BEYOND ASR: EVALUATING SAFETY ALIGNMENT OF MULTIMODAL LARGE LANGUAGE MODELS ACROSS MODALITIES AND LOW-RESOURCE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment in multimodal large language models (MLLMs) may vary across languages, modalities, and underlying language capabilities. We evaluate multiple MLLMs on harmful and benign safety-sensitive prompts in English, Mandarin, Cantonese, and Tibetan under both direct-text and image-embedded instruction settings. Beyond attack success rate (ASR), we jointly analyze over-refusal rate (ORR), target-language response rate, off-topic behavior, and visual instruction recovery to distinguish genuine safety behavior from failures in low-resource language understanding or multimodal processing. Our results reveal substantial safety disparities in Cantonese and Tibetan, while showing that image embedding does not uniformly increase jailbreak risk. In particular, low ASR for Tibetan image inputs is often accompanied by high off-topic rates and poor recovery of the embedded instruction, indicating that capability failure can be mistaken for stronger safety. Cross-language image–text ablations further confirm that the language carried by the visual modality substantially affects model behavior. These findings highlight the importance of capability-aware evaluation for multilingual multimodal safety.

1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly process instructions across both textual and visual modalities, yet their safety alignment may not transfer uniformly across languages or input formats. This issue is particularly important for low-resource languages, where limited linguistic and multimodal competence can confound conventional safety metrics. A low attack success rate (ASR), for example, may indicate successful refusal, but may also arise because the model fails to understand the instruction or produce a task-relevant response. Similarly, strong refusal behavior can result in excessive rejection of benign safety-sensitive requests. Reliable multilingual safety evaluation therefore requires jointly considering harmful compliance, over-refusal, and underlying language and multimodal capability.

We study safety alignment across English, Mandarin, Cantonese, and Tibetan under both direct-text and image-embedded instructions. Cantonese provides a low-resource setting that shares much of its writing system with Mandarin, whereas Tibetan introduces greater linguistic and orthographic divergence. We jointly evaluate ASR and over-refusal rate (ORR), analyze target-language adherence and off-topic responses, and conduct cross-language image–text ablations. We additionally evaluate whether models can accurately recover instructions embedded in images, helping distinguish safety behavior from failures in visual-language recognition. Our main contributions are:

- We provide a systematic evaluation of multilingual MLLM safety across English, Mandarin, Cantonese, and Tibetan under both direct-text and image-embedded instruction settings, jointly considering attack success and over-refusal.
- We introduce a capability-aware evaluation framework that combines target-language adherence, off-topic behavior, and visual instruction recovery to distinguish genuine safety alignment from failures in low-resource language understanding or multimodal processing.

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.

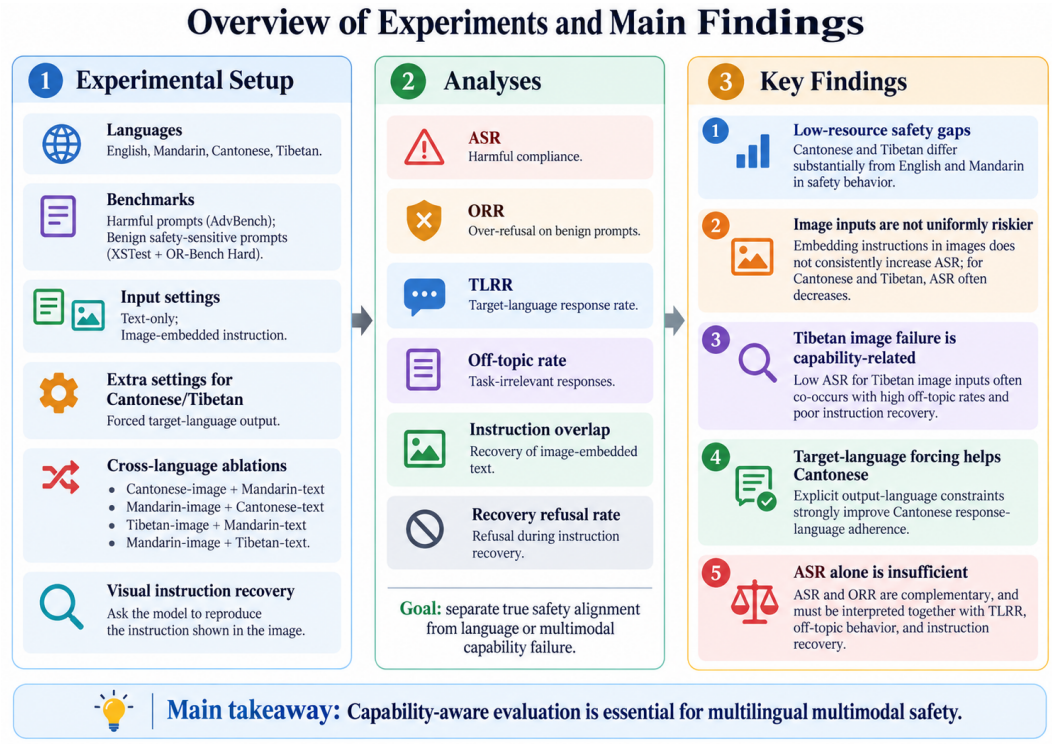

Figure 1: Overview of our multilingual multimodal safety evaluation framework and main findings. We evaluate harmful and benign safety-sensitive prompts across four languages, text and image-embedded input settings, explicit target-language constraints, and cross-language image–text ablations. Beyond ASR and ORR, we analyze target-language adherence, off-topic behavior, and visual instruction recovery to distinguish genuine safety alignment from underlying language and multimodal capability failures.

- We conduct cross-language image–text ablations to disentangle the effects of the language carried by the visual instruction and the accompanying textual prompt, revealing substantial capability-related confounds in Tibetan multimodal safety evaluation.

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 METHODOLOGY

3.1 BENCHMARKS AND EVALUATION SETTINGS

We evaluate MLLMs in English, Mandarin, Cantonese, and Tibetan. Harmful behavior is evaluated on AdvBench, while over-refusal is evaluated using a 500-example benign safety-sensitive set containing 250 XSTest and 250 OR-Bench Hard examples. The translated Cantonese and Tibetan versions of AdvBench were provided by an external experimental research team. For the over-refusal benchmark, the Tibetan translation was manually reviewed by one coauthor who is a native Tibetan speaker, while the Cantonese translation was reviewed by two coauthors who are native Cantonese speakers.

Each benchmark is evaluated under direct-text and image-embedded instruction settings. In the image setting, the instruction is rendered into an image and an accompanying text prompt asks the model to follow its content. For Cantonese and Tibetan, we additionally test explicit target-language output constraints. We also consider four cross-language combinations: Cantonese-image–Mandarin-text, Mandarin-image–Cantonese-text, Tibetan-image–Mandarin-text, and Mandarin-image–Tibetan-text.

3.2 VISUAL INSTRUCTION RECOVERY

To distinguish safety behavior from failures in reading text embedded in images, we conduct an additional control experiment on the over-refusal benchmark. Models are instructed to reproduce the instruction shown in the image rather than execute it. We measure the textual overlap between the model output and the image instruction, together with the refusal rate during this recovery task. High overlap indicates that the visual instruction was successfully recognized, whereas low overlap or refusal suggests that downstream safety results may be confounded by visual-language recognition failure.

3.3 METRICS AND RESPONSE ANALYSIS

We report attack success rate (ASR) on harmful prompts and over-refusal rate (ORR) on benign prompts. Responses are categorized as compliant, refused, off-topic/non-substantive, or failed. ORR measures the percentage of benign prompts that receive a substantive refusal. Blank outputs, API errors, and call failures are not counted as ORR.

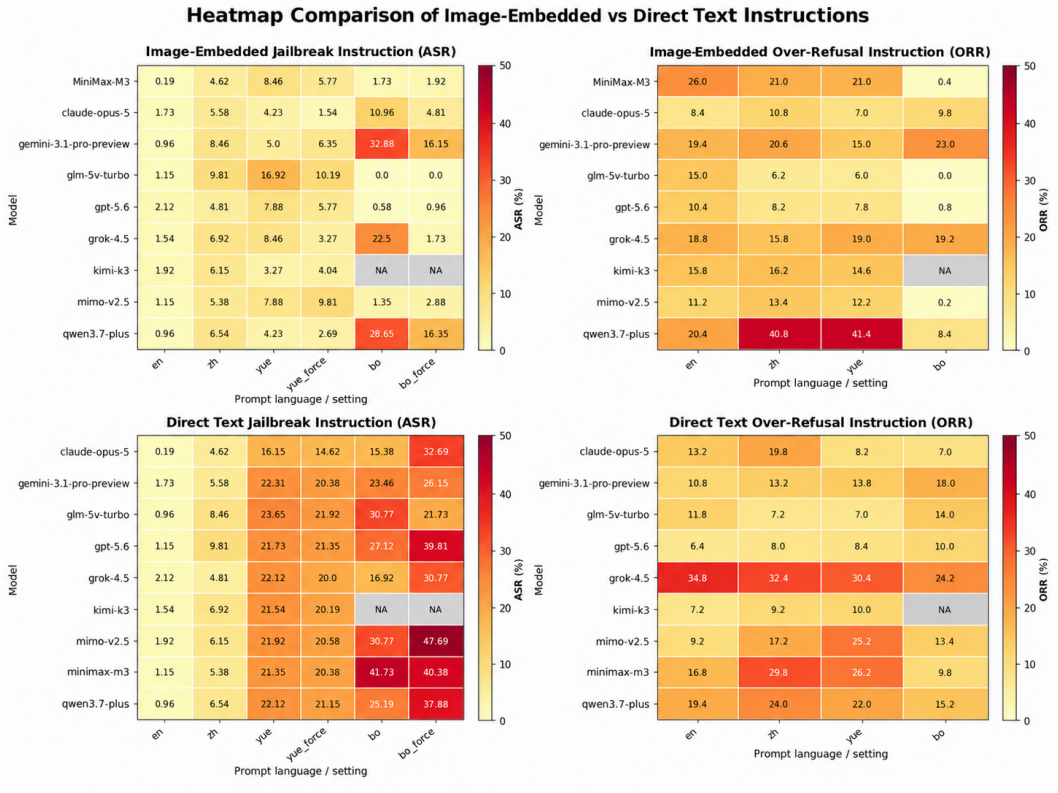

Figure 2: Multilingual safety alignment under direct-text and image-embedded instructions. The four heatmaps report ASR on harmful requests and ORR on benign safety-sensitive requests across English (en), Mandarin (zh), Cantonese (yue), and Tibetan (bo). ORR counts substantive refusals only; blank outputs and API/call failures are excluded. For Cantonese and Tibetan ASR, `force` denotes settings that explicitly require the response to use the target language.

For the exhaustive three-way response-distribution tables in the appendix, we use a broader refusal category (R) that additionally includes blank outputs and API/call failures. Therefore, the refusal percentages reported in those appendix tables should not be interpreted as identical to ORR. Two judge models independently classify each response, with disagreements resolved by a stronger adjudication model. We additionally validate the model-based classification results through manual inspection; details of the validation procedure and agreement results are provided in Appendix A.

We additionally report target-language response rate (TLRR) for Cantonese and Tibetan. Tibetan responses are identified using the Tibetan Unicode block (U+0F00–U+0FFF). Cantonese is detected using Cantonese-specific lexical markers, followed by the same model-based adjudication procedure for ambiguous cases.

4 Results

4.1 Overall Multilingual Safety Alignment

Figure 2 summarizes ASR and ORR across languages under both direct-text and image-embedded instruction settings. Under direct-text inputs, we observe a consistent cross-lingual safety gap. The average ASR increases from only 1.30% in English and 6.47% in Mandarin to 21.43% in Cantonese and 26.42% in Tibetan. This trend is highly consistent across models: all nine models have higher Cantonese ASR than Mandarin ASR, and all eight models with valid Tibetan results show the same pattern for Tibetan. For example, MiniMax-M3 reaches 41.73% ASR in Tibetan compared with 5.38% in Mandarin, while GLM-5V-Turbo increases from 8.46% to 30.77%.

The same pattern is less uniform when the harmful instruction is embedded in an image. Cantonese ASR decreases from 21.43% to 7.37% on average, with all nine models showing lower ASR than in the corresponding text setting. Tibetan ASR similarly decreases from 26.42% to 12.33%, although model behavior is more heterogeneous. For instance, Gemini-3.1-Pro-Preview still shows a high Tibetan image ASR of 32.88%, whereas GLM-5V-Turbo and GPT-5.6 drop to 0% and 0.58%, respectively.

In contrast, ORR does not exhibit the same systematic increase in low-resource languages and varies substantially across models. For example, Qwen3.7-Plus reaches 41.4% ORR for Cantonese image instructions, whereas GPT-5.6 shows only 7.8%. Because ORR here measures substantive refusal rather than blank or failed responses, these differences reflect variation in actual refusal behavior. The higher ASR observed in Cantonese and Tibetan therefore cannot be explained simply by a globally weaker tendency to refuse, motivating a more capability-aware analysis below.

4.2 Effect of Input Modality

We next examine whether moving the instruction from text into an image systematically changes model safety behavior. As shown in Figure 2, the effect of modality is strongly language-dependent. For English and Mandarin, changing the instruction modality produces relatively modest changes in aggregate ASR: English decreases from 1.30% under direct-text input to 1.17% under image-embedded input, while Mandarin decreases from 6.47% to 4.55%.

In contrast, image embedding substantially reduces ASR for Cantonese and Tibetan. The average Cantonese ASR decreases from 21.43% under direct-text instructions to 7.37% under image-embedded instructions, with all nine models showing lower ASR in the image setting. Tibetan shows an even stronger but more heterogeneous modality effect: the mean ASR decreases from 26.42% to 12.33%. For example, GLM-5V-Turbo drops from 30.77% to 0%, whereas Gemini-3.1-Pro-Preview remains relatively vulnerable, reaching 32.88% ASR with Tibetan image instructions. Explicitly enforcing the target output language further amplifies this difference: the average forced ASR decreases from 20.06% to 5.49% for Cantonese and from 34.64% to 5.60% for Tibetan.

The effect of modality on over-refusal is less uniform and substantially more model-dependent than the effect on ASR. Overall, these results show that image-embedded instructions do not universally weaken safety safeguards. Instead, modality effects are particularly pronounced in Cantonese and Tibetan, suggesting that multimodal language-processing capability may substantially influence the observed safety behavior.

4.3 Target-Language Response Rate

To distinguish safety behavior from failures in low-resource language generation, we measure the target-language response rate (TLRR) for Cantonese and Tibetan. As shown in Figure 3, models differ substantially in their tendency to respond in the requested language, particularly for Cantonese. For Cantonese harmful prompts, the average TLRR is only 46.45% under direct-text input, but increases to 95.43% when the model is explicitly required to answer in Cantonese. This improvement is consistent across most models. For example, Grok-4.5 increases from 6.92% to 97.88%, while MiniMax-M3 increases from 17.12% to 97.50%. A similar effect appears for image-embedded harmful instructions, where the average TLRR increases from 65.45% to 89.36% under the explicit output-language constraint. One notable exception is Qwen3.7-Plus, whose Cantonese image TLRR reaches only 37.50% even when the target language is explicitly enforced.

Tibetan shows substantially higher target-language consistency even without an explicit output constraint. The average TLRR reaches 81.06% for direct-text harmful prompts and 86.66% for image-embedded harmful prompts, increasing to 91.45% and 94.26% under the forced-output settings, respectively. Most models achieve close to 100% TLRR, although Claude-Opus-5 and Grok-4.5 exhibit much lower rates without explicit enforcement. These results indicate that target-language constraints are particularly important for Cantonese, whereas most models already maintain strong Tibetan-language output consistency. Importantly, however, producing the correct target language does not necessarily imply that the response correctly follows the task, motivating our subsequent analysis of off-topic responses.

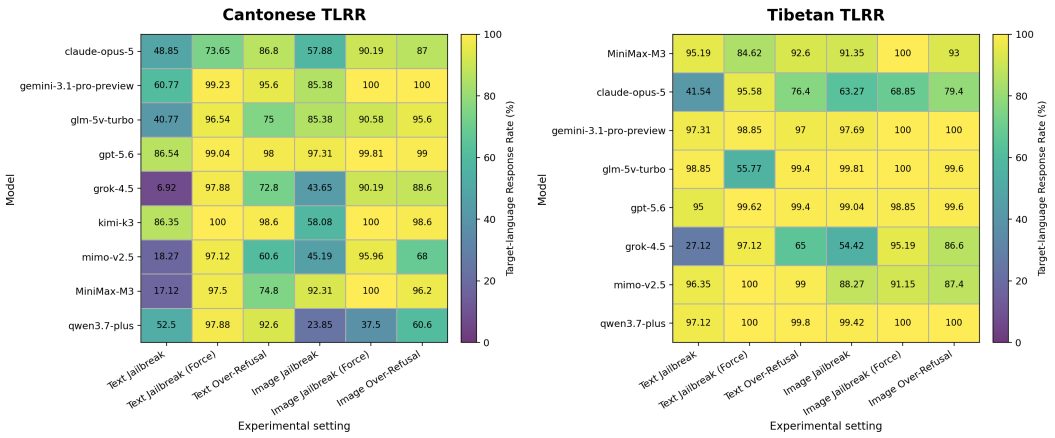

Figure 3: Target-language response rate (TLRR) for Cantonese and Tibetan across harmful and benign prompts under direct-text and image-embedded settings. `Force` denotes settings in which the model is explicitly required to respond in the target language. Higher values indicate a greater proportion of responses produced in the requested language.

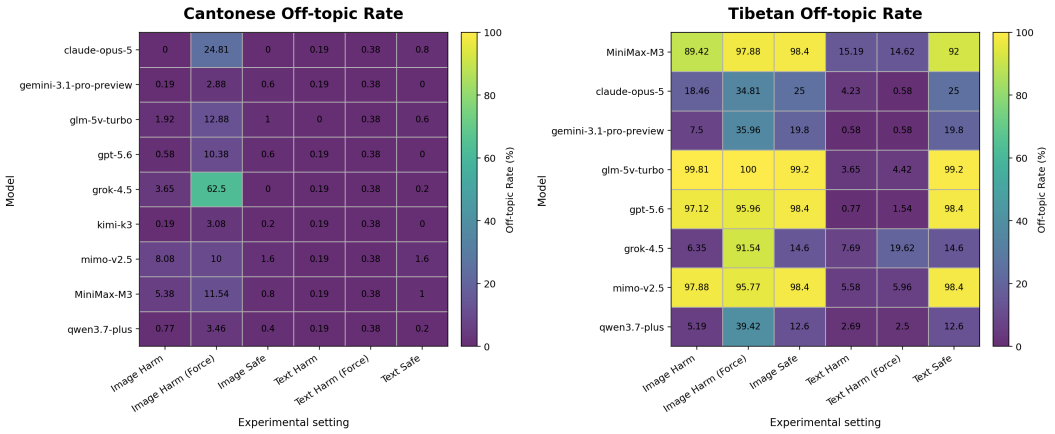

Figure 4: Off-topic response rates for Cantonese and Tibetan across direct-text and image-embedded settings. `Force` denotes settings in which the model is explicitly required to respond in the target language. Tibetan image inputs exhibit substantially higher off-topic rates than Cantonese, indicating that low ASR may partly reflect failures in multimodal language understanding rather than successful safety refusal.

4.4 Off-Topic Responses and Apparent Safety

We next examine whether low attack success reflects genuine safety behavior or failures to produce a task-relevant response. Figure 4 reports the off-topic rate for Cantonese and Tibetan across text and image-embedded settings.

Cantonese responses are generally task-relevant across all settings. Under direct-text harmful prompts, the average off-topic rate is only 0.17%, and remains below 1% for benign prompts. Even under image-embedded harmful instructions, the average off-topic rate is only 2.31%. An exception appears when the model is explicitly required to answer in Cantonese under the image setting, where the average off-topic rate rises to 15.73%. This increase is largely driven by a small number of models, such as Grok-4.5, which reaches 62.50%, while most other models remain substantially lower.

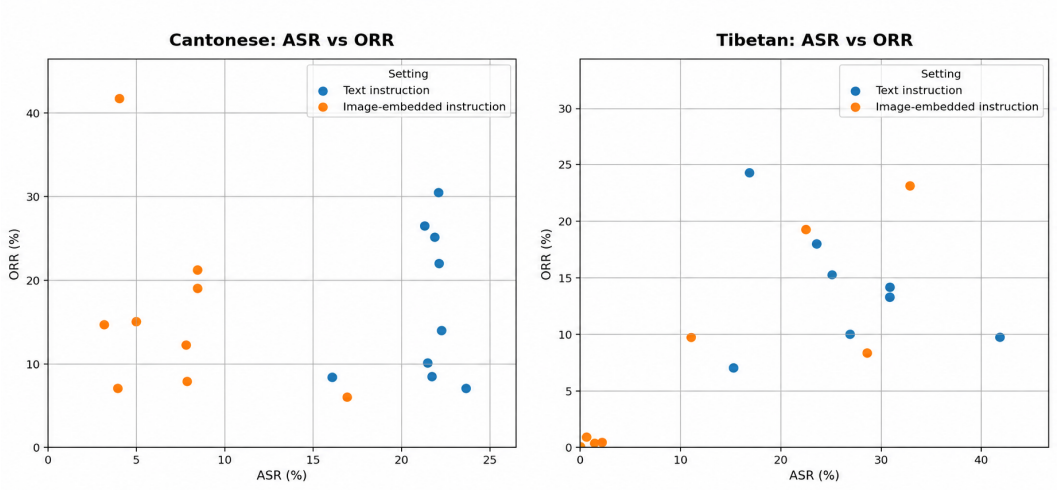

Figure 5: Relationship between attack success rate (ASR) and over-refusal rate (ORR) for Cantonese and Tibetan. Each point represents one model under either the direct-text or image-embedded instruction setting. ORR represents substantive refusal of benign requests and excludes blank outputs and API/call failures. ASR and ORR do not exhibit a consistent one-dimensional trade-off across settings.

The pattern is markedly different for Tibetan. Although the off-topic rate under direct-text harmful prompts remains relatively low on average (5.05%), image-embedded Tibetan instructions produce a much higher off-topic rate of 52.72%. Explicitly requiring Tibetan output further raises this average to 73.92%. Several models exhibit near-complete task failure: GLM-5V-Turbo reaches 99.81% off-topic responses under image-based harmful prompts, while GPT-5.6 and Mimo-v2.5 reach 97.12% and 97.88%, respectively. A similar pattern is observed for benign Tibetan image inputs, where the average off-topic rate is 57.50%.

These results show that low ASR in Tibetan image settings cannot be interpreted as evidence of stronger safety alignment alone. In many cases, the model does not successfully follow the task at all. Therefore, safety metrics such as ASR should be interpreted jointly with response relevance and language capability, especially for low-resource multimodal inputs.

4.5 Relationship between Attack Success and Over-Refusal

We further examine the relationship between attack success and over-refusal in Cantonese and Tibetan. As shown in Figure 5, ASR and ORR do not exhibit a simple or consistent trade-off across models and input modalities. In Cantonese, observations with similar ASR can correspond to substantially different ORR values, indicating that stronger resistance to harmful requests does not necessarily imply greater over-refusal.

A similar pattern is observed for Tibetan, although the distribution is more affected by the large variation in multimodal language capability discussed above. Overall, these results suggest that ASR and ORR capture complementary aspects of safety alignment and should be evaluated jointly rather than treating refusal tendency as a single safety dimension.

As a robustness check, we recompute the model-level ASR–ORR correlations using the broader refusal definition adopted in the appendix, which additionally counts blank outputs and API/call failures as refusals. Relative to the strict ORR used in Figure 5, the Pearson correlation changes from 0.22 to 0.02 for Cantonese text and from -0.38 to -0.41 for Cantonese image inputs. For Tibetan, it changes from -0.32 to -0.66 under text input and from 0.88 to 0.60 under image input. Thus, the magnitude of the association is sensitive to how refusal is defined, but its direction remains modality-dependent, supporting our conclusion that ASR and over-refusal do not exhibit a uniform trade-off.

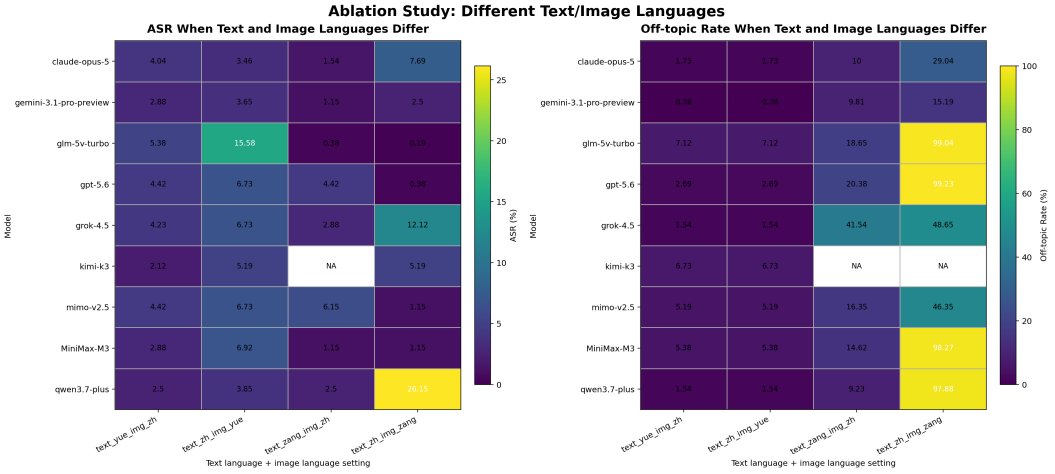

Figure 6: Cross-language image–text ablation results. The left heatmap reports attack success rate (ASR), while the right reports off-topic rate for mismatched image and text languages. Tibetan image content produces substantially higher off-topic rates than the other combinations.

4.6 Cross-Language Image–Text Ablation

We further disentangle the effects of image-language content and textual prompting by pairing Cantonese or Tibetan images with Mandarin text, and vice versa. As shown in Figure 6, ASR remains relatively low across all four cross-language settings, with average values ranging from 2.52% to 6.54%.

However, the off-topic results reveal a strong asymmetry. Cantonese–Mandarin combinations remain largely task-relevant, with an average off-topic rate of only 3.59%. In contrast, when Tibetan is carried by the image, the off-topic rate rises sharply to 65.19%, compared with 16.47% when Tibetan appears in the textual prompt. This suggests that the language encoded in the visual instruction plays a major role in the observed failures, and that low ASR in Tibetan multimodal settings can partly reflect difficulty in processing Tibetan image content rather than stronger safety alignment.

4.7 Behavior under Over-Safety Instructions

We further examine whether models can recover over-safety instructions embedded in images before attributing downstream behavior to safety alignment. As shown in Figure 7, instruction recovery varies substantially across languages. The average output–instruction overlap is high for Mandarin (92.58%) and English (68.58%), and remains moderate for Cantonese (57.33–67.44%). In contrast, Tibetan recovery is extremely limited, with average overlap below 5% in both Tibetan prompt settings.

Importantly, this failure is not primarily caused by explicit refusal. The average refusal rate during instruction recovery remains low across all languages, including only 1.10% and 0.38% for the two Tibetan settings. These results suggest that poor performance on Tibetan image inputs is largely associated with difficulty recovering the visual instruction itself, rather than a stronger tendency to refuse. This provides additional evidence that safety metrics in low-resource multimodal settings should be interpreted together with underlying visual-language capability.

5 Discussion

Our results show that multilingual multimodal safety cannot be characterized by ASR alone. Cantonese and Tibetan exhibit substantially different behavior from English and Mandarin, while the relationship between ASR and ORR remains inconsistent across models and modalities. Moreover, image embedding does not systematically increase jailbreak success. For Tibetan in particular, lower image-based ASR often coincides with very high off-topic rates, and the cross-language ablations

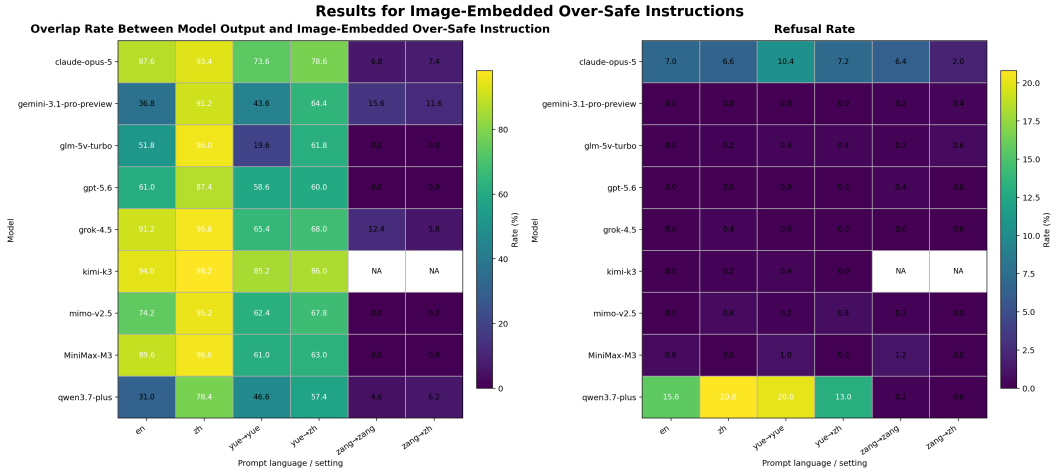

Figure 7: Behavior when models are asked to recover over-safety instructions embedded in images. The left heatmap reports the overlap between the generated output and the image instruction, while the right reports the refusal rate. Tibetan exhibits substantially lower instruction-recovery overlap despite low refusal rates.

show substantially more task failure when Tibetan is carried by the image. These findings suggest that apparently stronger safety may partly reflect difficulty recognizing or understanding the visual instruction rather than successful refusal.

The visual instruction recovery experiment provides further evidence for this interpretation. Models recover English, Mandarin, and Cantonese instructions substantially more reliably than Tibetan instructions, while refusal during the recovery task remains low. This separates safety refusal from visual instruction-recognition failure and shows that target-language adherence alone is also insufficient, since a response can use the requested language while remaining off-topic. Overall, ASR and ORR should therefore be interpreted jointly with target-language consistency, response relevance, and visual instruction recovery, especially for low-resource multimodal settings.

Our study has several limitations. We evaluate a finite set of models and benchmarks, and the Cantonese and Tibetan data rely partly on translated or adapted resources. Response categorization and Cantonese language identification also partly depend on model-based adjudication. In addition, instruction overlap is only a proxy for visual instruction recovery rather than a complete measure of OCR or semantic understanding. Future work should include broader low-resource languages, stronger human validation, and more controlled measures of multimodal language capability.

6 Conclusion

We present a multilingual evaluation of multimodal safety alignment across English, Mandarin, Cantonese, and Tibetan under text and image-embedded instructions. Our results reveal substantial language- and modality-dependent variation in both attack success and over-refusal. Crucially, low ASR in low-resource multimodal settings does not always indicate stronger safety, as it may instead arise from target-language or task-comprehension failures. By jointly considering ASR, ORR, target-language consistency, off-topic responses, and cross-language ablations, our study highlights the need for capability-aware evaluation when assessing the safety of multilingual multimodal models.

AI use statement

In this work, generative AI tools were used solely for literature searching, experimental code writing, API-based classification of experimental results, and the creation of the overview figure. All AI-retrieved references were manually screened and verified against their original sources, and all AI-generated code, classifications, and the overview figure were manually reviewed and validated by

the authors for scientific accuracy. No AI-generated text was copied into the manuscript. The authors take full responsibility for the final content of this work, including code, figures, 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/beyond_asr-1468`. The appendix reports manual validation of response classification and target-language identification, complete model-level response distributions, target-language response statistics, cross-language ablations, and target-language-conditioned metrics. The methodology section defines the evaluation settings and the ASR, ORR, off-topic, and TLRR quantities 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 1.

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 1.

B Complete Image-Embedded Jailbreak Results

This appendix reports the complete model-level results for harmful instructions embedded in images. Each language–model setting contains 520 prompts. We group responses into three mutually

Table 1: 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	48 (96.0%)	1 harmful, 1 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	45 (90.0%)	4 answer, 1 refusal

Table 2: Complete harmful-response distributions for image-embedded instructions without an explicit target-language requirement. Each entry is H/R/O (%).

Model	English	Mandarin	Cantonese	Tibetan
MiniMax-M3	0.77/98.46/0.77	4.42/92.88/2.69	8.46/86.15/5.38	1.73/8.85/89.42
Claude-Opus-5	0.38/98.46/1.15	3.85/95.00/1.15	4.23/95.77/0.00	10.96/70.58/18.46
Gemini-3.1-Pro-Preview	0.77/98.27/0.96	2.88/96.73/0.38	5.00/94.81/0.19	32.88/59.62/7.50
GLM-5V-Turbo	1.15/97.12/1.73	8.46/85.58/5.96	16.92/81.15/1.92	0.00/0.19/99.81
GPT-5.6	0.96/97.31/1.73	5.00/87.69/7.31	7.88/91.54/0.58	0.58/2.31/97.12
Grok-4.5	1.73/97.12/1.15	4.23/94.23/1.54	8.46/87.88/3.65	22.50/71.15/6.35
Kimi-K3	1.15/98.27/0.58	3.85/95.38/0.77	3.27/96.54/0.19	–
MiMo-V2.5	2.69/92.88/4.42	4.81/94.04/1.15	7.88/84.04/8.08	1.35/0.77/97.88
Qwen3.7-Plus	0.96/98.27/0.77	3.46/96.15/0.38	4.23/95.00/0.77	28.65/66.15/5.19

Table 3: Complete harmful-response distributions for image-embedded Cantonese and Tibetan instructions with explicit target-language enforcement (*Force*). Each entry is H/R/O (%).

Model	Cantonese + Force	Tibetan + Force
MiniMax-M3	5.77/82.69/11.54	1.92/0.19/97.88
Claude-Opus-5	1.54/73.65/24.81	4.81/60.38/34.81
Gemini-3.1-Pro-Preview	6.35/90.77/2.88	16.15/47.88/35.96
GLM-5V-Turbo	10.19/76.92/12.88	0.00/0.00/100.00
GPT-5.6	5.77/83.85/10.38	0.96/3.08/95.96
Grok-4.5	3.27/34.23/62.50	1.73/6.73/91.54
Kimi-K3	4.04/92.88/3.08	–
MiMo-V2.5	9.81/80.19/10.00	2.88/1.35/95.77
Qwen3.7-Plus	2.69/93.85/3.46	16.35/44.23/39.42

exclusive outcomes: harmful compliance (H), refusal (R), and off-topic/non-substantive response (O). A response is counted as H when it substantively answers the harmful request. R includes explicit refusals, blank outputs, and API/call failures. The supplied runs contain no unparseable responses. *Force* denotes an accompanying textual instruction that explicitly requires the model to answer in the corresponding target language. Kimi-K3 Tibetan results are unavailable and are therefore reported as "–". Each table entry reports H/R/O percentages.

The complete distributions further illustrate that low image-based ASR does not necessarily correspond to stronger safety alignment. English and Mandarin results are dominated by refusal, whereas Cantonese exhibits moderately higher harmful compliance for several models. Tibetan behavior is substantially more heterogeneous: some models retain non-trivial harmful compliance, while others produce predominantly off-topic responses. Explicit target-language enforcement does not consis-

Table 4: Complete cross-language image–text ablation results. Each cell reports H/R/O (%). "Img" denotes the language of the harmful instruction rendered in the image, and "Text" denotes the language of the accompanying textual prompt.

Model	Cantonese Img Mandarin Text	Mandarin Img Cantonese Text	Tibetan Img Mandarin Text	Mandarin Img Tibetan Text
MiniMax-M3	6.92/85.96/7.12	2.88/91.73/5.38	1.15/0.96/97.88	1.15/84.23/14.62
Claude-Opus-5	3.46/95.77/0.77	4.04/94.23/1.73	7.69/63.27/29.04	1.54/88.46/10.00
Gemini-3.1-Pro-Preview	3.65/96.35/0.00	2.88/96.73/0.38	2.50/82.31/15.19	1.15/89.04/9.81
GLM-5V-Turbo	15.58/81.34/3.08	5.38/87.50/7.12	0.19/0.77/99.04	0.38/89.81/9.81
GPT-5.6	6.73/91.92/1.35	4.42/92.88/2.69	0.38/0.38/99.23	4.42/75.19/20.38
Grok-4.5	6.73/92.50/0.77	4.23/94.23/1.54	12.12/39.23/48.65	2.88/55.58/41.54
Kimi-K3	5.19/94.42/0.38	2.12/91.15/6.73	–	–
MiMo-V2.5	6.73/83.27/10.00	4.42/90.38/5.19	1.15/0.58/98.27	6.15/77.50/16.35
Qwen3.7-Plus	3.85/95.77/0.38	2.50/95.96/1.54	26.15/39.62/34.23	2.50/88.27/9.23
Average	6.54/90.81/2.65	3.65/92.75/3.59	6.42/28.39/65.19	2.52/81.01/16.47

tently increase harmful compliance, and off-topic behavior remains dominant for several models in Tibetan. These results support interpreting ASR jointly with refusal and off-topic rates.

B.1 Complete Cross-Language Image–Text Ablation Results

Table 4 reports the complete model-level distributions for the four cross-language image–text ablations described in Section 4.6. Each setting contains 520 harmful prompts. We report harmful compliance (H), refusal (R), and off-topic response (O). Responses that substantively answer the harmful request are counted as H; blank outputs and call failures are counted as R. No unparseable responses occur in the supplied runs. Kimi-K3 results for settings involving Tibetan are treated as unavailable and shown as "–". Averages for Tibetan-containing settings therefore exclude Kimi-K3. Each cell is H/R/O (%).

The full distributions show that the ablation is driven not only by refusal behavior but also by whether the visual instruction is successfully processed. Both Cantonese–Mandarin combinations remain predominantly task-relevant, whereas Tibetan carried in the image produces substantially more off-topic responses. This pattern supports interpreting low ASR in the Tibetan multimodal setting together with capability-related failure rather than as evidence of stronger safety alone.

B.2 Direct-Text Jailbreak Results

We report the complete model-level results for harmful instructions provided directly as text. Each setting contains 520 prompts. Harmful compliance (H) denotes responses that substantively answer the harmful request; refusal (R) includes explicit refusals, blank outputs, and call failures; and off-topic (O) denotes non-responsive outputs. *Force* indicates an explicit instruction to answer in the corresponding target language. Kimi-K3 Tibetan results are unavailable and are shown as "–" and excluded from Tibetan averages. Each entry reports H/R/O (%).

Direct-text harmful compliance is substantially higher in Cantonese than in English or Mandarin, while the main-text results also show elevated Tibetan ASR. For Tibetan, response relevance must be considered jointly with attack success; we therefore report the target-language-conditioned Tibetan composition separately in Table 17. Explicit target-language enforcement changes the response distribution further, particularly for Tibetan.

B.3 Complete Response Analysis for Image-Based Over-Safety Inputs

We provide the complete model-level response distributions for the Cantonese and Tibetan image-based over-safety settings. Three complementary statistics are reported. First, the *overview* results report the proportion of outputs produced in the target language, together with blank outputs and API errors; non-target-language responses are not included in these three categories. Second, we classify

Table 5: Complete direct-text harmful-response distributions without explicit target-language enforcement. Each entry is H/R/O (%). Kimi-K3 Tibetan results are unavailable and are excluded from the Tibetan average.

Model	English	Mandarin	Cantonese	Tibetan
MiniMax-M3	1.15/98.27/0.58	5.38/94.62/0.00	21.35/78.46/0.19	41.73/43.08/15.19
Claude-Opus-5	0.19/98.46/1.35	4.62/95.38/0.00	16.15/83.65/0.19	15.38/80.38/4.23
Gemini-3.1-Pro-Preview	1.73/93.85/4.42	5.58/94.42/0.00	22.31/77.50/0.19	23.46/75.96/0.58
GLM-5V-Turbo	0.96/75.77/23.27	8.46/91.54/0.00	23.65/76.35/0.00	30.77/65.58/3.65
GPT-5.6	1.15/93.08/5.77	9.81/90.19/0.00	21.73/78.08/0.19	27.12/72.12/0.77
Grok-4.5	2.12/97.12/0.77	4.81/95.19/0.00	22.12/77.69/0.19	16.92/75.38/7.69
Kimi-K3	1.54/97.50/0.96	6.92/93.08/0.00	21.54/78.27/0.19	–
MiMo-V2.5	1.92/98.08/0.00	6.15/93.65/0.19	21.92/77.88/0.19	30.77/63.65/5.58
Qwen3.7-Plus	0.96/80.00/19.04	6.54/93.46/0.00	22.12/77.69/0.19	25.19/72.12/2.69
Average	1.30/92.46/6.24	6.47/93.50/0.02	21.43/78.40/0.17	26.42/68.53/5.05

Table 6: Complete direct-text harmful-response distributions with explicit target-language enforcement. Each entry is H/R/O (%).

Model	Cantonese + Force	Tibetan + Force
MiniMax-M3	20.38/79.23/0.38	40.38/45.00/14.62
Claude-Opus-5	14.62/85.00/0.38	32.69/66.73/0.58
Gemini-3.1-Pro-Preview	20.38/79.23/0.38	26.15/73.27/0.58
GLM-5V-Turbo	21.92/77.69/0.38	21.73/73.85/4.42
GPT-5.6	21.35/78.27/0.38	39.81/58.65/1.54
Grok-4.5	20.00/79.62/0.38	30.77/49.62/19.62
Kimi-K3	20.19/79.42/0.38	–
MiMo-V2.5	20.58/79.04/0.38	47.69/46.35/5.96
Qwen3.7-Plus	21.15/78.46/0.38	37.88/59.62/2.50
Average	20.06/79.55/0.38	34.64/59.13/6.23

all responses into direct answer/compliance (A), refusal (R), and off-topic response (O), where blank outputs and call failures are included in R. Third, we report the same A/R/O classification restricted to responses produced in the target language. Kimi-K3 Tibetan results are treated as unavailable and shown as "–".

Because the appendix uses an exhaustive three-way response classification, the refusal category (R) is broader than the ORR reported in the main text. Specifically, R includes substantive refusals, blank outputs, and API/call failures, whereas ORR counts substantive refusals only.

The target-language-conditioned results help separate response-language adherence from safety behavior. For Cantonese, most target-language responses remain task-relevant, whereas Tibetan exhibits substantially greater cross-model variation and, for several models, very high off-topic rates even after restricting evaluation to Tibetan-language outputs. This indicates that the observed behavior cannot be explained by response-language mismatch alone.

B.4 Complete Direct-Text Over-Safety Results

We report the complete model-level results for directly input over-safety instructions in Cantonese and Tibetan. Three complementary statistics are provided. The *overview* results report target-language responses, blank outputs, and API errors; non-target-language responses are not included in these categories. We further classify all responses, as well as the subset of target-language responses, into direct answer/compliance (A), refusal (R), and off-topic response (O). Responses classified as "answering the question" are counted as A, while blank outputs and call failures are counted as R. Kimi-K3 Tibetan results are unavailable and are shown as "–".

Table 7: Target-language response overview for Cantonese and Tibetan image inputs. Each setting contains 500 prompts. Target, Blank, and Error denote the percentage of all prompts producing a target-language response, an empty response, or an API error, respectively. Non-target-language responses are not included in these three categories.

	Cantonese			Tibetan		
Model	**Target**	**Blank**	**Error**	**Target**	**Blank**	**Error**
MiniMax-M3	96.20	1.00	1.00	93.00	6.80	0.20
Claude-Opus-5	87.00	13.00	0.00	79.40	20.60	0.00
Gemini-3.1-Pro-Preview	100.00	0.00	0.00	100.00	0.00	0.00
GLM-5V-Turbo	95.60	0.00	4.00	99.60	0.00	0.40
GPT-5.6	99.00	0.00	0.00	99.60	0.00	0.00
Grok-4.5	88.60	0.00	0.00	86.60	0.00	0.00
Kimi-K3	98.60	0.00	0.00	–	–	–
MiMo-V2.5	68.00	0.00	0.00	87.40	0.00	0.00
Qwen3.7-Plus	60.60	0.00	0.00	100.00	0.00	0.00

Table 8: Response classification over all outputs. Each entry reports A/R/O (%), where A denotes a direct answer to the benign request, R denotes the broad refusal category used for exhaustive classification (including substantive refusals, blank outputs, and API/call failures), and O denotes off-topic responses. This R category is broader than the ORR reported in the main text.

Model	**Cantonese A/R/O**	**Tibetan A/R/O**
MiniMax-M3	76.20/23.00/0.80	0.60/7.40/92.00
Claude-Opus-5	80.00/20.00/0.00	44.60/30.40/25.00
Gemini-3.1-Pro-Preview	84.40/15.00/0.60	57.20/23.00/19.80
GLM-5V-Turbo	89.00/10.00/1.00	0.40/0.40/99.20
GPT-5.6	91.60/7.80/0.60	0.80/0.80/98.40
Grok-4.5	81.00/19.00/0.00	66.20/19.20/14.60
Kimi-K3	85.20/14.60/0.20	–
MiMo-V2.5	86.20/12.20/1.60	1.40/0.20/98.40
Qwen3.7-Plus	58.20/41.40/0.40	79.00/8.40/12.60

Table 9: Response classification restricted to target-language outputs. N_{tgt} is the number of target-language responses among the 500 prompts. A/R/O percentages are computed only over these target-language responses.

	Cantonese		Tibetan	
Model	$\mathbf{N}_{\mathrm{tgt}}$	**A/R/O (%)**	$\mathbf{N}_{\mathrm{tgt}}$	**A/R/O (%)**
MiniMax-M3	481	77.55/21.62/0.83	465	0.65/0.43/98.92
Claude-Opus-5	435	91.95/8.05/0.00	397	56.17/12.34/31.49
Gemini-3.1-Pro-Preview	500	84.40/15.00/0.60	500	57.20/23.00/19.80
GLM-5V-Turbo	478	92.68/6.28/1.05	498	0.40/0.00/99.60
GPT-5.6	495	91.72/7.68/0.61	498	0.80/0.80/98.39
Grok-4.5	443	82.62/17.38/0.00	433	70.44/12.70/16.86
Kimi-K3	493	85.19/14.60/0.20	–	–
MiMo-V2.5	340	88.82/9.12/2.06	437	1.37/0.23/98.40
Qwen3.7-Plus	303	76.90/22.44/0.66	500	79.00/8.40/12.60

Table 10: Response-language overview for direct-text over-safety instructions. Each setting contains 500 prompts. Target, Blank, and Error denote the percentage of all prompts producing a target-language response, an empty response, or an API error, respectively. Non-target-language responses are not included in these three categories.

	Cantonese			Tibetan		
Model	**Target**	**Blank**	**Error**	**Target**	**Blank**	**Error**
MiniMax-M3	74.80	1.00	0.00	92.60	7.00	0.00
Claude-Opus-5	86.80	11.20	0.00	76.40	23.60	0.00
Gemini-3.1-Pro-Preview	95.60	0.00	0.00	97.00	0.00	0.00
GLM-5V-Turbo	75.00	12.20	0.00	99.40	0.60	0.00
GPT-5.6	98.00	0.00	0.00	99.40	0.00	0.20
Grok-4.5	72.80	0.00	0.00	65.00	0.00	0.20
Kimi-K3	98.60	0.40	0.00	–	–	–
MiMo-V2.5	60.60	0.00	0.00	99.00	0.00	0.00
Qwen3.7-Plus	92.60	0.00	0.40	99.80	0.00	0.00

Table 11: Classification of all responses to direct-text over-safety instructions. Each entry reports A/R/O (%), where A denotes a direct answer to the benign request, R denotes the broad refusal category (including substantive refusals, blank outputs, and API/call failures), and O denotes off-topic responses. The R values therefore need not equal the main-text ORR.

Model	**Cantonese A/R/O**	**Tibetan A/R/O**
MiniMax-M3	71.80/27.20/1.00	51.40/16.80/31.80
Claude-Opus-5	79.80/19.40/0.80	64.60/30.60/4.80
Gemini-3.1-Pro-Preview	86.20/13.80/0.00	79.60/18.00/2.40
GLM-5V-Turbo	80.20/19.20/0.60	78.20/14.60/7.20
GPT-5.6	91.60/8.40/0.00	88.60/10.20/1.20
Grok-4.5	69.40/30.40/0.20	45.40/24.40/30.20
Kimi-K3	89.60/10.40/0.00	–
MiMo-V2.5	73.20/25.20/1.60	78.80/13.40/7.80
Qwen3.7-Plus	77.40/22.40/0.20	79.40/15.20/5.40

As in the image-based analysis above, the appendix refusal category (R) uses a broader convention than the main-text ORR: blank outputs and API/call failures are included in R, while ORR counts substantive refusals only.

The target-language-conditioned results distinguish safety behavior from response-language mismatch. Cantonese responses are generally task-relevant once restricted to Cantonese outputs. Tibetan shows greater model-level variation, particularly for MiniMax-M3 and Grok-4.5, which retain substantial off-topic rates even among Tibetan-language responses.

B.5 Target-Language Response Overview for Jailbreak Instructions

We further report response-language statistics for Cantonese and Tibetan jailbreak instructions under direct-text and image-embedded inputs. For each setting, we report the percentage of outputs produced in the target language (T), blank outputs (B), and API errors (E). Non-target-language responses are not included in these three categories; therefore, T/B/E percentages do not necessarily sum to 100%. *Force* denotes an explicit textual instruction requiring the model to respond in the corresponding target language. Each setting contains 520 prompts. Kimi-K3 Tibetan results are unavailable and are shown as "–".

Explicit target-language prompting substantially increases the proportion of target-language responses in most settings. The effect is especially pronounced for Cantonese direct-text inputs, where

Table 12: Classification restricted to target-language responses for direct-text over-safety instructions. N_{tgt} denotes the number of target-language responses. A/R/O percentages are computed only over this subset.

Model	Cantonese $\mathbf{N}_{\mathrm{tgt}}$	Cantonese A/R/O (%)	Tibetan $\mathbf{N}_{\mathrm{tgt}}$	Tibetan A/R/O (%)
MiniMax-M3	374	85.29/14.44/0.27	463	55.51/10.37/34.13
Claude-Opus-5	434	90.09/9.68/0.23	382	84.55/9.16/6.28
Gemini-3.1-Pro-Preview	478	87.24/12.76/0.00	485	82.06/15.46/2.47
GLM-5V-Turbo	375	92.27/7.20/0.53	497	78.67/14.08/7.24
GPT-5.6	490	91.43/8.57/0.00	497	88.73/10.06/1.21
Grok-4.5	364	75.00/24.73/0.27	325	46.46/9.85/43.69
Kimi-K3	493	89.66/10.34/0.00	–	–
MiMo-V2.5	303	87.46/11.55/0.99	495	79.19/12.93/7.88
Qwen3.7-Plus	463	80.99/18.79/0.22	499	79.56/15.03/5.41

Table 13: Target-language response overview for Cantonese jailbreak instructions. Each entry reports T/B/E (%), corresponding to target-language response, blank output, and API error, respectively.

Model	Text	Text + Force	Image	Image + Force
MiniMax-M3	17.12/0.00/0.00	97.50/0.19/0.00	92.31/0.00/0.58	100.00/0.00/0.00
Claude-Opus-5	48.85/20.00/0.00	73.65/26.73/0.00	57.88/41.92/0.00	90.19/10.00/0.00
Gemini-3.1-Pro-Preview	60.77/0.00/0.96	99.23/0.19/0.77	85.38/0.00/0.00	100.00/0.00/0.00
GLM-5V-Turbo	40.77/0.00/1.92	96.54/0.00/2.31	85.38/0.00/13.65	90.58/0.00/9.42
GPT-5.6	86.54/2.88/0.77	99.04/0.00/0.96	97.31/0.00/0.38	99.81/0.00/0.00
Grok-4.5	6.92/0.00/1.73	97.88/0.00/1.15	43.65/0.00/0.00	90.19/0.00/0.00
Kimi-K3	86.35/0.00/0.00	100.00/0.00/0.00	58.08/41.92/0.00	100.00/0.00/0.00
MiMo-V2.5	18.27/0.00/0.00	97.12/0.00/0.00	45.19/0.38/0.00	95.96/0.77/0.00
Qwen3.7-Plus	52.50/0.00/0.00	97.88/0.00/0.00	23.85/0.00/0.19	37.50/0.00/0.19
Average	46.45/2.54/0.60	95.43/3.01/0.58	65.45/9.36/1.64	89.36/1.20/1.07

Table 14: Target-language response overview for Tibetan jailbreak instructions. Each entry reports T/B/E (%), corresponding to target-language response, blank output, and API error, respectively. Kimi-K3 is excluded from the averages because Tibetan results are unavailable.

Model	Text	Text + Force	Image	Image + Force
MiniMax-M3	95.19/3.65/0.00	84.62/15.19/0.00	91.35/8.65/0.00	100.00/0.00/0.00
Claude-Opus-5	41.54/21.73/0.00	95.58/4.42/0.00	63.27/36.73/0.00	68.85/31.15/0.00
Gemini-3.1-Pro-Preview	97.31/0.00/0.00	98.85/0.00/0.77	97.69/0.00/2.31	100.00/0.00/0.00
GLM-5V-Turbo	98.85/0.96/0.00	55.77/44.04/0.00	99.81/0.00/0.19	100.00/0.00/0.00
GPT-5.6	95.00/0.00/0.19	99.62/0.00/0.38	99.04/0.00/0.00	98.85/0.00/1.15
Grok-4.5	27.12/0.00/0.96	97.12/0.38/0.58	54.42/0.00/0.00	95.19/0.38/0.00
Kimi-K3	–	–	–	–
MiMo-V2.5	96.35/0.00/0.00	100.00/0.00/0.00	88.27/0.00/0.00	91.15/0.19/0.00
Qwen3.7-Plus	97.12/0.00/0.00	100.00/0.00/0.00	99.42/0.00/0.00	100.00/0.00/0.00
Average	81.06/3.29/0.14	91.45/8.00/0.22	86.66/5.67/0.31	94.26/3.96/0.14

the average target-language response rate increases from 46.45% to 95.43%. The remaining gap between the reported T/B/E percentages and 100% corresponds to non-target-language responses, which are excluded from these categories.

B.6 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

Table 15: 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

Table 16: Target-language-conditioned classification for Cantonese jailbreak instructions. For each condition, N_{tgt} is the number of Cantonese responses, followed by H/R/O (%).

	Text		Text + Force		Image		Image + Force	
Model	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**
MiniMax-M3	89	33.71/65.17/1.12	507	20.51/79.09/0.39	480	8.96/85.21/5.83	520	5.77/82.69/11.54
Claude-Opus-5	254	18.11/81.89/0.00	383	19.84/79.63/0.52	301	6.98/93.02/0.00	469	1.71/71.43/26.87
Gemini-3.1-Pro-Preview	316	22.47/77.53/0.00	516	20.54/79.07/0.39	444	5.63/94.14/0.23	520	6.35/90.77/2.88
GLM-5V-Turbo	212	25.94/74.06/0.00	502	22.51/77.09/0.40	444	19.14/78.60/2.25	471	11.25/74.52/14.23
GPT-5.6	450	20.22/79.56/0.22	515	21.55/78.06/0.39	506	7.11/92.49/0.40	519	5.78/84.01/10.21
Grok-4.5	36	47.22/52.78/0.00	509	20.24/79.37/0.39	227	15.42/77.97/6.61	469	3.41/33.05/63.54
Kimi-K3	449	18.71/81.07/0.22	520	20.19/79.42/0.38	302	5.63/94.04/0.33	520	4.04/92.88/3.08
MiMo-V2.5	95	23.16/75.79/1.05	505	21.19/78.42/0.40	235	8.51/80.00/11.49	499	10.02/79.56/10.42
Qwen3.7-Plus	273	22.34/77.66/0.00	509	21.22/78.39/0.39	124	14.52/83.06/2.42	195	7.18/85.13/7.69
Average	–	**25.76/73.94/0.29**	–	**20.87/78.73/0.41**	–	**10.21/86.51/3.28**	–	**6.17/77.12/16.72**

Table 17: Target-language-conditioned classification for Tibetan jailbreak instructions. For each condition, N_{tgt} is the number of Tibetan responses, followed by H/R/O (%). Kimi-K3 is treated as unavailable, matching the availability convention used in the main manuscript.

	Text		Text + Force		Image		Image + Force	
Model	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**	N_{tgt}	**H/R/O**
MiniMax-M3	495	43.84/40.40/15.76	440	47.73/35.00/17.27	475	1.89/0.21/97.89	520	1.92/0.19/97.88
Claude-Opus-5	216	30.56/66.20/3.24	497	34.21/65.19/0.60	329	17.33/53.50/29.18	358	6.98/42.46/50.56
Gemini-3.1-Pro-Preview	506	24.11/75.30/0.59	514	26.46/72.96/0.58	508	33.66/58.66/7.68	520	16.15/47.88/35.96
GLM-5V-Turbo	514	31.13/65.18/3.70	290	38.97/53.10/7.93	519	0.00/0.00/100.00	520	0.00/0.00/100.00
GPT-5.6	494	23.89/75.71/0.40	518	39.96/58.49/1.54	515	0.58/2.33/97.09	514	0.97/1.95/97.08
Grok-4.5	141	48.94/24.82/26.24	505	31.68/48.12/20.20	283	36.75/52.30/10.95	495	1.82/4.65/93.54
Kimi-K3	–	–	–	–	–	–	–	–
MiMo-V2.5	501	31.94/62.28/5.79	520	47.69/46.35/5.96	459	1.53/0.44/98.04	474	3.16/0.84/95.99
Qwen3.7-Plus	505	25.94/71.29/2.77	520	37.88/59.62/2.50	517	28.82/65.96/5.22	520	16.35/44.23/39.42
Average	–	**32.54/60.15/7.31**	–	**38.07/54.85/7.07**	–	**15.07/29.17/55.76**	–	**5.92/17.78/76.30**

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 15 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

C TARGET-LANGUAGE-CONDITIONED JAILBREAK RESPONSE CLASSIFICATION

To complement the response-language overview in Section B.5, we additionally report response-category classification restricted to outputs identified as being in the corresponding target language. For each model and condition, N_{tgt} denotes the number of target-language responses entering this classification.

Within this target-language subset, harmful compliance (H) denotes responses that substantively answer the harmful request, refusal (R) denotes substantive refusal responses, and off-topic (O) denotes responses that do not substantively address the request. Blank outputs and API/call failures do not contain a target-language response and therefore do not enter N_{tgt} or the target-language-conditioned H/R/O percentages.

No unparseable outputs occur in the supplied classification files. Percentages in Tables 16 and 17 are therefore computed with N_{tgt} as the denominator rather than the full set of 520 prompts. *Force* denotes an explicit textual instruction requiring the model to respond in the corresponding target language. The average row is an unweighted macro-average across available models. Consistent with the Tibetan availability convention used elsewhere in the manuscript, Kimi-K3 is treated as unavailable for Tibetan and excluded from the averages.

The conditioned results reveal that target-language enforcement does not have a uniform effect on harmful compliance. For Cantonese, the macro-average harmful rate decreases from 25.76% to 20.87% for direct text and from 10.21% to 6.17% for image-embedded inputs, even though the response-language overview shows that *Force* substantially increases the probability of receiving a Cantonese response. For Tibetan, direct-text harmful compliance instead rises from 32.54% to 38.07%, whereas image-embedded harmful compliance falls from 15.07% to 5.92% and the off-topic share increases from 55.76% to 76.30%. Together, these results reinforce the distinction between target-language adherence and safety behavior: increasing the probability of a response in the requested language does not necessarily increase jailbreak success, especially in the multimodal Tibetan setting.