Beyond Vision: How Language affects Visual Grounding in UAV Imagery

Jue Chen[1], Penghui Huang[2], Ran Ding[2], Zhentao Zou[2], Xue Yang[2], Xue Jiang[2], Yue Zhou*[1], Hongxin Yang[1], Jonathan Li[1,3]

[1]Hinton STAI Institute and Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

[2]Shanghai Jiao Tong University, Shanghai 200241, China

[3]Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Keywords: Visual grounding, Vision language model (VLM), Multi-modal large language model (MLLM).

Abstract

Visual Grounding (VG) is a core multimodal task that localizes image targets via natural language descriptions, and it is crucial for Unmanned Aerial Vehicle (UAV) applications. However, existing remote sensing (RS) VG datasets primarily rely on rule-driven explicit descriptions, which are inconsistent with real-world demands for interpreting implicit descriptions based on context, common sense, or domain knowledge. In addition, the cross-lingual robustness of Large Vision-Language Models (LVLMs) in implicit VG remains to be thoroughly investigated. This study evaluates the cross-lingual performance of Qwen2.5-VL-7B and InternVL3.5-8B across nine languages, incorporating analyses of text length dynamics, visual attention, and language structural effects. The results demonstrate that Qwen2.5-VL-7B exhibits outstanding performance in maintaining consistent task paradigm alignment (explicit VG outperforms implicit VG) and balanced text output, which benefits from the syntactic stability and low cognitive load of East Asian languages. In contrast, InternVL3.5-8B presents task paradigm misalignment, uncontrolled text expansion, and generative hallucinations. Furthermore, differences in language structures: East Asian languages depend on word order for semantic expression, whereas Western languages feature complex lexical morphology, significantly affect attention allocation and VG accuracy. This study provides key insights for optimizing cross-lingual vision-language alignment of LVLMs and advancing practical multimodal applications in UAV scenarios.

1. Introduction

VG refers to the core task of localizing targets in images based on natural language descriptions. As a cornerstone of advanced multimodal intelligence, it profoundly embodies the fundamental cognitive skill that underpins human language-visual association. Research in developmental psychology has confirmed that infants can establish correspondences between spoken language and visual objects during the early stages of language acquisition (Hollich et al., 2007). Furthermore, cognitive science regards this capability as a key hallmark of the transition from primary to advanced human intelligence—just as relevant studies have noted, While toddlers aged 1–2 can easily comprehend explicit references based on color and relative position, they struggle to process implicit references that require common sense or domain knowledge. This distinction precisely delineates the boundary between primary and advanced intelligence (Osina et al., 2017).

In the field of RS, the value of VG capability has become increasingly prominent with the rise of LVLMs. Existing LVLMs have demonstrated excellent performance across multi-granularity VG tasks: from region-level referring detection (Sun et al., 2022; Zhan et al., 2023) to pixel-level referring segmentation (Yuan et al., 2024), all achieving effective synergy between visual and linguistic modalities. However, the textual descriptions in current mainstream RS VG datasets—such as DIOR-RSVG (Zhan et al., 2023), RSVG-HR (Lan et al., 2024), and OPT-RSVG (Li et al., 2024)—are all rule-driven explicit references, relying solely on direct visual attributes like color, relative position, and size. This creates a significant disconnect from the practical needs of UAV applications in real-world scenarios. In practical UAV scenarios including disaster response, traffic monitoring, and security surveillance, task success often depends on the understanding of implicit references. Such references require reasoning based on scene context, common sense, or domain-specific knowledge (e.g., ”a vehicle making an illegal left turn” or ”people trapped by floodwaters”). Yet, the capabilities of existing LVLMs in this domain have not been fully explored.

To address the inherent limitations of global intelligent UAVs in multilingual VG, this study evaluates the cross-lingual robustness of Qwen2.5-VL-7B and InternVL3.5 8B in implicit VG tasks. Using Acc@0.5 (accuracy at a 0.5 intersection-over-union (IoU) threshold) as the core metric, We selected these nine languages (Chinese, English, Japanese, Russian, Korean, German, French, Spanish, and Portuguese) based on two core criteria: language speaker population and geographical diversity, covering major linguistic families and geographical regions worldwide. We systematically characterize the performance of these two LVLMs across explicit and implicit reference scenarios in these languages. Furthermore, we investigate the sensitivity of LVLMs to linguistic structural variations and output text length dynamics. In summary, our key contributions are as follows:

- Performance Metrics and Text Length Dynamics: Qwen2.5-VL-7B demonstrates excellent performance in both explicit and implicit reasoning across all languages, with balanced text length and accuracy unaffected by text length. In contrast, InternVL3.5-8B exhibits anomalies: implicit reasoning outperforms explicit reasoning (albeit with generally low overall accuracy), and it generates lengthy and disorganized text, exposing issues

Figure 1. Multi-scene drone image dataset: Samples from 6 scene categories

of generative hallucinations and flaws in the output mechanism.

- Visual Attention in Disaster Scenario (Chinese vs. English): Although Qwen2.5-VL-7B achieves comparable accuracy in Chinese and English, it performs better in Chinese: its attention can precisely focus on the target, maintaining consistency between text and visual elements. In English, however, attention becomes dispersed, and reasoning deviates from the task, leading to a mismatch between localization and the target—this confirms stronger robustness and ”text-visual” alignment in Chinese.

- Visual Attention Across East Asian vs. Western Languages: Qwen2.5-VL-7B delivers outstanding performance in East Asian languages, benefiting from its highly concentrated attention (locking onto the target with minimal interference from irrelevant elements). For Western languages, the complexity of lexical morphology scatters attention to secondary details, thereby reducing focus on the core target—this explains the gaps in its cross-linguistic performance.

2. Dataset

Our dataset consists of UAV-acquired imagery covering six scenarios: Traffic, Disaster, Security, Sport, Social Activity, and Productive Activity, each of which includes multiple subcategories (e.g., 8 distinct sport activities). As presented in Figure 1, we showcase representative annotated samples for each scenario, with target regions delineated by red bounding boxes. Specifically, for each scenario, we provide both implicit and explicit queries, which are translated into nine languages (English, Chinese, Japanese, Russian, Spanish, Korean, German, French, and Portuguese). This multilingual annotation framework ensures the dataset’s adaptability to diverse global application contexts.

Notably, scenario-specific commonsense variations render template-based generation or LVLM-assisted annotation (even with human validation) infeasible; thus, a fully manual annotation workflow was adopted to guarantee the reliability of the benchmark. A total of 873 test samples were annotated, each incorporating explicit and implicit questions, and their real-world relevance was validated by domain experts. These test samples are stratified across scenarios according to the practical deployment frequency of UAVs: 42 Security samples, 288 Traffic samples, 91 Social Activity samples, 53 Disaster samples, 242 Productive Activity samples, and 157 Sport samples. To support model training and validation, a corresponding training set (1,892 samples in total) was constructed, with 60, 1,048, 99, 38, 496, and 249 samples allocated to the Security, Traffic, Social Activity, Disaster, Productive Activity, and Sport scenarios, respectively. In total, the benchmark dataset comprises 2,765 samples (873 test samples + 1,892 training samples) across all six scenarios, ensuring comprehensive coverage of scenario-specific visual and linguistic characteristics.

3. Experiments

3.1 Multilingual VG capability

To evaluate the multilingual VG capability of LVLMs, we analyzed their accuracy (Acc@0.5%) across explicit and implicit tasks in nine languages, with Qwen2.5-VL-7B and InternVL3.5-8B results shown in Table 1.

Qwen2.5-VL-7B exhibits performance characteristics consistent with the inherent properties of VG tasks across all tested languages: specifically, its accuracy in explicit VG tasks is consistently higher than that in implicit VG tasks, with corresponding Reduction Ratios all showing negative values (marked by downward arrows). This result is highly aligned with the fundamental cognitive principle that ”feature-driven tasks are more tractable in terms of computation and cognition”. Meanwhile, the magnitude of performance attenuation from explicit to implicit tasks of this model demonstrates significant cross-linguistic heterogeneity: the attenuation magnitude is the largest in Portuguese (48.6%, labeled ”Max” with four downward arrows), reflecting a prominent performance advantage of explicit tasks over implicit tasks; in contrast, the attenuation magnitude is the smallest in Korean (14.4%), which narrows the performance gap between the two task types.

Language	Qwen2.5-VL-7B			InternVL3.5-8B		
	Explicit Acc	Implicit Acc	Reduction Ratio	Explicit Acc	Implicit Acc	Reduction Ratio
English	59.60%	47.39%	-20.5% ↓↓	5.82%	4.97%	-14.6% ↓
Chinese	62.17%	47.46%	-23.7% ↓↓	12.45%	11.36%	-8.8% ↓
Japanese	55.43%	35.33%	-36.3% ↓↓↓	6.92%	6.26%	-9.5% ↓
Korean	53.90%	46.16%	-14.4% ↓	10.03%	7.36%	-26.6% ↓↓
German	46.34%	36.16%	-22.0% ↓↓	28.75%	27.57%	-4.1% ↓
French	55.81%	36.68%	-34.3% ↓↓↓	31.31%	32.38%	+3.4% ↑
Spanish	53.01%	34.13%	-35.6% ↓↓↓	15.39%	19.52%	+26.8% ↑↑
Portuguese	51.94%	26.68%	-48.6% ↓↓↓↓ (Max)	20.65%	27.54%	+33.4% ↑↑ (Max)
Russian	52.34%	41.72%	-20.3% ↓↓	40.09%	48.77%	+21.7% ↑↑

Table 1. Cross-lingual Performance Comparison: Accuracy and Reduction Ratios for Explicit vs. Implicit VG

By contrast, InternVL3.5-8B presents obvious task paradigm misalignment: although it still maintains the trend of ”explicit accuracy being higher than implicit accuracy” in some languages (e.g., Chinese, Japanese) with corresponding negative Reduction Ratios, implicit accuracy surpasses explicit accuracy in French, Spanish, Portuguese, and Russian (with positive Reduction Ratios, marked by upward arrows). This anomalous pattern indicates that the model has fundamental deviations in its understanding of the core VG task paradigms. Furthermore, the cross-linguistic performance fluctuation of this model is more drastic: the performance improvement of implicit tasks relative to explicit tasks is the largest in Portuguese (+33.4%, labeled ”Max” with two upward arrows), while the attenuation magnitude from explicit to implicit tasks is the smallest in German (4.1%), where the accuracies of the two tasks are nearly equal.

From the perspective of absolute accuracy, Qwen2.5-VL-7B significantly outperforms InternVL3.5-8B across all languages and task paradigms: for instance, the Chinese explicit accuracy of Qwen2.5-VL-7B (62.17%) is approximately five times that of InternVL3.5-8B (12.45%); even in Korean, where InternVL3.5-8B performs relatively best, its explicit accuracy (28.75%) is still much lower than the corresponding value of Qwen2.5-VL-7B (53.90%), and the performance gap between the two models is consistent and significant.

In summary, Table 1 clearly reveals the core differences in cross-linguistic vision-language alignment (VLA) capabilities between the two models: Qwen2.5-VL-7B maintains stable task paradigm alignment and excellent cross-linguistic performance, whereas InternVL3.5-8B suffers from defects such as task paradigm misalignment, extremely strong cross-linguistic performance fluctuation, and low overall accuracy. These quantitative findings provide critical empirical references for subsequent research on optimizing the cross-linguistic VLA mechanism of LVLMs.

3.2 Cross-Lingual Analysis of Text Length

After Table 1 clarified the performance differences and paradigm characteristics of the two models in cross-lingual vision-language tasks, Table 2 provides an interpretable output mechanism basis for the aforementioned performance from the perspective of generated text length. It records the average, maximum, and minimum token counts of Qwen2.5-VL-7B and InternVL3.5-8B in implicit and explicit tasks across 9 languages, whose data features can form a complementary association with the performance results in Table 1.

For Qwen2.5-VL-7B, the text lengths in implicit and explicit tasks under the same language are highly consistent: for instance, the average length of implicit tasks in Chinese is 121.99 tokens, which is comparable to the 109.72 tokens of explicit tasks, and no significant discrepancy is observed in Japanese, Korean, or other languages. This length stability echoes the ”task paradigm alignment” performance of the model in Table 1, reflecting the unified output logic of the model for the two task types without deviations in output style due to task differences. In contrast, the text length of InternVL3.5-8B exhibits obvious fluctuations and imbalances: in most languages, the length of its generated text is significantly longer than that of Qwen2.5-VL-7B. For example, the average length of implicit tasks in Russian (2247.64 tokens) is more than 5 times that of Qwen2.5-VL-7B, and the maximum length of implicit tasks in Korean (9915 tokens) is nearly 10 times that of Qwen2.5-VL-7B, showing ”uncontrolled expansion”. Meanwhile, its minimum text length is generally around 100 tokens, which contrasts with the extremely short texts of Qwen2.5-VL-7B (e.g., 5 tokens for both implicit and explicit tasks in Chinese), further highlighting the differences in output strategies between the two models. More importantly, the length fluctuations of InternVL3.5-8B are directly related to its task paradigm deviation: taking Portuguese as an example, the average length of its implicit tasks (2783.36 tokens) is much longer than that of explicit tasks (20.65 tokens), while in Table 1, the implicit accuracy of this language exceeds the explicit accuracy (+33.4%). Excessively long implicit texts may introduce redundant information, interfering with the focus on the core task objectives, which explains the cause of paradigm deviation from the perspective of output characteristics.

The aforementioned length features are deeply associated with the performance differences presented in Table 1: the length stability of Qwen2.5-VL-7B essentially reflects its accurate alignment with the vision-language task paradigm (the target boundary of implicit/explicit tasks). It not only corresponds to the stable cross-lingual accuracy of the model but also explains the cognitive regularity that ”explicit task accuracy is always higher than implicit task accuracy”. Additionally, its length distribution is balanced, with no significant negative correlation between text length and accuracy, thus avoiding performance deviations caused by uncontrolled length. In contrast,

Figure 2. Cross-Language Comparison of VG Performance for Qwen2.5-VL-7B

the length imbalance of InternVL3.5-8B is not only an external manifestation of paradigm deviation but also reveals the risk of generative hallucination in its cross-lingual processing: for languages where its accuracy is lower than 5% in Table 1 (e.g., Russian, French), the model exhibits abnormally lengthy generation in Table 2 (2247.64 tokens for implicit tasks in Russian and 3098.85 tokens for implicit tasks in French). The co-occurrence of ”high linguistic perplexity and low accuracy” indicates a severe hallucination issue—when the model fails to parse task instructions in linguistically unfamiliar contexts, it will generate semantically vacuous and redundant texts to cover the breakdown of the reasoning chain, forming an irrational strategy of ”replacing alignment with length”. This also corroborates the anomaly of ”implicit accuracy exceeding explicit accuracy in some languages” in Table 1, further demonstrating that InternVL3.5-8B not only has cognitive biases in cross-lingual VLA but also suffers from fundamental bottlenecks in the generative mechanism, failing to maintain rigorous reasoning logic in linguistically unfamiliar contexts.

3.3 Cross-Lingual Analysis of Visual Attention

Visual attention heatmaps enable interpretability analysis of attention distribution in multimodal models, and their generation principle and process are as follows: First, high-dimensional attention information is extracted from all attention layers across each generation stage of the model. Subsequently, dimensionality reduction is performed on this high-dimensional information to derive the average attention weight for each token. Based on predefined image tokens, query text tokens, and generated text token intervals, the corresponding attention weights of each token type are aggregated, and eventually, a feature matrix is constructed to characterize the attention distribution features of these three token types.

To intuitively visualize the spatial patterns of attention, the feature matrix is further converted into a heatmap and fused with the original image, yielding an image annotated with attention distribution features. Specifically, OpenCV is employed to read the original image, which is then converted to the RGB color space. The input feature matrix, compatible with both PyTorch tensors and NumPy arrays, is processed into a two-dimensional structure and resized to a uniform 23×23 dimension. To eliminate scale differences between feature values, normalization is applied to the feature matrix. Thereafter, the preprocessed feature matrix is scaled to match the size of the original image using a predefined method, and a specified colormap (e.g., jet colormap) is applied to map the matrix to a color heatmap. Using a transparency adjustment parameter, the heatmap is weighted and fused with the original image—this fusion ensures that the scene details of the original image are retained while clearly highlighting regions of interest. Finally, the fused image is converted to the BGR format and saved to a specified path, with both the fused image and the heatmap data output simultaneously.

The entire process incorporates an exception handling mechanism designed to address potential errors that may occur during image reading and processing. This mechanism plays a critical role in ensuring the stability and reliability of the proposed method, thereby supporting its applicability in subsequent experimental validations and practical scenarios.

To further unpack the underlying reasoning mechanisms behind the cross-lingual performance heterogeneity of Qwen2.5-VL-7B in implicit VG tasks, this study generated visual attention heatmaps to systematically compare the model’s reasoning logic, attention distribution when processing implicit queries in Chinese versus English.

Taking the earthquake disaster scenario test case illustrated in Figure 2 as an example: this scenario contains two types of structures (a traditional-style building and an under-construction modern building), with the ground-truth region for VG corresponding to “the under-construction modern building most severely damaged by the earthquake” (coordinate interval [376.6, 145.4, 643.1, 614.2]). When the Chinese implicit query was input, the model exhibited a coherent reasoning chain well-adapted to the disaster scenario: it first parsed the building types and states in the image (the traditional building had been repaired, while the modern building was under construction), then matched the semantic description of “most severely damaged”, and finally locked onto the under-construction modern building. The corresponding attention heatmap showed distinct

Language	Qwen2.5-VL-7B						InternVL3.5-8B					
	Implicit (I)			Explicit (E)			Implicit (I)			Explicit (E)		
	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.
Chinese	121.99	3051	**5**	109.72	271	**5**	602.56	1840	116	697.93	1697	113
English	475.01	2302	15	451.47	2420	125	893.25	4865	244	1092.44	4813	207
Japanese	147.17	384	46	137.49	907	60	340.45	2163	78	375.37	3539	67
Russian	439.26	1391	112	385.98	913	67	**2247.64**	4885	407	**2311.87**	5123	495
Korean	199.57	1004	36	176.56	737	81	322.34	**9915**	90	311.66	**9677**	98
German	552.52	1985	177	455.88	1934	164	1183.79	5095	252	1542.53	4916	246
French	368.98	1207	95	352.43	1187	**24**	**3098.85**	5141	559	**3173.00**	4942	669
Spanish	407.75	1049	136	344.33	1554	118	2185.72	4968	380	2398.34	4842	417
Portuguese	390.42	1039	129	277.66	634	100	**2783.36**	5044	435	**2809.01**	4817	424

Table 2. Cross-language text length index analysis

Note: Metrics represent token counts of generated text. Gray-highlighted values denote core indicators: extremely short minimum lengths (Qwen2.5-VL-7B) and extremely fluctuating lengths (InternVL3.5-8B). Avg. = Average Length; Max. = Maximum Length; Min. = Minimum Length.

concentration characteristics, with the highlighted area highly overlapping the ground-truth region—achieving precise alignment between the semantic query and the visual region. This performance aligns mechanistically with the relatively high implicit task accuracy (47.46%) of Qwen2.5-VL-7B for Chinese in Table 1.

In contrast, when processing the English implicit query : although the model also executed a stepwise reasoning process (identifying building structures → assessing the state of the traditional building → assessing the state of the modern building), a critical deviation emerged in the reasoning chain: it erroneously concluded that “neither building showed significant severe damage”, leading to a final localized region that completely deviated from the ground truth. The corresponding attention heatmap exhibited a markedly dispersed distribution, with the highlighted area failing to concentrate on the target building, and the alignment between visual attention and the semantic query was substantially degraded. This phenomenon explains why, while the implicit task accuracy for English (47.39%) in Table 1 is close to that for Chinese, there is in fact a latent gap in the robustness of the reasoning process and the reliability of the results.

The above comparison indicates that Qwen2.5-VL-7B exhibits stronger adaptability to the semantic and syntactic cues of Chinese in its implicit reasoning capability: this adaptability is not only reflected in accuracy values (the difference in implicit accuracy between Chinese and English in Table 1 is small) but also in the coherence of the reasoning logic and the concentration of visual attention. This characteristic may stem from the deep alignment with Chinese semantic expressions and syntactic structures in domain-specific scenarios (e.g., disaster contexts) during the model’s training phase; it also provides a vision-language reasoning mechanism-level explanation for the overall superior performance of this model in East Asian languages, as observed in Table 1.

3.4 Linguistic Structural Analysis of Cross-Linguistic Attention

Taking ”earthquake-damaged pavements (cracked areas)” as the target of implicit VG tasks, Figure 3 systematically illustrates the cross-lingual performance heterogeneity of Qwen2.5-VL-7B between the East Asian language group (Chinese, Japanese, Korean) and the Western language group (English, Portuguese, German) through visual attention heatmap analysis. This visualization result forms complementary validation with the quantitative data in Table 1—where East Asian languages generally exhibit superior performance in both explicit and implicit tasks—with the core mechanism attributable to differences in the dimensionality of structural complexity between the two language groups.

In the experiment, the ground-truth region of the task corresponds to the core of the pavement cracks in the figure. From the perspective of attention distribution characteristics: queries in the East Asian language group correspond to a focused attention pattern—the highlighted weight in the model’s heatmap is highly concentrated on the ground-truth region of cracks, with minimal attention interference from irrelevant visual elements (e.g., surrounding undamaged pavements, background structures). In contrast, queries in the Western language group exhibit a dispersed attention pattern: attention weights not only cover the cracked areas but also spread significantly to secondary visual details around the target, leading to a substantial reduction in the concentration of attention on the core target.

The essence of this difference stems from variations in the dimensionality of grammatical complexity between the two language groups:The grammatical complexity of East Asian languages is primarily reflected in the semantic-word order interface. Taking Chinese as an example, the semantic contrast between ”Cats chase mice” and ”Mice chase cats” is determined entirely by word order, while the forms of words such as ”cat”, ”mouse”, and ”chase” remain unchanged regardless of their syntactic roles (subject or object)—exhibiting high stability in lexical morphology. In this context, when processing queries in East Asian languages, the model does not require additional parsing of lexical morphological variants; instead, it can directly establish a mapping relationship between the core semantics of ”earthquake-damaged pavements” and the visual ground-truth region, thereby forming a focused attention distribution.

In contrast, the grammatical complexity of Western languages is concentrated at the lexical morphology level. For instance, in inflectional languages, nouns require case marking changes according to their roles (subject or object), and verbs undergo conjugation based on tense and person. Taking German as an example, the form of pavement-related nouns changes with their syntactic roles in sentences. When processing such languages, the model must allocate computational resources to

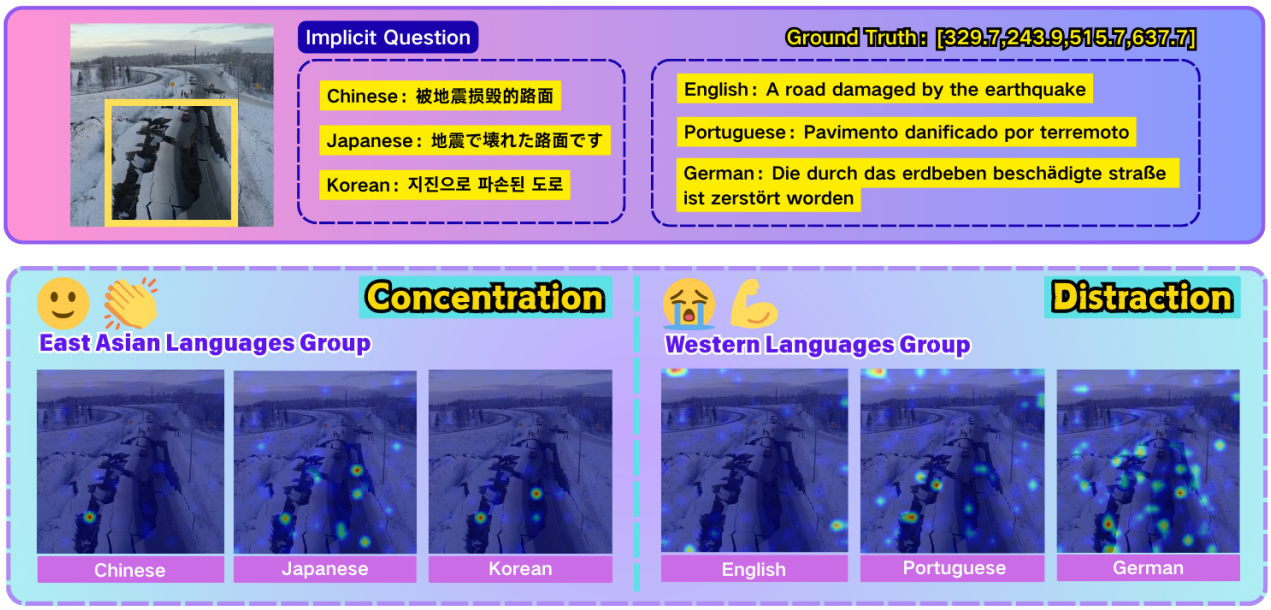

Figure 3. East Asian vs. Western Language Performance on Implicit VG with Qwen2.5-VL-7B

parse the correspondence between lexical morphological variants and syntactic functions. This process often relies on auxiliary matching with secondary visual details (e.g., the shape and spatial position of cracks), which in turn causes attention resources to spread to non-core visual elements and weakens the focus on the target region.

These mechanistic observations provide an attention-level explanation for the superior performance of East Asian languages in explicit and implicit VG tasks (as shown in Table 1): the structural characteristics of East Asian languages reduce the model’s cognitive load and enhance the accuracy of semantic-visual alignment. In contrast, the morphological complexity of Western languages increases the redundancy of information processing, which in turn affects task performance and the target focus of attention.

4. Conclusion

This study focuses on the core requirements of cross-lingual VLA for LVLMs, with key contributions as follows: An explicit/implicit VG benchmark dataset is constructed (873 UAV scenario samples, 9 languages), addressing gaps in task paradigm consistency and output mechanism interpretability. Cross-lingual VLA heterogeneity among LVLMs is revealed: Qwen2.5-VL-7B shows stable paradigms, balanced text length, and concentrated attention for East Asian languages, while InternVL3.5-8B has critical flaws (paradigm misalignment, uncontrolled text expansion, generative hallucinations). East Asian languages outperform morphologically complex Western languages in both VG tasks by leveraging syntactic stability to reduce cognitive load. Targeted optimizations are proposed: mitigating hallucinations via length-accuracy constraints, language-specific preprocessing for Western languages, and integrating explicit/implicit task pairs and UAV data into benchmarks. These contributions support LVLMs’ cross-lingual VLA optimization and UAV multimodal application advancement.

References

Hollich, G., Golinkof, R., Hirsh-Pasek, K., 2007. Young children associate novel words with complex objects rather than salient parts. *Developmental Psychology.*, 43, 1051.

Lan, M., Rong, F., Jiao, H.and Gao, Z., Zhang, L., 2024. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing.*, 62, 1–13.

Li, K., Wang, D., Xu, H.and Zhong, H., Wang, C., 2024. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing.*, 62, 1–13.

Osina, M., Saylor, M., Ganea, P., 2017. Out of reach, out of mind? infants’ comprehension of references to hidden inaccessible objects. *Child development.*, 88, 1572–1580.

Sun, Y., Feng, S., Li, X., Ye, Y., Kang, J., Huang, X., 2022. *Visual grounding in remote sensing images*. 404–412.

Yuan, Z., Mou, L., Hua, Y.and Zhu, X., 2024. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing.*, 62.

Zhan, Y., Xiong, Z., Yuan, Y., 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing.*, 61, 1–13.