	1.63	1.89	0.54	**1.45**
	DR (%) ↓	17.10	23.19	15.84	16.91	45.64	22.06	25.19	9.84	22.89	8.54	**20.72**
	NDR (%) ↓	4.34	9.13	5.49	7.54	8.06	8.68	6.34	4.16	7.11	4.68	**6.55**
LLM.Int8()	Acc/F1 (%) ↑	76.52	55.47	62.04	69.31	45.97	53.35	54.21	64.35	36.33	85.36	60.29
	Acc/F1 Drop (%) ↓	0.11	-0.23	-0.13	0.15	0.12	0.10	-0.16	-0.26	-0.11	0.49	**0.01**
	DR (%) ↓	2.89	6.35	2.88	3.87	7.46	5.35	5.14	3.00	9.44	0.69	**4.71**
	NDR (%) ↓	0.62	1.85	0.88	1.59	0.90	1.31	1.07	1.04	2.33	0.53	**1.21**
LLaVA-Next *(Baseline)*	Acc/F1 (%) ↑	80.06	65.32	64.26	70.25	64.82	44.23	60.74	67.10	36.67	86.41	63.98
AWQ (W4A16)	Acc/F1 (%) ↑	79.80	64.57	63.53	69.61	64.40	43.89	60.22	66.49	36.89	86.57	63.60
	Acc/F1 Drop (%) ↓	0.26	0.74	0.73	0.64	0.42	0.33	0.52	0.60	-0.22	-0.16	**0.39**
	DR (%) ↓	5.84	9.07	5.54	7.24	11.38	10.70	7.27	5.22	16.33	1.33	**7.99**
	NDR (%) ↓	1.25	3.56	2.21	3.12	2.00	2.91	2.08	2.08	3.11	0.59	**2.29**
GPTQ (W4A16)	Acc/F1 (%) ↑	79.62	64.54	63.83	69.26	63.72	42.23	58.70	65.81	36.11	86.72	63.05
	Acc/F1 Drop (%) ↓	0.44	0.78	0.44	0.99	1.10	1.99	2.03	1.29	0.56	-0.30	**0.93**
	DR (%) ↓	11.82	6.95	6.38	21.33	15.26	1.51	8.78	13.86	9.45	6.92	**10.22**
	NDR (%) ↓	1.60	4.60	2.46	3.92	2.78	5.25	3.84	2.36	5.22	0.69	**3.27**
LLM.Int8()	Acc/F1 (%) ↑	79.83	65.25	64.13	70.10	64.29	42.49	60.18	67.18	35.44	85.99	63.49
	Acc/F1 Drop (%) ↓	0.23	0.06	0.14	0.15	0.53	1.73	0.55	-0.09	1.22	0.43	**0.50**
	DR (%) ↓	4.17	5.99	3.74	3.32	9.10	7.37	4.61	2.73	8.89	1.38	**5.13**
	NDR (%) ↓	0.94	2.10	1.28	1.34	1.98	3.09	1.55	0.99	2.00	0.84	**1.61**
Qwen2.5-VL *(Baseline)*	Acc/F1 (%) ↑	82.56	82.51	60.41	76.20	82.84	42.10	70.21	83.85	50.67	86.17	71.75
AWQ (W4A16)	Acc/F1 (%) ↑	82.12	82.25	59.98	82.30	81.66	38.38	70.28	82.99	49.00	85.31	71.43
	Acc/F1 Drop (%) ↓	0.43	0.26	0.43	-6.10	1.18	3.72	-0.06	0.86	1.67	0.86	**0.32**
	DR (%) ↓	8.60	5.54	8.46	12.25	10.42	15.60	12.90	4.39	23.56	1.52	**10.32**
	NDR (%) ↓	1.61	2.36	2.48	2.03	1.72	6.58	2.43	1.46	5.22	1.10	**2.70**
GPTQ (W4A16)	Acc/F1 (%) ↑	82.04	82.25	59.92	85.57	81.85	38.38	69.23	82.56	49.00	85.86	71.67
	Acc/F1 Drop (%) ↓	0.51	0.26	0.48	-9.37	0.99	3.72	0.98	1.29	1.67	0.31	**0.08**
	DR (%) ↓	8.81	5.54	34.58	14.28	10.22	13.67	12.57	4.48	22.22	1.30	**12.77**
	NDR (%) ↓	1.67	2.36	15.32	1.09	1.60	5.71	2.59	1.57	4.56	0.79	**3.73**
LLM.Int8()	Acc/F1 (%) ↑	82.54	82.64	60.26	79.57	82.65	41.66	70.31	83.42	49.89	85.96	71.89
	Acc/F1 Drop (%) ↓	0.02	-0.13	0.14	-3.37	0.19	0.44	-0.09	0.43	0.78	0.21	**-0.14**
	DR (%) ↓	3.72	2.40	3.36	6.59	4.60	5.31	6.16	2.13	12.00	0.60	**4.69**
	NDR (%) ↓	0.67	0.84	0.88	0.89	0.56	1.72	1.25	0.72	2.22	0.38	**1.01**

Table 2: **Instance-Level Instability in Quantization Methods.** This table presents Acc/F1, Acc/F1 Drop, DR, and NDR for various quantization methods. Most methods exhibit high DR and NDR values, indicating significant instance-level instability, similar to token reduction techniques. Only the LLM.int8() method (Dettmers et al., 2022) is a notable exception, maintaining relatively low DR and NDR. Red values indicate the largest NDR per baseline model within each benchmark column.

stability extends to quantisation methods as shown in table 2. For instance, aggressive W4A16 quantization methods like GPTQ and AWQ applied to LLaVA-1.5 resulted in average aggregate performance drops of only ∼0.9-1.5%, yet induced high average Deviation Ratios (DR) of ∼19-21% and average Negative Deviation Ratios (NDR) of ∼5.7-6.6%. Individual benchmarks exhibited even more severe divergence, with DRs occasionally exceeding 40% and NDRs surpassing 8%. While less aggressive techniques like LLM.int8() showed markedly lower DR/NDR values (e.g., LLaVA-1.5 average DR 4.71%, NDR 1.21%), the trend for commonly used aggressive quantization is a significant and concerning level of instance-level prediction change. Table 2 only includes the results of Qwen2.5-VL (Bai et al., 2025) for the Qwen-VL model series since it is the improved version of Qwen2-VL. We show the results of Qwen2-VL (Wang et al., 2024) separately in Appendix E.

In summary, these results underscore a critical, largely overlooked deficiency in current VLM acceleration practices. To better view the overall distribution of relation between Acc/F1 Drop and DR/NDR values, we visualize the data in Appendix D. **The substantial DR and NDR values with minimal changes in aggregate metrics, provide compelling evidence that accelerated models can**

Method	Metric	GQA			VQAv2		
		No Pertb.	Vision Pertb.	Text Pertb.	No Pertb.	Vision Pertb.	Text Pertb.
LLaVA-1.5 (Baseline)							
FastV	DR (%) ↓	10.85	12.59	8.78	7.88	9.94	8.98
	NDR (%) ↓	4.41	4.75	2.76	2.17	2.86	2.11
HiRED	DR (%) ↓	13.26	16.08	47.06	14.72	17.62	15.86
	NDR (%) ↓	4.47	4.97	4.95	3.50	4.34	3.41
PyramidDrop	DR (%) ↓	11.34	12.12	11.30	8.97	10.17	10.80
	NDR (%) ↓	4.72	4.77	3.30	2.29	2.62	2.38
SparseVLM	DR (%) ↓	13.97	15.25	13.42	12.10	13.70	14.25
	NDR (%) ↓	5.81	5.77	3.85	3.33	3.65	3.34
VisionZip	DR (%) ↓	14.46	14.84	14.61	13.10	14.84	16.03
	NDR (%) ↓	6.15	5.55	4.36	3.64	3.68	3.76
LLaVA-Next (Baseline)							
FastV	DR (%) ↓	6.34	8.71	6.45	5.77	8.13	6.44
	NDR (%) ↓	2.31	3.23	1.63	1.44	2.39	1.46
HiRED	DR (%) ↓	17.47	55.84	27.28	15.40	33.38	36.77
	NDR (%) ↓	7.60	25.55	6.75	4.43	4.23	3.51
PyramidDrop	DR (%) ↓	10.44	10.69	12.47	8.18	9.44	10.58
	NDR (%) ↓	3.96	3.60	3.14	1.97	2.09	2.23
SparseVLM	DR (%) ↓	13.48	14.43	16.22	11.83	14.28	14.92
	NDR (%) ↓	5.20	5.13	3.81	3.28	3.81	3.55
VisionZip	DR (%) ↓	14.03	15.14	19.11	12.28	16.34	16.97
	NDR (%) ↓	6.01	5.84	4.32	3.44	4.24	4.12

Table 3: Instance-level instability of token reduction methods under input perturbation. This table reports Divergence Ratio (DR) and Negative Divergence Ratio (NDR) across three input states: "No Pertb." (original inputs), "Vision Pertb." (e.g., image noise, blur), and "Text Pertb." (e.g., text misspellings, paraphrasing). Red signifies higher DR/NDR under perturbation than without; blue signifies lower. The table illustrates that most methods suffer greater instance-level instability when inputs are perturbed.

Method	Metric	GQA			VQAv2		
		No Pertb.	Vision Pertb.	Text Pertb.	No Pertb.	Vision Pertb.	Text Pertb.
LLaVA-1.5 (Baseline)							
AWQ	DR (%) ↓	14.84	17.20	14.89	16.32	19.13	19.01
	NDR (%) ↓	5.17	5.54	3.78	3.93	4.77	4.00
GPTQ	DR (%) ↓	15.84	18.79	18.16	17.10	20.01	21.38
	NDR (%) ↓	5.49	6.03	4.66	4.34	5.07	4.85
LLM.Int8()	DR (%) ↓	2.88	2.98	4.05	2.89	3.11	4.78
	NDR (%) ↓	0.88	0.92	0.82	0.62	0.63	0.96
Qwen25-vl (Baseline)							
AWQ	DR (%) ↓	8.46	13.10	25.30	8.60	13.56	17.57
	NDR (%) ↓	2.48	3.33	3.12	1.61	2.76	3.28
GPTQ	DR (%) ↓	34.58	14.37	24.85	8.81	15.05	19.09
	NDR (%) ↓	15.32	3.79	2.99	1.67	3.20	3.34
LLM.Int8()	DR (%) ↓	3.36	7.38	14.44	3.72	7.79	10.14
	NDR (%) ↓	0.88	1.89	1.43	0.67	1.64	1.59

Table 4: Instance-level instability of quantisation methods under input perturbation. Most quantisation methods demonstrate increased instance-level instability under input perturbations.

indeed become unreliable for specific instances they previously handled correctly.

4.2 Instance-Level Instability Under Input Perturbations

To further demonstrate the risk of instance-level instability under practical settings, we conducted experiments involving perturbations to both text and vision inputs to VLMs, representing common real-world inputs disturbances. The detailed results are shown in Table 3 and Table 4. We only show the DR and NDR values in the tables. Acc/F1 and Acc/F1 Drop values are listed in appendix C. **The clear takeaway is that these perturbations generally exacerbate the Divergence Ratios (DR) and Negative Divergence Ratios (NDR) already observed in non-perturbed conditions.** For instance, applying vision perturbation to LLaVA-1.5 with AWQ quantization on VQAv2 increased its DR from 16.32% to 19.13% and its NDR from 3.93% to 4.77%. Text perturbation on the same model and benchmark also increased DR to 19.01% and NDR, albeit slightly, to 4.00%. Similarly, for token reduction, LLaVA-1.5 with the HIRED method on GQA saw vision perturbation elevate DR from 13.26% to 16.08% and NDR from 4.47% to 4.97%; text perturbation in this case markedly increased DR to 47.06% and NDR to 4.95%. This observed pattern of increased instability under noisy conditions was generally consistent across different types of acceleration methods, including both quantization and token reduction. **Consequently, the levels of instance-level instability likely aggravate when these accelerated models are deployed in dynamic, real-world environments where input data is rarely pristine.**

4.3 Instance-Level Prediction Instability in the Medical Domain

In this section, we apply VisionZip (Yang et al., 2024c), PyramidDrop (Xing et al., 2024), and LLM.int8() (Dettmers et al., 2022) to LLaVA-Med (Li et al., 2023a). We firstly verify the generalisation of these acceleration methods by evaluating them on the biomedical multimodal conversation test set introduced by Li et al. (2023a). We then conduct a case study by measuring the DR and NDR values on a medical VQA dataset VQA-RAD (Lau et al., 2018), revealing similarly high DR and NDR values as shown on general domain benchmarks as discussed in Section 4.1.

Generalisation of Acceleration Methods to Medical Domain. Table 5 summarizes the performance of various acceleration methods compared to the baseline model LLaVA-Med on the biomedical multimodal conversation test set. Results indicate that all examined acceleration methods (VisionZip, PyramidDrop and LLM.int8()) maintained

Method	Question Types		Domains					Overall
	Conversation	Description	Chest-Xray	MRI	Histology	Gross	CT	
LLaVA-Med *(Baseline)*	63.91	49.19	65.14	48.38	64.91	61.74	59.88	60.10
VisionZip	65.08	46.59	64.18	49.57	68.45	60.92	57.38	60.29
PyramidDrop	64.12	47.12	64.15	48.46	63.51	64.86	57.76	59.72
LLM.Int8()	63.96	50.20	64.47	47.82	64.75	64.07	60.65	60.39

Table 5: Evaluation of VisionZip (Yang et al., 2024c), PyramidDrop (Xing et al., 2024), and LLM.int8() (Dettmers et al., 2022) applied to LLaVA-Med (Li et al., 2023a) on its biomedical multimodal conversation test set. The results confirm the negligible overall performance impact of extending these acceleration techniques to the medical domain.

Method	VQA-RAD	
LLaVA-Med *(Baseline)*	Open (Recall) (%) ↑	30.29
	Closed (Acc) (%) ↑	59.35
VisionZip	Open (Recall) (%) ↑	30.89
	Closed (Acc) (%) ↑	58.66
	Recall/Acc Drop (%) ↓	0.15
	DR (%) ↓	29.85
	NDR (%) ↓	5.12
PyramidDrop	Open (Recall) (%) ↑	30.38
	Closed (Acc) (%) ↑	58.81
	Recall/Acc Drop (%) ↓	0.27
	DR (%) ↓	26.20
	NDR (%) ↓	4.54
LLM.Int8()	Open (Recall) (%) ↑	31.24
	Closed (Acc) (%) ↑	58.20
	Recall/Acc Drop (%) ↓	0.26
	DR (%) ↓	25.80
	NDR (%) ↓	4.80

Table 6: Evaluation of VisionZip (Yang et al., 2024c), PyramidDrop (Xing et al., 2024), and LLM.int8() (Dettmers et al., 2022) on LLaVA-Med (Li et al., 2023a) using the VQA-RAD (Lau et al., 2018) dataset (comprising open-ended and closed-ended questions). While aggregate performance loss was minimal, all three acceleration methods exhibited significant instance-level deviations.

almost identical performance to the baseline across diverse medical imaging modalities. This demonstrates minimal overall performance impact from generalising acceleration methods to medical context.

High Risk Instance-Level Instability in Medical Domain. Despite minimal overall performance drop, significant instance-level deviations were observed on the VQA-RAD benchmark as shown in Table 6. Deviation Ratio (DR) values were notably high, ranging between 25.80%-29.85% across the evaluated methods, suggesting that accelerated models frequently altered their predictions compared to the baseline model. More critically, Negative Deviation Ratios (NDR), representing detrimental prediction changes, were considerably higher in the medical domain (4.54%-5.12%) compared to general domain benchmarks. This indicates heightened instability risks when deploying accelerated VLMs in high-stake medical applications, where unstable outputs such as misdiagnoses could have severe consequences.

5 Conclusion

This work addressed the critical yet often overlooked issue of instance-level prediction stability in accelerated Vision-Language Models (VLMs), a factor vital for their trustworthy deployment in sensitive real-world applications. Our comprehensive empirical investigation revealed a stark reality: despite minimal impact on aggregate performance metrics, common acceleration techniques induced significant instability. This concerning trend of instability was consistently observed across diverse models and methods, further exacerbated by input perturbations, and confirmed in a medical VLM case study, exposing a crucial vulnerability in current VLM acceleration practices. We therefore conclude with an urgent imperative for incorporating rigorous instance-level stability checks to ensure these models are genuinely faithful and trustworthy for real-world deployment.

6 Limitations

We acknowledge certain limitations in this study. Our findings regarding instance-level instability primarily stem from experiments on academic benchmarks conducted in controlled laboratory settings. While we employed input perturbation techniques to approximate real-world data variability, these simulations may not fully capture the diverse complexities inherent in live industrial environments. Consequently, caution is warranted when directly generalizing our specific quantitative findings to all industrial applications. Further research on industrial cases is recommended to validate and extend these insights across broader operational conditions.

References

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2025. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *AAAI*, pages 1773–1781. AAAI Press.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Frank Bodendorf. 2025. A data-driven use case planning and assessment approach for AI portfolio management. *Electron. Mark.*, 35(1):22.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV (81)*, volume 15139 of *Lecture Notes in Computer Science*, pages 19–35. Springer.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip H. S. Torr, and Volker Tresp. 2023. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *CoRR*, abs/2208.07339.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. 2024. Accuracy is not all you need. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323.

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

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 235–251. Springer.

Weronika Lajewska, Momchil Hardalov, Laura Aina, Neha Anna John, Hang Su, and Lluís Màrquez. 2025. Understanding and improving information preservation in prompt compression for llms. *CoRR*, abs/2503.19114.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1).

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Zhen Li, Yupeng Su, Runming Yang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. 2025. Quantization meets reasoning: Exploring LLM low-bit quantization degradation for mathematical reasoning. *CoRR*, abs/2501.03035.

Q. Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *CoRR*, abs/2306.03100.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In *MLSys*. mlsys.org.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Zichen Tang, Junlin Huang, Rudan Yan, Yuxin Wang, Zhenheng Tang, Shaohuai Shi, Amelie Chi Zhou, and Xiaowen Chu. 2024. Bandwidth-aware and overlap-weighted compression for communication-efficient federated learning. In *ICPP*, pages 866–875. ACM.

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. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Laura Weidinger, Inioluwa Deborah Raji, Hanna M. Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative AI systems. *CoRR*, abs/2503.05336.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? *CoRR*, abs/2502.11501.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *CoRR*, abs/2410.17247.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

Ge Yang, Changyi He, Jinyang Guo, Jianyu Wu, Yifu Ding, Aishan Liu, Haotong Qin, Pengliang Ji, and Xianglong Liu. 2024b. Llmcbench: Benchmarking large language model compression for efficient deployment. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2024c. Visionzip: Longer is better but not necessary in vision language models. *CoRR*, abs/2412.04467.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, Lifang He, Brian D. Davison, Quanzheng Li, Yong Chen, Hongfang Liu, and Lichao Sun. 2023. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *CoRR*, abs/2305.17100.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. In *ICLR*. OpenReview.net.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *CoRR*, abs/2410.04417.

A Benchmark Details

Benchmark	Split	Number of Samples
VQAv2	Validation	214354
AI2D	Test	3088
GQA	Test-DeV	12578
MMBench	English-Dev	4377
MMMU	Validation	900
OKVQA	Validation	5046
POPE	Test	9000
ScienceQA	Training	2017
TextVQA	Validation	5000
VizWiz	Validation	4319
VQA-RAD	Train+Test	2248
Total Samples		**253972**

Table 7: Summary of benchmark datasets, splits, and their respective sample sizes.

To comprehensively evaluate our methods, we utilize a diverse array of ten established benchmarks, as detailed in Table 7. This selection spans various visual and multimodal understanding tasks, including Visual Question Answering (VQAv2, GQA, AI2D, OKVQA, TextVQA, ScienceQA, VizWiz), multimodal reasoning (MMMU), and general multimodal capabilities (MMBench, POPE). The evaluation is conducted on standard splits such as validation, test, or development sets, encompassing a significant total of 251,679 samples. Notably, VQAv2 contributes the largest portion with 214,354 validation samples, ensuring a robust assessment across different challenge domains and scales. For evaluation in the medical domain, we utilize the VQA-RAD benchmark, employing both its training and test sets. This dataset comprises 1299 closed-ended (yes/no) questions, for which we assess exact-match accuracy, and 949 open-ended questions, evaluated using recall, defined as the ratio of ground truth tokens present in the prediction.

Text Perturbation Methods	Severity
OCR	5
Punct	1
Typos	5
Keyboard	5
Spelling Error	5
char random insert	5
char random replace	5
char random swap	5
char random delete	5
Passive	1
Tense	1
Formal	1
Casual	1
Active	1
Double Neg	1
InsertAdv	1
AppendIrr	1
Random Insert	5
Drop NN	1
Drop Rand NN	1
DropVB	1
Drop VB & NN	1
Only NN	1
Only VB	1
Only NN & VB	1
Drop Rand VB	1
Drop First	1
Drop Last	1
Drop First and Last	1
Shuffle Order	1
Random Delete	5
SwapSyn Word Embd	5
SwapSyn WordNet	5
Back Trans	1
Random Swap	5
35 methods	**87 levels of severity**

Table 8: Summary of text perturbation methods.

B Hyper-Parameter Settings

For all the token reduction methods, we choose the signature or best-performing hyper-parameter settings as reported in the corresponding papers.Specifically, for VisionZip (Yang et al., 2024c), the number of retained tokens was set to 192. For PyramidDrop (Xing et al., 2024), we use pruning layers at indices [8, 16, 24] and corresponding pruning ratios of [0.5, 0.25, 0.125]. For SparseVLM(Zhang et al., 2024), the number of retained tokens is set to 192. For FastV(Chen et al., 2024), we utilize settings of K=3 and R=0.5. Finally, HiRed(Arif et al., 2025) was configured with a token budget of 20%. These settings were consistently applied across relevant experiments.

Image Perturbation Methods	Severity
Impulse	5
Gaussian	5
Shot	5
Speckle	5
Zoom	5
Defocus	5
Motion	5
Frosted Glass	5
Gaussian Blur	5
JPEG	5
Contrast	5
Elastic	5
Saturate	5
Spatter	5
Pixelate	5
Snow	5
Frost	5
Fog	5
Brightness	5
Blank	1
20 methods	**96 levels of severity**

Table 9: Summary of image perturbation methods.

C Input Perturbation Details

To evaluate robustness, we utilize a comprehensive suite of input perturbation techniques proposed by Chen et al. (2023). The specifics of these perturbations are detailed for text in Table 8 and for images in Table 9. Accounting for various severity levels, these amount to 87 distinct configurations for text inputs and 96 for image inputs. We randomly apply these varied perturbations to the text and image inputs of the VQAv2 (Goyal et al., 2017) and GQA(Hudson and Manning, 2019) datasets. Importantly, to ensure a fair and consistent comparison across experiments, the exact same perturbed inputs are used for all tested acceleration methods.

D Data Visualisation

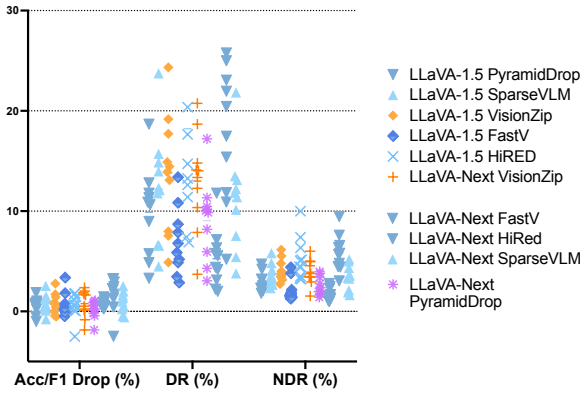

Figure 2: Statistical Distribution of Metrics (Acc/F1 Drop, DR, NDR) for Token Reduction Strategies Across All Benchmarks and Implemented Methods.

To better view the distribution of Acc/F1 Loss together with DR and NDR values, we plot a scatter diagram for Token Reduction Methods and Quantisation Methods, respectively. As shown in figure 2, it reveals a consistent trend across various models and methods. In both diagrams, the "Acc/F1 Drop (%)" remains notably low, generally appearing under 5% and often close to or below 2%. In stark contrast, the "DR (%)" and "NDR (%)" values are substantially higher, frequently ranging between 10% and 30%. This significant disparity underscores that while the accuracy or F1 score experiences minimal degradation, the other metrics, DR and NDR, show much more pronounced changes.

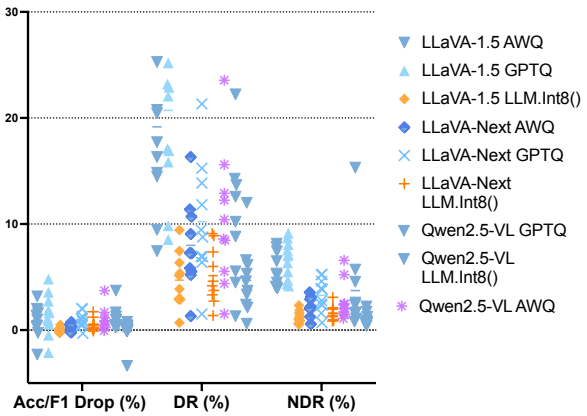

Figure 3: Statistical Distribution of Metrics (Acc/F1 Drop, DR, NDR) for Quantisation Strategies Across All Benchmarks and Implemented Methods.

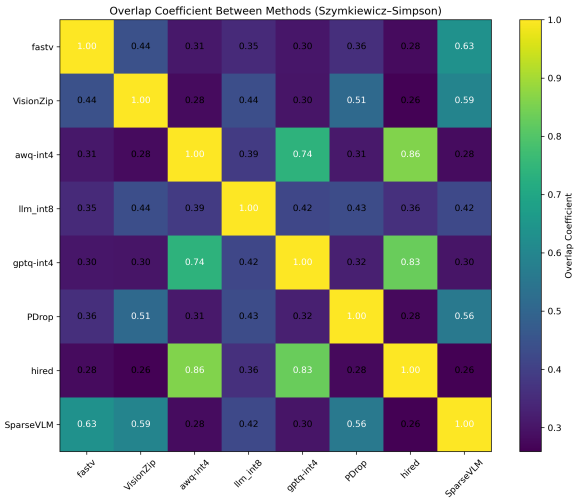

Figure 4: Overlap ratios of negatively diverged instances among acceleration methods for LLaVA-1.5 (Liu et al., 2024a).

Acceleration Methods Divergence Direction We further investigate the "divergence direction" of acceleration methods by examining the ex-

Method	Metric	VQAv2	AI2D	GQA	ScienceQA Img	TextVQA	OKVQA	VizWiz	MMBench	MMMU	POPE	Average
Qwen2-VL *(Baseline)*	Acc/F1 (%) ↑	78.69	70.14	59.83	59.00	79.34	44.02	66.05	71.22	40.67	86.00	64.03
AWQ (W4A16)	Acc/F1 (%) ↑	77.48	69.17	58.64	68.62	78.02	23.45	60.74	67.61	40.11	86.81	61.46
	Acc/F1 Drop (%) ↓	1.21	0.97	1.19	-9.62	1.33	20.57	5.31	3.61	0.56	-0.81	2.57
	DR (%) ↓	13.05	13.05	13.67	26.97	14.46	44.61	27.23	17.81	35.78	1.92	21.72
	NDR (%) ↓	3.13547	5.505	4.524	4.1646009	3.12	25.06936	10.257	3.5574036	9.44444	0.656	7.36637
GPTQ (W4A16)	Acc/F1 (%) ↑	77.58	68.85	58.71	55.23	78.34	35.88	65.06	65.38	37.33	85.96	61.19
	Acc/F1 Drop (%) ↓	1.11	1.30	1.13	3.77	1.01	8.14	0.99	5.84	3.33	0.04	2.84
	DR (%) ↓	12.15	12.82	12.56	30.09	13.82	33.04	27.11	14.85	33.44	1.56	19.92
	NDR (%) ↓	2.9386	5.44	4.269	12.493803	2.9	11.95006	5.0938	5.4285054	11.1111	0.767	6.60597

Table 10: Instance-Level Instability of quantisation methods (Lin et al., 2024; Frantar et al., 2022) in Qwen2-VL model (Wang et al., 2024).

Method	Metric	GQA			VQAv2		
		No Pertb.	Vision Pertb.	Text Pertb.	No Pertb.	Vision Pertb.	Text Pertb.
LLaVA-1.5 (Baseline)							
FastV	Acc/F1 (%) ↑	60.089	54.285	40.396	75.862	65.851	59.048
	Acc/F1 Drop (%) ↓	0.018	0.018	0.010	0.008	0.011	0.005
HiRED	Acc/F1 (%) ↑	61.226	55.478	61.226	76.515	66.529	59.506
	Acc/F1 Drop (%) ↓	0.007	0.006	-0.198	0.001	0.004	0.001
PyramidDrop	Acc/F1 (%) ↑	60.049	54.412	40.197	75.927	66.238	59.037
	Acc/F1 Drop (%) ↓	0.019	0.017	0.012	0.007	0.007	0.005
SparseVLM	Acc/F1 (%) ↑	59.374	53.967	39.959	75.259	65.695	58.548
	Acc/F1 Drop (%) ↓	0.025	0.021	0.015	0.014	0.013	0.010
VisionZip	Acc/F1 (%) ↑	59.135	54.158	39.649	74.926	65.867	58.372
	Acc/F1 Drop (%) ↓	0.028	0.019	0.018	0.017	0.011	0.012
LLaVA-Next (Baseline)							
FastV	Acc/F1 (%) ↑	63.873	55.470	41.398	79.632	66.836	62.249
	Acc/F1 Drop (%) ↓	0.004	0.010	0.002	0.004	0.011	0.004
HiRED	Acc/F1 (%) ↑	61.329	39.831	39.831	77.571	77.571	77.571
	Acc/F1 Drop (%) ↓	0.029	0.166	0.018	0.025	-0.097	-0.150
PyramidDrop	Acc/F1 (%) ↑	63.381	55.987	40.968	79.460	67.702	62.181
	Acc/F1 Drop (%) ↓	0.009	0.005	0.007	0.006	0.002	0.004
SparseVLM	Acc/F1 (%) ↑	62.888	55.359	41.024	78.424	66.455	61.245
	Acc/F1 Drop (%) ↓	0.014	0.011	0.006	0.016	0.014	0.014
VisionZip	Acc/F1 (%) ↑	61.894	54.556	40.579	78.340	66.296	61.099
	Acc/F1 Drop (%) ↓	0.024	0.019	0.011	0.017	0.016	0.015

Table 11: Performance and performance drop of token reduction methods under input perturbation.

tent to which they are affected by the same instances. A high degree of overlap in these instances suggests that different methods diverge in a predictable, controllable manner. This shared divergence would simplify the development of universal solutions to mitigate instability. Conversely, minimal overlap—indicating highly separated divergences—would imply more unpredictable behavior, posing greater uncertainty for the practical deployment of these methods. To explore this, we analyzed results from LLaVA-1.5 (Liu et al., 2024a), measuring the overlap of affected instances across various acceleration techniques. The findings are presented in Figure 4, which demonstrates that most pairings exhibit more "highly separated" divergences.

E Qwen2-VL Results

We conduct experimetns on Qwen2-VL (Wang et al., 2024) 3B model with AWQ(Lin et al., 2024) and GPTQ(Frantar et al., 2022) quantisation methods, detailed in table 10. It reveals varied performance impacts across different benchmarks.

Method	Metric	GQA			VQAv2		
		No Pertb.	Vision Pertb.	Text Pertb.	No Pertb.	Vision Pertb.	Text Pertb.
LLaVA-1.5 (Baseline)							
AWQ	Acc/F1 (%) ↑	60.916	0.551	0.408	76.231	0.663	0.594
	Acc/F1 Drop (%) ↓	1.002	0.010	0.006	0.405	0.007	0.002
GPTQ	Acc/F1 (%) ↑	60.860	0.553	0.403	75.770	0.661	0.588
	Acc/F1 Drop (%) ↓	1.057	0.008	0.011	0.866	0.008	0.008
LLM.Int8()	Acc/F1 (%) ↑	62.045	0.561	0.414	76.522	0.669	0.594
	Acc/F1 Drop (%) ↓	-0.127	0.000	0.000	0.114	0.001	0.002
Qwen25-vl (Baseline)							
AWQ	Acc/F1 (%) ↑	59.978	0.505	0.367	82.121	0.662	0.632
	Acc/F1 Drop (%) ↓	0.429	0.006	0.004	0.435	0.005	0.001
GPTQ	Acc/F1 (%) ↑	59.922	0.500	0.365	82.045	0.656	0.624
	Acc/F1 Drop (%) ↓	0.485	0.012	0.006	0.511	0.011	0.010
LLM.Int8()	Acc/F1 (%) ↑	60.264	0.506	0.370	82.540	0.665	0.635
	Acc/F1 Drop (%) ↓	0.143	0.005	0.001	0.015	0.002	-0.001

Table 12: Performance and performance drop of quantisation methods under input perturbation.

On average, AWQ quantization leads to a 2.57% drop in Acc/F1 score, an outcome notably influenced by an unexpected 9.62% performance increase on the ScienceQA Img benchmark, alongside a significant 20.57% performance decrease on OKVQA. GPTQ quantization results in a slightly higher average Acc/F1 drop of 2.84%, with its most pronounced performance reductions observed on OKVQA (8.14% drop) and MMBench (5.84% drop). While the average changes in Acc/F1 scores are relatively contained, both quantization techniques generally cause substantial increases in DR (%) and NDR (%) values across the evaluated benchmarks.

F Input Perturbation Impacts on Acc/F1 and Acc/F1 Drop

Table 11 and table 12 detail the performance of various acceleration techniques—quantization (AWQ, GPTQ, LLM.Int8()) and token reduction (FastV, HIRED, PyramidDrop, SparseVLM, VisionZip)—on models like LLaVA-1.5, LLaVA-Next, and Qwen2.5-vl, across GQA and VQAv2 datasets under no, vision, and text perturbations. A consistent trend across both sets of methods is the remarkably low impact on Acc/F1 scores; the Acc/F1 Drop (%) is generally minimal, often well

below 1% and frequently in the hundredths of a percent, irrespective of the specific acceleration technique or perturbation type applied.

G Qualifying Examples

In this section, we present qualifying examples: specific test instances showing how applying acceleration methods to a Vision Language Model (VLM) can cause prediction divergence.

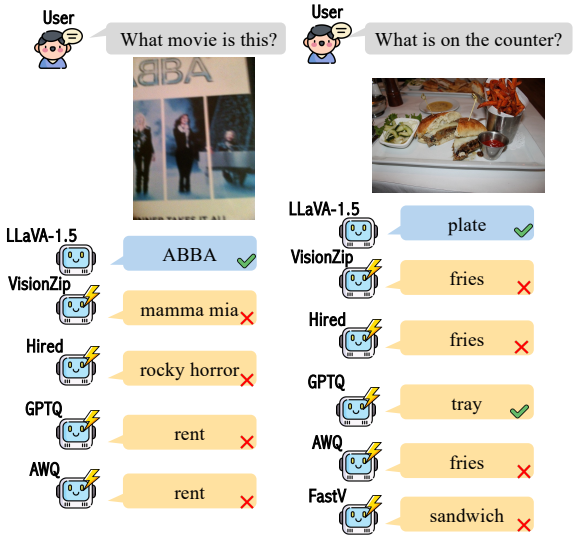

Figure 5: Acceleration Instances Divergence qualifying examples for LLaVA-1.5 (Liu et al., 2024a).

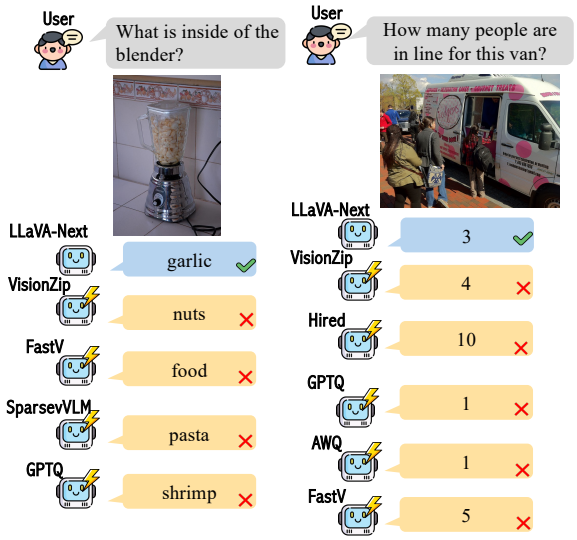

Figure 6: Acceleration Instances Divergence qualifying examples for LLaVA-Next (Liu et al., 2024b).

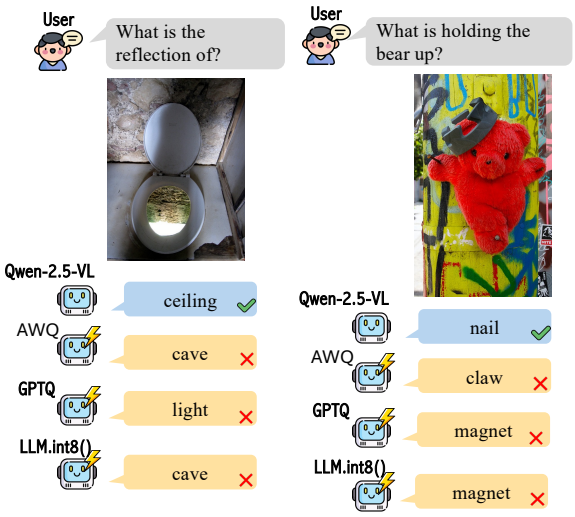

Figure 7: Acceleration Instances Divergence qualifying examples for Qwen2.5-VL (Bai et al., 2025).

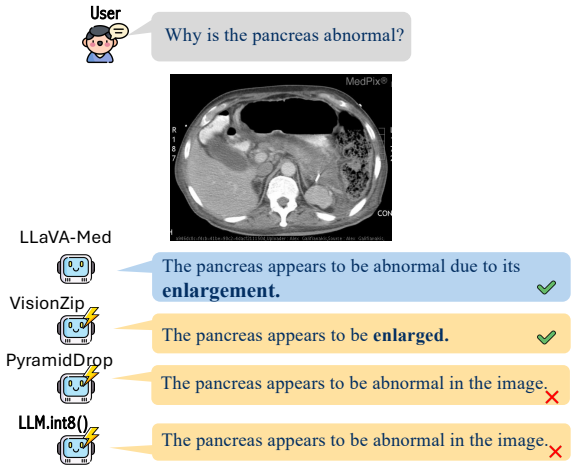

Figure 8: Acceleration Instances Divergence qualifying examples for LLaVA-Med (Li et al., 2023a).

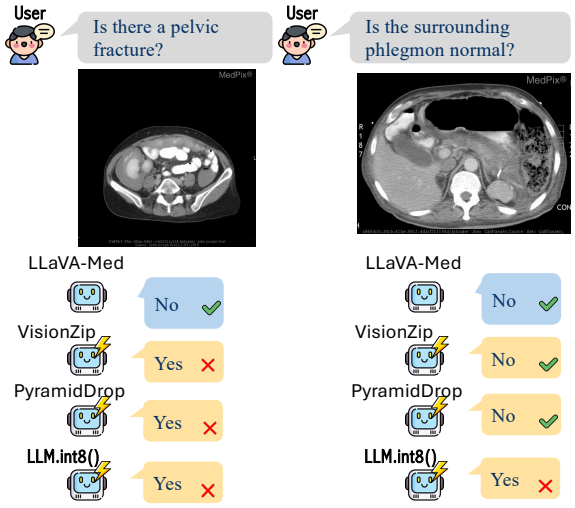

Figure 9: Acceleration Instances Divergence qualifying examples for LLaVA-Med (Li et al., 2023a).