Evaluating Uncertainty in Vision–Language Models under Structured Linguistic Degradation

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) are increasingly considered for deployment in assistive communication contexts, where users with language impairments routinely produce degraded linguistic input. However, their robustness to such input remains largely unexplored. We present a preemptive evaluation framework grounded in aphasia-inspired text degradation, comprising five perturbation categories namely phonemic paraphrasia, semantic substitution, agrammatism, circumlocution, and disfluency at two severity levels. We evaluate two contrastive (CLIP ViT-B/32 and ViT-L/14) and two generative (Qwen2-VL-2B and Qwen2-VL-7B) VLMs on 1000 Microsoft-COCO image-caption pairs using a suite of proposed evaluation metrics that focus on embedding space alignment, decision confidence, and token level prediction. We find that all metrics worsen monotonically with degradation severity and that lexical degradations are consistently most damaging while structural degradations are largely tolerated. Additionally, we observe that semantic substitution most undermines the decision confidence whereas phonemic paraphrasia disrupts the alignment and token level prediction the most, thus demonstrating the necessity for a more robust evaluation framework before deploying VLMs in assistive communication context.

1 Introduction

Vision Language Models (VLMs) such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), LLaVA (Liu et al., 2023), Qwen (Bai et al., 2023), and many others demonstrate strong performances across a wide range of multimodal tasks, including image-text retrieval, multimodal reasoning, image captioning, and visual question answering, among others. These capabilities have motivated their prospective adoption in assistive technology contexts such as Augmentative and Alternative Communication (AAC) devices and speech rehabilitation tools (Mulye et al.), where users with language impairments can benefit from them. The use of VLMs in these domains is still in the conception phase, but rapid advances could lead to their accelerated adoption in the future. The multimodality of VLMs can be used to rectify the user's language, produce image descriptions of varying text complexity depending on the user's comprehension ability, and so on. With the foreseeable dependence on VLMs in such critical fields, it is imperative to preemptively test their robustness in such real-world scenarios, where their input can often be noisy, incomplete, or of lower quality. In this work, we present an evaluation framework that does exactly this.

To this end, we subject selected VLMs to text perturbations to understand how the models process them. We identify *Aphasia*, a language disorder resulting from a stroke or brain injury that impairs a person's ability to communicate, to simulate the input text perturbations. Aphasic speech is characterized by lexical omissions, syntactic simplifications, and semantic substitutions that are replicable and categorizable. These errors can be simulated using Large Language Models (LLMs) and offer a novel replicable systematic framework to perturb the input text to the VLMs.

In this paper, we aim to answer the following questions, **RQ1:** How do the performances of contrastive and generative VLMs vary under controlled aphasia-inspired linguistic perturbations? **RQ2:** Do the uncertainty metrics computed for these models faithfully track the severity and type of the degradations?

The main contributions of this paper are: (i) We introduce a novel aphasia-inspired language degradation approach that enables controlled simulation of linguistic impairments. (ii) We propose a set of metrics to capture the variations in VLM performance variation for perturbed inputs for different

VLM architectures. (iii) We conduct a systematic comparison of how different VLM architecture respond to structured degradation of input text, offering insights into their potential for real-world application. The code will be made available at the time of publication.

2 Related Work

Visual and Linguistic Robustness in VLMs Several recent works focus on perturbing the visual input to the VLM to study aspects such as spatial fragility, resampling artefacts, and sensitivity to common visual corruptions such as noise, blur, and weather (Lee et al., 2024; Saxena et al., 2026; Velayuthan and Tavakkoli, 2026; Wang et al., 2026). Additional research work has focused on perturbing the input text to the VLM. (Khan and Fu, 2024) exhibit that state-of-the-art VLMs produce inconsistent answers across semantically equivalent rephrasing of the same question and suggest using this neighborhood consistency as an uncertainty metric for selective Visual Question Answer task. (Joong Lee, 2026) similarly exploit controlled linguistic transformations to observe the VLM behaviour. Our work complements these efforts by deliberately degrading the textual input along independently controllable dimensions of lexical, phonemic, syntactic, pragmatic, and disfluency-level corruption, and measure both alignment quality and uncertainty at each severity.

Uncertainty Quantification in VLMs Uncertainty quantification for VLMs has gained attention in recent years but remains largely evaluated on clean and well curated inputs. (Kostumov et al., 2024) introduce a benchmark that incorporates conformal-prediction-based uncertainty in VLM evaluation. (Groot and Valdenegro-Toro, 2024) deomnstrate that both VLMs and LLMs exhibit substantial calibration error and tend to be overconfident when asked to verbalize their confidence. (Upadhyay et al., 2023) equips pre-trained contrastive encoders such as CLIP and BLIP with a probabilistic adapter that estimates embedding-level uncertainty. (Ye et al., 2024) extend the uncertainty quantification benchmarking to the language models. These works however, focus on clean inputs, and none examines how the uncertainty metric they compute responds to systematically degraded language.

Aphasia and Artificial Intelligence Recent works (Azevedo et al., 2024; Baalakumar et al., 2025; Bailey et al., 2026; Mulye et al.) explore the use of generative Artificial Intelligence as an assistive technology for rehabilitation and management of aphasia. This serves as motivation to take a closer look at how VLMs might respond to text that resembles aphasia speech

We emphasize that our contributions in this work is an evaluation protocol for VLMs and not a clinical study. The artifacts introduced in the VLM text input are used as a guideline for structured linguistic degradation and we quantify the performance of VLMs using automated metrics that we will discuss in the next sections.

3 Methodology

In this section, we provide the implementation details of our experiments.

3.1 Overview

We start from a curated image captioning dataset that contains images and corresponding clean human written ground truth captions. For every clean caption, we generate a set of degraded variants spanning five linguistic axes and three severity levels using a Large Language Model (LLM). We then use each triplet (image, clean caption, and degraded caption) with two contrastive and two generative VLMs. Based on this, we compute a unified family of uncertainty metrics. The remainder of this section describes each part of our approach in detail. An overview of the experiments conducted in this paper has been provided in Figure 1.

3.2 Dataset and Aphasia-Inspired degradation

We use the validation split of MS-COCO 2017 (Lin et al., 2014) dataset, a widely adopted image-captioning benchmark in which each image is annotated with multiple human written captions in English. We select MS-COCO for the following reasons: (i) it contains everyday scenes that would be helpful for a real-world VLM application, (ii) its captions are created by humans, fluent and well-formed, thus providing a clean baseline against which the degraded variants can be compared. From the validation split of the dataset, we draw a sample of $1,000$ images, pairing each image with one of its ground truth captions. Each clean caption is then subjected to two levels of perturbations with increasing severity. This yields

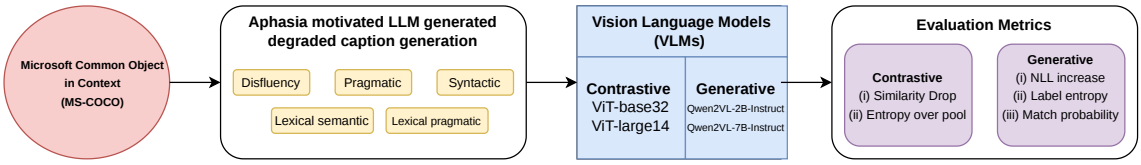

Figure 1: An illustration of the our proposed approach to compute the uncertainty metrics for Vision Language Models (VLMs) when provided with aphasia inspired degraded inputs.

$1,000 \times 5 \times 2 = 10,000$ degraded captions. Including the clean baseline captions, the full curated degradation dataset contains $1{,}000 \times 5 \times 3 = 15{,}000$ (image, caption) pairs.

Severity levels. For each linguistic perturbation that we describe subsequently, we define three severity levels: S_0 (clean, c), S_1 (mild perturbation; severity parameter $\sigma = 0.3$), and S_2 (severe perturbation; $\sigma = 0.9$). The severity parameter controls how many tokens of c the generator may transform: at S_1 a small fraction of eligible tokens is altered, while at S_2 the transformation is applied aggressively across the caption. S_0 serves as the clean baseline caption in our experiments.

Linguistic Perturbations. The five degradations are set-up to simulate aphasia error patterns (Damasio, 1992). We define them as below for our experiments:

- **Disfluency:** inserting filled pauses, repetitions, self-revisions or incomplete phrases. At S_1, one filled pause *(um, uh, er)* and one word repetition are inserted while the original caption remains mostly intact. At S_2, multiple filled pauses, several word and phrase repetitions and potentially one incomplete phrase. (*the man walks* $\rightarrow$ S_1: *the um... the man walks, walks*; S_2: *the... uh... the man, the man is... um...*)

- **Lexical-phonemic:** substituting characters or syllables to produce sound-like distortions of content words. At S_1, 1-2 sounds in exactly one word are changed while keeping the word recognisable. At S_2, many words are heavily distorted phonemically such that they barely resemble the original word. (*table* $\rightarrow$ S_1: *tabel*; S_2: *toble*)

- **Lexical-semantic:** replacing content words with semantically related substitutions. At S_1, exactly one word is replaced with closely related and the sentence the meaning of the sentence does not change drastically. At S_2, as many words as possible are replaced with semantically related but incorrect words, making the sentence hard to understand. (*outdoor spectacle* $\rightarrow$ S_1: *cat*; S_2: outside show)

- **Pragmatic:** replacing direct referents with indirect descriptions. At S_1, exactly one noun is replaced with a circumlocution of 3-5 words. At S_2, most content words are replaced with circumlocutions and hesitation markers, making the sentence very difficult to understand. (*jacket* $\rightarrow$ S_1: *thing that looks like a shirt*; S_2: *thing that you wear out*)

- **Syntactic:** omitting articles, auxiliaries, and inflections to yield telegraphic output. At S_1, only the articles *(a, an, the)* are removed. At S_2, only the single most important noun and verb are retained and everything else is removed. (*a man is walking the dog* $\rightarrow$ S_1: *man is walking dog*; S_2: *man walk dog*)

Perturbation Generation. We use Llama-3.2-3B-Instruct[1], an instruction-tuned LLM to create the aforementioned perturbations. For each (c, d, s) triple, where c is the clean caption, d is the type of perturbation, and s the severity $s \in \{S_1, S_2\}$, the LLM is provided an in-context learning (ICL) prompt containing the (i) description of the target aphasic phenomenon, (ii) two illustrative examples explicitly marked as not to be copied, (iii) the input caption to transform, and (iv) the type and severity instructions. Generation uses temperature (τ) of 0.7 and raw outputs are post-processed to obtain the degraded captions.

3.3 VLM Evaluation

Contrastive VLMs. We use two variants of the OpenAI CLIP (Radford et al., 2021), namely, (i) ViT-B/32 (CLIP-VIT-BASE-PATCH32), (ii) ViT-L/14 (CLIP-VIT-LARGE-PATCH14).

[1] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

We compute the embeddings of the images and the captions (clean and degraded). We $L2$ normalize and compute the cosine similarity between the image and each caption, yielding $\mathrm{sim}(I, c)$ and $\mathrm{sim}(I, \tilde{c}_k)$ where I is the image, c is the clean caption and $\tilde{c}_k$ is the kth degraded caption. These scores form the basis for the similarity drop metric that we discuss later.

To measure the relative alignment of the degraded caption with respect to the clean caption and unrelated alternatives, we construct a caption pool $\mathcal{P} = \{c, \tilde{c}_k, r_1, \ldots, r_8\}$ for each degraded caption variant, where $r_1, \ldots, r_8$ are distractor captions drawn uniformly at random from the clean captions of other images in the dataset. The same eight distractors are reused across all five degradation types for a given image to ensure within-image comparisons are fair. We compute a softmax distribution over the raw dot products between the image and all pool captions. The resulting distribution forms the basis for the pool-softmax entropy metric H_{pool}, which captures the degree to which the clean caption dominates the pool as degradation increases. Since, the pool contains both the clean and degraded caption as competitors, a decrease in H_{pool} reflects the growing gap between their respective similarity scores. As the degraded caption loses alignment with the image, probability mass concentrates increasingly on the clean caption. We use H_{pool} to track the relative change in discriminability across severity levels and degradation types within each model. Cross-model comparisons of absolute entropy values are not made, as the two CLIP variants operate at different embedding scales.

Generative VLMs. We evaluate two variants from the Qwen2-VL family (Wang et al., 2024) namely, Qwen2-VL-2B-Instruct and Qwen2-VL-7B-Instruct.

To measure how plausible a model finds a caption as a description of the image, we compute the average per-token negative log likelihood (NLL) assigned to the caption text conditioned on the image. The score reflects the model's intrinsic assessment of caption-image fit independent of the evaluation framing. The same procedure is applied to the clean caption (c) and each degraded variant ($\tilde{c}_k$) under identical prompt scaffolding, with the loss averaged per token to control for caption length differences. The resulting NLL increase (Δ_{NLL}) therefore isolates the effect of degradation from both image-specific difficulty and caption length variation.

Additionally, we prompt the model with an explicit three-option matching question, asking it to respond with exactly one label from $\{\text{YES}, \text{NO}, \text{UNSURE}\}$ given the image and caption. Rather than relying only on the decoded output string, we extract the probability distribution over the three label tokens at the first generated position and re-normalize these probabilities to sum to one. This gives a three-way distribution that captures not just which label the model chooses, but with a measure of confidence as well.

3.4 Uncertainty Metrics

Existing work on uncertainty quantification has addressed output-level uncertainty in both language and vision-language models. For language models, sampling-based approaches measure disagreement across multiple outputs (Kuhn et al., 2023), whereas methods such as (Kadavath et al., 2022) prompt the model to state its own confidence. For vision-language models, calibration studies have examined whether confidence scores align with empirical accuracy across visual recognition tasks (Tu et al., 2024). All of aforementioned approaches target output-level uncertainty given well-formed inputs, and do not measure how the model behaviour changes under systematic input degradation.

In contrast, our metrics address the complementary problem of input-level variation. Rather than measuring output uncertainty in isolation, we anchor our metrics to clean-input baseline, so that the observed changes directly reflect the effect of degradation.

Contrastive Metrics. We compute two metrics for the CLIP models.

1. **Similarity drop (Δ_{sim}):** We define this as the relative change in cosine similarity between the image and the caption when the caption is degraded.

$$\Delta_{\text{sim}} = \frac{\mathrm{sim}(I, c) - \mathrm{sim}(I, \tilde{c})}{\mathrm{sim}(I, c)}. \quad (1)$$

   Normalization accounts for differences in the cosine similarity computation for images of differing complexity, ensuring that Δ_{sim} reflects the proportional alignment loss attributable solely to caption degradation.

2. **Pool-softmax entropy (H_{pool}):** We define this metric as the normalized Shannon entropy of the distribution that CLIP places over a caption pool $\mathcal{P} = \{c, \tilde{c}, r_1, \ldots, r_8\}$,

$$H_{\text{pool}} = \frac{-\sum_i p_i \log p_i}{\log |\mathcal{P}|}, \quad (2)$$

where p_i is the softmax probability assigned to the i-th pool caption. Since, the pool contains both the clean caption and the degraded caption, H_{pool} reflects the degree to which the clean caption dominates the pool relative to the degraded variant. Decreasing entropy under degradation indicates that the clean caption is the clear winner. We track the relative change in H_{pool} across severity levels and degradation types within each model.

Generative metrics. We compute three metrics for the QWEN models.

3. **NLL increase (Δ_{NLL}):** We define this as the increase in the average per-token negative log-likelihood (NLL) when the caption is degraded,

$$\Delta_{\text{NLL}} = \text{NLL}(\tilde{c} \mid I) - \text{NLL}(c \mid I), \quad (3)$$

where each NLL is the average per-token cross-entropy loss from the caption scoring, computed under identical prompting. A high positive value indicates that the degraded caption is less plausible to the model as a description of the image compared to the clean caption.

4. **Label entropy (H_{label}):** This is defined as the normalised Shannon entropy of the three-way label distribution,

$$H_{\text{label}} = \frac{-\sum_{l \in \mathcal{L}} p_l \log p_l}{\log |\mathcal{L}|}, \quad (4)$$

where $\mathcal{L} = \{\text{YES}, \text{NO}, \text{UNSURE}\}$. Low values indicate a confident and committed judgment whereas, high values indicate genuine uncertainty across labels.

5. **Match probability (p_{match}):** We define this as the normalized probability assigned to the YES label at the first generated token position under the label-decision prompt,

$$p_{\text{match}} = \frac{p(\text{YES})}{p(\text{YES}) + p(\text{NO}) + p(\text{UNSURE})}, \quad (5)$$

where each $p(\cdot)$ is the softmax probability of the corresponding label token in the full vocabulary distribution. Values close to 1.0 indicate that the model confidently endorses the caption as a correct image description whereas, values close to 0.0 indicate higher uncertainty. Unlike Δ_{NLL}, which captures token-level predictive surprise, p_{match} captures the model's explicit judgment of caption-image correspondence.

4 Results and Discussions

We present the results from our evaluation and analyse the model behaviour along two dimensions namely, degradation severity and degradation type. Tables 1–3 report values for contrastive metrics and Tables 4–6 report values for generative metrics. The corresponding plots are shown in Figure 2 (contrastive) and Figure 3 (generative).

Degradation severity is faithfully tracked across both paradigms. Both CLIP models show monotonically increasing Δ_{sim} with severity ($0.000 \rightarrow 0.064 \rightarrow 0.164$ for ViT-B/32 and $0.000 \rightarrow 0.089 \rightarrow 0.217$ for ViT-L/14). For H_{pool}, we note a corresponding decrease ($0.331 \rightarrow 0.147$ for ViT-B/32 and $0.301 \rightarrow 0.047$ for ViT-L/14), confirming that aphasia-inspired degradation is detectable through embedding-space alignment. The inverse relationship between the two metrics is by design, as the degraded caption loses alignment with the image, clean caption increasingly dominates the pool competition and thus reducing the entropy. The generative models show a similar trend. Δ_{NLL} increases monotonically ($0.000 \rightarrow 0.774 \rightarrow 1.328$ for VL-2B and $0.000 \rightarrow 0.869 \rightarrow 1.389$ for 7B). The p_{match} decreases proportionally by 42–44% from S_0 to S_2 for both models. As both contrastive and generative metrics independently reproduce this pattern, this strengthens confidence in our analysis.

Model scale amplifies sensitivity. We note that ViT-L/14 consistently produces larger similarity drops than ViT-B/32 across all conditions, most markedly for lexical types (0.134 vs 0.089 for lexical-semantic). This suggests that larger encoders, having learnt finer-grained image-text alignment, are proportionally *more* vulnerable to content-word perturbations. This is a counterintuitive finding with direct implications for assistive deployment of larger VLMs. For the generative metrics, despite different baseline NLL values

Model	Type (Sev)	Clean	Degr.	Sim. Drop	Pool Entr.
ViT-B	disfl. (S0)	0.304 ± 0.034	0.304 ± 0.034	0.000 ± 0.000	0.331 ± 0.073
	disfl. (S1)	0.304 ± 0.034	0.295 ± 0.032	0.026 ± 0.040	0.256 ± 0.103
	disfl. (S2)	0.304 ± 0.034	0.271 ± 0.029	0.105 ± 0.071	0.148 ± 0.118
	lex-p (S0)	0.304 ± 0.034	0.304 ± 0.034	0.000 ± 0.000	0.331 ± 0.073
	lex-p (S1)	0.304 ± 0.034	0.277 ± 0.037	0.084 ± 0.101	0.194 ± 0.127
	lex-p (S2)	0.304 ± 0.034	0.237 ± 0.036	0.213 ± 0.127	0.155 ± 0.125
	lex-s (S0)	0.304 ± 0.034	0.304 ± 0.034	0.000 ± 0.000	0.331 ± 0.073
	lex-s (S1)	0.304 ± 0.034	0.275 ± 0.037	0.092 ± 0.099	0.189 ± 0.138
	lex-s (S2)	0.304 ± 0.034	0.248 ± 0.039	0.177 ± 0.136	0.128 ± 0.138
	prag. (S0)	0.304 ± 0.034	0.304 ± 0.034	0.000 ± 0.000	0.331 ± 0.073
	prag. (S1)	0.304 ± 0.034	0.274 ± 0.033	0.092 ± 0.100	0.179 ± 0.127
	prag. (S2)	0.304 ± 0.034	0.248 ± 0.031	0.177 ± 0.115	0.151 ± 0.120
	syn. (S0)	0.304 ± 0.034	0.304 ± 0.034	0.000 ± 0.000	0.331 ± 0.073
	syn. (S1)	0.304 ± 0.034	0.295 ± 0.033	0.026 ± 0.038	0.286 ± 0.094
	syn. (S2)	0.304 ± 0.034	0.257 ± 0.032	0.146 ± 0.112	0.154 ± 0.120
ViT-L	disfl. (S0)	0.257 ± 0.037	0.257 ± 0.037	0.000 ± 0.000	0.301 ± 0.036
	disfl. (S1)	0.257 ± 0.037	0.249 ± 0.037	0.030 ± 0.063	0.133 ± 0.110
	disfl. (S2)	0.257 ± 0.037	0.226 ± 0.035	0.115 ± 0.088	0.065 ± 0.094
	lex-p (S0)	0.257 ± 0.037	0.257 ± 0.037	0.000 ± 0.000	0.301 ± 0.036
	lex-p (S1)	0.257 ± 0.037	0.225 ± 0.042	0.120 ± 0.132	0.071 ± 0.102
	lex-p (S2)	0.257 ± 0.037	0.186 ± 0.040	0.270 ± 0.159	0.033 ± 0.072
	lex-s (S0)	0.257 ± 0.037	0.257 ± 0.037	0.000 ± 0.000	0.301 ± 0.036
	lex-s (S1)	0.257 ± 0.037	0.222 ± 0.044	0.134 ± 0.144	0.053 ± 0.090
	lex-s (S2)	0.257 ± 0.037	0.187 ± 0.045	0.267 ± 0.171	0.026 ± 0.069
	prag. (S0)	0.257 ± 0.037	0.257 ± 0.037	0.000 ± 0.000	0.301 ± 0.036
	prag. (S1)	0.257 ± 0.037	0.222 ± 0.039	0.132 ± 0.135	0.053 ± 0.092
	prag. (S2)	0.257 ± 0.037	0.195 ± 0.036	0.232 ± 0.140	0.052 ± 0.091
	syn. (S0)	0.257 ± 0.037	0.257 ± 0.037	0.000 ± 0.000	0.301 ± 0.036
	syn. (S1)	0.257 ± 0.037	0.249 ± 0.037	0.031 ± 0.051	0.159 ± 0.112
	syn. (S2)	0.257 ± 0.037	0.203 ± 0.035	0.199 ± 0.148	0.058 ± 0.092

Table 1: Contrastive evaluation results broken down by degradation type, severity level, and model. Clean and degraded similarity scores are reported alongside similarity drop and pool-softmax entropy (mean ± standard deviation).

Model	Sev.	Sim. Drop	Pool Entr.
ViT-B	S0	0.000 ± 0.000	0.331 ± 0.073
	S1	0.064 ± 0.087	0.221 ± 0.126
	S2	0.164 ± 0.120	0.147 ± 0.125
ViT-L	S0	0.000 ± 0.000	0.301 ± 0.036
	S1	0.089 ± 0.122	0.094 ± 0.110
	S2	0.217 ± 0.155	0.047 ± 0.086

Table 2: Contrastive metrics aggregated by severity level (mean ± standard deviation).

Model	Degr.	Sim. Drop	Pool Entr.
ViT-B	disfl.	0.044 ± 0.065	0.245 ± 0.125
	lex-ph	0.099 ± 0.128	0.227 ± 0.134
	lex-sem	0.089 ± 0.121	0.216 ± 0.147
	prag.	0.090 ± 0.114	0.220 ± 0.135
	syn.	0.057 ± 0.093	0.257 ± 0.123
ViT-L	disfl.	0.048 ± 0.079	0.166 ± 0.131
	lex-ph	0.130 ± 0.162	0.135 ± 0.140
	lex-sem	0.134 ± 0.169	0.126 ± 0.142
	prag.	0.121 ± 0.147	0.135 ± 0.140
	syn.	0.077 ± 0.125	0.173 ± 0.132

Table 3: Contrastive metrics aggregated by degradation type (mean ± standard deviation), averaged across severity levels.

(2.457 for VL-2B vs 1.894 for VL-7B at S_0), both Qwen models produce near-identical Δ_{NLL} at S_2 (1.328 for VL-2B vs 1.389 for VL-7B), indicating that NLL increase is scale-invariant within the same architecture family. We observe a substantially different p_{match} (0.858 for 7B vs 0.524 for 2B) at S_0. This means that H_{label} is only informative for Qwen2-VL-7B, where it increases meaningfully with severity ($0.298 \rightarrow 0.463 \rightarrow 0.643$). Across degradation types, disfluency perturbations produces the highest decision-level uncertainty (0.546) and syntactic the lowest (0.364), consistent with the p_{match} ordering as seen in Table 6. As for Qwen2-VL-2B, the saturated baseline entropy ($\sim$0.843), leaves no scope for degradation detection, thus establishing a model-scale dependency in decision-level uncertainty quantification.

Lexical degradations are most damaging whereas structural degradations most tolerated. Across both types (contrastive and generative) of evaluation, we note a consistent pattern. For ViT-L/14, lexical-phonemic and lexical-semantic produce the largest Δ_{sim} (0.130 and 0.134) and lowest H_{pool} (0.135 and 0.124), while disfluency and syntactic produce the smallest drops (0.048 and 0.077) and highest entropies (0.166 and 0.173). The same ordering holds for ViT-B/32, though with smaller absolute differences between types. For generative metrics, lexical-phonemic produces the highest Δ_{NLL} (1.389 for 2B, 1.298 for 7B) while disfluency and syntactic remain lowest. This can be attributed to corruption of vocabulary used by both CLIP and Qwen by semantic and phonemic perturbation, which is not the case for disfluency and agrammatism. Notably, ViT-B/32 cannot distinguish lexical-semantic from pragmatic on Δ_{sim} (0.089 vs 0.090), while ViT-L/14 can (0.134 vs 0.121), suggesting larger models offer finer-grained discrimination between degradation types.

Token-level metrics miss two important failure modes. Two anomalies are visible in the Table 4 namely, mild syntactic degradation produces near-zero or negative Δ_{NLL} for both models at S_1, and pragmatic S_2 falls below S_1. In both cases though, p_{match} correctly captures the underlying degradation where Δ_{NLL} does not, underscoring the necessity of considering decision-level metrics alongside token-level metrics.

Different metrics capture different failure modes. We note that lexical-phonemic consistently produces the highest Δ_{NLL} and syntactic the highest p_{match} across both Qwen models. However, p_{match} behaves differently for lexical-semantic. Despite moderate Δ_{NLL} (0.679 for 2B and 0.788 for 7B), it produces the lowest p_{match} of all types (0.311 for 2B and 0.464 for 7B), suggesting that semantic substitution damages caption-image matching confidence more than token-level surprise alone

Model	Type (Sev)	Cap. NLL	NLL Inc.	Label Entr.	P-Match
VL-2B	disfl.(S0)	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	disfl.(S1)	3.248 ± 0.737	0.791 ± 0.632	0.890 ± 0.044	0.436 ± 0.120
	disfl.(S2)	2.732 ± 0.464	0.276 ± 0.820	0.883 ± 0.047	0.229 ± 0.098
	lex-ph(S0)	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	lex-ph(S1)	3.798 ± 0.979	1.341 ± 0.894	0.866 ± 0.063	0.437 ± 0.158
	lex-ph(S2)	5.284 ± 0.832	2.827 ± 0.879	0.898 ± 0.047	0.352 ± 0.134
	lex-sem(S0)	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	lex-sem(S1)	3.044 ± 0.938	0.587 ± 0.753	0.801 ± 0.111	0.286 ± 0.220
	lex-sem(S2)	3.906 ± 0.894	1.449 ± 0.889	0.715 ± 0.128	0.122 ± 0.161
	prag.(S0)	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	prag.(S1)	3.725 ± 0.775	1.268 ± 0.878	0.869 ± 0.061	0.421 ± 0.159
	prag.(S2)	3.317 ± 0.466	0.860 ± 0.896	0.889 ± 0.049	0.360 ± 0.133
	syn.(S0)	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	syn.(S1)	2.339 ± 0.877	-0.117 ± 0.692	0.832 ± 0.061	0.542 ± 0.124
	syn.(S2)	3.683 ± 1.656	1.226 ± 1.760	0.853 ± 0.066	0.461 ± 0.157
VL-7B	disfl.(S0)	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	disfl.(S1)	2.827 ± 0.622	0.933 ± 0.547	0.477 ± 0.232	0.788 ± 0.218
	disfl.(S2)	2.370 ± 0.404	0.476 ± 0.696	0.862 ± 0.146	0.420 ± 0.237
	lex-ph(S0)	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	lex-ph(S1)	3.178 ± 0.850	1.284 ± 0.754	0.515 ± 0.283	0.682 ± 0.317
	lex-ph(S2)	4.504 ± 0.832	2.610 ± 0.800	0.719 ± 0.238	0.519 ± 0.303
	lex-sem(S0)	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	lex-sem(S1)	2.570 ± 0.824	0.676 ± 0.710	0.547 ± 0.244	0.392 ± 0.394
	lex-sem(S2)	3.583 ± 0.862	1.689 ± 0.862	0.559 ± 0.187	0.143 ± 0.268
	prag.(S0)	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	prag.(S1)	3.249 ± 0.700	1.355 ± 0.790	0.475 ± 0.277	0.719 ± 0.310
	prag.(S2)	3.007 ± 0.440	1.113 ± 0.758	0.585 ± 0.250	0.665 ± 0.293
	syn.(S0)	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	syn.(S1)	1.988 ± 0.771	0.094 ± 0.560	0.303 ± 0.255	0.854 ± 0.236
	syn.(S2)	2.952 ± 1.453	1.058 ± 1.555	0.491 ± 0.294	0.672 ± 0.341

Table 4: Generative evaluation results broken down by degradation type, severity level, and model (mean ± standard deviation).

Model	Sev.	Cap. NLL	NLL Inc.	Label Entr.	P-Match
VL-2B	S0	2.457 ± 0.785	0.000 ± 0.000	0.843 ± 0.055	0.524 ± 0.121
	S1	3.231 ± 1.014	0.774 ± 0.939	0.852 ± 0.078	0.424 ± 0.180
	S2	3.784 ± 1.285	1.328 ± 1.395	0.847 ± 0.101	0.305 ± 0.181
VL-7B	S0	1.894 ± 0.672	0.000 ± 0.000	0.298 ± 0.247	0.858 ± 0.234
	S1	2.763 ± 0.885	0.869 ± 0.820	0.463 ± 0.272	0.687 ± 0.341
	S2	3.283 ± 1.140	1.389 ± 1.221	0.643 ± 0.264	0.484 ± 0.350

Table 5: Generative metrics aggregated by severity level (mean ± standard deviation).

would indicate. The disfluency perturbation exhibits the opposite pattern. A low Δ_{NLL} (0.356 for 2B, 0.470 for 7B) yet the highest H_{label} (0.872 for 2B and 0.546 for 7B) meaning the model processes disfluent text fluently but remains genuinely uncertain whether it matches the image. These observations show that the metrics we define suffer from limitations, supporting a multi-metric evaluation design.

Model	Degr.	Cap. NLL	NLL Inc.	Label Entr.	P-Match
VL-2B	disfl.	2.812 ± 0.752	0.356 ± 0.682	0.872 ± 0.053	0.396 ± 0.168
	lex-ph	3.846 ± 1.445	1.389 ± 1.363	0.869 ± 0.060	0.438 ± 0.156
	lex-sem	3.135 ± 1.058	0.679 ± 0.898	0.786 ± 0.116	0.311 ± 0.239
	prag.	3.166 ± 0.870	0.710 ± 0.896	0.867 ± 0.058	0.435 ± 0.154
	syn.	2.826 ± 1.321	0.370 ± 1.249	0.843 ± 0.061	0.509 ± 0.140
VL-7B	disfl.	2.364 ± 0.692	0.470 ± 0.637	0.546 ± 0.317	0.689 ± 0.299
	lex-ph	3.192 ± 1.326	1.298 ± 1.240	0.511 ± 0.309	0.686 ± 0.319
	lex-sem	2.682 ± 1.052	0.788 ± 0.947	0.468 ± 0.258	0.464 ± 0.426
	prag.	2.717 ± 0.852	0.823 ± 0.865	0.453 ± 0.284	0.747 ± 0.292
	syn.	2.278 ± 1.132	0.384 ± 1.067	0.364 ± 0.281	0.794 ± 0.288

Table 6: Generative metrics aggregated by degradation type (mean ± standard deviation), averaged across severity levels.

5 Conclusion

We presented a framework for evaluating VLM robustness to aphasia-inspired input text degradation across contrastive and generative model design using custom metrics. We show that models reliably become less accurate and less confident as captions become more degraded. Phonemic distortions and semantic substitutions consistently cause the greatest loss in image-caption alignment whereas structural changes such as hesitation or article removal are largely tolerated. Crucially, we observe that a model may handle degraded text fluently at the surface level yet silently fail to match it to the correct image. This failure only becomes visible when measuring explicit decision confidence rather than token-level prediction. These findings suggest that evaluating VLM robustness requires more than a single metric, and highlight the importance of testing such models before deploying them in assistive communication contexts where such models are provided with impaired input.

6 Limitations

Our framework relies on LLM-generated perturbations rather than authentic aphasia speech, which is inherently more variable than any systematic simulation. The generated degradations have not been validated by speech language therapists and may not fully capture the nuances of actual impairment. The evaluation is limited to two architecture families which may not generalise to other model designs. Results are reported as means with standard deviations. The formal significance testing is left for future work given the exploratory nature of this evaluation. Finally, H_{label} is uninformative for Qwen2-VL-2B due to poor baseline calibration, and sampling-based uncertainty metrics were uninformative for both models, suggesting that some proposed metrics require well-calibrated models to be effective.

References

Nancy Azevedo, Eva Kehayia, Gonia Jarema, Guylaine Le Dorze, Christel Beaujard, and Marc Yvon. 2024. How artificial intelligence (ai) is used in aphasia rehabilitation: A scoping review. *Aphasiology*, 38(2):305–336.

NN Baalakumar, S Prabakar, K Mythili, N Naveen, M Sabharigirivasan, and S Mathan. 2025. Ai based assistive device for aphasia elderly patients. In

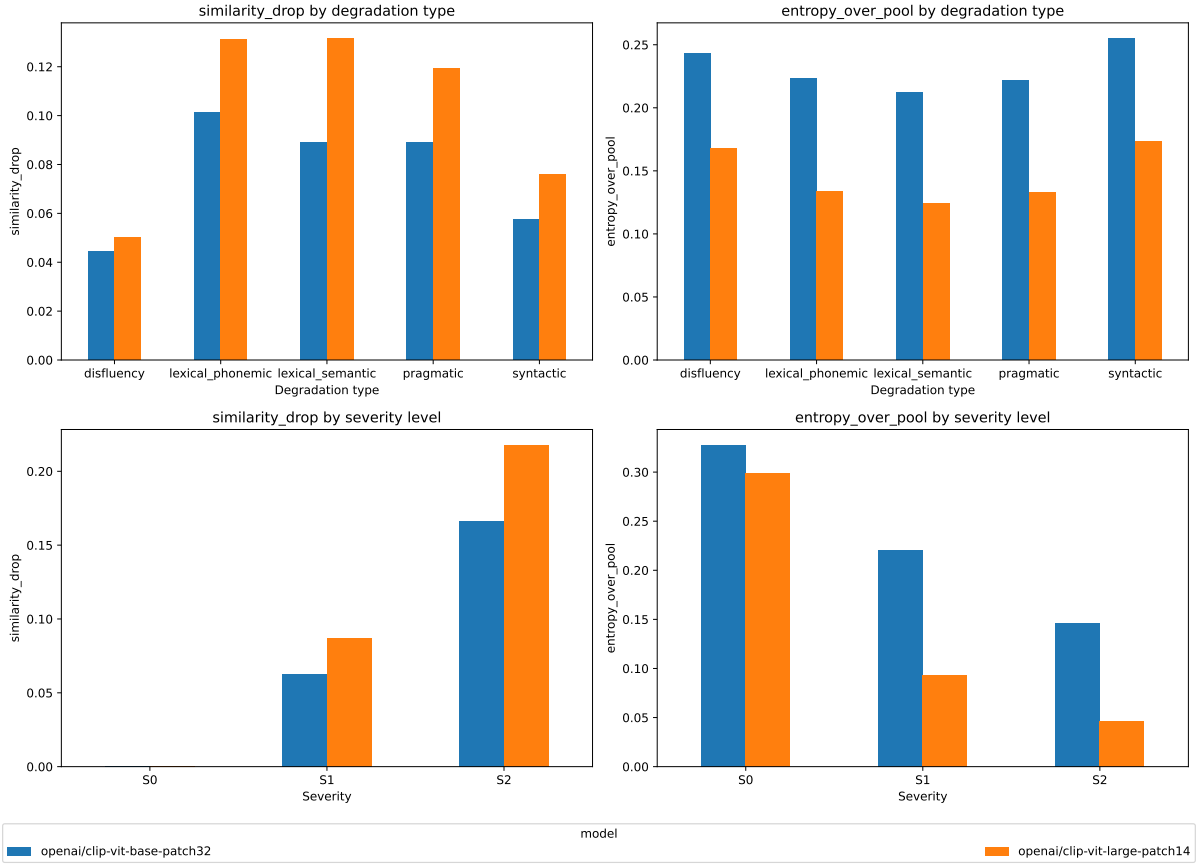

Figure 2: Contrastive metrics for CLIP models based on (i) degradation type and (ii) severity level.

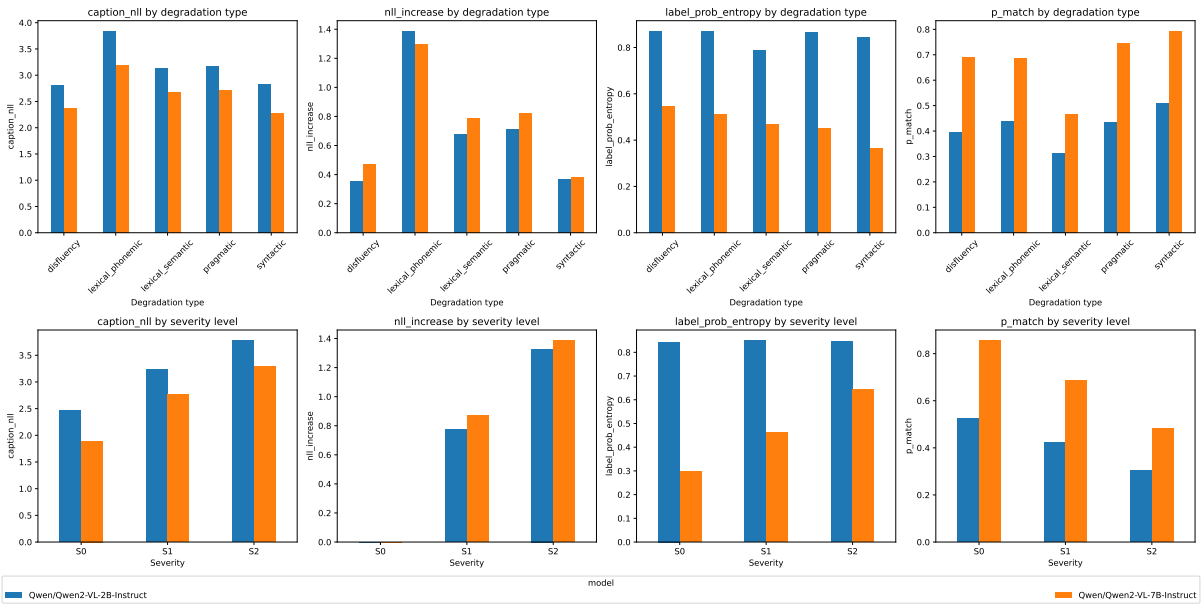

Figure 3: Generative metrics for Qwen models based on (i) degradation type and (ii) severity level.

2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN), pages 1862–1867. IEEE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Dallin J Bailey, Francesca Herget, Derek Hansen, Forrest Burton, Grant Pitt, Tyson Harmon, and David Wingate. 2026. Generative ai applied to aac for aphasia: a pilot study of aphasia-gpt. *Aphasiology*, 40(1):150–165.

Antonio R Damasio. 1992. Aphasia. *New England Journal of Medicine*, 326(8):531–539.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171.

Jae Joong Lee. 2026. Language-guided invariance probing of vision–language models. *Pattern Recognition Letters*, 202:108–113.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. 2024. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and 1 others. 2024. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Mihir Mulye, Stefan Conrad, and Stefan Knecht. Using multi-modal generative models for creating speech therapy material. In *Artificial Intelligence for Healthcare, and Hybrid Models for Coupling Deductive and Inductive Reasoning: First International Joint Conference, HC@ AIxIA+ HYDRA 2025, Bologna, Italy, October 25–26, 2025, Proceedings*, page 21. Springer Nature.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. 2024. An empirical study into what matters for calibrating vision-language models. *arXiv preprint arXiv:2402.07417*.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1899–1910.

Purushoth Velayuthan and Alireza Tavakkoli. 2026. Evaluating robustness of vision-language models under noisy conditions. In *2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0724–0729. IEEE.

Chenyu Wang, Tianle Chen, HM Ahmad, Kayhan Batmanghelich, and Wenchao Li. 2026. Vlm-uqbench: A benchmark for modality-specific and cross-modality uncertainties in vision language models. *arXiv preprint arXiv:2602.09214*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via

uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.