ORBIT: OFFSET-BASED ACTIVATION STEERING FOR CROSS-LINGUAL MULTIMODAL FACT VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fact verification faces a fundamental accessibility bottleneck: while large proprietary models exhibit cross-lingual robustness, small vision-language models suffer a severe, disproportionate accuracy collapse on low-resource languages. Closing this gap traditionally requires prohibitive scale or extensive language-specific fine-tuning. In this work, we propose Offset Recovery via Bilingual Inference-time Technique (ORBIT), an algorithmic intervention that mitigates this collapse at inference time. ORBIT estimates span-specific offset vectors from parallel claim–evidence instances and applies them to claim and evidence representations, targeting the cross-modal grounding failure induced by low-resource evidence. Experiments across 81 configurations reveal that ORBIT substantially recovers both accuracy and F1-score in low-resource settings, exposes calibration offset norms as a diagnostic proxy, and incurs lower overhead than translation-based pipelines.

1 INTRODUCTION

Misinformation today is rarely confined to a single modality or language. A false narrative from a conflict zone may spread with a Spanish caption and reach Vietnamese fact-checkers, but the evidence needed to refute it may exist only in low-resource local reports. Nevertheless, such misinformation often persists in linguistically underserved communities. Multimodal fact verification, which involves assessing a claim against associated text and image evidence, emerges as a crucial defense against such complex content (Yao et al., 2023).

However, current multimodal fact verification benchmarks are predominantly monolingual and English-centric. Datasets such as Factify (Mishra et al., 2022), MOCHEG (Yao et al., 2023), VERITE (Papadopoulos et al., 2024), and MM-FakeBench (Liu et al., 2025) implicitly assume that claims, evidence, and models all operate in the same high-resource language. This assumption breaks down in real-world scenarios: a model may verify a Javanese claim against English documents, or assess an English claim using Tigrinya evidence. We investigate a fundamental question: *When visual evidence remains constant, how does variation in the language of textual evidence affect verification accuracy?*

Prior work in cross-lingual natural language inference documents the brittleness of text-only models in low-resource scenarios (Conneau et al., 2018; Hu et al., 2020). However, these findings do not directly transfer to the multimodal setting. Vision-language models (VLMs) are trained to ground language in visual perception, a process heavily skewed toward English image captions and Western visual contexts (Bugliarello et al., 2022; Thapliyal et al., 2022). Consequently, while visual encoding may remain robust, cross-modal alignment degrades substantially for low-resource language evidence, giving rise to language-conditioned visual grounding errors (Zhou et al., 2021; Pfeiffer et al., 2022).

While recent large proprietary models demonstrate improved robustness to cross-lingual setups on multilingual benchmarks (Google DeepMind, 2026b; OpenAI, 2026; Anthropic, 2026), this performance comes at the cost of accessibility and transparency. For fact-checking organizations operating with limited computational resources in linguistically diverse regions, small vision-language models (SVLMs) – typically under 10B parameters (Ahmed et al., 2026) – represent the most viable option for local deployment. Despite this practical importance, their cross-lingual behavior remains least understood precisely where SVLMs are most needed. It remains unknown whether SVLMs merely

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

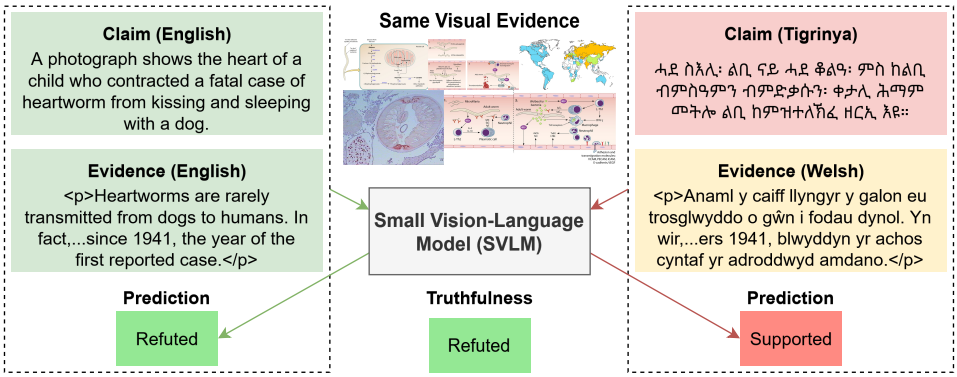


Figure 1: Overview of cross-lingual multimodal fact verification. Identical visual evidence and semantically equivalent text yield divergent predictions as claim and evidence languages vary.

underperform as scaled-down replicas, or suffer from a severe low-resource collapse rooted in the breakdown of cross-modal alignment.

To systematically investigate this question, we construct the controlled diagnostic testbed **MOCHEG-XL** by extending MOCHEG (Yao et al., 2023) via machine translation into eight target languages, yielding nine languages including English. We select MOCHEG because it provides gold textual evidence paired with each claim–image instance, and remains challenging even for strong recent baselines (Chen et al., 2024; Lee et al., 2024; Braun et al., 2025; Pang et al., 2025). We emphasize that this dataset is designed as an **analytical instrument** rather than a human-verified benchmark. By holding visual content constant while varying the language of claims and evidence across 81 distinct configurations, we isolate and quantify performance degradation attributable solely to linguistic shift. This controlled setup reveals cross-lingual brittleness that is otherwise confounded by content variation in naturally multilingual data.

Observing a disproportionate and severe collapse in low-resource languages, we introduce **Offset Recovery via Bilingual Inference-time Technique (ORBIT)**, a lightweight inference-time intervention designed to restore low-resource cross-modal grounding. ORBIT estimates span-specific offset vectors from parallel claim–evidence instances and applies them separately to claim and evidence representations, directly targeting the multimodal grounding failure induced by low-resource evidence. Operating without gradient updates, it avoids catastrophic forgetting and overfitting risks (Sung et al., 2022; Vu et al., 2022) while bypassing the prohibitive latency of translation-based pipelines.

Our contributions are threefold:

1. We conduct a systematic empirical analysis quantifying cross-lingual degradation in SVLMs for multimodal fact verification. We expose a disproportionate and severe collapse in low-resource languages, even when visual content is held constant.
2. We propose ORBIT, a training-free activation steering method that recovers low-resource performance at modest inference-time cost. We additionally show that calibration offset norms serve as a lightweight diagnostic proxy for predicting cross-lingual collapse without full evaluation.
3. We will publicly release our translated evaluation configurations (MOCHEG-XL) as a diagnostic toolkit to facilitate future research on cross-lingual robustness of vision-language models.

2 Related Work

Multimodal fact verification has largely been developed and evaluated under the assumption of monolingual, English-centric settings. Existing benchmarks assume linguistic homogeneity: claims, evidence, and models operate in the same high-resource language. We show that language shift and visual grounding are not independent failures; rather, they compound. Prior work treats cross-lingual

transfer and multimodal reasoning as separate problems; consequently, these lines of work cannot diagnose their interaction.

Early multimodal fact verification is largely shaped by binary out-of-context detection benchmarks, with Factify (Mishra et al., 2022) and NewsCLIPings (Luo et al., 2021) representative of an approach that flags image misuse without requiring evidence-based reasoning about *why* a claim is false. While MOCHEG (Yao et al., 2023) and subsequent efforts (Papadopoulos et al., 2024; Liu et al., 2025) move toward fine-grained, evidence-grounded verdicts, they preserve the same blind spot of linguistic homogeneity. We isolate the effect of language shift by extending MOCHEG cross-lingually, holding content constant while varying language.

Do cross-lingual gaps persist when images provide a shared grounding signal? Cross-lingual NLP benchmarks (Conneau et al., 2018; Hu et al., 2020) reveal systematic performance drops across languages. While recent inference-time methods like INCLINE (Wang et al., 2025) attempt to map source languages to a target language space using learned alignment matrices, they operate strictly within the text modality. Studies on vision-language tasks (Pfeiffer et al., 2022; Zhou et al., 2021) demonstrate that models suffer language-conditioned grounding errors with non-English text (Thapliyal et al., 2022). These findings, however, are drawn from high-resource multilingual settings and confined to generative or retrieval tasks. Low-resource grounding failures in structured verification tasks remain unexplored. Fact verification is fundamentally different, as it requires fine-grained tri-modal alignment among claim semantics, evidence text, and image content to output a discrete verdict. We provide the first controlled framework to isolate and quantify cross-lingual grounding failures in multimodal fact verification.

The practical obstacle is that fine-tuning on target-language data is often challenging due to data scarcity. Labeled cross-lingual multimodal data is scarce, and gradient-based updates risk catastrophic forgetting (Sung et al., 2022; Vu et al., 2022). Parameter-efficient methods (Houlsby et al., 2019; Hu et al., 2022; Lee et al., 2025) reduce trainable parameters but still require supervision. Activation steering offers a training-free alternative by adding precomputed direction vectors to hidden states during inference. In language models, it improves truthfulness (Li et al., 2023); in vision-language models, it mitigates hallucinations (Wu et al., 2025); and in cross-lingual settings, approaches like CLAS (Pokharel et al., 2026) modulate text representations by dynamically scaling shared versus language-specific neurons. Yet all these applications assume monolingual or text-only inputs, leaving multimodal cross-lingual transfer unaddressed. We bridge this gap with a highly efficient inference-time intervention. Rather than dynamic neuron-level masking, which requires per-token computation, we estimate a static, per-language correction vector from parallel evidence pairs for each target language. Each vector applies an additive shift to realign that language's representations toward the anchor space, bypassing auxiliary decoding while preserving efficiency.

3 Cross-Lingual Collapse

We isolate linguistic variation from visual content to measure its effect on multimodal fact verification. We use Qwen3-VL 4B[1] (Bai et al., 2025), the only open-source 4B-parameter SVLM in our candidate pool to exceed 50% accuracy on MOCHEG, and evaluate nine typologically diverse languages. We treat English (en) as the reference baseline, as it is the source language of MOCHEG (Yao et al., 2023). The remaining eight languages are categorized by pretraining data availability and typological coverage:

1. **High-resource languages:** Spanish (es), Russian (ru), Vietnamese (vi), Japanese (ja).
2. **Low-resource languages:** Javanese (jw), Welsh (cy), Māori (mi), Tigrinya (ti).

Translations for the eight target languages are generated using the Google Cloud Translation API rather than open-source encoder–decoder models such as NLLB-200 (Costa-Jussà et al., 2022). NLLB-200 is trained with a maximum sequence length of 512 tokens, which MOCHEG evidence passages frequently exceed (see Appendix A). Prior work establishes that autoregressive neural translation suffers from exposure bias and hallucination as sequence length increases (Wang & Sennrich, 2020). We therefore rely on the Google Cloud Translation API for its greater stability on

[1] Qwen/Qwen3-VL-4B-Instruct

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	55.00	53.64	52.99	51.72	51.84	47.71	43.24	38.70	36.98
es	54.55	54.01	52.54	51.19	50.53	45.74	42.51	37.59	35.54
ru	54.01	52.78	53.03	50.74	49.06	44.80	41.89	38.21	35.46
vi	54.05	51.11	50.70	50.70	48.57	45.29	40.70	35.59	35.42
ja	54.46	51.43	50.04	50.29	49.96	44.59	40.34	37.92	35.18
jw	48.32	46.23	44.55	45.25	43.94	45.17	37.10	33.82	33.09
cy	46.40	44.51	42.14	43.20	41.85	37.22	42.51	33.78	33.33
mi	42.79	41.11	40.21	39.97	39.72	36.12	34.48	35.38	32.76
ti	40.09	39.68	36.90	37.35	36.81	35.26	34.32	33.33	35.01

Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	53.28	51.77	51.50	50.76	51.32	48.06	43.78	38.51	35.89
es	52.17	52.24	51.04	50.25	50.11	46.11	43.04	37.46	34.15
ru	51.81	51.39	51.47	49.81	48.68	45.20	42.38	38.11	34.25
vi	52.18	50.16	49.57	49.66	48.45	45.70	41.18	35.06	33.70
ja	52.15	50.08	48.94	49.49	49.46	44.92	40.88	37.69	33.55
jw	47.40	46.02	44.43	45.38	44.35	45.41	36.84	32.11	30.37
cy	46.18	44.82	42.40	43.66	42.38	36.50	42.53	31.29	28.87
mi	43.11	41.37	40.46	40.15	39.90	35.03	32.87	35.02	27.58
ti	40.36	39.60	36.60	36.99	35.49	33.19	31.98	29.51	32.60

F1-score (%)

Performance levels: High ($>50\%$) Medium (40–50%) Low ($<40\%$)

Figure 2: Cross-lingual performance of Qwen3-VL across 81 claim–evidence configurations. Low-resource textual evidence (jw, cy, mi, ti) causes severe degradation, with Māori and Tigrinya near random. Same-language alignment benefit persists only for high-resource languages, which remain stable across both monolingual and cross-lingual settings.

long-form text. Translation quality metrics based on LLM-as-a-judge (Zheng et al., 2023) evaluation are provided in Appendix B.

By holding visual content constant while varying claim and evidence languages across 81 distinct configurations, we isolate performance degradation attributable solely to linguistic shift. Figure 2 reports accuracy and macro F1-score across all language pairs. The results show that evidence language is the dominant driver of collapse. When evidence shifts into a low-resource language, accuracy collapses across every claim language. For instance, while English claims achieve 55.00% accuracy with English evidence, using Javanese evidence causes a 7.29-point drop. This indicates that visual encoding remains robust, but cross-modal alignment deteriorates when evidence language shifts.

At the extreme end of the resource spectrum, performance collapses entirely. Māori and Tigrinya evidence reduce the model to near-random guessing: accuracy falls to 33–39% and F1 to 29–38%, regardless of claim language. This is not mere underperformance; it is a complete breakdown of cross-modal grounding: the model fails to map low-resource textual features into the shared vision–language space. This breakdown also explains why the same-language advantage disappears under low-resource evidence. Same-language claim–evidence pairs outperform mismatched ones for high-resource languages, yet this diagonal advantage vanishes entirely for Javanese, Welsh, Māori, and Tigrinya. Without adequate multilingual pretraining signal, the model cannot maintain coherent representations. The effect becomes uniform: the claim language no longer matters, because the evidence encoding has already failed.

These findings expose a fundamental vulnerability: visual encoding remains robust, yet cross-lingual textual grounding collapses under low-resource evidence. Although this analysis focuses on Qwen3-VL, we also evaluate Gemma 4 E2B[2] (Gemma Team et al., 2026) and LFM2-VL 3B[3] (Amini et al., 2025); their zero-shot results (Appendix C) exhibit the same evidence-language collapse, though the severity varies across models. We further show in Section 5.3 that the severity of this collapse can be predicted from calibration offset norms alone, offering a lightweight diagnostic before full evaluation. This divergence motivates targeted, inference-time intervention; the next section details our approach to realigning hidden representations without gradient updates.

4 ORBIT

We propose **ORBIT**, a training-free intervention that tackles the systematic degradation Section 3 reveals. By estimating and applying correction vectors to the model's hidden states during inference,

[2]google/gemma-4-E2B-it
[3]LiquidAI/LFM2-VL-3B

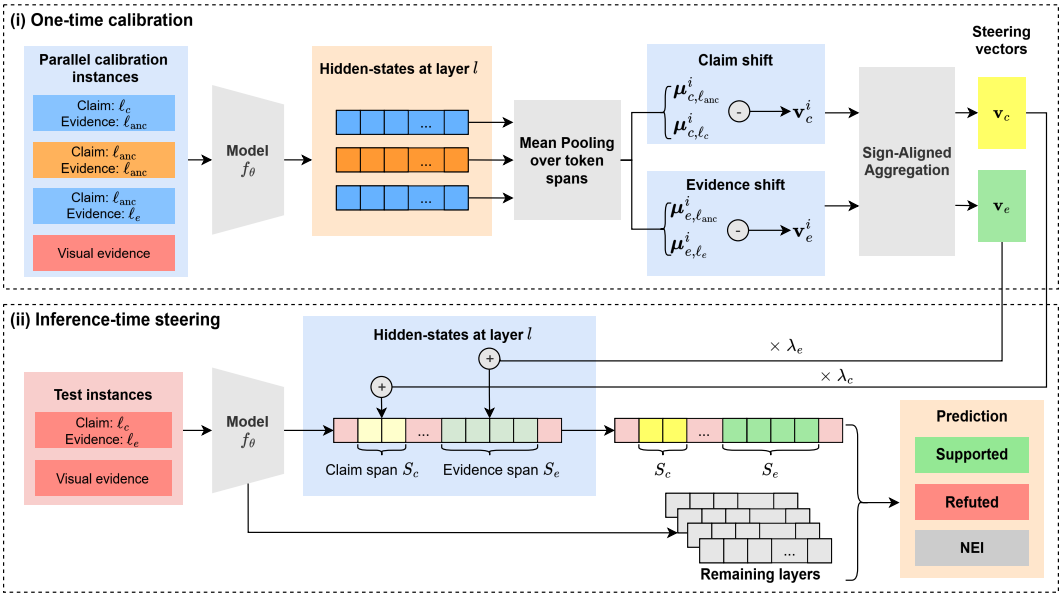


Figure 3: Overview of ORBIT. (i) One-time calibration aggregates per-sample claim and evidence shifts from parallel instances into per-language steering vectors $\mathbf{v}_c$ and $\mathbf{v}_e$. (ii) Inference-time steering adds these vectors to hidden states at layer l, realigning the claim span S_c and evidence span S_e toward the anchor space without parameter updates.

this approach realigns low-resource language text representations toward their anchor-language counterparts.

Consider a multimodal fact verification instance (c, e, I) where c is the claim in language ℓ_c, e is the textual evidence in language ℓ_e, and I is the visual evidence. ORBIT operates without access to task labels. Let $\mathbf{h}^{(l)}_{\ell_c,\ell_e} \in \mathbb{R}^{T\times d}$ denote the hidden states at layer l, where T is the sequence length and d is the hidden dimension. ORBIT estimates two language-specific correction vectors to realign representations toward an anchor language ℓ_{anc} during inference: $\mathbf{v}_c \in \mathbb{R}^d$ for the claim component and $\mathbf{v}_e \in \mathbb{R}^d$ for the evidence component. These corrections are applied independently without updating model parameters.

4.1 Estimating Steering Vectors

ORBIT operates in a one-time calibration phase using a small set of parallel multimodal instances, where semantically equivalent claim and evidence texts in the anchor and target languages are paired with the same visual evidence, extending cross-lingual alignment techniques to multimodal settings.

Per-sample shift estimation. For each calibration instance i with parallel claims $(c^i_{\ell_{\text{anc}}}, c^i_{\ell_c})$, parallel evidence $(e^i_{\ell_{\text{anc}}}, e^i_{\ell_e})$, and shared visual content I^i, we extract hidden states at layer l for $(a, b) \in \{(\ell_{\text{anc}}, \ell_{\text{anc}}), (\ell_c, \ell_{\text{anc}}), (\ell_{\text{anc}}, \ell_e)\}$:

$$\mathbf{h}^{(l),i}_{a,b} = f^{(l)}_\theta(c^i_a, e^i_b, \mathrm{I}^i). \tag{1}$$

For each configuration, we compute mean representations over claim and evidence token spans S_c and S_e[4]:

$$\boldsymbol{\mu}^i_{c,a,b} = \frac{1}{|S_c|}\sum_{t\in S_c} \mathbf{h}^{(l),i}_{a,b}[t], \tag{2}$$

$$\boldsymbol{\mu}^i_{e,a,b} = \frac{1}{|S_e|}\sum_{t\in S_e} \mathbf{h}^{(l),i}_{a,b}[t]. \tag{3}$$

[4]Spans are identified via explicit delimiter tokens in the prompt (see Appendix D).

To disentangle language-specific shifts, we hold one textual component fixed in ℓ_{anc} and measure the representation change when the other translates: claim from ℓ_{anc} to ℓ_c, or evidence from ℓ_{anc} to ℓ_e:

$$\mathbf{v}_c^i = \boldsymbol{\mu}^i_{c,\ell_{\text{anc}},\ell_{\text{anc}}} - \boldsymbol{\mu}^i_{c,\ell_c,\ell_{\text{anc}}}, \tag{4}$$

$$\mathbf{v}_e^i = \boldsymbol{\mu}^i_{e,\ell_{\text{anc}},\ell_{\text{anc}}} - \boldsymbol{\mu}^i_{e,\ell_{\text{anc}},\ell_e}. \tag{5}$$

Both shifts use the fully anchor-language configuration $(\ell_{\text{anc}}, \ell_{\text{anc}})$ as the reference point to ensure consistency.

Sign-aligned aggregation. Following Zou et al. (2025), we prevent destructive cancellation by aligning vector signs before averaging. While the original RepE formulation uses binary task labels to determine sign direction, our calibration phase contains no task labels. We therefore adapt this step to a label-free setting by estimating a reference direction from the first k calibration samples and aligning each vector to it using dot-product similarity. For each component $x \in \{c, e\}$:

1. Compute reference direction from the first k samples:

$$\mathbf{v}_{x,\text{ref}} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{v}_x^i. \tag{6}$$

2. Determine alignment signs:

$$s_x^i = \text{sign}\big(\langle \mathbf{v}_x^i, \mathbf{v}_{x,\text{ref}} \rangle\big). \tag{7}$$

3. Aggregate with sign correction:

$$\mathbf{v}_x = \frac{1}{N} \sum_{i=1}^{N} s_x^i \cdot \mathbf{v}_x^i. \tag{8}$$

4.2 Applying Steering Vectors at Inference

At test time, given an instance with claim in language ℓ_c and evidence in language ℓ_e, we apply both precomputed steering vectors via forward hooks at layer l:

$$\tilde{\mathbf{h}}^{(l)}[t] = \mathbf{h}^{(l)}[t] + \lambda_c \cdot \mathbf{v}_c, \quad \forall t \in S_c, \tag{9}$$

$$\tilde{\mathbf{h}}^{(l)}[t] = \mathbf{h}^{(l)}[t] + \lambda_e \cdot \mathbf{v}_e, \quad \forall t \in S_e, \tag{10}$$

where the per-component strengths λ_c, λ_e are derived automatically from the relative vector norms to avoid manual tuning. Let $n_c = \|\mathbf{v}_c\|$ and $n_e = \|\mathbf{v}_e\|$. We set:

$$\lambda_c = \lambda \cdot \frac{n_e}{n_c + n_e}, \quad \lambda_e = \lambda \cdot \frac{n_c}{n_c + n_e}, \tag{11}$$

with a uniform fallback $\lambda_c = \lambda_e = \lambda/2$ when $n_c + n_e < 10^{-8}$. This inverse-norm weighting ensures the smaller vector receives proportionally stronger scaling, preventing under-steering of weakly shifted components. Steering vectors are broadcast across batch and sequence dimensions; modified representations propagate through remaining layers without parameter updates.

5 Experiments

5.1 Setup

Dataset. We evaluate on MOCHEG-XL, our machine-translated extension of MOCHEG (Yao et al., 2023) spanning nine languages. We use 2,442 test instances and split the 1,490 development instances into 372 calibration and 1,118 validation samples. Each instance contains a claim, textual evidence, visual evidence, and a three-way label: *Supported*, *Refuted*, or *NEI (Not Enough Information)*.

Model and configuration. We evaluate Qwen3-VL 4B, Gemma 4 E2B, and LFM2-VL 3B. Steering vectors are estimated via sign-aligned aggregation in Section 4.1, with key hyperparameters selected mainly from Qwen3-VL validation experiments, while LFM2-VL and Gemma 4 are evaluated to demonstrate generality. Full implementation and hyperparameter details are provided in Appendix F. We assess significance using paired bootstrap resampling over test instances with 10,000 resamples, matching the aggregation level of the reported tier means.

Method	High-resource evidence						Low-resource evidence				
	en	es	ru	vi	ja	mean	jw	cy	mi	ti	mean
Qwen3-VL 4B											
ΔAccuracy											
Pivot	$-1.48_{\pm 2.40}$	$-1.99_{\pm 1.88}$	$-0.50_{\pm 2.03}$	$-1.52_{\pm 2.01}$	$+0.61_{\pm 1.78}$	-0.98	$+0.99_{\pm 1.34}$	$-1.28_{\pm 1.89}$	$+0.42_{\pm 1.74}$	$+\mathbf{2.64}_{\pm 2.54}$	+0.69
CoT	$-1.77_{\pm 2.03}$	$-0.34_{\pm 2.20}$	$+0.05_{\pm 1.61}$	$-0.08_{\pm 1.78}$	$+0.84_{\pm 1.66}$	-0.26	$+1.91_{\pm 1.66}$	$+0.64_{\pm 2.18}$	$+0.62_{\pm 1.24}$	$+0.20_{\pm 1.56}$	+0.84
ORBIT	$+\mathbf{0.49}_{\pm 1.21}$	$+\mathbf{0.80}_{\pm 0.82}$	$+\mathbf{1.44}_{\pm 0.92}$	$+\mathbf{1.37}_{\pm 0.62}$	$+\mathbf{2.09}_{\pm 1.22}$	**+1.24**	$+\mathbf{2.70}_{\pm 1.27}$	$+\mathbf{2.22}_{\pm 1.12}$	$+\mathbf{1.80}_{\pm 0.99}$	$+1.76_{\pm 0.96}$	**+2.12**
ΔF1-score											
Pivot	$-3.08_{\pm 2.45}$	$-3.98_{\pm 2.13}$	$-2.51_{\pm 2.38}$	$-3.28_{\pm 2.32}$	$-1.50_{\pm 2.70}$	-2.87	$-0.40_{\pm 1.53}$	$-2.13_{\pm 2.70}$	$-1.80_{\pm 2.27}$	$+\mathbf{3.05}_{\pm 1.44}$	-0.32
CoT	$-3.20_{\pm 3.81}$	$-2.02_{\pm 3.56}$	$-1.67_{\pm 2.65}$	$-1.77_{\pm 3.08}$	$-1.25_{\pm 2.20}$	-1.98	$+0.40_{\pm 2.50}$	$-0.37_{\pm 2.98}$	$+0.17_{\pm 2.37}$	$-2.96_{\pm 2.46}$	-0.69
ORBIT	$-\mathbf{1.63}_{\pm 2.01}$	$-\mathbf{1.09}_{\pm 1.37}$	$-\mathbf{0.53}_{\pm 1.15}$	$-\mathbf{0.44}_{\pm 1.20}$	$+\mathbf{0.16}_{\pm 1.42}$	**-0.71**	$+\mathbf{1.70}_{\pm 1.14}$	$+\mathbf{1.84}_{\pm 0.73}$	$+\mathbf{2.56}_{\pm 1.31}$	$+2.68_{\pm 1.21}$	**+2.20**
Gemma 4 E2B											
ΔAccuracy											
Pivot	$-1.32_{\pm 1.31}$	$+0.10_{\pm 1.63}$	$+1.05_{\pm 0.93}$	$+0.41_{\pm 1.45}$	$+1.44_{\pm 1.43}$	+0.34	$+1.09_{\pm 1.09}$	$+0.80_{\pm 1.96}$	$-0.27_{\pm 1.90}$	$+\mathbf{3.23}_{\pm 2.11}$	+1.21
CoT	$-1.87_{\pm 1.56}$	$+0.77_{\pm 1.28}$	$+0.94_{\pm 1.11}$	$+0.72_{\pm 1.46}$	$+0.74_{\pm 1.33}$	+0.26	$+0.41_{\pm 1.68}$	$+0.12_{\pm 1.67}$	$+0.06_{\pm 1.94}$	$+0.73_{\pm 1.94}$	+0.33
ORBIT	$+\mathbf{1.48}_{\pm 0.90}$	$+\mathbf{2.59}_{\pm 1.20}$	$+\mathbf{2.54}_{\pm 0.63}$	$+\mathbf{2.61}_{\pm 1.06}$	$+\mathbf{2.32}_{\pm 0.63}$	**+2.31**	$+\mathbf{2.45}_{\pm 0.80}$	$+\mathbf{1.45}_{\pm 1.01}$	$+\mathbf{1.39}_{\pm 1.08}$	$+1.34_{\pm 1.17}$	**+1.66**
ΔF1-score											
Pivot	$-1.16_{\pm 1.29}$	$-0.20_{\pm 1.47}$	$+\mathbf{1.07}_{\pm 1.14}$	$+0.34_{\pm 1.42}$	$+\mathbf{1.42}_{\pm 2.10}$	+0.37	$+1.77_{\pm 1.44}$	$+0.61_{\pm 2.39}$	$-2.32_{\pm 2.59}$	$+\mathbf{4.61}_{\pm 3.43}$	+1.17
CoT	$-1.81_{\pm 1.65}$	$+0.43_{\pm 1.12}$	$+0.39_{\pm 0.92}$	$+0.26_{\pm 1.40}$	$+0.31_{\pm 1.19}$	-0.08	$+1.00_{\pm 1.59}$	$+1.36_{\pm 2.19}$	$+1.27_{\pm 2.32}$	$+2.14_{\pm 3.07}$	+1.44
ORBIT	$+\mathbf{0.58}_{\pm 1.15}$	$+\mathbf{1.30}_{\pm 2.04}$	$+0.57_{\pm 1.37}$	$+\mathbf{0.65}_{\pm 1.77}$	$+0.58_{\pm 1.44}$	**+0.74**	$+\mathbf{2.86}_{\pm 0.75}$	$+\mathbf{2.40}_{\pm 1.25}$	$+\mathbf{2.26}_{\pm 1.22}$	$+4.05_{\pm 0.81}$	**+2.89**
LFM2-VL 3B											
ΔAccuracy											
Pivot	$+0.65_{\pm 1.84}$	$+0.09_{\pm 1.89}$	$+\mathbf{2.78}_{\pm 1.94}$	$+3.77_{\pm 1.66}$	$+1.40_{\pm 2.39}$	+1.74	$+2.57_{\pm 1.14}$	$+1.13_{\pm 0.86}$	$+1.39_{\pm 1.06}$	$+\mathbf{5.56}_{\pm 1.16}$	+2.66
CoT	$-2.31_{\pm 1.61}$	$-1.41_{\pm 2.09}$	$+1.35_{\pm 1.62}$	$+3.53_{\pm 2.06}$	$+1.47_{\pm 2.57}$	+0.53	$+2.98_{\pm 1.48}$	$+2.18_{\pm 0.78}$	$+2.52_{\pm 1.01}$	$+2.13_{\pm 1.37}$	+2.45
ORBIT	$+\mathbf{2.11}_{\pm 0.82}$	$+\mathbf{2.28}_{\pm 1.16}$	$+2.65_{\pm 0.74}$	$+\mathbf{5.15}_{\pm 1.31}$	$+\mathbf{2.78}_{\pm 0.76}$	**+3.00**	$+\mathbf{3.06}_{\pm 1.34}$	$+\mathbf{4.42}_{\pm 1.92}$	$+\mathbf{3.64}_{\pm 1.18}$	$+4.30_{\pm 1.01}$	**+3.85**
ΔF1-score											
Pivot	$+0.85_{\pm 2.65}$	$+0.34_{\pm 2.65}$	$+3.12_{\pm 2.97}$	$+4.92_{\pm 2.08}$	$+1.24_{\pm 3.31}$	+2.09	$+3.75_{\pm 2.07}$	$+0.67_{\pm 1.62}$	$-0.33_{\pm 2.00}$	$+3.96_{\pm 2.15}$	+2.01
CoT	$-4.16_{\pm 2.97}$	$-2.37_{\pm 3.85}$	$+0.21_{\pm 3.23}$	$+5.23_{\pm 3.37}$	$+3.23_{\pm 4.21}$	+0.43	$+3.39_{\pm 3.16}$	$+1.27_{\pm 1.68}$	$+1.40_{\pm 1.96}$	$-2.32_{\pm 1.30}$	+0.93
ORBIT	$+\mathbf{2.98}_{\pm 1.09}$	$+\mathbf{3.40}_{\pm 1.21}$	$+\mathbf{3.64}_{\pm 0.99}$	$+\mathbf{5.69}_{\pm 1.68}$	$+\mathbf{3.66}_{\pm 1.03}$	**+3.87**	$+\mathbf{4.33}_{\pm 2.18}$	$+\mathbf{5.30}_{\pm 2.47}$	$+\mathbf{4.58}_{\pm 1.74}$	$+\mathbf{4.01}_{\pm 0.61}$	**+4.55**

Table 1: Comparison of ORBIT with translation-based baselines across evidence languages. Pivot refers to Pivot NLLB and CoT refers to Self-translate CoT. Values report mean ± standard deviation for ΔAccuracy and ΔF1-score relative to the zero-shot baseline, averaged over claim languages. Bold indicates the best result per language and metric. Full per-configuration results are provided in Appendix E.

Baselines. We compare ORBIT with two translation-based frameworks. **Self-translate CoT** (Chain-of-Thought) prompts the base model to translate non-English text into English before verification. **Pivot NLLB** translates all input text using NLLB-200 distilled 600M[5] (Costa-Jussà et al., 2022) before SVLM prediction. We use the distilled 600M variant to keep the comparison within a resource-constrained deployment setting; higher-capacity NLLB variants or commercial MT systems are treated as upper-bound pipelines outside our small-model deployment scope.

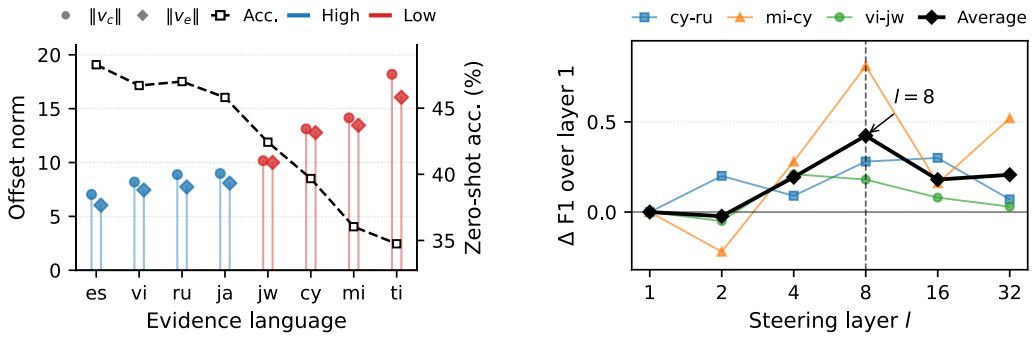
5.2 Main Results

Table 1 reports changes in accuracy and macro F1-score relative to the zero-shot baseline. For each evidence language, we average results over all claim languages; the standard deviation (SD) reflects how sensitive each method is to claim-language variation. We prioritize per-evidence aggregations, as they directly isolate the source of cross-lingual degradation; the complete 81-configuration matrix is provided in Appendix E.

Low-resource recovery. ORBIT substantially recovers low-resource evidence, pulling collapsed representations toward the English anchor. Gains are model-dependent: LFM2-VL benefits most across both tiers, Gemma 4 improves steadily, while Qwen3-VL shows more modest recovery and incurs a high-resource F1 drop. For Qwen3-VL, paired bootstrap resampling confirms these low-resource gains remain statistically significant despite a smaller magnitude (Appendix G), while the model's high-resource F1 degradation is driven primarily by the *NEI* class (Appendix H). This occurs because overshooting erases the representational ambiguity required for abstention and pushes borderline instances toward *Supported* or *Refuted*.

Evidence-language asymmetry. The gains follow the same asymmetry observed in the zero-shot collapse: evidence language, rather than claim language, is the dominant source of performance variation. Evidence passages are longer and carry the factual content needed to ground the image,

[5] facebook/nllb-200-distilled-600M



(a) (b)

Figure 4: Diagnostic analyses on Qwen3-VL: (a) Offset norms and zero-shot accuracy by evidence language. Accuracy is averaged over all claim languages for each fixed evidence language. (b) Layer-selection ablation on the 1,118-instance validation split. ΔF1 is measured relative to layer 1.

whereas claims act as compact verification hypotheses. Consequently, ORBIT yields consistent accuracy improvements across claim languages (full results in Appendix E), while the largest recoveries occur when the evidence language is low-resource. The claim language itself matters less than the evidence language.

Comparison with translation-based baselines. ORBIT consistently outperforms both Self-translate CoT and Pivot NLLB across all evaluated architectures, with the margin widening for low-resource evidence. Despite differing in whether translation is handled internally or by an external model, both baselines share a translate-then-verify paradigm that falls below zero-shot across all evidence languages. Translation noise on long passages disrupts cross-modal grounding and incurs a severe latency overhead. By operating directly on hidden representations, ORBIT bypasses auxiliary decoding to deliver more robust grounding recovery. However, the impact of static per-language steering is model-dependent. Although LFM2-VL and Gemma 4 achieve consistent gains across both resource tiers, uniform vectors may over-steer already well-aligned high-resource representations in Qwen3-VL. This contrast exposes a clear trade-off between low-resource recovery and high-resource preservation.

5.3 Calibration Norms as a Diagnostic Proxy

Figure 4a shows that evidence offset norms from 372 calibration samples strongly predict zero-shot accuracy collapse in Figure 2, yielding a Pearson correlation of -0.986 and enabling lightweight triage without full-test evaluation. Claim norms exceed evidence norms, yet evidence-side steering drives recovery because evidence carries the factual content required for visual grounding. This strong correlation holds because offset norms directly quantify the representational gap between anchor and target language hidden states prior to inference. Appendix I reports a representative calibration-size sensitivity check on Qwen3-VL with Vietnamese claims and low-resource evidence languages.

5.4 Layer Selection

We evaluate steering layers $l \in \{1, 2, 4, 8, 16, 32\}$ on the validation split for Qwen3-VL (Figure 4b). Performance peaks at $l = 8$, but ORBIT remains stable across layers, with mean F1 varying by less than 0.5 points. This suggests that the method does not rely on a finely tuned intervention depth. Mid-layer steering is slightly stronger, likely because multilingual representations are already partially composed but not yet fully specialized toward the final verdict. Sensitivity is higher for severely misaligned pairs such as mi–cy ($+0.81$ from layer 1 to 8), while vi–jw changes only modestly ($+0.18$).

Method	Inference Time	ΔAccuracy $_{\text{low-res}}$
Zero-shot baseline	**1057.28s**	–
Pivot NLLB	2007.45s	+0.69
Self-translate CoT	20888.31s	*+0.84*
ORBIT (Ours)	*1258.62s*	**+2.12**

Table 2: Mean inference time on the test set using Qwen3-VL. ORBIT introduces moderate overhead versus the zero-shot baseline but remains significantly faster and more resource-efficient than translation-based pipelines.

5.5 Efficiency Analysis

Table 2 compares inference time of Qwen3-VL across methods on the test set. Self-translate CoT is prohibitively slow, taking nearly 20$\times$ longer than zero-shot because it must autoregressively translate long evidence passages before verification. Pivot NLLB is faster but still requires a separate translation model, increasing memory use and pipeline complexity. ORBIT avoids external components and applies steering through a lightweight forward hook, adding 19% overhead over zero-shot while achieving the largest low-resource accuracy gain. This makes ORBIT the fastest non-baseline intervention and more practical for resource-constrained deployment.

6 Conclusion

We identify and quantify a severe, resource-asymmetric collapse in small vision-language models for cross-lingual multimodal fact verification. While high-resource languages degrade modestly, low-resource evidence languages drive accuracy to near-random levels despite intact visual encoding. To address this without gradient updates, we propose ORBIT, a training-free activation steering method that realigns target-language representations toward the model's high-resource anchor space using per-language correction vectors estimated from parallel bilingual pairs. ORBIT partially recovers low-resource performance while preserving capability on high-resource languages, establishing that inference-time representation adjustment can bridge cross-lingual gaps when training resources are limited.

Limitations

ORBIT exhibits a model-dependent trade-off. Static per-language steering vectors applied uniformly cannot simultaneously preserve well-aligned high-resource representations and correct severely misaligned low-resource ones. As shown in Section 5.2, this trade-off primarily affects Qwen3-VL, where high-resource evidence incurs F1 penalties – especially on the NEI class. In contrast, low-resource evidence benefits from realignment across all models. Future work should explore instance-specific modulation based on estimated representation quality rather than coarse resource-tier heuristics, though reliably estimating such quality at inference time without ground-truth supervision remains non-trivial.

Our evaluation is intentionally controlled: MOCHEG-XL is machine-translated and serves as a diagnostic instrument rather than a human-verified benchmark. Consequently, it does not capture noisier evidence, code-switching, culturally specific visuals, or multilingual multi-source settings found in real-world fact-checking. We pair each claim with single-language evidence and do not yet aggregate conflicting cross-lingual signals. Steering vectors estimated from machine-translated instances inherit translation noise, especially for severely low-resource languages. This warrants investigation into higher-quality calibration sources and robustness tests under systematic noise.

AI use statement

In this work, we used ChatGPT 5.5, Claude Sonnet 4.5, and Gemini 3.1 Pro for language polishing, grammar checking, sentence-level clarity improvements, and coding assistance. We have reviewed

all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work addresses cross-lingual inequity in multimodal fact verification using publicly available artifacts under their respective licenses. MOCHEG-XL is a machine-translated diagnostic extension designed for controlled analysis of linguistic shift, not a human-verified benchmark. Upon acceptance, we will release MOCHEG-XL and ORBIT implementation under an open license to support reproducible, equitable multilingual research.

REFERENCES

Sheikh Iftekhar Ahmed, Muhammad Zubair Hasan, Abrar Jahin Niloy, Syed Mostofa Monsur, and Mark V. Albert. Scaling down, powering up: A survey on the advancements of small vision-language models. *Information Fusion*, 127:103805, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103805. URL https://www.sciencedirect.com/science/article/pii/S156625352500867X.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, et al. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Anthropic. Claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7. Accessed: April 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. DEFAME: Dynamic evidence-based FAct-checking with multimodal experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5383–5417. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/braun25b.html.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2370–2392. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bugliarello22a.html.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. MetaSumPerceiver: Multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8742–8757, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.474. URL https://aclanthology.org/2024.acl-long.474/.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958830.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Google DeepMind. Gemini 3.1 flash-lite, 2026a. URL https://deepmind.google/models/gemini/flash-lite/. Accessed: May 2026.

Google DeepMind. Gemini 3.1 pro model evaluation. Technical report, 2026b. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_3-1_pro_model_evaluation.pdf. Accessed: April 2026.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Jaeyoung Lee, Ximing Lu, Jack Hessel, Faeze Brahman, Youngjae Yu, Yonatan Bisk, Yejin Choi, and Saadia Gabriel. How to train your fact verifier: Knowledge transfer with multimodal open models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13060–13077, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.764. URL https://aclanthology.org/2024.findings-emnlp.764/.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15100–15119, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.734. URL https://aclanthology.org/2025.acl-long.734/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Xuannan Liu, Zekun Li, Pei Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=D6zn6ozJs7.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.545. URL https://aclanthology.org/2021.emnlp-main.545/.

Shreyash Mishra, Suryavardan S, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, and Asif Ekbal. Factify: A multi-modal fact

verification dataset, 2022. URL https://ceur-ws.org/Vol-3199/paper18.pdf. DeFactify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022.

OpenAI. Gpt-5.4 thinking: Deployment and safety evaluation. Technical report, 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf. Accessed: April 2026.

Hui Pang, Chaozhuo Li, Litian Zhang, Senzhang Wang, and Xi Zhang. Beyond text: Fine-grained multi-modal fact verification with hypergraph transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6389–6397, Apr. 2025. doi: 10.1609/aaai.v39i6.32684. URL https://ojs.aaai.org/index.php/AAAI/article/view/32684.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024. ISSN 2192-662X. doi: 10.1007/s13735-023-00312-6. URL https://doi.org/10.1007/s13735-023-00312-6.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2497–2511, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.196. URL https://aclanthology.org/2022.findings-acl.196/.

Rhitabrat Pokharel, Ameeta Agrawal, and Tanay Nagar. Cross-lingual activation steering for multilingual language models. *arXiv preprint arXiv:2601.16390*, 2026.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5227–5237, June 2022.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 715–729, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.45. URL https://aclanthology.org/2022.emnlp-main.45/.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1), 1993.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9279–9300, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.630. URL https://aclanthology.org/2022.emnlp-main.630/.

Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3544–3552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.326. URL https://aclanthology.org/2020.acl-main.326/.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5418–5433, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.270. URL https://aclanthology.org/2025.acl-long.270/.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in LVLMs via representation

engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.725. URL https://aclanthology.org/2025.emnlp-main.725/.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2733–2743, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591879. URL https://doi.org/10.1145/3539618.3591879.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4155–4165, June 2021.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2025.

APPENDICES

A MOCHEG STATISTICS

Statistic	Dev	Test
Number of claims	1,490	2,442
Evidence Distribution		
Number of text paragraphs	4,067	6,268
Number of images	1,178	2,007
Label Distribution		
Refuted	488	825
Supported	501	817
NEI	501	800

Table 3: Detailed statistics of the MOCHEG dataset splits used in our experiments.

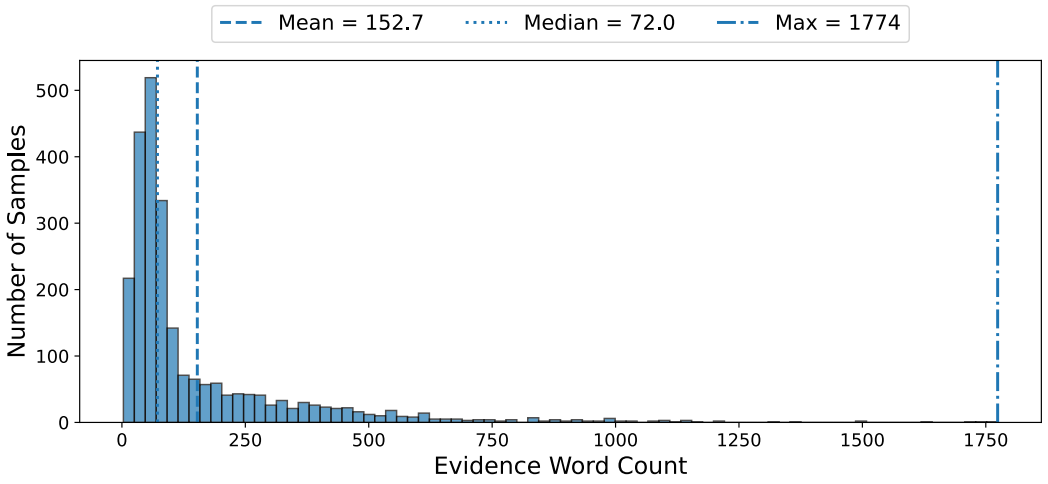


Figure 5: Word count distribution of textual evidence in the MOCHEG testing set. URLs are counted as single tokens, and words with hyphens or apostrophes are preserved as single units.

B TRANSLATION QUALITY

We evaluate translation quality using GPT-5.4 Mini (OpenAI, 2026) and Gemini 3.1 Flash-Lite (Google DeepMind, 2026a) as LLM judges (Zheng et al., 2023).

B.1 PROMPT

```
Evaluate whether the translated text preserves the meaning of the
    English text.

Consider the overall meaning and important factual details. Allow
    natural paraphrasing, reordering, and language-specific phrasing.
    Mark as not adequate if the translation omits important information,
    adds unsupported information, changes facts, reverses meaning, or is
    too incomplete/unclear to verify.

Return ONLY valid JSON:
```

```
{{
  "adequate": true/false,
  "score": 0-100,
  "reason": "brief explanation",
  "major_issues": []
}}

Use "major_issues" only for important meaning changes. Leave it empty if
    the translation is adequate.

English text:
{english_text}

Translated text:
{translated_text}
```

B.2 RESULTS

Model	Metric	High-Resource				Low-Resource			
		es	ru	vi	ja	jw	cy	mi	ti
GPT 5.4 Mini	Claim	96.86	94.42	95.72	91.55	89.49	95.59	87.61	92.06
	Evidence	96.56	93.44	93.07	90.33	83.81	83.00	76.58	76.08
Gemini 3.1 Flash-Lite	Claim	98.10	93.40	95.31	93.53	86.75	95.50	86.10	93.62
	Evidence	98.38	92.89	94.21	93.33	81.50	87.61	75.26	81.92

Table 4: Average LLM-as-a-judge evaluation scores on MOCHEG-XL across high-resource and low-resource languages.

Our evaluation protocol mitigates potential LLM-as-a-judge bias in low-resource languages by explicitly prioritizing factual alignment over syntactic fluency. Instead of evaluating the low-resource output in isolation, the prompt (Appendix B.1) uses the original English text as a reference to verify core entities and key facts. Furthermore, two distinct LLM judges produce highly consistent scores across all target languages (Table 4). This agreement indicates that the evaluations reflect actual semantic preservation rather than shared hallucinations. Consequently, while low-resource translation may introduce minor grammatical noise, the core facts remain intact. This ensures MOCHEG-XL remains a valid diagnostic instrument for isolating cross-modal grounding failures, rather than merely reflecting translation errors.

C BASE RESULTS

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	48.08	46.60	45.17	45.21	45.99	39.72	37.10	38.04	35.34
es	49.63	48.03	45.66	45.45	44.92	39.68	36.32	36.77	34.81
ru	48.24	45.78	46.03	45.13	44.76	40.25	35.30	36.12	34.03
vi	47.58	45.13	44.88	46.72	45.29	39.68	35.01	36.40	34.40
ja	48.24	46.15	44.27	44.39	46.89	38.78	34.11	35.22	33.09
jw	44.14	41.65	41.89	40.91	41.85	41.44	33.42	35.67	32.88
cy	41.69	37.96	37.96	37.22	38.21	36.40	38.12	34.73	32.96
mi	43.00	39.23	38.86	39.23	39.89	36.04	33.78	39.35	33.58
ti	43.24	38.86	38.57	39.35	38.78	36.61	33.70	34.44	36.86

Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	46.43	45.18	44.25	44.27	45.41	38.52	35.82	36.94	33.02
es	48.34	46.25	45.54	45.67	45.03	39.15	35.27	35.67	32.03
ru	46.97	45.29	44.92	45.23	44.89	39.60	33.81	34.80	31.22
vi	46.45	44.86	45.12	45.90	45.46	38.76	33.15	35.06	30.80
ja	46.94	45.78	44.31	44.50	46.64	38.03	32.28	33.71	29.59
jw	43.97	41.88	42.17	41.34	42.22	40.71	31.52	34.27	29.62
cy	41.93	38.19	38.04	37.35	38.24	35.54	37.63	32.97	29.79
mi	43.23	39.63	39.26	39.62	40.18	35.18	32.07	39.09	30.67
ti	43.54	39.19	38.60	39.33	38.68	35.11	30.92	31.65	35.52

F1-score (%)

Performance levels: High (> 45%) Medium (35–45%) Low (< 35%)

Figure 6: Cross-lingual performance of Gemma 4 across 81 claim–evidence configurations.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	46.81	44.88	39.72	34.64	40.66	32.43	31.90	30.96	30.71
es	45.54	46.56	39.35	35.54	41.36	33.29	32.47	31.61	30.79
ru	43.04	42.26	41.89	34.32	39.93	32.39	30.88	30.88	30.10
vi	41.56	38.82	35.71	36.49	38.17	31.24	30.47	30.18	29.44
ja	44.92	44.02	40.09	36.00	44.80	32.60	31.74	31.20	31.00
jw	39.60	39.56	35.09	33.91	36.73	31.45	30.26	29.77	28.58
cy	38.66	37.02	34.44	32.96	35.95	31.33	30.34	30.84	27.93
mi	39.19	37.80	34.11	32.39	36.65	31.70	31.16	32.19	27.85
ti	37.31	36.53	32.72	32.23	34.56	32.15	30.92	30.22	27.15

Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	45.26	42.93	37.77	31.29	38.41	28.64	27.59	26.56	27.18
es	43.63	44.74	37.30	32.51	38.67	30.18	28.22	27.54	27.84
ru	41.06	39.98	40.62	30.82	37.48	28.72	26.58	26.53	26.66
vi	38.62	35.46	31.72	33.61	34.24	26.27	24.57	24.51	25.04
ja	43.96	42.78	38.51	32.93	43.69	29.51	27.76	27.14	27.14
jw	35.77	35.50	30.89	29.22	32.17	28.08	25.13	24.78	24.77
cy	35.01	33.11	29.96	27.59	31.26	25.93	26.20	25.52	24.04
mi	35.42	33.47	29.40	27.02	31.61	26.75	25.61	28.16	23.84
ti	33.10	32.10	27.19	26.32	29.48	25.95	23.63	23.08	21.95

F1-score (%)

Performance levels: High ($>40\%$) Medium (30–40%) Low ($<30\%$)

Figure 7: Cross-lingual performance of LFM2-VL across 81 claim–evidence configurations.

Corroborating the trends observed for Qwen3-VL in Section 3, both Gemma 4 and LFM2-VL exhibit a severe evidence-language collapse (Figures 6 and 7). When evidence shifts to low-resource languages (jw, cy, mi, ti), performance across all claim languages degrades to near-random levels. Crucially, the same-language diagonal advantage present in high-resource settings disappears entirely. These findings confirm that cross-modal grounding failure under low-resource conditions is a generalized vulnerability of SVLMs, rather than an isolated architectural artifact.

D Prediction Prompt

```
Your task is to verify the truthfulness of the following claim using the
    provided text evidence and the accompanying image evidence.

IMPORTANT: The marker strings are for processing only; ignore them.

Classify the claim into exactly one of the following categories:
- Supported
- Refuted
- NEI (Not Enough Information)

Claim:
<CLAIM_START>
{claim_text}
<CLAIM_END>

Text Evidence:
<EVID_START>
{evidence_text}
<EVID_END>

Output only the label (Supported, Refuted, or NEI).
```

E Full Results

Figures 8–16 report the complete per-configuration results across all 81 claim–evidence language pairs. Each cell shows the change in accuracy or macro F1-score relative to the zero-shot baseline. These full matrices support the aggregate trends in Table 1: evidence language drives most performance variation, ORBIT produces the clearest recovery under low-resource evidence, and translation-based baselines are less stable across the full language grid.

F Implementation Details

F.1 ORBIT Hyperparameters and Inference Settings

Steering vectors are estimated from calibration samples using sign-aligned aggregation. We use the first $k = 16$ samples to compute the reference direction. Hyperparameters are selected on the Qwen3-VL validation split and applied uniformly across all models to demonstrate generality. Steering is applied at layer 8, selected from $\{1, 2, 4, 8, 16, 32\}$. We apply a single steering strength $\lambda = 0.05$ across all evidence languages, tuned on the validation split from $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Images are resized so that the shortest side is 480 pixels while preserving the original aspect ratio. Generation uses greedy decoding with a maximum of 16 new tokens. All experiments are run on a single NVIDIA A100 SXM4 40GB GPU.

F.2 Pivot NLLB Implementation

For Pivot NLLB, we back-translate the non-English MOCHEG-XL claims and evidence into English, rather than using the original English MOCHEG inputs. Since our NLLB pipeline uses a 512-token maximum input length, long evidence passages are segmented with NLTK sentence tokenization before translation. Consecutive sentences are grouped into chunks of approximately 80 words, a conservative threshold chosen to keep low-resource language inputs below the token limit because words may split into multiple subword tokens.

Each chunk is translated independently with beam search using four beams, a repetition penalty of 1.15, and a no-repeat trigram constraint. Translated chunks are concatenated in their original order to reconstruct the English evidence passage. Empty evidence fields are left unchanged.

F.3 Self-translate CoT Prompt

For the Self-translate CoT baseline, the base model is first prompted to translate non-English text into English before verification. We use the following translation prompt:

```
Translate the following text into English.
- Preserve the original meaning exactly.
- Keep names, dates, numbers, and units unchanged.
- Do NOT add explanations, notes, or any extra text.

Text:
{text}

Output ONLY the English translation.
```

G Bootstrap Significance Analysis

Evidence	Accuracy 95% CI	Macro F1 95% CI
jw	[+1.82, +3.59]	[+0.89, +2.52]
cy	[+1.41, +3.02]	[+1.08, +2.61]
mi	[+0.98, +2.61]	[+1.78, +3.33]
ti	[+1.04, +2.46]	[+1.93, +3.43]
Tier mean	[**+1.50**, **+2.73**]	[**+1.63**, **+2.78**]

Table 5: Paired bootstrap 95% confidence intervals for ORBIT (Qwen3-VL) gains under low-resource evidence. All intervals exclude zero, indicating statistically significant improvements.

Among the three evaluated models, Qwen3-VL exhibits the smallest low-resource gains (Table 1). Because it also serves as the primary backbone for our in-depth diagnostic and ablation analyses (e.g., Sections 5.3, 5.4), we focus our rigorous significance testing on this most conservative case.

While we do not exhaustively bootstrap all architectures, the larger point estimates for Gemma 4 and LFM2-VL on the same large test set ($n = 2,442$) strongly suggest their improvements are similarly robust.

We assess significance via paired bootstrap resampling (Efron, 1979; Tibshirani & Efron, 1993), which estimates uncertainty directly from the data without assuming any particular score distribution. In each of 10,000 iterations, we draw test instances with replacement, compute the paired difference (ORBIT minus zero-shot) for each instance, and record the aggregated tier-level gain. The 2.5th and 97.5th percentiles of the resulting distribution form the 95% confidence interval. As shown in Table 5, all intervals for both accuracy and macro F1 exclude zero across every low-resource evidence language, confirming that ORBIT's recovery is statistically significant even under the weakest-model setting.

H Classification Report

Evidence Language	Supported	Refuted	NEI	ΔMacro-F1
en	1.47	2.01	-8.38	-1.63
es	1.31	2.68	-7.25	-1.09
ru	1.97	3.00	-6.55	-0.53
vi	1.93	3.61	-6.88	-0.44
ja	2.34	5.89	-7.76	0.16
jw	2.72	7.86	-5.47	1.70
cy	2.07	7.35	-3.89	1.84
mi	1.37	8.87	-2.57	2.56
ti	1.40	7.05	-0.43	2.68

Table 6: Class-wise ΔF1 on Qwen3-VL by evidence language. Values report ORBIT minus zero-shot F1, averaged over all claim languages for each evidence language.

Table 6 decomposes ORBIT's macro F1 changes on Qwen3-VL by individual class performance. F1 drops on high-resource evidence stem almost entirely from the NEI class, while Supported and Refuted verdicts improve modestly. Uniform steering overshoots well-aligned representations and collapses the ambiguity required for abstention. Conversely, low-resource evidence benefits across all three classes, as ORBIT pulls severely misaligned representations back toward viable decision boundaries.

Tables 8, 9, and 10 show the full per-configuration breakdown for each class on Qwen3-VL. The pattern holds across all claim languages: high-resource evidence columns exhibit systematic NEI degradation, while low-resource columns show predominantly positive changes for Supported and Refuted, with NEI recovering from collapse.

I Calibration Size Sensitivity

We fix the claim language to Vietnamese and evaluate low-resource evidence settings on Qwen3-VL: vi–jw, vi–cy, vi–mi, and vi–ti. Vietnamese is chosen because it is a high-resource non-English language in our setup, providing a representative cross-lingual claim condition without relying on the English source language as the claim input. This also controls claim-side variation, so the analysis focuses on how calibration size affects steering under low-resource evidence shift. We vary the calibration size $N \in \{32, 64, 128, 256, 372\}$ while keeping all other hyperparameters fixed. Table 7 reports the per-setting accuracy and macro F1 changes.

N	**vi–jw**		**vi–cy**		**vi–mi**		**vi–ti**	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
32	48.52	46.67	45.57	45.22	40.82	41.17	40.48	40.70
64	48.34	46.51	45.30	44.96	40.82	41.17	40.66	40.85
128	48.61	46.84	46.02	45.69	40.91	41.26	40.66	40.88
256	48.52	46.76	45.75	45.44	41.00	41.34	40.48	40.65
372	48.61	46.84	45.48	45.13	40.73	41.07	40.48	40.69

Table 7: Calibration-size sensitivity for ORBIT on Qwen3-VL with Vietnamese claims and selected low-resource evidence languages.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.25	+1.93	+1.52	+2.01	+3.04	+2.71	+2.42	+2.22
es	-0.24	0.00	+1.28	+1.11	+1.39	+3.07	+2.18	+1.97	+2.79
ru	-0.37	+0.58	+0.94	+1.11	+2.42	+3.53	+2.34	+0.74	+1.40
vi	-0.65	+1.28	+1.32	+1.97	+3.00	+3.56	+2.34	+2.83	+2.01
ja	-0.82	+0.74	+1.68	+0.91	+2.67	+3.40	+3.04	+1.32	+1.85
jw	+1.35	+2.05	+2.95	+2.33	+3.69	+2.71	+1.68	+2.50	+1.93
cy	+2.63	+2.05	+2.38	+2.05	+3.07	+3.40	+3.24	+1.19	+0.49
mi	+1.19	+0.50	+0.49	+0.62	+0.74	+0.58	+0.90	+1.52	+0.12
ti	+0.86	-0.20	+0.01	+0.74	-0.12	+0.66	+0.09	0.00	+2.05

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	-2.18	-0.75	-0.87	-1.04	+1.27	+2.07	+2.91	+2.51
es	-2.83	-2.63	-1.31	-1.43	-1.71	+1.36	+1.55	+2.36	+3.10
ru	-3.27	-2.39	-2.12	-1.58	-0.68	+1.46	+1.75	+1.07	+1.84
vi	-3.44	-1.47	-1.40	-1.06	+0.06	+1.68	+1.85	+3.58	+2.32
ja	-3.95	-2.30	-1.42	-2.05	-1.00	+1.32	+2.40	+1.92	+2.44
jw	-1.47	+0.05	+0.84	+0.40	+1.91	+0.63	+2.41	+4.20	+3.16
cy	+0.64	+0.92	+1.29	+1.19	+2.51	+4.43	+1.82	+2.48	+1.20
mi	+0.39	+0.21	+0.16	+0.48	+0.81	+1.52	+1.62	+2.03	+0.23
ti	+0.87	+0.00	-0.05	+0.93	+0.54	+1.11	+0.07	+0.05	+3.60

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 8: Per-configuration ORBIT gains (Δ = **ORBIT** − **Zero-shot**) on Qwen3-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.49	+0.20	+1.06	+0.12	+1.11	+0.25	-0.29	-1.19
es	-0.33	-0.57	-0.29	+0.74	+0.90	+2.13	+0.61	+1.39	+0.74
ru	-0.82	-0.25	-0.49	+0.70	+2.25	+3.77	+0.90	+0.16	+0.78
vi	-0.16	+1.56	+2.29	+0.86	+2.74	+3.64	+2.25	+2.74	+0.45
ja	-1.64	+1.56	+0.70	-0.53	+1.64	+3.32	+1.11	-0.41	+0.16
jw	+0.08	+2.01	+2.38	+2.38	+2.33	+0.04	+3.60	+2.29	+2.74
cy	-2.66	-1.19	-0.37	-1.02	+0.37	+2.70	-3.52	+0.37	+0.33
mi	-2.54	-1.52	-1.39	-1.11	-0.29	+1.47	+2.33	+0.16	+0.74
ti	-6.06	-5.16	-2.62	-3.77	-2.54	-1.02	-1.80	-0.82	-2.99

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.71	-0.10	+0.76	-0.85	+0.37	-0.21	-0.88	-5.26
es	+0.27	-0.49	-0.46	+0.45	-0.01	+1.37	+0.09	+1.56	-2.10
ru	-1.91	-2.28	-2.27	-1.27	-0.39	+1.67	-0.01	+0.04	-2.24
vi	-1.72	-0.96	+0.41	-1.07	-0.03	+1.65	+1.27	+3.05	-2.40
ja	-2.48	+0.08	-1.13	-2.14	-0.70	+1.73	+0.38	-0.41	-2.76
jw	-0.89	+0.51	+0.70	+0.80	+0.28	-1.20	+3.54	+3.17	+0.28
cy	-3.09	-2.17	-1.43	-2.02	-1.15	+2.69	-4.06	+0.94	-2.04
mi	-3.68	-2.70	-2.70	-2.16	-1.51	+0.90	+1.83	-1.52	-1.79
ti	-12.12	-10.90	-8.04	-9.32	-6.91	-5.60	-6.17	-4.38	-8.34

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 9: Per-configuration Self-translate CoT gains (Δ = **Self-translate CoT** − **Zero-shot**) on Qwen3-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	-0.29	-0.04	+0.61	+0.98	+2.21	-0.78	+0.45	+4.56
es	+0.53	-1.88	-0.78	-1.06	+1.06	+2.33	-0.90	+1.43	+5.13
ru	-0.74	-2.62	-2.33	-1.97	+0.82	+1.97	-1.27	+0.16	+4.44
vi	-1.56	-2.46	-1.23	-1.39	+0.45	+0.70	-0.70	+3.19	+3.94
ja	-1.68	-1.56	+0.20	-1.84	+0.25	+1.39	+0.04	-0.37	+3.62
jw	+0.12	+0.66	+2.01	+0.33	+2.38	-0.82	+1.39	+2.46	+3.62
cy	-4.79	-4.42	-2.09	-3.81	-1.06	+1.64	-5.53	+0.29	-0.81
mi	-5.16	-5.00	-3.24	-5.24	-2.74	-1.47	-2.13	-1.97	-1.91
ti	+1.45	-0.36	+2.96	+0.67	+3.37	+1.00	-1.67	-1.83	+1.16

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	-2.35	-2.63	-1.51	-0.63	+0.31	-1.64	-1.40	+4.37
es	-1.28	-3.97	-2.95	-2.95	-0.46	+0.45	-2.00	-0.79	+4.86
ru	-2.38	-4.78	-4.39	-3.65	-0.97	-0.09	-2.64	-2.18	+3.80
vi	-3.24	-4.61	-3.03	-2.66	-1.22	-1.09	-1.77	+0.88	+3.61
ja	-2.85	-3.34	-1.79	-3.36	-1.20	-0.27	-1.23	-2.95	+3.17
jw	-1.72	-1.41	-0.04	-1.56	-0.05	-2.31	+0.62	+0.67	+3.64
cy	-6.37	-6.61	-4.31	-5.87	-4.28	+0.53	-7.50	-2.77	+1.65
mi	-6.96	-7.40	-5.60	-7.75	-7.08	-3.06	-4.67	-6.67	+1.98
ti	+0.14	-1.33	+2.16	-0.18	+2.41	+1.91	+1.64	-0.96	+0.36

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 10: Per-configuration Pivot NLLB gains (Δ = **Pivot NLLB** − **Zero-shot**) on Qwen3-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+1.92	+2.62	+2.50	+2.25	+3.40	+1.39	+1.56	+0.90
es	+0.41	+1.15	+1.35	+1.68	+2.33	+2.42	+1.97	+2.25	+1.47
ru	+1.19	+1.76	+1.92	+1.39	+2.01	+2.25	+1.68	+1.15	+1.68
vi	+1.84	+1.72	+2.38	+1.43	+2.21	+2.46	+2.29	+2.70	+0.57
ja	+1.39	+2.01	+3.48	+2.95	+0.94	+2.87	+1.72	+1.76	+2.25
jw	+2.33	+3.03	+2.66	+3.81	+2.54	+3.28	+0.20	-0.41	+0.12
cy	+2.29	+5.00	+3.15	+4.14	+2.58	+2.05	+3.15	+0.86	+0.41
mi	+2.38	+3.40	+2.42	+3.60	+3.24	+2.62	+0.16	+2.58	+0.74
ti	+0.04	+3.32	+2.87	+2.01	+2.78	+0.70	+0.49	+0.08	+3.89

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.06	+0.52	+0.09	+0.12	+3.56	+2.28	+1.96	+3.35
es	-1.33	-1.01	-1.33	-1.18	-0.07	+2.33	+2.58	+2.87	+4.48
ru	-0.01	-0.50	-1.29	-1.19	-0.08	+2.31	+2.89	+2.40	+4.69
vi	+0.35	-0.13	-0.31	-1.76	-0.09	+2.79	+4.07	+3.78	+4.25
ja	+0.58	+0.30	+1.04	+0.88	-2.11	+3.57	+3.48	+3.26	+5.71
jw	+1.27	+1.84	+0.94	+2.03	+0.97	+1.38	+1.90	+0.77	+3.26
cy	+2.09	+4.92	+2.44	+3.37	+2.05	+3.32	-0.22	+2.50	+3.38
mi	+1.94	+2.81	+0.78	+1.95	+2.00	+3.59	+1.56	-0.08	+3.58
ti	-0.24	+3.42	+2.36	+1.65	+2.43	+2.94	+3.05	+2.88	+3.78

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 11: Per-configuration ORBIT gains (Δ = **ORBIT** − **Zero-shot**) on Gemma 4 across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+1.29	+1.95	+1.78	+1.70	+2.18	-0.65	+0.42	+1.15
es	-3.05	-1.08	+1.09	-0.02	+0.80	+1.32	+0.09	+0.33	+1.07
ru	-2.23	+0.39	+1.01	+0.97	+1.25	+1.36	+0.29	+0.46	+1.36
vi	-0.71	+1.58	+0.97	-0.43	+0.60	+1.40	+0.70	+1.89	+2.18
ja	-0.67	-0.18	+1.78	+1.21	+0.11	+2.10	+1.44	+2.34	+2.30
jw	+0.47	+2.89	+2.44	+3.38	+3.46	-1.51	+2.55	+1.52	+2.83
cy	-2.39	+0.84	+0.11	+1.42	+0.19	-1.80	-3.56	-1.59	-0.73
mi	-1.86	+1.87	+0.43	-0.26	-0.30	+0.54	+0.42	-3.84	-0.08
ti	-4.52	-0.63	-1.29	-1.57	-1.16	-1.88	-0.20	-0.98	-3.47

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.85	+0.98	+1.33	+1.00	+2.69	+0.74	+1.76	+2.80
es	-3.07	-0.84	-0.08	-1.38	-0.23	+1.31	+1.04	+1.46	+2.81
ru	-1.99	-0.08	+0.98	+0.06	+0.49	+1.81	+1.89	+1.87	+3.02
vi	-0.63	+0.64	-0.35	-0.60	-0.15	+1.94	+2.53	+3.37	+5.03
ja	-0.09	-0.48	+0.96	+0.64	-0.14	+2.74	+3.23	+3.83	+4.47
jw	+0.09	+2.10	+1.70	+2.61	+2.95	-0.96	+4.29	+2.99	+4.98
cy	-2.28	+0.84	+0.22	+1.48	+0.36	-1.14	-3.51	-0.71	-0.20
mi	-1.66	+1.89	+0.49	-0.19	-0.11	+1.61	+1.35	-3.62	+0.86
ti	-4.85	-1.02	-1.39	-1.64	-1.39	-1.01	+0.70	+0.50	-4.55

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 12: Per-configuration Self-translate CoT gains (Δ = **Self-translate CoT** − **Zero-shot**) on Gemma 4 across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	-0.29	+1.88	+0.45	+1.64	+2.24	-0.08	-1.11	+4.15
es	-2.01	-2.54	0.00	-0.29	+1.35	+1.74	+0.49	+0.20	+4.56
ru	-1.35	-0.25	+0.12	+0.41	+1.68	+1.33	+1.64	+1.39	+4.39
vi	-1.47	+0.20	+1.02	-2.38	+0.20	+1.13	+2.33	+0.20	+4.56
ja	-1.64	-0.66	+2.01	+1.72	+0.78	+2.73	+2.95	+1.88	+5.01
jw	+0.33	+3.32	+1.92	+2.78	+3.97	-0.63	+2.95	+0.94	+5.01
cy	-1.60	+0.49	+0.57	+1.15	+1.60	+0.11	-3.24	-0.45	+0.95
mi	-3.48	-0.90	-0.12	-0.45	-1.02	+0.02	+0.29	-4.55	+0.42
ti	+0.66	+1.52	+2.01	+0.25	+2.79	+1.11	-0.16	-0.94	-0.03

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.19	+2.15	+0.89	+2.75	+3.25	-0.05	-2.62	+6.05
es	-1.89	-1.46	-0.42	-0.72	+1.79	+2.20	+0.05	-1.58	+6.84
ru	-1.01	-0.26	+0.77	-0.02	+2.05	+2.12	+1.88	-0.10	+6.63
vi	-0.99	+0.27	+0.87	-1.70	+0.50	+2.13	+2.47	-1.54	+7.30
ja	-1.19	-0.44	+2.06	+1.74	+1.54	+3.59	+2.94	-0.01	+7.45
jw	+0.24	+3.17	+1.87	+2.65	+4.01	+0.04	+3.33	-0.78	+7.00
cy	-1.50	+0.22	+0.55	+0.78	+0.11	+0.12	-4.17	-3.67	+0.43
mi	-3.56	-1.53	-0.67	-1.33	-3.13	-0.28	-1.43	-8.49	-0.23
ti	+0.62	+1.67	+2.46	+0.76	+3.15	+2.71	+0.50	-2.11	+0.03

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 13: Per-configuration Pivot NLLB gains (Δ = **Pivot NLLB** − **Zero-shot**) on Gemma 4 across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+1.27	+3.32	+5.32	+2.70	+2.74	+3.77	+3.81	+3.89
es	+1.88	+1.27	+2.95	+5.08	+2.46	+2.29	+3.93	+3.97	+3.85
ru	+1.31	+0.98	+1.80	+4.63	+2.21	+2.83	+3.56	+2.38	+3.85
vi	+3.07	+3.60	+2.91	+6.84	+2.70	+3.07	+4.38	+3.64	+3.52
ja	+3.07	+3.28	+2.01	+3.85	+1.35	+3.15	+3.40	+2.21	+2.62
jw	+2.42	+1.97	+3.81	+5.98	+3.97	+5.53	+6.06	+4.87	+4.55
cy	+2.74	+3.60	+3.15	+5.77	+3.15	+4.18	+8.44	+4.87	+5.24
mi	+1.39	+3.40	+1.97	+6.27	+3.03	+3.15	+4.71	+4.95	+5.57
ti	+1.02	+1.19	+1.92	+2.58	+3.48	+0.57	+1.56	+2.01	+5.57

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+2.51	+4.36	+6.24	+3.58	+5.18	+5.70	+5.96	+4.50
es	+2.77	+2.13	+4.11	+5.81	+3.74	+4.40	+5.92	+5.98	+4.41
ru	+1.81	+2.29	+1.87	+5.26	+2.55	+4.77	+4.74	+3.58	+3.62
vi	+4.87	+5.51	+4.61	+7.62	+4.12	+4.73	+5.18	+4.75	+2.86
ja	+3.46	+4.22	+2.89	+4.92	+1.60	+5.69	+5.57	+4.41	+3.44
jw	+3.40	+3.25	+4.79	+6.41	+5.03	+6.44	+7.11	+5.87	+4.20
cy	+3.71	+4.39	+4.18	+6.40	+4.33	+5.09	+8.65	+5.27	+4.23
mi	+1.65	+4.13	+2.98	+6.77	+3.60	+3.80	+5.33	+4.92	+4.82
ti	+2.19	+2.19	+3.01	+1.73	+4.35	-1.14	-0.47	+0.44	+4.00

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 14: Per-configuration ORBIT gains (Δ = **ORBIT** − **Zero-shot**) on LFM2-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+2.14	+4.22	+5.74	+5.04	+4.47	+2.05	+3.16	+2.05
es	-1.76	-2.91	+1.84	+5.53	+4.18	+3.08	+1.44	+2.26	+0.56
ru	-1.11	-0.61	-1.02	+5.94	+3.28	+3.69	+2.59	+3.45	+1.14
vi	-1.27	+0.49	+2.38	+3.04	+2.26	+3.32	+2.63	+3.41	+1.50
ja	-2.50	-1.72	+0.94	+4.14	-1.51	+3.98	+3.08	+1.93	+0.48
jw	-0.16	-0.82	+2.42	+3.49	+2.55	+4.31	+2.71	+3.90	+2.69
cy	-2.62	-1.39	+0.82	+1.85	+0.21	+2.75	+2.67	+2.05	+3.35
mi	-3.89	-2.83	+1.23	+2.26	-0.77	+1.24	+1.89	+0.83	+2.69
ti	-5.16	-5.08	-0.70	-0.22	-2.00	+0.01	+0.58	+1.69	+4.66

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+4.67	+6.54	+8.67	+7.69	+6.68	+1.65	+2.92	-2.94
es	-2.53	-3.54	+1.14	+7.99	+7.29	+4.70	+1.05	+1.79	-4.02
ru	-2.22	-1.10	-2.91	+8.50	+5.80	+4.66	+1.93	+2.79	-3.03
vi	-1.97	+0.26	+2.27	+4.68	+5.54	+4.98	+3.03	+3.11	-2.50
ja	-4.26	-2.80	-0.40	+6.81	+0.06	+4.85	+2.14	+1.07	-3.49
jw	-0.78	-1.37	+1.69	+5.77	+5.72	+4.79	+2.66	+3.25	-1.20
cy	-5.16	-3.12	-0.94	+3.16	+1.71	+3.20	+0.99	+1.11	-1.12
mi	-6.45	-4.75	-0.84	+3.24	+0.03	+0.28	+0.65	-1.62	-2.60
ti	-9.89	-9.57	-4.66	-1.79	-4.80	-3.65	-2.70	-1.86	+0.05

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 15: Per-configuration Self-translate CoT gains (Δ = **Self-translate CoT** − **Zero-shot**) on LFM2-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.25	+2.83	+5.82	+3.20	+3.93	+1.36	+1.40	+6.02
es	-0.25	-2.95	+2.83	+5.25	+1.36	+2.29	+0.46	+0.46	+5.07
ru	+1.56	+0.29	+0.42	+5.66	+2.38	+3.60	+2.05	+2.38	+6.38
vi	+2.13	+3.36	+5.37	+3.28	+3.73	+3.72	+2.01	+2.46	+6.02
ja	-0.33	-1.64	+1.89	+3.53	-2.94	+1.80	+0.83	+1.69	+4.54
jw	+3.24	+0.98	+4.84	+3.81	+3.69	+2.90	+1.81	+2.55	+6.18
cy	-0.04	+0.45	+1.11	+1.52	+0.17	+2.00	+1.15	+0.74	+4.83
mi	-2.66	-1.52	+0.66	+1.19	-1.59	+0.32	-0.65	-0.65	+3.56
ti	+1.55	+1.55	+5.11	+3.84	+2.58	+2.61	+1.15	+1.52	+7.41

Δ Accuracy (%)

Claim	Evidence Language								
	en	es	ru	vi	ja	jw	cy	mi	ti
en	–	+0.44	+2.64	+7.08	+3.04	+5.72	+1.15	+0.02	+5.14
es	-0.05	-2.88	+2.86	+6.15	+1.49	+3.51	-0.05	-1.56	+3.57
ru	+1.53	+0.67	-0.30	+7.02	+1.96	+5.07	+1.40	+0.64	+4.98
vi	+2.80	+4.43	+6.45	+3.86	+4.61	+5.91	+2.67	+1.99	+5.52
ja	-1.31	-2.41	+1.11	+4.18	-4.70	+2.35	-0.27	-0.76	+3.21
jw	+4.76	+2.73	+6.59	+5.82	+4.80	+3.51	+1.48	+1.15	+4.56
cy	-0.76	-0.63	+0.73	+2.10	-1.01	+2.31	-0.91	-1.86	+1.62
mi	-3.24	-2.36	+0.52	+1.60	-2.56	-0.26	-2.09	-4.24	-0.05
ti	+3.04	+3.06	+7.48	+6.48	+3.48	+5.65	+2.65	+1.70	+7.04

Δ F1-score (%)

Improvement levels: Positive (>0) Zero or Negative (≤ 0)

Figure 16: Per-configuration Pivot NLLB gains (Δ = **Pivot NLLB** − **Zero-shot**) on LFM2-VL across all 81 claim–evidence language pairs.

Claim	Evidence Language								
	en	**es**	**ru**	**vi**	**ja**	**jw**	**cy**	**mi**	**ti**
en	–	+2.14	+3.24	+3.68	+3.70	+3.51	+2.86	+3.53	+1.71
es	+0.76	+1.02	+3.16	+2.40	+2.36	+3.53	+2.82	+1.65	+1.71
ru	+0.88	+1.75	+2.77	+2.40	+3.76	+4.64	+3.73	+1.61	+2.19
vi	+1.50	+3.25	+3.84	+4.15	+4.04	+4.77	+3.30	+4.46	+1.70
ja	+0.49	+1.39	+3.45	+1.18	+4.38	+3.01	+3.23	+0.64	+1.66
jw	+2.96	+2.82	+2.65	+3.15	+3.40	+3.52	+2.45	+2.07	+0.93
cy	+3.32	+1.48	+3.14	+2.42	+3.23	+4.15	+1.67	+0.12	+1.29
mi	+0.43	-0.54	-1.01	-0.87	-1.51	-1.04	+0.13	+0.35	-0.53
ti	+1.38	-1.54	-3.51	-1.10	-2.34	-1.62	-1.58	-2.08	+1.95

Table 8: ORBIT ΔF1-score on Qwen3-VL for *Supported* class relative to zero-shot baseline.

Claim	Evidence Language								
	en	**es**	**ru**	**vi**	**ja**	**jw**	**cy**	**mi**	**ti**
en	–	+0.86	+2.53	+2.19	+4.35	+6.55	+9.04	+9.81	+9.33
es	+0.72	+1.59	+1.28	+2.73	+4.29	+7.33	+7.26	+9.01	+8.84
ru	+0.00	+1.80	+0.79	+2.49	+4.53	+9.84	+8.63	+8.46	+6.82
vi	-0.28	+1.54	+1.27	+2.28	+5.13	+8.09	+7.71	+11.30	+7.63
ja	+0.01	+1.60	+1.54	+2.96	+4.44	+8.26	+9.15	+10.90	+7.25
jw	+2.85	+4.23	+5.84	+6.08	+8.56	+6.54	+8.81	+12.79	+8.70
cy	+5.04	+5.54	+5.81	+5.80	+8.88	+11.19	+9.02	+7.86	+2.81
mi	+3.75	+3.70	+4.60	+4.06	+7.33	+8.20	+5.40	+7.99	+1.12
ti	+3.99	+3.22	+3.30	+3.90	+5.51	+4.75	+1.14	+1.75	+10.98

Table 9: ORBIT ΔF1-score on Qwen3-VL for *Refuted* class relative to zero-shot baseline.

Claim	Evidence Language								
	en	**es**	**ru**	**vi**	**ja**	**jw**	**cy**	**mi**	**ti**
en	–	-9.54	-8.02	-8.49	-11.16	-7.36	-4.80	-3.00	-0.06
es	-9.97	-10.51	-8.38	-9.42	-11.79	-5.82	-5.20	-3.80	+0.92
ru	-10.69	-10.73	-9.91	-9.62	-10.32	-8.21	-5.57	-4.01	-0.22
vi	-11.56	-9.19	-9.30	-9.61	-8.99	-6.86	-4.79	-2.77	-0.42
ja	-12.36	-9.90	-9.24	-10.28	-11.81	-8.36	-5.38	-4.94	-1.38
jw	-10.22	-6.90	-5.97	-8.05	-6.24	-8.15	-4.03	-2.26	-0.16
cy	-6.45	-4.26	-5.08	-4.65	-4.58	-2.04	-5.22	-0.53	-0.49
mi	-3.02	-2.53	-3.10	-1.76	-3.39	-2.60	-0.65	-2.26	+0.10
ti	-2.77	-1.68	+0.06	-0.02	-1.55	+0.19	+0.64	+0.47	-2.13

Table 10: ORBIT ΔF1-score on Qwen3-VL for *NEI* class relative to zero-shot baseline.