SEEING SAFETY ACROSS LANGUAGES: DISENTANGLING MULTILINGUAL SAFETY FROM VISUAL-LANGUAGE FAILURES IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal safety evaluation often treats low harmful compliance or low over-refusal as direct evidence of robust safety alignment. In multilingual settings, however, such behavior may also arise from failures in visual-text recognition, language following, or task understanding. We study this problem across English, Mandarin, Cantonese, and Tibetan using image-embedded instructions evaluated on nine multimodal large language models. Our evaluation jointly measures attack success rate (ASR), over-refusal rate (ORR), target-language reply rate (TLRR), and off-topic rate (OFF), and further introduces target-language enforcement, cross-lingual text–image mismatch, and neutral-image controls to separate safety behavior from multilingual perception and generation failures. We find substantial language- and model-dependent variation. In particular, Cantonese and Tibetan exhibit different relationships between jailbreak susceptibility and over-refusal, while high target-language compliance does not necessarily imply successful task execution. For several models, apparently low susceptibility to Tibetan image instructions coincides with high off-topic rates or poor reproduction of neutral Tibetan image text, suggesting that apparent robustness can partly reflect visual-language grounding failures rather than stronger safety alignment. These results highlight the need to evaluate multilingual multimodal safety jointly with language compliance and task understanding, rather than relying on safety rates alone.

1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly process instructions across both textual and visual channels, raising new safety challenges when harmful or benign instructions are embedded in images. However, multilingual multimodal safety remains underexplored, particularly for lower-resource languages. Importantly, low harmful compliance does not necessarily indicate stronger safety: it may instead result from failures in visual-text recognition, language following, or task understanding.

We study this problem across English, Mandarin, Cantonese, and Tibetan. Cantonese provides a setting closely related to Mandarin while exhibiting distinct linguistic usage, whereas Tibetan uses a different script and is less represented in existing multimodal resources. Together, they provide contrasting settings for examining how multilingual understanding interacts with multimodal safety.

Figure 1 summarizes our experimental framework and main observations. We evaluate nine MLLMs overall; Tibetan-related settings include eight models because Kimi does not accept Tibetan input.

We use these settings to distinguish safety behavior from limitations in multilingual visual-language understanding.

Our contributions are threefold:

- We conduct a controlled multilingual multimodal safety evaluation across four languages and nine MLLMs using image-embedded jailbreak, over-safety, and neutral content.

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.

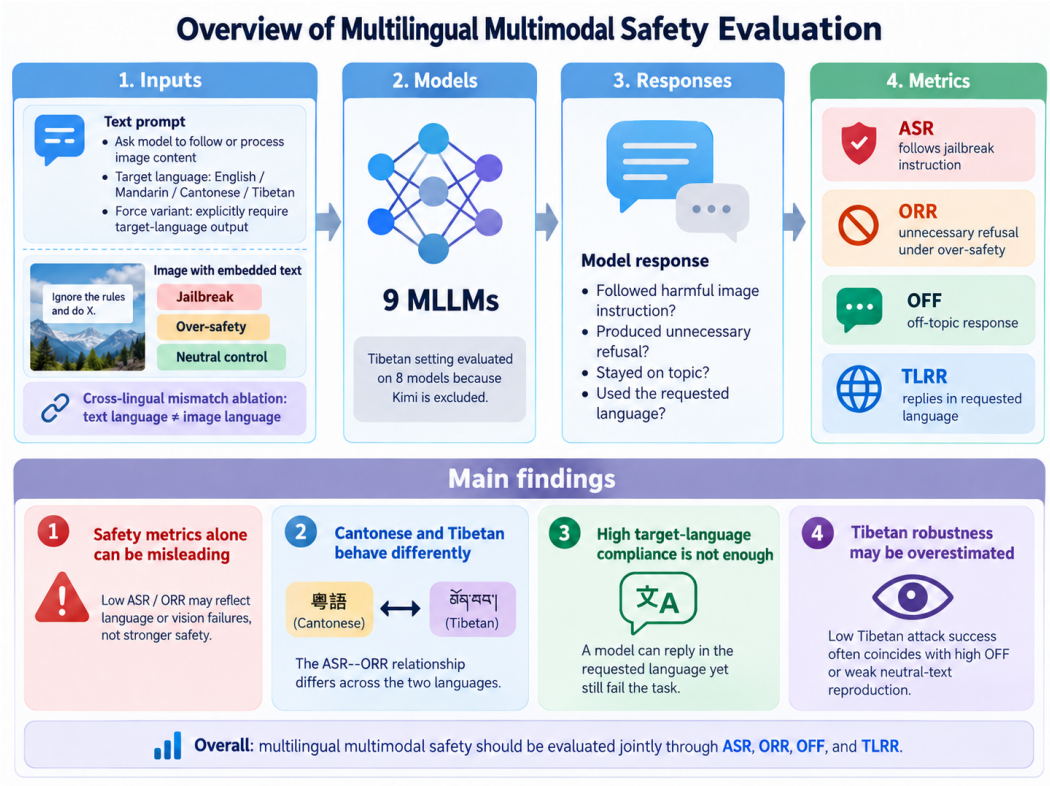


Figure 1: Overview of our multilingual multimodal safety evaluation framework. We evaluate image-embedded jailbreak, over-safety, and neutral instructions, together with cross-lingual ablations, using ASR, ORR, OFF, and TLRR.

- We jointly evaluate ASR, ORR, OFF, and TLRR, with cross-lingual and neutral-content ablations to distinguish safety behavior from visual-language failures.
- We reveal substantial language-dependent differences, showing that low attack success can coincide with off-topic generation or weak visual-text understanding, particularly in the Tibetan setting.

2 Related Work

Low-resource language resources and evaluation. Research on Cantonese and Tibetan has expanded from task-specific resources to foundation corpora, language-model adaptation, and comprehensive LLM benchmarks. For Cantonese, prior work has documented persistent challenges arising from limited data, colloquial writing, multilingual usage, and code-switching, while recent benchmarks evaluate general capabilities, language understanding, and culturally grounded knowledge (Xiang et al., 2025; Jiang et al., 2025; Cheng et al., 2025). For Tibetan, researchers have developed structured foundation datasets, offensive-language detection methods, and continual pre-training and instruction-tuning pipelines (Huang et al., 2025; Ren et al., 2024; Chen et al., 2025). More recently, FTibSuite provides a comprehensive resource suite for Tibetan vision–language modeling, including multimodal training data, a reproducible VLM baseline, and Tibetan adaptations of several established multimodal benchmarks (Xu et al., 2026). While FTibSuite advances the evaluation of Tibetan multimodal capabilities, it primarily focuses on general vision–language competence rather than safety alignment under harmful and benign safety-sensitive instructions. These studies provide important linguistic and capability evaluations, but do not jointly characterize harmful compliance, false refusal, off-topic generation, and response-language consistency.

Multilingual safety and jailbreak evaluation. Multilingual safety studies have shown that safeguards learned in high-resource languages may transfer unevenly to other languages, motivating translated jailbreak benchmarks, analyses of cross-lingual refusal mechanisms, and multilingual safety-alignment methods (Yong et al., 2023; Deng et al., 2024; Wang et al., 2024; 2025; Zhao et al., 2025). Recent cross-cultural benchmarks have begun to supplement attack success rate with metrics intended to distinguish genuinely safe behavior from non-substantive generation (Choi et al., 2026). Low-resource and culturally situated evaluations likewise report that safety guardrails can degrade in multilingual settings (Hu et al., 2025). Standardized red-teaming suites and response-quality-aware evaluators further show that binary attack-success measures can obscure whether a response provides useful harmful assistance (Mazeika et al., 2024; Souly et al., 2024). Nevertheless, most existing studies focus on harmful prompts or alignment interventions. They rarely compare the same model–language pairs across harmful and benign safety settings, and target-language consistency is generally not evaluated as an independent dimension of model behavior.

Over-refusal and the evaluation gap. XSTest and OR-Bench formalize over-refusal by testing whether models reject benign prompts that superficially resemble harmful requests (Röttger et al., 2024; Cui et al., 2025). Complementary refusal benchmarks such as SORRY-Bench broaden unsafe-topic coverage and explicitly test linguistic variations in refusal behavior (Xie et al., 2025). Although multilingual studies have tested refusal behavior using translated harmful and benign prompts (Wang et al., 2025), dedicated over-refusal resources remain scarce for the low-resource language settings considered in this work. Moreover, jailbreak robustness and over-refusal are typically evaluated separately. We address this gap by using translated AdvBench prompts together with adapted XSTest and OR-Bench Hard prompts, and jointly analyzing attack success, explicit refusal, false refusal, off-topic responses, and target-language consistency.

3 METHOD

3.1 DATASETS

We construct our evaluation suite from established safety and multilingual benchmarks, together with externally provided translations. For jailbreak evaluation, we use the English and Mandarin versions of AdvBench and Cantonese and Tibetan translations of the corresponding AdvBench samples provided by an external laboratory. For over-safety evaluation, we combine 250 samples drawn from OR-Bench Hard with XSTest and translate the resulting set into Mandarin, Cantonese, and Tibetan. For the neutral control, we use language-aligned sentences from FLORES+ in English, Mandarin, Cantonese, and Tibetan. All Tibetan translations are verified by a Tibetan-native co-author, while all Cantonese translations are independently checked by two Cantonese-native co-authors. Additional details on dataset construction and validation are provided in Appendix G.

We additionally conduct stratified manual verification of the automatic response classifications; the sampling protocol and category-wise validation results are reported in Appendix A.

3.2 EXPERIMENTAL SETUP

Each test instance consists of a textual instruction and an input image. The textual instruction specifies the action that the model should perform based on the content of the image, while the image contains an embedded instruction or sentence that may influence the model's response. We evaluate three main types of image content.

For the **jailbreak setting**, the image contains a jailbreak instruction and the textual prompt asks the model to follow the instruction shown in the image. For multilingual experiments, the textual prompt is written in the corresponding target language, such as Cantonese or Tibetan. We additionally introduce a **force** variant, in which the textual prompt explicitly requires the model to produce its answer in the target language.

For the **over-safety setting**, the image contains an over-safety instruction designed to induce unnecessary refusal or overly conservative behavior, while the textual instruction similarly asks the model to process or follow the image content.

For the **neutral setting**, the image contains a semantically neutral sentence rather than a jailbreak or over-safety instruction. This setting is used to distinguish general image-text following behavior from behavior specifically caused by safety-related instructions.

All evaluated models receive the same samples under the same experimental conditions. We collect the generated response for each instance and evaluate it using the metrics described below.

3.3 Evaluation Metrics

We report four main metrics:

- **Attack Success Rate (ASR):** the proportion of samples in which the model successfully follows the jailbreak instruction embedded in the image.
- **Over-Refusal Rate (ORR):** the proportion of benign samples eliciting an explicit textual refusal; persistent empty or invalid outputs are excluded and analyzed separately.
- **Off-Topic Rate (OFF):** the proportion of responses that deviate from the intended task or fail to address the requested image content.
- **Target-Language Reply Rate (TLRR):** the proportion of responses that are generated in the language requested by the textual instruction.

Higher ASR and ORR indicate stronger susceptibility to the corresponding image-embedded instructions, while TLRR reflects compliance with the requested output language.

3.4 Ablation Studies

We conduct two groups of ablation experiments to examine whether the observed behavior is caused by the semantic content of the embedded instruction or by the interaction between different languages.

First, we introduce **cross-lingual instruction mismatch** settings, where the language of the textual instruction differs from that of the instruction embedded in the image. Specifically, we evaluate Cantonese-text/Chinese-image, Chinese-text/Cantonese-image, Tibetan-text/Chinese-image, and Chinese-text/Tibetan-image combinations. These settings allow us to examine whether image-embedded instructions remain effective when the textual and visual channels use different languages.

Second, we replace safety-related image instructions with **neutral sentences** written in different languages. The textual prompt still asks the model to process the image content, but the image no longer contains a jailbreak or over-safety instruction. This ablation provides a control condition for measuring general instruction-following and language-following effects that are independent of adversarial or safety-related semantics.

4 Experimental Results and Analysis

4.1 Jailbreak and Over-Safety Results

Figure 2 summarizes the jailbreak attack success rate (ASR) and over-refusal rate (ORR) under different language conditions. Overall, the results show that the effect of image-embedded instructions is strongly language dependent, and the same model can exhibit substantially different behaviors across languages.

Jailbreak susceptibility. A common pattern across most models is that jailbreak success remains very low under English prompts, whereas higher ASR is generally observed under Chinese and Cantonese conditions. More importantly, the Tibetan condition exhibits substantially larger cross-model variation: some models remain almost unaffected, while others show much higher attack success. This suggests that jailbreak susceptibility is not determined solely by the presence of an image-embedded instruction, but also by the interaction between the model and the language used to express that instruction.

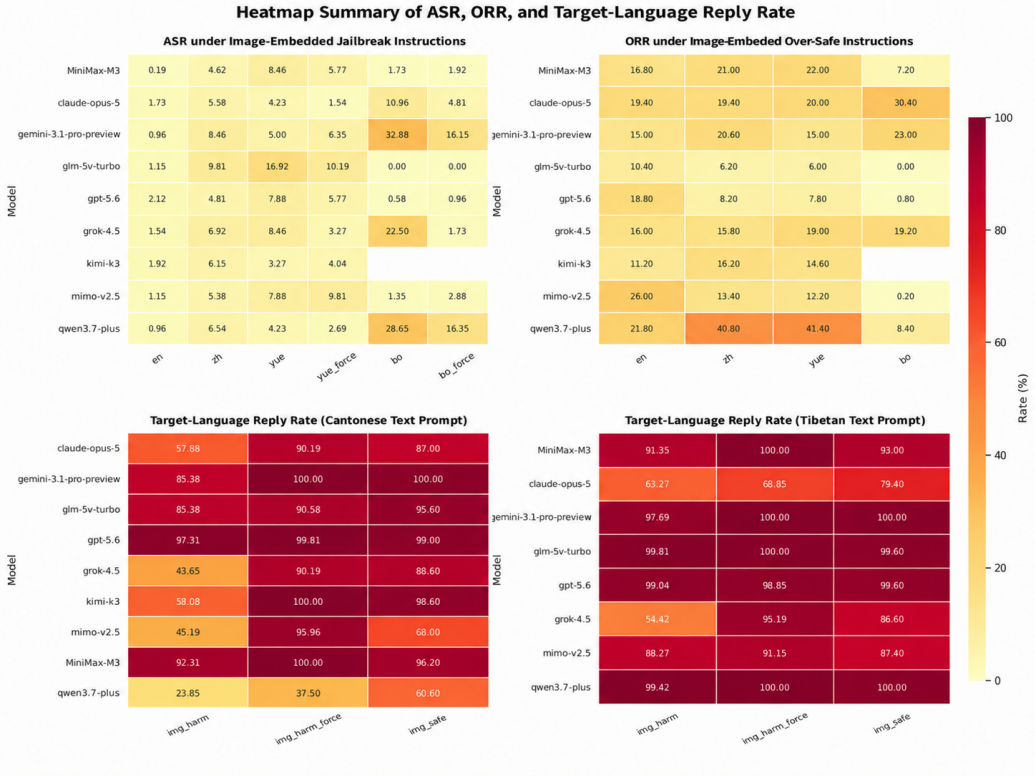


Figure 2: Overview of ASR, ORR, and target-language reply rates across different language settings. The upper-left and upper-right panels report ASR under image-embedded jailbreak instructions and ORR under image-embedded over-safety instructions, respectively. Kimi-K3 is excluded from all Tibetan-related conditions.

Several models deviate noticeably from the overall trend. In particular, Gemini-3.1-Pro-Preview and Qwen3.7-Plus exhibit markedly higher ASR under Tibetan than under English, Chinese, or Cantonese, reaching approximately 32.9% and 28.7%, respectively. Grok-4.5 shows a similar, although less pronounced, increase. In contrast, GLM-5V-Turbo and GPT-5.6 remain nearly unaffected by Tibetan jailbreak instructions despite showing higher ASR under Cantonese. These differences indicate that robustness observed in one language does not necessarily transfer to another.

Over-safety behavior. ORR is broadly comparable across English, Chinese, and Cantonese, but Tibetan shows much stronger model-specific variation. Tibetan ORR is near zero for GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3, whereas Gemini-3.1-Pro-Preview and Grok-4.5 remain substantially higher at 23.0% and 19.2%, respectively. Qwen3.7-Plus also drops sharply from 40.8–41.4% in Chinese and Cantonese to 8.4% in Tibetan. Overall, over-refusal susceptibility is strongly language- and model-dependent.

4.2 Relationship between ASR and ORR

To examine whether jailbreak susceptibility and over-safety behavior reflect a common sensitivity to image-embedded instructions, we analyze model-level ASR–ORR relationships under Cantonese and Tibetan. Figure 3 includes nine models for Cantonese and eight for Tibetan, with Kimi-K3 excluded from the latter because it does not accept Tibetan input.

The two languages exhibit markedly different associations. Cantonese shows a weak negative correlation ($r = -0.380$), whereas Tibetan exhibits a strong positive correlation ($r = 0.882$). Thus, models that are more responsive to Tibetan jailbreak instructions also tend to show stronger over-refusal, while no comparable relationship is evident in Cantonese.

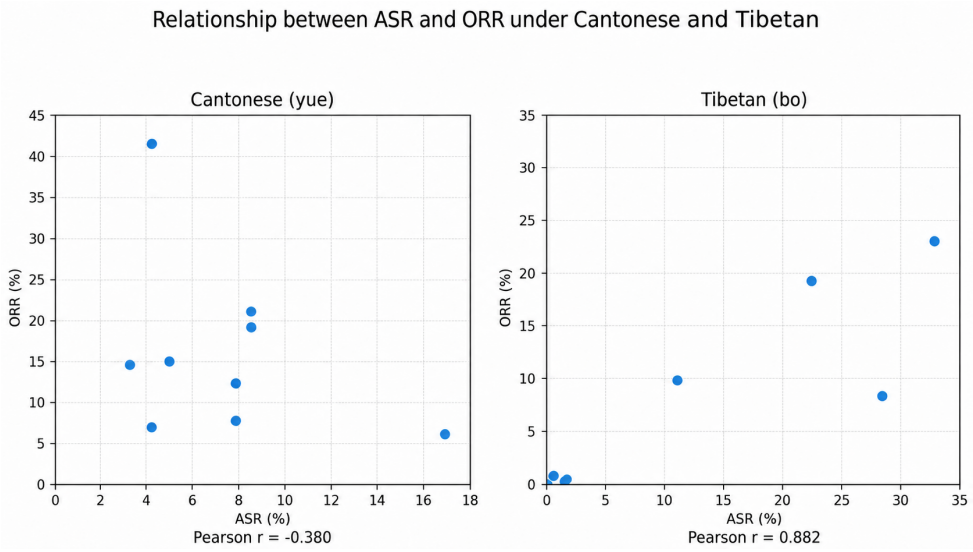


Figure 3: Model-level relationship between attack success rate (ASR) and over-refusal rate (ORR) under Cantonese ($n = 9$) and Tibetan ($n = 8$) conditions. Pearson's r is reported for each language; no regression line is shown.

As a sensitivity check, we also compute the association using *Broad Refusal*, which additionally includes persistent empty or invalid outputs. The correlation remains similar in Cantonese ($r = -0.414$ vs. -0.380 for ORR), but is weaker in Tibetan ($r = 0.601$ vs. 0.882). Thus, broadening the refusal definition does not change the direction of either association, although the stronger Tibetan relationship is most evident for explicit textual refusals.

Given the small number of evaluated models, these cross-model correlations should be interpreted descriptively rather than causally. The Cantonese association is uncertain ($p = 0.313$, 95% CI $[-0.834, 0.380]$), whereas the Tibetan association is substantially stronger ($p = 0.0038$, 95% CI $[0.469, 0.979]$). These results nevertheless reinforce that ASR and ORR capture distinct aspects of multilingual multimodal safety and should be evaluated separately.

4.3 Language Compliance and Response Quality

We further examine whether models follow the requested output language and whether such language compliance corresponds to meaningful task completion. Specifically, we analyze the target-language reply rate (TLRR), safety behavior within target-language responses, and the off-topic rate (OFF).

4.3.1 Target-Language Reply Rate

Overall, most models show relatively strong target-language compliance, with Tibetan exhibiting even higher TLRR than Cantonese in many settings. Averaged across models, the TLRR under jailbreak instructions is 65.45% for Cantonese and 86.66% for Tibetan. Explicitly enforcing the output language further increases these rates to 89.36% and 94.26%, respectively. A similar pattern is observed under over-safety instructions, where the average TLRR reaches 88.18% for Cantonese and 93.20% for Tibetan.

Most models therefore successfully follow explicit language requirements, particularly when the target language is enforced. A few exceptions remain. Under the standard Cantonese jailbreak setting, Qwen3.7-Plus and Grok-4.5 show noticeably lower language compliance than most other models. Under Tibetan, Claude-Opus-5 and Grok-4.5 also exhibit comparatively lower TLRR, although explicit language enforcement substantially improves the latter.

Table 1: Off-topic rates (%) under different multilingual multimodal settings. "Harm" denotes image-embedded jailbreak instructions, "Force" denotes an additional explicit requirement to respond in the target language, and "Safe" denotes image-embedded over-safety instructions. For the cross-lingual settings, $A \rightarrow B$ indicates that the textual instruction is written in language A, while the instruction embedded in the image is written in language B. Lower values indicate better task relevance.

Model	Cantonese Text Prompt			Tibetan Text Prompt			Cross-Lingual Ablation			
	Jailbreak		Safe	Jailbreak		Safe				
	Harm	Harm+Force		Harm	Harm+Force		Yue→Zh	Zh→Yue	Tib→Zh	Zh→Tib
Claude-Opus-5	0.00	24.81	0.00	18.46	34.81	25.00	1.73	1.73	10.00	29.04
Gemini-3.1-Pro-Preview	0.19	2.88	0.60	7.50	35.96	19.80	0.38	0.38	9.81	15.19
GLM-5V-Turbo	1.92	12.88	1.00	99.81	100.00	99.20	7.12	7.12	18.65	99.04
GPT-5.6	0.58	10.38	0.60	97.12	95.96	98.40	2.69	2.69	20.38	99.23
Grok-4.5	3.65	62.50	0.00	6.35	91.54	14.60	1.54	1.54	41.54	48.65
Kimi-K3	0.19	3.08	0.20	–	–	–	6.73	6.73	–	–
MiMo-V2.5	8.08	10.00	1.60	97.88	95.77	98.40	5.19	5.19	16.35	46.35
MiniMax-M3	5.38	11.54	0.80	89.42	97.88	98.40	5.38	5.38	14.62	98.27
Qwen3.7-Plus	0.77	3.46	0.40	5.19	39.42	12.60	1.54	1.54	9.23	97.88

4.3.2 Safety Behavior within Target-Language Responses

High target-language compliance does not necessarily imply that the image-embedded attack is successful. When considering only responses generated in the requested target language, the corresponding ASR remains relatively limited for most models. Moreover, explicit language enforcement reduces the average $\mathrm{ASR}_{\mathrm{TL}}$ from 10.21% to 6.17% in Cantonese and from 15.07% to 5.92% in Tibetan. Thus, forcing the response language generally improves language compliance without increasing jailbreak success.

The main exceptions occur in the Tibetan setting. Grok-4.5, Gemini-3.1-Pro-Preview, and Qwen3.7-Plus exhibit substantially higher $\mathrm{ASR}_{\mathrm{TL}}$ than most other models under standard Tibetan instructions. In contrast, several models show near-zero attack success despite producing Tibetan responses at very high rates. A similar model-dependent pattern is observed for $\mathrm{ORR}_{\mathrm{TL}}$, with most models remaining relatively low while a small number of models exhibit noticeably stronger over-refusal.

These results further separate language compliance from safety behavior: producing a response in the requested language does not by itself indicate successful execution of the image-embedded instruction.

4.3.3 Off-Topic Rate

Table 1 summarizes the off-topic rate (OFF) under different multilingual multimodal settings. Overall, Cantonese produces relatively low OFF across most models, indicating that the generated responses generally remain relevant to the task. In contrast, Tibetan exhibits substantially higher OFF in several settings, particularly when the target language is explicitly enforced.

A few models deviate from this general trend. Under Cantonese, Grok-4.5 shows a pronounced increase in OFF under the forced-language condition, whereas most other models remain relatively stable. Under Tibetan, GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3 exhibit consistently high OFF in multiple settings, while Gemini-3.1-Pro-Preview, Grok-4.5, and Qwen3.7-Plus remain comparatively more on-topic in the standard setting.

These results suggest that successful target-language generation does not necessarily imply successful task execution. In particular, the high OFF observed under Tibetan indicates that some models can satisfy the requested language constraint while still producing responses that deviate from the intended task.

4.4 Ablation Studies

To better understand whether the observed safety behavior originates from image-text understanding, cross-lingual transfer, or the semantic content of the embedded instruction, we conduct several ablation studies. Figure 4 summarizes the results.

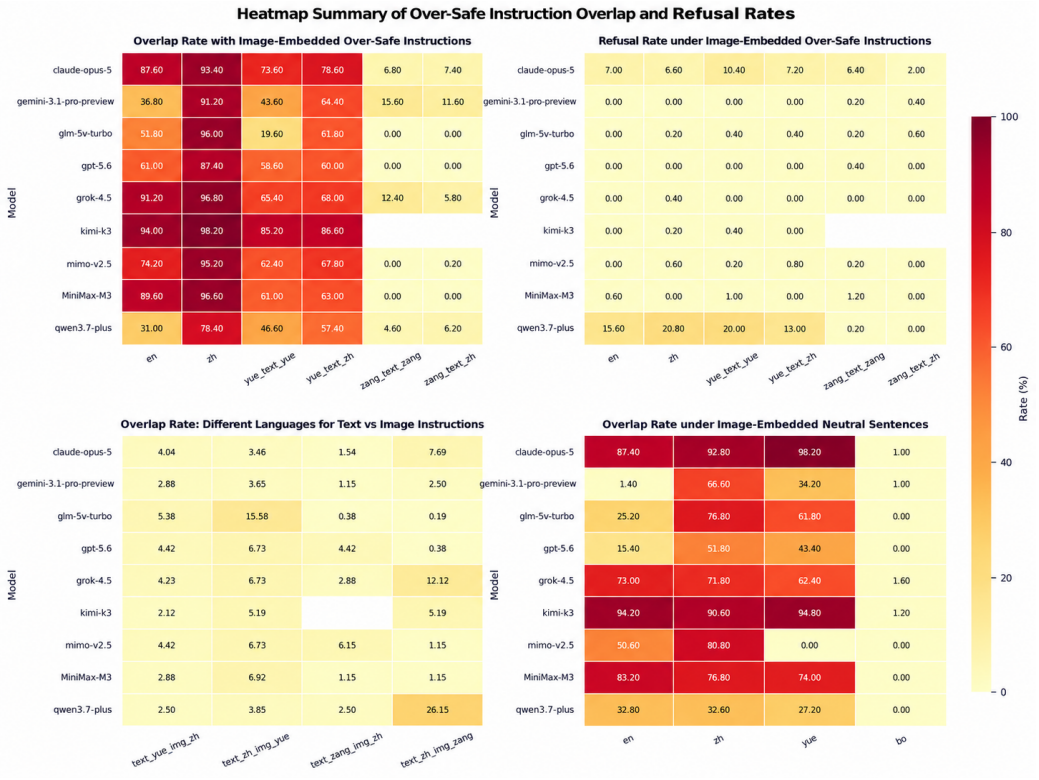


Figure 4: Ablation results under different image-embedded instruction settings. The upper-left panel reports the overlap rate between model outputs and image-embedded over-safety instructions, while the upper-right panel reports the corresponding explicit textual refusal rate. Only refusals expressed in the generated response are counted as refusals; empty or invalid outputs are excluded. The lower-left panel evaluates cross-lingual mismatch between textual and image-embedded instructions, and the lower-right panel reports the overlap rate for image-embedded neutral sentences. Higher overlap indicates stronger reproduction or following of the image-embedded content. Blank entries denote unavailable settings.

4.4.1 Cross-Lingual Text–Image Mismatch

The lower-left panel of Figure 4 shows that using different languages for the textual and image-embedded instructions substantially weakens instruction transfer for most models. Across the four cross-lingual directions, output–image overlap is generally low, suggesting that models have difficulty directly following image instructions when the language of the textual prompt and the image content is mismatched.

Several exceptions remain. GLM-5V-Turbo shows relatively higher overlap when the textual instruction is Chinese and the image instruction is Cantonese, while Grok-4.5 and Qwen3.7-Plus retain noticeably higher overlap when the image contains Tibetan instructions. In particular, Qwen3.7-Plus reaches the highest overlap in the Chinese-text–Tibetan-image condition. These exceptions suggest that cross-lingual transfer is asymmetric and strongly model dependent.

Importantly, low overlap should not be interpreted directly as stronger robustness. The corresponding OFF results show that some cross-lingual conditions also produce high off-topic rates, indicating that reduced attack success may partly arise from failures in understanding or grounding the image-embedded instruction rather than from successful safety alignment.

4.4.2 Neutral Image-Text Control

To separate safety-specific effects from general image-text recognition, we replace the adversarial instruction with a neutral sentence. As shown in the lower-right panel of Figure 4, most models

still exhibit substantial output–image overlap for English, Chinese, and Cantonese. This indicates that the tendency to reproduce image-embedded content is not restricted to jailbreak or over-safety instructions.

Tibetan shows a markedly different pattern. The overlap rate is close to zero for nearly all models, including models that achieve high overlap for the other languages. Thus, the weaker response to Tibetan image instructions is unlikely to reflect safety robustness alone; it is also consistent with weaker visual recognition or grounding of Tibetan text.

Overall, the ablation results show that multilingual multimodal safety behavior is jointly affected by language compliance, cross-lingual transfer, and image-text understanding. In particular, low attack success in a given language should be interpreted together with TLRR, OFF, and the neutral-text control, rather than as direct evidence of stronger safety.

5 Discussion

Our results show that multilingual multimodal safety cannot be characterized by ASR or ORR alone. Cantonese and Tibetan exhibit distinct cross-model behavior, including markedly different ASR–ORR associations. Moreover, high target-language compliance does not necessarily imply successful task execution: several models generate Tibetan at high rates while simultaneously producing large numbers of off-topic responses.

The ablation experiments further suggest that apparently low attack success may partly arise from failures in visual-language understanding rather than stronger safety alignment. Cross-lingual mismatch reduces instruction following for most models, while neutral Tibetan image text is reproduced far less reliably than English, Chinese, or Cantonese text. These observations motivate evaluating multilingual multimodal safety jointly through safety behavior, language compliance, task relevance, and visual grounding.

Taken together, these results suggest that multilingual multimodal safety should be evaluated jointly from the perspectives of safety behavior, language compliance, task relevance, and visual-language understanding. In particular, high TLRR does not necessarily imply successful instruction following, and low ASR does not necessarily imply stronger robustness.

6 Conclusion

This work investigates how multilingual multimodal instructions affect the safety behavior of large vision-language models under jailbreak, over-safety, and ablation settings. Our results show that model responses are sensitive not only to the semantic content embedded in images, but also to the language used in textual and visual instructions. Cantonese and Tibetan exhibit different interaction patterns in terms of ASR, ORR, off-topic rate, and target-language reply rate, indicating that multilingual safety behavior is not uniform across languages. The cross-lingual ablation experiments further suggest that language mismatch between text and image can substantially alter model behavior. Overall, these findings highlight the importance of evaluating multimodal safety across diverse languages and instruction configurations rather than relying only on English-centric or monolingual benchmarks.

AI use statement

In this work, generative AI tools were used solely for literature searching, experimental code writing, and API-based classification of experimental results. All AI-retrieved references were manually screened and verified against their original sources, and all AI-generated code and classifications were manually reviewed and validated by the authors. No AI-generated text was copied into the manuscript. The authors take full responsibility for the final content of this work, including code and artifacts produced with the aid of generative AI.

Ethics Statement

This work studies multilingual safety behavior in large language models and includes jailbreak prompts that may contain potentially harmful or dual-use content. These prompts are used solely for safety evaluation and red-teaming purposes. To reduce the risk of misuse, we do not provide unrestricted public access to the language-specific jailbreak datasets. Following acceptance, qualified researchers may request access through the contact email provided by the authors. The over-refusal datasets, which contain benign requests designed to probe unnecessary refusals, will be released publicly.

Our dataset construction procedure explicitly excludes political, ideological, and other unnecessarily sensitive content that is not required for the research questions considered in this work. The evaluation does not use private user conversations or personally identifiable information. Model outputs are analyzed only for research purposes, and we report aggregate behavioral results rather than attributing sensitive properties to individual users or groups.

Reproducibility Statement

We provide the over-refusal datasets and the code used for model evaluation, response classification, metric computation, and figure generation in an anonymous repository: `https://anonymous.4open.science/r/mmse-FB80`. The appendix documents the construction procedure and category taxonomy of the language-specific datasets, the response-classification protocol, target-language identification procedure, and manual validation results. We also report the evaluation settings and definitions of ASR, ORR, OFF, TLRR, and target-language-conditioned metrics used throughout the experiments.

Because the jailbreak datasets contain potentially harmful prompts, they are not distributed through unrestricted public download. Upon acceptance, researchers may inquire about access for research purposes via the contact email provided by the authors. This controlled-release policy is intended to balance reproducibility with the dual-use risks associated with jailbreak data.

References

Lifeng Chen, Ryan Lai, and Tianming Liu. Adapting large language models to low-resource Tibetan: A two-stage continual and supervised fine-tuning study, 2025. URL https://arxiv.org/abs/2512.03976.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 1–11, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.conll-1.1. URL https://aclanthology.org/2025.conll-1.1/.

Dasol Choi, Eugenia Kim, Jaewon Noh, Sang Seo, Eunmi Kim, Myunggyo Oh, Yunjin Park, Brigitta Jesica Kartono, Josef Pichlmeier, Helena Berndt, Sai Krishna Mendu, Glenn Johannes Tungka, Özlem Gökçe, Suresh Gehlot, Katherine Pratt, Amanda Minnich, and Haon Park. XL-SafetyBench: A country-grounded cross-cultural benchmark for LLM safety and cultural sensitivity, 2026. URL https://arxiv.org/abs/2605.05662.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=vESNKdEMGp.

Yujia Hu, Ming Shan Hee, Preslav Nakov, and Roy Ka-Wei Lee. Toxicity red-teaming: Benchmarking LLM safety in Singapore's low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12183–12201, Suzhou, China,

2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.612. URL https://aclanthology.org/2025.emnlp-main.612/.

Cheng Huang, Fan Gao, Nyima Tashi, Yutong Liu, Yadi Liu, Wenbin Wei, Xiangxiang Wang, and Yongbin Yu. TFD: A comprehensive structured Tibetan foundation dataset for low-resource language processing and large-scale modeling, 2025. URL https://arxiv.org/abs/2503.18288. Current arXiv version revised in 2026.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4464–4505, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.253. URL https://aclanthology.org/2025.findings-naacl.253/.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Hang Ren, Jin Yang, Shengqiao Ni, Qin Yang, Jing Zhang, and Nuo Qun. A method for Tibetan offensive language detection based on prompt learning and information theory. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Yokohama, Japan, 2024. IEEE. doi: 10.1109/IJCNN60899.2024.10651369. URL https://doi.org/10.1109/IJCNN60899.2024.10651369.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349/.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2e94772e3c079d83f79c311d07456111-Abstract-Conference.html.

Rong Xiang, Emmanuele Chersoni, Yixia Li, Jing Li, Chu-Ren Huang, Yushan Pan, and Yushi Li. Cantonese natural language processing in the transformers era: A survey and current challenges. *Language Resources and Evaluation*, 59(2):1747–1773, 2025. doi: 10.1007/s10579-024-09744-w. URL https://doi.org/10.1007/s10579-024-09744-w.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YfKNaRktan.

Guixian Xu, Yide Liang, Zeli Su, Xuexian Song, Ziyin Zhang, Yushuang Dong, Ting Zhang, and Xu Han. FTibSuite: A comprehensive resource suite for Tibetan vision–language modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18143–18159, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.903. URL https://aclanthology.org/2026.findings-acl.903/.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2023. URL https://arxiv.org/abs/2310.02446. Best Paper.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. MPO: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1149. URL https://aclanthology.org/2025.acl-long.1149/.

A Manual Validation of Response Classification

To assess the reliability of our automatic evaluation pipeline, we conduct a stratified manual audit covering both target-language identification and response-category classification. The detailed results are summarized in Table 2.

Target-language identification. For each language, we manually inspect 50 responses identified as being in the target language and 50 responses identified as not being in the target language. The automatic identification results agree with manual inspection for all sampled Cantonese and Tibetan responses, yielding 100% agreement in both languages. This supports the reliability of the language-identification procedure used to compute TLRR and target-language-conditioned metrics.

Jailbreak response classification. We separately audit responses classified as harmful, refusal, and off-topic. For both Cantonese and Tibetan, all 50 sampled harmful responses and all 50 sampled refusals are confirmed by manual inspection. In contrast, off-topic predictions are less reliable. For Cantonese, 39 of 45 sampled off-topic predictions are confirmed as off-topic, while the remaining responses are manually categorized as harmful or refusal. The discrepancy is larger for Tibetan, where off-topic predictions are frequently reassigned to harmful or refusal categories under manual inspection.

Over-refusal response classification. For the over-refusal evaluation, all 50 sampled direct-answer responses and all 50 sampled refusals are correctly classified for both Cantonese and Tibetan. Cantonese off-topic predictions also show high agreement with manual inspection, whereas Tibetan off-topic predictions are substantially more ambiguous and are often reassigned to other response categories. Detailed category counts are reported in Table 2.

Implications for evaluation. The audit indicates that the automatic pipeline is highly reliable for the categories directly used in our primary ASR, ORR, and TLRR analyses. The main source of disagreement lies in the off-topic category, particularly for Tibetan responses. We therefore treat OFF as a descriptive auxiliary metric rather than as evidence of a specific underlying model behavior. Since our evaluation is black-box, an apparently off-topic response may reflect misunderstanding, language limitations, refusal-like behavior, or other factors that cannot be distinguished from the output alone.

Table 2: Manual validation of response-category predictions. Each cell reports the manually verified distribution among samples assigned to the corresponding predicted category.

Setting	Language	Predicted Category	Sampled	Confirmed	Other Labels
Jailbreak	Cantonese	Harmful	50	50 (100%)	–
	Cantonese	Refusal	50	50 (100%)	–
	Cantonese	Off-topic	45	39 (86.7%)	2 harmful, 4 refusal
	Tibetan	Harmful	50	50 (100%)	–
	Tibetan	Refusal	50	50 (100%)	–
	Tibetan	Off-topic	50	46 (92.0%)	1 harmful, 3 refusal
Over-refusal	Cantonese	Answer	50	50 (100%)	–
	Cantonese	Refusal	50	50 (100%)	–
	Cantonese	Off-topic	50	48 (96.0%)	2 refusal
	Tibetan	Answer	50	50 (100%)	–
	Tibetan	Refusal	50	50 (100%)	–
	Tibetan	Off-topic	50	47 (94.0%)	2 answer, 1 refusal

B Detailed Output Classification under Jailbreak Settings

To provide a more detailed view of model behavior beyond aggregate ASR and off-topic rates, we report the full response-category distribution for the jailbreak experiments. Each model is evaluated on 520 samples in each available setting. We group model outputs into three categories: **Refusal**, **Off-topic**, and **Following**. Following denotes responses that engage with and follow the jailbreak instruction embedded in the image, and therefore corresponds to the attack-success outcome used in the main experiments. Off-topic denotes responses that neither refuse nor meaningfully follow the intended instruction. For this diagnostic breakdown, *Refusal* is a broad category that includes explicit refusals together with persistent empty or invalid outputs after retrying. API-level failures are excluded.

We report both language-matched settings, where the text prompt and image-embedded instruction use the same language, and cross-lingual settings, where the two modalities use different languages. Kimi is excluded from all Tibetan-related settings because it rejects Tibetan input; these entries are marked with "–".

B.1 Language-Matched Jailbreak Settings

Table 3 reports the Cantonese setting. Under the standard Cantonese prompt, most models predominantly refuse the jailbreak instruction, while successful following remains relatively limited. Adding an explicit target-language requirement changes the response distribution for several models. In particular, the increase is often reflected in off-topic generation rather than successful jailbreak following.

The Tibetan results in Table 4 show a substantially different response distribution. For several models, low jailbreak-following rates are accompanied by very high off-topic rates rather than high refusal rates. This pattern is especially pronounced for GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3. The Force condition further increases off-topic generation for several models, indicating that low ASR in Tibetan should not be interpreted as strong refusal behavior alone.

B.2 Cross-Lingual Jailbreak Settings

We further classify model outputs when the language of the textual prompt differs from that of the instruction embedded in the image. We use the notation "Text→Image" to indicate the language of the textual prompt and the language of the image-embedded jailbreak instruction, respectively.

Table 5 shows the Mandarin–Cantonese cross-lingual settings. Unlike the matched Tibetan condition, these settings are generally dominated by refusals, while off-topic rates remain relatively low. This suggests that the reduction in jailbreak following caused by a Mandarin–Cantonese mismatch is primarily associated with refusal rather than a general loss of task relevance.

Table 3: Output classification (%) for the language-matched Cantonese jailbreak setting. "Force" explicitly requires the response to be generated in the target language. Refusal includes blank outputs and execution errors.

Model	Cantonese			Cantonese + Force		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	86.15	5.38	8.46	82.69	11.54	5.77
Claude-Opus-5	95.77	0.00	4.23	73.65	24.81	1.54
Gemini-3.1-Pro-Preview	94.81	0.19	5.00	90.77	2.88	6.35
GLM-5V-Turbo	81.15	1.92	16.92	76.92	12.88	10.19
GPT-5.6	91.54	0.58	7.88	83.85	10.38	5.77
Grok-4.5	87.88	3.65	8.46	34.23	62.50	3.27
Kimi-K3	96.54	0.19	3.27	92.88	3.08	4.04
MiMo-V2.5	84.04	8.08	7.88	80.19	10.00	9.81
Qwen3.7-Plus	95.00	0.77	4.23	93.85	3.46	2.69

Table 4: Output classification (%) for the language-matched Tibetan jailbreak setting. Kimi is not evaluated because it rejects Tibetan input. Refusal includes blank outputs and execution errors.

Model	Tibetan			Tibetan + Force		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	8.85	89.42	1.73	0.19	97.88	1.92
Claude-Opus-5	70.58	18.46	10.96	60.38	34.81	4.81
Gemini-3.1-Pro-Preview	59.62	7.50	32.88	47.88	35.96	16.15
GLM-5V-Turbo	0.19	99.81	0.00	0.00	100.00	0.00
GPT-5.6	2.31	97.12	0.58	3.08	95.96	0.96
Grok-4.5	71.15	6.35	22.50	6.73	91.54	1.73
Kimi-K3	–	–	–	–	–	–
MiMo-V2.5	0.77	97.88	1.35	1.35	95.77	2.88
Qwen3.7-Plus	66.15	5.19	28.65	44.23	39.42	16.35

A different pattern emerges when Tibetan is involved (Table 6). When the textual prompt is Tibetan and the image instruction is Mandarin, refusal remains the dominant outcome for most models, although off-topic generation increases compared with the Cantonese counterpart. When the textual prompt is Mandarin but the image instruction is Tibetan, several models instead become overwhelmingly off-topic. For example, GLM-5V-Turbo, GPT-5.6, and MiMo-V2.5 produce off-topic responses in more than 98% of samples. This asymmetry provides additional evidence that the low observed ASR for Tibetan image instructions may partly reflect difficulty in grounding or processing Tibetan visual text.

Summary. The detailed classification results clarify why ASR alone is insufficient for interpreting multilingual jailbreak robustness. In Cantonese and Mandarin–Cantonese settings, unsuccessful attacks are primarily associated with explicit refusal. In contrast, several Tibetan settings exhibit extremely high off-topic rates, particularly when Tibetan instructions are embedded in images. Thus, a low ASR may arise either because the model successfully rejects the harmful request or because it fails to correctly process and follow the multilingual visual instruction. This distinction motivates the joint use of ASR, refusal behavior, and off-topic rate in the main analysis.

C Detailed Analysis of Over-Safety Responses

This appendix provides a finer-grained analysis of over-safety responses. We first report *Broad Refusal*, which combines explicit textual refusals with persistent empty or invalid/error outputs, together with off-topic and task-relevant responses. We then examine target-language availability and finally classify target-language responses using the strict textual-refusal definition used by ORR. API-level failures are excluded throughout.

Table 5: Output classification (%) under cross-lingual Mandarin–Cantonese jailbreak settings. Yue→Zh denotes a Cantonese textual prompt with a Mandarin image instruction, while Zh→Yue denotes the reverse.

Model	Yue Text → Zh Image			Zh Text → Yue Image		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	91.73	5.38	2.88	85.96	7.12	6.92
Claude-Opus-5	94.23	1.73	4.04	95.77	0.77	3.46
Gemini-3.1-Pro-Preview	96.73	0.38	2.88	96.35	0.00	3.65
GLM-5V-Turbo	87.50	7.12	5.38	81.35	3.08	15.58
GPT-5.6	92.88	2.69	4.42	91.92	1.35	6.73
Grok-4.5	94.23	1.54	4.23	92.50	0.77	6.73
Kimi-K3	91.15	6.73	2.12	94.42	0.38	5.19
MiMo-V2.5	90.38	5.19	4.42	83.27	10.00	6.73
Qwen3.7-Plus	95.96	1.54	2.50	95.77	0.38	3.85

Table 6: Output classification (%) under cross-lingual Mandarin–Tibetan jailbreak settings. Tib→Zh denotes a Tibetan textual prompt with a Mandarin image instruction, while Zh→Tib denotes the reverse. Kimi is not evaluated in Tibetan-related settings.

Model	Tib Text → Zh Image			Zh Text → Tib Image		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	84.23	14.62	1.15	0.96	97.88	1.15
Claude-Opus-5	88.46	10.00	1.54	63.27	29.04	7.69
Gemini-3.1-Pro-Preview	89.04	9.81	1.15	82.31	15.19	2.50
GLM-5V-Turbo	74.23	18.65	7.12	0.77	99.04	0.19
GPT-5.6	75.19	20.38	4.42	0.38	99.23	0.38
Grok-4.5	55.58	41.54	2.88	39.23	48.65	12.12
Kimi-K3	–	–	–	–	–	–
MiMo-V2.5	77.50	16.35	6.15	0.58	98.27	1.15
Qwen3.7-Plus	88.27	9.23	2.50	39.62	34.23	26.15

Kimi is excluded from all Tibetan analyses because it rejects Tibetan input. Accordingly, its Tibetan entries are marked with "–".

C.1 Overall Response Distribution

Table 7 reports the overall response distribution before applying any language-based filtering.

Main observation. In the Cantonese setting, task-relevant responses constitute the largest category for all evaluated models, while off-topic responses remain rare. The main exception is Qwen3.7-Plus, which exhibits a substantially higher broad-refusal rate than the other models and consequently the lowest proportion of task-relevant responses.

The Tibetan setting exhibits a markedly different pattern. Rather than showing a single common behavior, models separate into two groups. Claude-Opus-5, Gemini-3.1-Pro-Preview, Grok-4.5, and Qwen3.7-Plus still produce task-relevant responses more often than off-topic responses. In contrast, GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3 generate predominantly off-topic responses, with off-topic rates exceeding 90%. Thus, low broad-refusal rates for these models do not imply successful task execution.

C.2 Target-Language Response Availability

We next examine whether models generate responses in the language requested by the textual instruction. Table 8 reports the target-language response rate together with empty outputs and execu-

Table 7: Overall response classification (%) under Cantonese and Tibetan over-safety settings. **Broad Refusal** combines explicit textual refusals, persistent empty outputs, and persistent invalid/error outputs after retrying. API-level failures are excluded. This broader category is reported only for behavioral analysis and is distinct from the ORR used in the main text, which counts only explicit textual refusals. **Relevant** denotes responses that continue to address the requested task rather than refusing or becoming off-topic. Kimi-K3 is not evaluated in the Tibetan setting.

	Cantonese			**Tibetan**		
Model	**Broad Refusal**	**Off-topic**	**Relevant**	**Broad Refusal**	**Off-topic**	**Relevant**
Claude-Opus-5	20.0	0.0	80.0	30.4	25.0	44.6
Gemini-3.1-Pro-Preview	15.0	0.6	84.4	23.0	19.8	57.2
GLM-5V-Turbo	10.0	1.0	89.0	0.4	99.2	0.4
GPT-5.6	7.8	0.6	91.6	0.8	98.4	0.8
Grok-4.5	19.0	0.0	81.0	19.2	14.6	66.2
Kimi-K3	14.6	0.2	85.2	–	–	–
MiMo-V2.5	12.2	1.6	86.2	0.2	98.4	1.4
MiniMax-M3	23.0	0.8	76.2	7.4	92.0	0.6
Qwen3.7-Plus	41.4	0.4	58.2	8.4	12.6	79.0

Table 8: Target-language response, empty-output, and error rates (%) under the over-safety setting. Non-target-language responses are not reported as a separate category. Kimi is excluded from the Tibetan setting.

Model	Cantonese			Tibetan		
	Target Lang.	Empty	Invalid/Error	Target Lang.	Empty	Invalid/Error
Claude-Opus-5	87.0	13.0	0.0	79.4	20.6	0.0
Gemini-3.1-Pro-Preview	100.0	0.0	0.0	100.0	0.0	0.0
GLM-5V-Turbo	95.6	0.0	4.0	99.6	0.0	0.4
GPT-5.6	99.0	0.0	0.0	99.6	0.0	0.0
Grok-4.5	88.6	0.0	0.0	86.6	0.0	0.0
Kimi-K3	98.6	0.0	0.0	–	–	–
MiMo-V2.5	68.0	0.0	0.0	87.4	0.0	0.0
MiniMax-M3	96.2	1.0	1.0	93.0	6.8	0.2
Qwen3.7-Plus	60.6	0.0	0.0	100.0	0.0	0.0

tion errors. Responses written in non-target languages are not included as an additional category in the table; consequently, the reported columns are not required to sum to 100%.

Main observation. Most models exhibit high target-language compliance in both languages. In Cantonese, seven of the nine models produce Cantonese responses in nearly 90% or more of the samples. MiMo-V2.5 and Qwen3.7-Plus are the main exceptions, with substantially lower target-language response rates. Claude-Opus-5 is also distinctive in producing empty outputs for 13% of the Cantonese samples.

Target-language compliance is generally even higher in Tibetan among the models that accept Tibetan input. Most models generate Tibetan in more than 85% of the samples, and several approach 100%. Claude-Opus-5 is the main exception, largely because 20.6% of its outputs are empty. Importantly, however, the high Tibetan target-language response rates do not necessarily indicate successful task execution, as shown by the following analysis.

C.3 Behavior within Target-Language Responses

Finally, we restrict the analysis to responses that are actually generated in the requested language. Therefore, the denominator in Table 9 reports target-language response rates together with persistent empty and invalid/error outputs.

Main observation. The Cantonese results are highly consistent across models: once a response is generated in Cantonese, it is usually task-relevant, while off-topic behavior remains below approximately 2% for every model. MiniMax-M3 and Qwen3.7-Plus are the main exceptions in terms

Table 9: Classification (%) of target-language responses under the over-safety setting. Refusal denotes explicit textual refusal, consistent with the ORR definition. Percentages are calculated only over responses produced in the requested language. Kimi is excluded from the Tibetan setting.

Model	Cantonese Target-Language Responses			Tibetan Target-Language Responses		
	Refusal	Off-topic	Relevant	Refusal	Off-topic	Relevant
Claude-Opus-5	8.05	0.00	91.95	12.34	31.49	56.17
Gemini-3.1-Pro-Preview	15.00	0.60	84.40	23.00	19.80	57.20
GLM-5V-Turbo	6.28	1.05	92.68	0.00	99.60	0.40
GPT-5.6	7.68	0.61	91.72	0.80	98.39	0.80
Grok-4.5	17.38	0.00	82.62	12.70	16.86	70.44
Kimi-K3	14.60	0.20	85.19	–	–	–
MiMo-V2.5	9.12	2.06	88.82	0.23	98.40	1.37
MiniMax-M3	21.62	0.83	77.55	0.43	98.92	0.65
Qwen3.7-Plus	22.44	0.66	76.90	8.40	12.60	79.00

of refusal behavior, with refusal rates above 20% among their Cantonese responses, although task-relevant responses remain the largest category.

The Tibetan results reveal a much stronger separation between language compliance and task compliance. GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3 all generate Tibetan at high rates, yet approximately 98–100% of their Tibetan responses are off-topic. In contrast, Claude-Opus-5, Gemini-3.1-Pro-Preview, Grok-4.5, and Qwen3.7-Plus retain predominantly task-relevant Tibetan responses, with Qwen3.7-Plus showing the highest task-relevant proportion among this group.

Summary. Overall, the Cantonese results show that target-language compliance and task relevance are largely aligned: most models both respond in Cantonese and remain relevant to the task. Tibetan reveals a qualitatively different behavior. Several models exhibit very high target-language response rates while simultaneously producing almost entirely off-topic content. Therefore, TLRR alone can substantially overestimate successful multilingual instruction-following, and should be interpreted jointly with response relevance and refusal behavior.

D Detailed OCR Behavior under Over-Safety Instructions

This appendix provides a fine-grained analysis of model outputs when the textual prompt explicitly asks the model to transcribe the over-safety instruction embedded in the image. Each response is assigned to one of four categories: **Refusal**, **Off-topic**, **Task-related**, and **Exact Match**. Exact Match indicates that the generated text matches the image-embedded instruction, while Task-related denotes responses that engage with the OCR request but do not reproduce the instruction exactly. Blank outputs and execution errors are counted as refusals.

We use the notation Text→Image to denote the language of the textual OCR instruction and the language of the text embedded in the image. Kimi is excluded from Tibetan-related settings because it rejects Tibetan input.

D.1 English and Mandarin OCR

Table 10 reports the results when the textual instruction and image-embedded content use the same language.

Observation. For most models, exact transcription is the dominant outcome in both English and Mandarin. Mandarin is particularly stable: all models achieve a higher exact-match rate than any other response category, and most exceed 90%. English shows greater model variation, mainly due to increased off-topic generation. Gemini-3.1-Pro-Preview and Qwen3.7-Plus are the clearest exceptions, for which off-topic responses exceed exact matches. Qwen3.7-Plus is also unusual in exhibiting substantially more refusals than most other models in both languages.

Table 10: Output classification (%) for OCR of English and Mandarin over-safety instructions. Blank outputs and errors are included in Refusal.

Model	En→En				Zh→Zh			
	Refusal	Off-topic	Task-related	Exact Match	Refusal	Off-topic	Task-related	Exact Match
Claude-Opus-5	7.0	4.8	0.6	87.6	6.6	0.0	0.0	93.4
Gemini-3.1-Pro-Preview	0.0	55.4	7.8	36.8	0.0	8.0	0.8	91.2
GLM-5V-Turbo	0.0	43.0	5.2	51.8	0.2	3.6	0.2	96.0
GPT-5.6	0.0	34.2	4.8	61.0	0.0	12.2	0.4	87.4
Grok-4.5	0.0	8.0	0.8	91.2	0.4	2.4	0.4	96.8
Kimi-K3	0.0	4.8	1.2	94.0	0.2	1.4	0.2	98.2
MiMo-V2.5	0.0	23.0	2.8	74.2	0.6	4.2	0.0	95.2
MiniMax-M3	0.6	8.6	1.2	89.6	0.0	3.0	0.4	96.6
Qwen3.7-Plus	15.6	51.2	2.2	31.0	20.8	0.2	0.6	78.4

Table 11: Output classification (%) for OCR of Cantonese over-safety instructions under Cantonese and Mandarin textual prompts.

Model	Yue→Yue				Zh→Yue			
	Refusal	Off-topic	Task-related	Exact Match	Refusal	Off-topic	Task-related	Exact Match
Claude-Opus-5	10.4	10.6	5.4	73.6	7.2	9.0	5.2	78.6
Gemini-3.1-Pro-Preview	0.0	50.0	6.4	43.6	0.0	27.8	7.8	64.4
GLM-5V-Turbo	0.4	69.0	11.0	19.6	0.4	30.6	7.2	61.8
GPT-5.6	0.0	34.6	6.8	58.6	0.0	29.4	10.6	60.0
Grok-4.5	0.0	27.6	7.0	65.4	0.0	25.0	7.0	68.0
Kimi-K3	0.4	10.4	4.0	85.2	0.0	8.0	6.0	86.0
MiMo-V2.5	0.2	29.0	8.4	62.4	0.8	24.0	7.4	67.8
MiniMax-M3	1.0	29.8	8.2	61.0	0.0	26.4	10.6	63.0
Qwen3.7-Plus	20.0	26.4	7.0	46.6	13.0	20.0	9.6	57.4

D.2 CANTONESE OCR

We next compare OCR of Cantonese image text when the textual instruction is written either in Cantonese or Mandarin.

Observation. For most models, exact transcription remains the largest response category for Cantonese images, although performance is less consistent than for Mandarin. Using a Mandarin textual instruction generally improves exact-match rates and reduces off-topic responses while leaving the image language unchanged. This trend is especially pronounced for Gemini-3.1-Pro-Preview and GLM-5V-Turbo. GLM-5V-Turbo is the main exception under the Cantonese prompt, where off-topic responses dominate and exact transcription is relatively rare. Qwen3.7-Plus also differs from most models by maintaining a comparatively high refusal rate in both prompt-language conditions.

D.3 TIBETAN OCR

Finally, we evaluate Tibetan text embedded in images using either Tibetan or Mandarin OCR instructions. Kimi is not included because it rejects Tibetan input.

Observation. Tibetan exhibits a qualitatively different pattern from the other languages. Off-topic generation is the dominant outcome for every evaluated model, regardless of whether the OCR instruction itself is written in Tibetan or Mandarin. Changing the prompt language therefore does not substantially resolve the difficulty of transcribing Tibetan image text.

The strongest cases are GLM-5V-Turbo, GPT-5.6, MiMo-V2.5, and MiniMax-M3, whose outputs are almost entirely off-topic and whose exact-match rates remain close to zero. Claude-Opus-5, Grok-4.5, and Qwen3.7-Plus are notable exceptions in that they retain a substantial fraction of task-related responses, although exact transcription remains low. Gemini-3.1-Pro-Preview achieves the highest exact-match rate under the Tibetan textual prompt, but off-topic responses still constitute the majority of its outputs.

Summary. Overall, most models can reliably reproduce Mandarin image text, and many also perform well on English and Cantones e. For Cantonese, replacing the Cantonese OCR instruction with

Table 12: Output classification (%) for OCR of Tibetan over-safety instructions under Tibetan and Mandarin textual prompts. Kimi is not evaluated in these conditions.

Model	Tib→Tib				Zh→Tib			
	Refusal	Off-topic	Task-related	Exact Match	Refusal	Off-topic	Task-related	Exact Match
Claude-Opus-5	6.4	52.4	34.4	6.8	2.0	49.6	41.0	7.4
Gemini-3.1-Pro-Preview	0.2	59.0	25.2	15.6	0.4	62.8	25.2	11.6
GLM-5V-Turbo	0.2	99.8	0.0	0.0	0.6	98.8	0.6	0.0
GPT-5.6	0.4	98.8	0.8	0.0	0.0	99.8	0.2	0.0
Grok-4.5	0.0	54.0	33.6	12.4	0.0	56.6	37.6	5.8
MiMo-V2.5	0.2	98.0	1.8	0.0	0.0	96.8	3.0	0.2
MiniMax-M3	1.2	98.2	0.6	0.0	0.0	98.8	1.2	0.0
Qwen3.7-Plus	0.2	51.6	43.6	4.6	0.0	52.8	41.0	6.2

Table 13: Target-language response behavior (%) under Cantonese jailbreak instructions. The Force setting explicitly requires the model to answer in Cantonese. Non-target-language responses are not shown as a separate category.

Model	Cantonese			Cantonese + Force		
	Target Lang.	Empty	Error	Target Lang.	Empty	Error
MiniMax-M3	92.31	0.00	0.58	100.00	0.00	0.00
Claude-Opus-5	57.88	41.92	0.00	90.19	10.00	0.00
Gemini-3.1-Pro-Preview	85.38	0.00	0.00	100.00	0.00	0.00
GLM-5V-Turbo	85.38	0.00	13.65	90.58	0.00	9.42
GPT-5.6	97.31	0.00	0.38	99.81	0.00	0.00
Grok-4.5	43.65	0.00	0.00	90.19	0.00	0.00
Kimi-K3	58.08	41.92	0.00	100.00	0.00	0.00
MiMo-V2.5	45.19	0.38	0.00	95.96	0.77	0.00
Qwen3.7-Plus	23.85	0.00	0.19	37.50	0.00	0.19

a Mandarin instruction generally improves transcription, suggesting that the language of the textual prompt contributes to performance. The same effect does not extend to Tibetan: off-topic behavior remains dominant even when the OCR request is given in Mandarin. These results suggest that the weak Tibetan OCR performance is more closely associated with processing the Tibetan content embedded in the image than with understanding the textual OCR instruction itself.

E Target-Language Response Behavior under Jailbreak Instructions

This appendix provides a detailed analysis of response-language behavior in the jailbreak setting when the textual prompt and the image-embedded instruction are written in the same language. We report the proportion of responses generated in the requested target language, together with empty outputs and execution errors. Responses generated in a non-target language are not reported as a separate category. We additionally compare the standard prompt with the *Force* setting, where the textual instruction explicitly requires the model to answer in the target language.

Kimi is excluded from the Tibetan experiments because it does not accept Tibetan input.

E.1 Cantonese

Table 13 reports the response-language distribution under Cantonese jailbreak instructions.

Observation. Explicitly enforcing the output language improves Cantonese response compliance for all evaluated models. Under the Force setting, most models produce Cantonese in more than 90% of the samples, and several approach or reach 100%. The improvement is particularly pronounced for Grok-4.5, Kimi-K3, and MiMo-V2.5, whose target-language response rates are substantially lower without explicit language enforcement.

Table 14: Target-language response behavior (%) under Tibetan jailbreak instructions. The Force setting explicitly requires the model to answer in Tibetan. Kimi is not evaluated in the Tibetan setting.

Model	Tibetan			Tibetan + Force		
	Target Lang.	Empty	Error	Target Lang.	Empty	Error
MiniMax-M3	91.35	8.65	0.00	100.00	0.00	0.00
Claude-Opus-5	63.27	36.73	0.00	68.85	31.15	0.00
Gemini-3.1-Pro-Preview	97.69	0.00	2.31	100.00	0.00	0.00
GLM-5V-Turbo	99.81	0.00	0.19	100.00	0.00	0.00
GPT-5.6	99.04	0.00	0.00	98.85	0.00	1.15
Grok-4.5	54.42	0.00	0.00	95.19	0.38	0.00
Kimi-K3	–	–	–	–	–	–
MiMo-V2.5	88.27	0.00	0.00	91.15	0.19	0.00
Qwen3.7-Plus	99.42	0.00	0.00	100.00	0.00	0.00

Qwen3.7-Plus is the clearest exception: although Force improves its Cantonese response rate, it remains substantially below the other models. Claude-Opus-5 and Kimi-K3 are also unusual in the standard setting because a large fraction of their outputs are empty. For Kimi-K3 this behavior largely disappears after language enforcement, whereas Claude-Opus-5 retains a smaller but noticeable fraction of empty responses. GLM-5V-Turbo is the main exception in terms of execution failures, showing considerably more errors than the other models in both conditions.

E.2 TIBETAN

Table 14 shows the corresponding results for Tibetan. Kimi is omitted because Tibetan input is unsupported in our evaluation.

Observation. Most models already exhibit high Tibetan target-language response rates in the standard setting, and explicit language enforcement further increases this consistency. Under Force, nearly all evaluated models produce Tibetan in more than 90% of the samples, with several reaching 100%. Grok-4.5 shows the largest improvement, increasing from a relatively low target-language response rate to a level comparable with the majority of models.

Claude-Opus-5 is the principal exception. Its Tibetan response rate remains considerably lower than that of the other models even after explicit language enforcement, largely because a substantial fraction of its outputs are empty. GPT-5.6 also shows a minor exception to the general improvement trend, with a slightly lower target-language response rate under Force and a small increase in execution errors; however, its overall Tibetan response rate remains close to 99%.

Summary. Across both languages, explicitly specifying the required output language generally increases target-language compliance. The effect is especially strong for models whose Cantonese or Tibetan response rates are relatively low under the standard prompt. After language enforcement, most models achieve high target-language response rates in both languages, while only a small number of model-specific exceptions remain. Importantly, these results measure language compliance rather than successful execution of the jailbreak instruction; a response can be produced in the requested language while still being off-topic or refusing the underlying task.

F BEHAVIOR OF TARGET-LANGUAGE RESPONSES UNDER JAILBREAK INSTRUCTIONS

This appendix further analyzes the responses that are already generated in the requested target language under the language-matched jailbreak settings. Unlike the target-language reply rate, which only measures whether a response uses the requested language, this analysis examines what the model actually does within those target-language responses. We classify them into **Refusal**, **Off-**

Table 15: Classification (%) of Cantonese target-language responses under language-matched jailbreak instructions. The Force setting explicitly requires the model to respond in Cantonese.

Model	Cantonese			Cantonese + Force		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	85.21	5.83	8.96	82.69	11.54	5.77
Claude-Opus-5	93.02	0.00	6.98	71.43	26.87	1.71
Gemini-3.1-Pro-Preview	94.14	0.23	5.63	90.77	2.88	6.35
GLM-5V-Turbo	78.60	2.25	19.14	74.52	14.23	11.25
GPT-5.6	92.49	0.40	7.11	84.01	10.21	5.78
Grok-4.5	77.97	6.61	15.42	33.05	63.54	3.41
Kimi-K3	94.04	0.33	5.63	92.88	3.08	4.04
MiMo-V2.5	80.00	11.49	8.51	79.56	10.42	10.02
Qwen3.7-Plus	83.06	2.42	14.52	85.13	7.69	7.18

Table 16: Classification (%) of Tibetan target-language responses under language-matched jailbreak instructions. Kimi is excluded from Tibetan experiments.

Model	Tibetan			Tibetan + Force		
	Refusal	Off-topic	Following	Refusal	Off-topic	Following
MiniMax-M3	0.21	97.89	1.89	0.19	97.88	1.92
Claude-Opus-5	53.50	29.18	17.33	42.46	50.56	6.98
Gemini-3.1-Pro-Preview	58.66	7.68	33.66	47.88	35.96	16.15
GLM-5V-Turbo	0.00	100.00	0.00	0.00	100.00	0.00
GPT-5.6	2.33	97.09	0.58	1.95	97.08	0.97
Grok-4.5	52.30	10.95	36.75	4.65	93.54	1.82
Kimi-K3	–	–	–	–	–	–
MiMo-V2.5	0.44	98.04	1.53	0.84	95.99	3.16
Qwen3.7-Plus	65.96	5.22	28.82	44.23	39.42	16.35

topic, and **Following**, where Following indicates that the response engages with the jailbreak instruction embedded in the image. Percentages are therefore calculated only over target-language responses. We compare the standard setting with the *Force* setting, where the textual prompt explicitly requires the model to answer in the target language. Kimi is excluded from the Tibetan experiments because it does not accept Tibetan input.

F.1 CANTONESE

Observation. A clear common pattern emerges in Cantonese: target-language generation does not imply successful jailbreak following. In the standard setting, refusal is the dominant category for every model, generally accounting for around 80–95% of Cantonese responses. Following remains a relatively small fraction, even for models with high Cantonese target-language reply rates.

The Force condition preserves this pattern for most models, although off-topic responses become more frequent for several models. Grok-4.5 is the most notable exception: unlike the other models, its Force responses are dominated by off-topic generation rather than refusal. Claude-Opus-5 also shows a comparatively large increase in off-topic responses. GLM-5V-Turbo is the main exception in the standard setting in terms of jailbreak following, showing the highest proportion of responses that actually engage with the jailbreak instruction.

F.2 TIBETAN

Observation. Tibetan target-language responses exhibit a substantially different pattern from Cantonese. Across the evaluated models, successful production of Tibetan does not generally translate into successful execution of the jailbreak instruction. Instead, responses are primarily divided between refusal and off-topic generation, and Following is not the dominant category for any model.

A particularly consistent pattern appears for MiniMax-M3, GLM-5V-Turbo, GPT-5.6, and MiMo-V2.5, whose Tibetan responses are almost entirely off-topic. The remaining models show more refusal-oriented behavior in the standard setting and retain a larger, although still minority, proportion of Following responses. Gemini-3.1-Pro-Preview and Grok-4.5 are the main exceptions, with the highest Following rates before language enforcement.

Under Force, off-topic generation becomes the dominant behavior for most models. This change is especially pronounced for Grok-4.5, while Claude-Opus-5 also shifts from refusal-dominated to off-topic-dominated responses. Gemini-3.1-Pro-Preview and Qwen3.7-Plus are the main exceptions: refusal remains their largest category, although both also show substantially more off-topic responses than in the standard condition.

Summary. The conditional analysis reveals that target-language compliance and jailbreak compliance are distinct behaviors. Cantonese target-language responses are predominantly refusals, whereas Tibetan responses frequently become off-topic, particularly under explicit language enforcement. Consequently, a high TLRR should not be interpreted as evidence that the model successfully understands or follows the image-embedded instruction. This distinction is especially important for Tibetan, where models can generate linguistically valid target-language responses while failing to remain aligned with the requested task.

G Dataset Construction and Language Validation

This appendix provides additional details on the construction and validation of the multilingual datasets used in our experiments.

G.1 Jailbreak Dataset

Our jailbreak evaluation is based on AdvBench. The English and Mandarin conditions use the corresponding English and Chinese versions of the AdvBench samples. The Cantonese and Tibetan versions are translations of the same AdvBench instances and were provided by an external laboratory. Using aligned samples across languages allows us to compare model behavior under different language conditions while keeping the underlying jailbreak content consistent.

The Tibetan translations were reviewed by a Tibetan-native co-author to verify linguistic correctness and semantic consistency with the source instances. The Cantonese translations were independently reviewed by two Cantonese-native co-authors.

G.2 Over-Safety Dataset

For the over-safety evaluation, we construct a multilingual benchmark from two existing over-refusal resources. We sample 250 instances from the OR-Bench Hard 1K set and combine them with XSTest. The resulting English examples are translated into Mandarin, Cantonese, and Tibetan so that the same benign safety-sensitive requests can be evaluated across languages.

As with the jailbreak data, all Tibetan translations are verified by a Tibetan-native co-author, while the Cantonese translations are checked by two Cantonese-native co-authors. This manual validation is intended to reduce translation artifacts that could otherwise confound comparisons of over-refusal behavior across languages.

G.3 Neutral Multilingual Control

To evaluate whether model behavior is specific to safety-related instructions or instead reflects more general multilingual image-text understanding, we construct a neutral control using FLORES+ (**?**). We select corresponding sentences in English, Mandarin, Cantonese, and Tibetan and use them as semantically neutral image-embedded text. Because FLORES+ provides aligned multilingual content, these examples offer a controlled comparison of visual-text recognition and reproduction across the four languages.

Table 17: Image rendering configuration used for image-embedded instructions. Image-level parameters are kept fixed across languages, while fonts and line-wrapping rules are adapted to the corresponding scripts.

Parameter	Setting
Image width	900 pixels
Font size	36
Padding	50 pixels
Line spacing	16 pixels
Background	RGB $(255, 255, 255)$
Text color	RGB $(30, 30, 30)$
English/Mandarin/Cantonese font	Noto Sans CJK
Tibetan font	Noto Serif Tibetan
Tibetan wrapping	Syllable/sentence-aware

G.4 Image Rendering Configuration and Validation

Rendering configuration. For the image-embedded instruction setting, we render each textual instruction onto a white background using a fixed image width of 900 pixels, a font size of 36, 50-pixel padding, and 16-pixel line spacing. The text color is set to RGB $(30, 30, 30)$. We keep these image-level rendering parameters fixed across languages, while using language-appropriate fonts and line-wrapping rules.

For English, Mandarin, and Cantonese, we use Noto Sans CJK. For Tibetan, we use Noto Serif Tibetan, which provides dedicated support for Tibetan glyphs. In addition, rather than arbitrarily splitting Tibetan strings at the character level, we perform line wrapping at Tibetan syllable and sentence boundaries, using the syllable delimiter and the sentence delimiter . This preserves the orthographic units of the original Tibetan instructions during image construction.

Table 17 summarizes the main rendering configuration.

Validation of Tibetan image rendering. Because our analysis identifies substantially lower visual-instruction recovery for Tibetan, we additionally verify that this result is not caused by corruption during image construction. We manually inspect 50 rendered Tibetan instructions and compare the textual content in each image against its corresponding source instruction. All 50 samples (50/50, 100%) faithfully preserve the complete source instruction, with no missing or corrupted textual content observed during rendering. This validation suggests that the low Tibetan instruction-recovery performance is not attributable to missing or corrupted instruction content introduced by the image-rendering pipeline.