

Cross-Lingual Activation Steering for Visual-Prior Conflicts

Yufei Wang Xiang Lorraine Li Adriana Kovashka

Department of Computer Science, University of Pittsburgh

yuw384@pitt.edu

xianglli@pitt.edu

kovashka@cs.pitt.edu

Abstract

Multimodal large language models (MLLMs) must often arbitrate between visual evidence and world-knowledge priors when an image conflicts with commonsense expectations. Existing work has demonstrated that activation steering vectors can bias model outputs toward either source of information, but it remains unclear whether this arbitration mechanism generalizes across languages. To investigate this, we construct a multilingual extension of the Visual CounterFact color split and derive activation steering directions that contrast image-grounded answers with prior-based answers. Across 12 languages, we find that English-derived and shared multilingual steering directions consistently shift model outputs on non-English prompts, suggesting that visual-prior arbitration has a substantial cross-lingual component. At the same time, we find that performance gains on common knowledge questions can come at the cost of image-grounded accuracy, a trade-off that is particularly pronounced for lower-resource languages. Code and data are available at <https://github.com/elffiona/cross-lingual-visual-prior-steering>.

1 Introduction

Multimodal large language models (MLLMs) must sometimes choose between visual evidence and language or world-knowledge priors. Prior work studies this problem with counterfactual images, where an object attribute in the image conflicts with common knowledge (Golovanevsky et al., 2025; Liu et al., 2025; Ortu et al., 2026). Related work on visual generics and exceptions examines a similar tension between conceptual knowledge and atypical visual instances, but focuses on diagnosing model behavior on typical and exceptional images rather than intervening on visual-prior arbitration or studying its cross-lingual transfer (Frank and Allaway, 2025). For example, a model may see a

blue strawberry but also know that strawberries are usually red. This visual-prior conflict matters for reliability because either source can mislead the model, depending on the question.

Activation steering offers one way to test and control this arbitration. Pixels Versus Priors builds steering vectors that move English MLLM outputs between prior-based and image-grounded answers on Visual CounterFact (Golovanevsky et al., 2025). Recent hallucination-mitigation work also uses hidden-state or attention-head steering in MLLMs (Yin et al., 2026; Su et al., 2025). Other work shows that MLLMs can underuse visual information in scientific diagrams or commonsense conflict settings (Wang and Kovashka, 2025; Liu et al., 2025). This line of work leaves a multilingual transfer gap: whether visual-prior steering directions transfer across prompt languages.

It is unclear whether visual-prior arbitration has a shared structure across languages. Such structure would mean that the same internal direction helps resolve the conflict when the prompt language and answer words change. Text-only work offers related but distinct evidence: refusal directions transfer across languages, and English-derived activation steering can outperform target-language steering in multilingual LLMs (Wang et al., 2025; Schut et al., 2025). We test whether similar transfer appears in MLLMs, where the model must arbitrate between visual evidence and common knowledge.

We ask whether visual-prior steering directions in a multilingual MLLM are language-specific or transferable cross-lingually. We translate the Visual CounterFact color split into 11 other languages and use Qwen3-VL-8B-Instruct (Bai et al., 2025) as our main evaluation setting. We then compare steering directions built from different languages to test whether the same intervention can shift behavior across languages. We additionally evaluate Qwen3-VL-32B-Instruct (Bai et al., 2025) and Gemma-3-4B (Gemma Team, 2025) to examine whether the

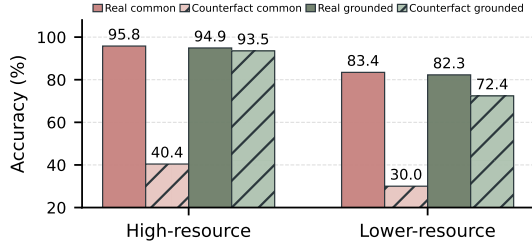


Figure 1: Direct Qwen3-VL-8B accuracy macro-averaged by language-resource group.

observed transfer behavior is specific to the main model setting.

Our contributions include: (1) showing that multilingual counterfactual color questions expose a consistent gap between common-knowledge and image-grounded answers; (2) finding that steering directions from different languages point to a similar internal axis; and (3) both English-derived directions and a shared multilingual direction can shift target-language model behavior, suggesting that this arbitration is not purely language-local.

2 Cross-Lingual Visual-Prior Conflict Evaluation

The color split of Visual CounterFact (Golovanovsky et al., 2025) provides a controlled setting for studying conflicts between visual evidence and common-knowledge priors: color directly changes the visual attribute in conflict without adding comparison or spatial reasoning. Each example pairs a real image with a counterfactual image whose object color conflicts with the object’s typical color. The dataset also features two types of questions, *common knowledge* (“What color are most bananas?”) and *image-grounded* (“What color is *this* banana?”). While the original setting focuses on English, we extend this split to multilingual visual-prior conflicts by translating both questions and answer labels into 11 target languages: Chinese, Japanese, Korean, Spanish, German, French, Arabic, Hindi, Vietnamese, Thai, and Bengali. These languages cover several language families and writing systems. We translate the examples with HY-MT1.5-7B, a recent multilingual machine translation model (Zheng et al., 2025). We run GPT-5.5-assisted quality control to review and correct clear errors (OpenAI, 2026) (e.g., translating the color “orange” as the fruit “orange”). This pass also keeps target objects at comparable granularity across languages, for example by preserving a breed-level object such as “Samoyed” rather than replacing

it with the broader word “dog.” See Appx. A.10. Our quality control process also inspects the translated color terms, including potential differences in how color boundaries are lexicalized across languages. The color split uses basic color categories consistently across common-knowledge and image-grounded questions, and we do not identify cases where such lexical differences make the intended answer label ambiguous or inconsistent. As an additional native-speaker check, we inspect 100 randomly sampled Simplified Chinese examples and do not identify translation errors that change the intended question or answer. Finally, we use a fixed 80/20 random split of base datapoints for steering-vector construction and evaluation.

We first evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) on the full multilingual set under four conditions: real images with common questions, real images with grounded questions, counterfactual images with common questions, and counterfactual images with grounded questions. Fig. 1 groups languages into higher- and lower-resource based on Qwen3-VL’s baseline real-grounded accuracy: “High” averages English, Chinese, Japanese, Korean, Spanish, German, and French; “Low” averages Arabic, Hindi, Vietnamese, Thai, and Bengali. The full per-language table is in Appx. A.2. Consistent with the English-only findings of Visual CounterFact, counterfactual images sharply reduce accuracy on common questions across both groups. Beyond this shared visual-prior conflict, grounded-question accuracy reveals an additional resource-sensitive gap: high-resource languages remain nearly unchanged between real and counterfactual images, whereas lower-resource languages drop from 82.3 to 72.4.

Motivated by the counterfactual-common failures above, we ask whether this arbitration can be shifted cross-lingually. We construct language-specific steering directions that contrast image-grounded and common-knowledge behavior.

3 Steering Vector Construction and Intervention Design

3.1 Direction Construction

As shown in Fig. 2b, we define two steering directions for each language and layer. To compute these directions, we use train-split examples where Qwen3-VL answers both counterfactual-common and grounded image questions correctly. For each eligible example, we query the model with the same

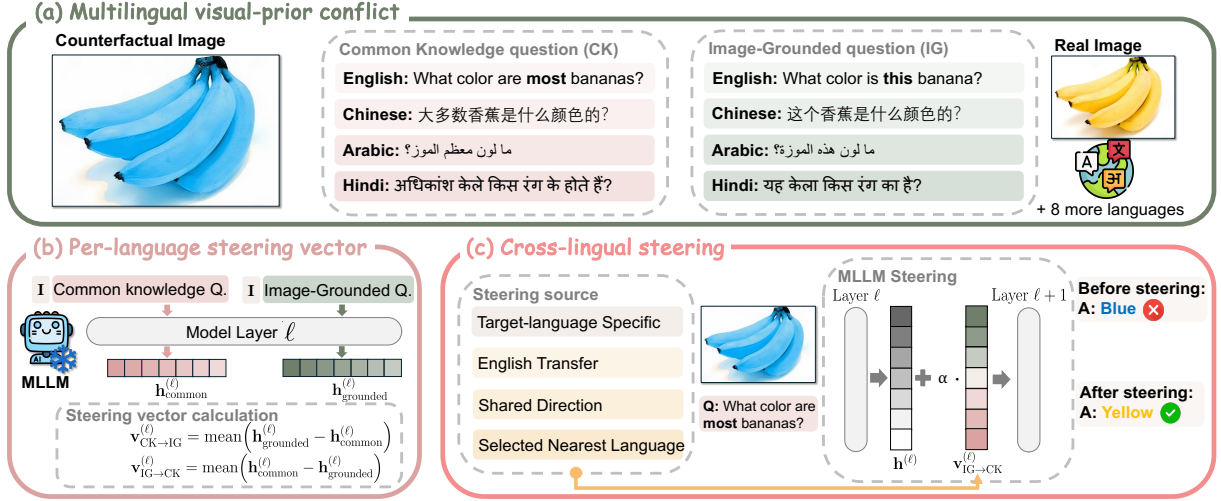


Figure 2: Overview of our cross-lingual steering setup. (a) Multilingual questions expose visual-prior conflicts (English from (Golovanevsky et al., 2025)). (b) Hidden-state differences between CK/IG questions define per-language steering vectors. (c) Different vector sources are applied to steer model behavior.

counterfactual image with the common-knowledge question and the image-grounded question, and extract hidden states from the final question token at layer ℓ . Let $\mathbf{h}_{\text{common}}^{(\ell)}$ and $\mathbf{h}_{\text{grounded}}^{(\ell)}$ denote the corresponding hidden vectors; for simplicity, we omit the language index. We then define:

$$\mathbf{v}_{\text{CK} \rightarrow \text{IG}}^{(\ell)} = \text{mean}(\mathbf{h}_{\text{grounded}}^{(\ell)} - \mathbf{h}_{\text{common}}^{(\ell)}),$$

$$\mathbf{v}_{\text{IG} \rightarrow \text{CK}}^{(\ell)} = \text{mean}(\mathbf{h}_{\text{common}}^{(\ell)} - \mathbf{h}_{\text{grounded}}^{(\ell)}).$$

The direction $\mathbf{v}_{\text{CK} \rightarrow \text{IG}}^{(\ell)}$ steers the model away from common-knowledge priors and toward image-grounded answers. The reverse, $\mathbf{v}_{\text{IG} \rightarrow \text{CK}}^{(\ell)}$, steers the model away from image-grounded answers and toward prior-based answers. This gives one steering vector for each layer, language, and direction.

We consider four activation sources from each transformer layer: full block output, attention output, MLP output, and `resid_pre`. The `resid_pre` source uses the input to each full block, so it is defined from the second layer onward. At inference, we add the steering vector back to the final prompt token during the prefill pass. We use a scalar weight α and only add at selected layers.

3.2 Intervention Layer Selection

Prior work motivates targeted intervention site selection rather than applying steering uniformly across all layers (Meng et al., 2022; Li et al., 2023; Todd et al., 2024; Chuang et al., 2024). We therefore probe layer-wise probability margins between the correct answer and the visually suggested

wrong answer under counterfactual images, aiming to identify layers most associated with arbitration between visual evidence and prior knowledge.

For each language, we select the longest consecutive layer range whose margin exceeds a fixed threshold, excluding final layers where very large margins may reflect output-dominated representations rather than useful steering sites. See Appx. A.6 for margin details and Appx. A.7 for the selected windows. Most languages select overlapping late-middle layers, with Arabic as the main exception. We then use layers 26–32 as a shared intervention window across languages.

4 Steering Experiments

4.1 Experimental Setup

We test whether activation steering shifts Qwen3-VL’s visual-prior arbitration across languages and whether the resulting directions transfer cross-lingually. Our main configuration, selected from ablations (See Appx. A.8), uses `resid_pre` vectors with $\alpha = 0.25$ over layers 26–32, which provides strong steering at a moderate intervention strength over a shared cross-lingual layer window. We apply $\mathbf{v}_{\text{IG} \rightarrow \text{CK}}$ for common-knowledge questions and $\mathbf{v}_{\text{CK} \rightarrow \text{IG}}$ for image-grounded questions.

4.2 Cross-Lingual Steering-Vector Transfer

A central goal of our evaluation is to test whether steering directions transfer across languages. We therefore compare steering-vector sources in this multilingual setup. The **language-specific** setting

Steering source	Common questions				Grounded questions			
	High		Low		High		Low	
	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.
Unsteered base	40.0	–	29.1	–	94.7	–	72.9	–
Lang-specific	61.9	66.8	42.8	51.7	86.0	95.5	60.0	75.4
English vector	68.7	73.2	50.1	57.6	86.4	96.0	48.7	75.2
Shared-SVD	64.8	69.7	47.5	54.5	88.0	95.4	54.9	75.2
LOO-SVD	65.2	70.0	46.9	54.3	88.0	95.2	54.5	75.2

Table 1: Main steering results on counterfactual images. Acc.=steer-all accuracy; Rec.=diagnostic recovery-only accuracy. High=higher-resource languages; Low=lower-resource languages.

uses a vector constructed from the target language’s own training examples. The **English-transfer** setting applies the English vector to non-English prompts, testing whether an English-derived arbitration direction aligns with other languages.

We also construct a **Shared-SVD** vector to capture a single multilingual steering direction. For each layer, we stack the language-specific steering vectors and compute an uncentered singular value decomposition (SVD). We use the top singular direction as the shared vector, sign-align it to English, and scale it by the average projection of the language-specific vectors onto that direction. For `resid_pre` vectors in layers 26–32, this direction explains 85.9% of cross-language vector energy on average, suggesting strong cross-lingual alignment in the steering space (see Appx. A.9). To ensure shared-vector gains do not simply come from including the target language, we evaluate **leave-one-language-out Shared-SVD** (LOO-SVD), where the target-language vector is excluded before re-computing the shared direction.

Finally, for lower-resource target languages, we test targeted source-language transfer. The **activation-nearest** source chooses the high-resource language with closest steering vector by cosine similarity, while the **linguistically nearest** source chooses the closest high-resource language in `URIEL/lang2vec` feature space (Littell et al., 2017), following prior work on linguistic similarity for cross-lingual transfer (Eronen et al., 2023). This selects Spanish for Arabic, Korean for Bengali, and Chinese for Hindi, Thai, and Vietnamese.

4.3 Main Results

Target-language steering improves visual-prior arbitration. Tab. 1 reports two metrics. Acc. is steer-all accuracy: all held-out examples of a condition are steered, without conditioning on whether

Target	Act.	CK	IG	Ling.	CK	IG
Arabic	fr	53.5/54.5	72.7/93.9	es	56.6/57.6	63.6/ 94.9
Hindi	de	39.4/53.5	34.3/68.7	zh	47.5/61.6	32.3/ 68.7
Vietnamese	de	51.5/61.6	59.6/79.8	zh	57.6/67.7	62.6/79.8
Thai	es	47.5/55.6	80.8/84.8	zh	47.5/54.5	80.8/85.9
Bengali	de	38.4/46.5	16.2/46.5	ko	36.4/45.5	16.2/45.5
Macro	–	46.1/54.3	52.7/74.7	–	49.1/57.4	51.1/ 74.9

Table 2: Lower-resource transfer. Cells are Acc./Rec. Act.=activation-nearest, Ling.=linguistic-nearest. de=German, es=Spanish, fr=French, ko=Korean, zh=Chinese.

the unsteered model was correct. Rec. is a diagnostic recovery-only score: it preserves examples the unsteered model answered correctly and applies steering only to examples it originally answered incorrectly. We first evaluate language-specific steering. On common-knowledge questions, accuracy improves by +21.9 points for high-resource languages and +13.7 points for lower-resource languages over the unsteered baseline, with recovery scores of 66.8% and 51.7%, respectively. In contrast, image-grounded questions start from much stronger unsteered baselines. Applying the same intervention to all examples lowers steer-all accuracy, especially for lower-resource languages (72.9% $\rightarrow$ 60.0%), but recovery-only scores remain high (95.5% for high-resource and 75.4% for lower-resource languages). Overall, the constructed directions effectively shift Qwen3-VL-8B’s behavior under visual-prior conflict, with the largest gains appearing on common-knowledge questions, where the unsteered model is often misled by counterfactual visual evidence.

English vectors provide a strong transfer baseline. On common questions, English transfer gives the largest gains: +28.7 points for high-resource languages and +21.0 points for lower-resource languages over the unsteered baseline, with recovery scores of 73.2% and 57.6%. Compared with language-specific steering, this adds another +6.8 and +7.3 accuracy points for high- and lower-resource languages, respectively. This suggests an English-derived arbitration direction aligns well with multilingual hidden states.

Shared multilingual directions are competitive. Shared-SVD provides a second transfer setting. On common questions, it improves accuracy by +24.8 points for high-resource and +18.4 points for low-resource languages, with recovery scores of 69.7% and 54.5%. LOO-SVD closely matches full

Model	Steering source	Common questions				Grounded questions			
		High		Low		High		Low	
		Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.
Qwen3-VL-32B	Unsteered base	53.0	—	39.8	—	94.4	—	86.3	—
	Lang-specific	67.5	69.4	38.8	44.6	91.2	95.3	72.1	86.3
	English vector	69.2	71.9	47.3	55.2	88.9	95.5	70.5	86.9
	Shared-SVD	69.4	71.7	47.3	54.7	90.6	95.1	71.5	87.1
	LOO-SVD	69.4	71.7	47.5	54.9	90.6	95.1	71.5	87.1
Gemma-3-4B	Unsteered base	20.9	—	17.6	—	98.5	—	96.6	—
	Lang-specific	38.0	43.1	38.2	43.0	76.1	99.5	69.3	96.6
	English vector	46.1	50.7	38.8	44.0	67.0	99.8	68.7	96.6
	Shared-SVD	43.3	47.3	39.6	43.6	77.6	99.7	69.9	96.6
	LOO-SVD	43.6	47.6	39.6	43.8	77.1	99.7	69.7	96.6

Table 3: Additional model steering results on counterfactual images. Metrics and language groups follow Tab. 1. Qwen3-VL-32B uses layers 49–57 with $\alpha = 0.1$; Gemma-3-4B uses layers 18–30 with $\alpha = 0.2$.

Shared-SVD, with gains of +25.2 and +17.8 points and recovery of 70.0% and 54.3%. Together, these results provide evidence for a shared cross-lingual component in the steering directions.

Targeted high-resource transfer helps lower-resource languages. For lower-resource targets (Tab. 2), linguistically nearest improves common-question accuracy by +20.0 points over the unsteered baseline and reaches 57.4% recovery. This is stronger than activation-nearest transfer, which gives a +17.0 point accuracy gain and 54.3% recovery, and nearly matches English transfer.

Gains are strongest for Arabic, Hindi, and Vietnamese; Thai is nearly tied, while activation-nearest is slightly better for Bengali. This suggests that high-resource-to-lower-resource steering is feasible, but the best donor remains target-dependent.

4.4 Evaluation on Additional Models

To examine whether the observed behavior is specific to Qwen3-VL-8B, we rerun the full workflow on Qwen3-VL-32B (Bai et al., 2025) and Gemma-3-4B (Gemma Team, 2025), covering both increased model size within the Qwen family and a different model family. For each model, we first select model-correct training examples, i.e., examples for which the unsteered model answers both counterfactual-common and image-grounded questions correctly, and construct steering vectors from this model-specific subset. We then repeat layer probing and select model-specific intervention layers and steering strength.

Tab. 3 reports results for the two additional models. In both cases, English-transfer and shared multilingual directions improve counterfactual-common accuracy on non-English prompts, while steer-all intervention can reduce image-grounded

accuracy. For Qwen3-VL-32B, lower-resource common accuracy increases from 39.8% to 47.3% with English-transfer and 47.5% with LOO-SVD. For Gemma-3-4B, it increases from 17.6% to 38.8% and 39.6%, respectively. Together with the Qwen3-VL-8B results, these findings suggest that the observed cross-lingual transfer behavior is not specific to a single model size or family.

Within the Qwen family, the larger 32B model also exhibits a milder image-grounded trade-off. Under English-transfer, lower-resource grounded accuracy decreases from 86.3% to 70.5%, compared with 72.9% to 48.7% for Qwen3-VL-8B. The image-grounded accuracy drop is larger for Gemma-3-4B despite its higher unsteered grounded accuracy. We report 95% Wilson confidence intervals for pooled non-English counterfactual-common results in Appx. A.5.

5 Conclusion

We study whether visual-prior steering directions transfer across languages in MLLMs. Our multilingual extension of the Visual CounterFact color split shows that visual-prior conflicts previously observed in English persist across languages, with a larger drop on image-grounded questions for lower-resource languages. English-derived and shared multilingual directions can shift model behavior on non-English prompts, indicating that this arbitration is not purely language-local. The same qualitative cross-lingual transfer pattern is also observed with Qwen3-VL-32B and Gemma-3-4B, suggesting that the effect is not specific to the Qwen3-VL-8B setting.

Limitations

We evaluate three model settings: Qwen3-VL-8B, Qwen3-VL-32B, and Gemma-3-4B. These settings cover a larger model size within the Qwen family and a different model family, and show a consistent qualitative pattern of cross-lingual steering transfer. However, the model coverage remains limited. The effective intervention layers and steering strengths are model-specific, and the larger Qwen3-VL-32B exhibits a milder image-grounded trade-off than Qwen3-VL-8B, while Gemma-3-4B is more sensitive to steer-all intervention. Evaluation across additional MLLM families and training mixtures would further test the extent to which the observed arbitration direction generalizes.

Our study also focuses on the color split of Visual CounterFact. Color provides a controlled setting for visual-prior conflict, but it does not cover other forms of arbitration, such as size, action, counting, spatial relations, or culturally specific commonsense priors. Extending the analysis to broader conflict types would clarify whether the observed steering directions reflect a general visual-prior arbitration mechanism.

Our multilingual evaluation set is translation-based. Although we apply quality control to fix clear translation errors, remaining artifacts may affect question naturalness, answer-label ambiguity, or object granularity across languages. We do not conduct systematic native-speaker validation for all target languages, which remains a limitation. This is especially relevant for lower-resource languages, where both translation quality and the model’s language understanding may be weaker.

Finally, our steering setup is a controlled diagnostic intervention rather than an autonomous deployment method. It uses the benchmark question type to select the steering direction, and steer-all intervention can reduce accuracy on already-correct image-grounded examples. Future work could explore selective gating or confidence-based mechanisms for applying such interventions in open-ended multilingual settings.

Acknowledgment

This work was partly supported by National Science Foundation Grant No. 2506946.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. [DoLa: Decoding by contrasting layers improves factuality in large language models](#). In *ICLR*.
- Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. [Zero-shot cross-lingual transfer language selection using linguistic similarity](#). *Information Processing & Management*, 60(3):103250.
- Stella Frank and Emily Allaway. 2025. [ViSaGE: Understanding visual generics and exceptions](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32549–32558, Suzhou, China. Association for Computational Linguistics.
- Gemma Team. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.
- Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025. [Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfactuals](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24837–24852, Suzhou, China. Association for Computational Linguistics.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *NeurIPS*.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jentsen Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2025. [Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17825–17846, Vienna, Austria. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *NeurIPS*.

OpenAI. 2026. [GPT-5.5 system card](#). Updated April 24, 2026.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. 2026. [When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14109–14130, San Diego, California, United States. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual LLMs think in English?](#) In *ICLR 2025 Workshop on Building Trust in LLMs and LLM Applications: From Guardrails to Explainability to Regulation*. Poster.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. [Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974, Vienna, Austria. Association for Computational Linguistics.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *ICLR*.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025. [Refusal direction is universal across safety-aligned languages](#). In *NeurIPS*. NeurIPS 2025 Poster.

Yufei Wang and Adriana Kovashka. 2025. [Probing logical reasoning of MLLMs in scientific diagrams](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10706–10718, Suzhou, China. Association for Computational Linguistics.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026. [Dynamic multimodal activation steering for hallucination mitigation in large vision-language models](#). In *ICLR*. ICLR 2026 Poster.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025. [HY-MT1.5 technical report](#). Preprint, arXiv:2512.24092.

A Appendix

A.1 Appendix Overview

This appendix provides additional details, extended results, and analyses to support the main paper. It is organized as follows:

- **Full baseline (Sec. A.2):** Per-language direct Qwen3-VL accuracy for real and counterfactual images under common and grounded questions.

- **Full steering table (Sec. A.3):** Per-language version of the grouped main steering table, including language-specific, English-vector, Shared-SVD, and LOO-SVD steering.

- **Grounded flip breakdown (Sec. A.4):** Transition counts showing recovery and retention effects for lower-resource grounded questions.

- **Confidence intervals (Sec. A.5):** Wilson intervals for the main non-English counterfactual-common steering result across all three evaluated models.

- **Layer probes (Sec. A.6):** Probing implementation details and full multilingual layer-wise margin plots.

- **Steering windows (Sec. A.7):** Language-specific intervention windows compared with the shared 26–32 layer window used in the main experiments.

- **Steering ablations (Sec. A.8):** Activation-source, steering-strength, and intervention-window ablations for the main steering configuration.

- **Shared-SVD construction (Sec. A.9):** Details of the rank-1 shared multilingual steering direction and its explained-energy diagnostic.

- **Translation QC (Sec. A.10):** Translation setup, answer-label handling, color-term inspection, native-speaker spot check, and per-language QC corrections.

- **AI Assistance (Sec. A.11):** Use of AI assistants.

A.2 Full Baseline Accuracy

Tab. 4 reports the full direct-evaluation accuracy for every language and condition. The main text uses group averages for readability, while this table preserves the per-language variation behind those averages.

A.3 Full Steering Results

Tab. 5 reports the per-language version of the grouped steering results in Tab. 1. This view preserves the language-level variation behind the high- and lower-resource averages.

Language	Real image		Counterfactual image	
	Common	Grounded	Common	Grounded
English	98.4	96.6	60.9	94.9
Chinese	97.2	95.9	63.9	92.9
Japanese	97.2	94.9	7.5	95.5
Korean	95.7	95.1	15.8	95.5
Spanish	96.1	95.5	44.2	92.9
German	94.7	95.9	54.0	88.4
French	91.3	90.3	36.7	94.5
Arabic	94.7	95.1	21.1	92.7
Hindi	76.7	72.4	32.5	66.3
Vietnamese	87.4	86.4	36.5	77.5
Thai	95.7	95.5	28.4	86.0
Bengali	62.7	61.9	31.6	39.6
Higher-resource avg.	95.8	94.9	40.4	93.5
Lower-resource avg.	83.4	82.3	30.0	72.4

Table 4: Full direct Qwen3-VL-8B accuracy on the translated Visual CounterFact color split. Values are percentages; bold marks the highest language value in each condition.

A.4 Grounded Flip Breakdown

For lower-resource image-grounded questions under Qwen3-VL-8B, English-transfer and language-specific steering recover similar numbers of originally wrong examples (11/134 vs. 12/134). Their difference in steer-all accuracy instead comes primarily from retention: English-transfer flips more originally correct examples to wrong (131/361 vs. 76/361). Thus, the lower grounded accuracy of English-transfer in Tab. 1 reflects reduced preservation of already-correct image-grounded answers rather than a weaker ability to recover originally wrong examples. Tab. 6 reports the full transition counts.

A.5 Confidence Intervals

Tab. 7 reports 95% Wilson confidence intervals for the pooled non-English counterfactual-common setting, which is the main cross-lingual transfer result.

A.6 Layer-Probe Diagnostics

Fig. 3 and Fig. 4 show the layer-wise margins used to inspect where the model separates the correct answer from the prior-wrong answer. At each layer, we project the hidden state through the output embedding and score each candidate color answer by its average answer-token log probability. We take the best-scoring correct candidate and the best-scoring prior-wrong candidate, compute their difference, and average this margin over baseline-correct examples. Positive values therefore indicate that the correct answer has higher average log probability. For language-specific window selec-

tion, we use counterfactual-common examples and choose the longest consecutive range with mean margin greater than 1.0 among layers no later than 32. We use these plots as diagnostics for choosing intervention layers.

A.7 Selected Steering Layers Across Languages

Fig. 5 summarizes the language-specific intervention windows selected from the layer-wise margin probes in Appendix A.6. Most languages select late-middle to late layers that overlap with the shared 26–32 window, while Arabic is the clearest exception with an earlier selected range.

The selected windows are diagnostic and need not be the unique best intervention range. We use the shared 26–32 window in the main transfer experiments because it performs well in the layer-window ablation and keeps all source-language and shared-vector interventions fixed across languages, making cross-lingual comparisons depend on the steering direction rather than on language-specific intervention sites.

A.8 Steering Ablations

Tab. 8, 9, and 10 report the ablations used to choose the main steering configuration. They isolate activation source, steering strength, and intervention window, respectively.

All three tables evaluate counterfactual-common questions. Unless otherwise specified, results are macro-averaged over all 12 languages. Tab. 8 fixes $\alpha = 0.25$ and the shared 26–32 layer window, showing that `resid_pre` and full-block outputs produce the strongest recovery, while attention and MLP outputs are less effective. Tab. 9 fixes `resid_pre` and the shared window, showing that increasing α reduces retention and does not improve recovery in this range. We therefore use $\alpha = 0.25$ as a moderate setting that improves recovery while limiting changes to originally correct answers. Tab. 10 fixes `resid_pre` and $\alpha = 0.25$, comparing the shared window against language-specific windows for each language. The shared 26–32 layer range performs better on average, possibly because language-specific ranges contain more language-processing information unrelated to visual-prior arbitration.

A.9 Shared-SVD Direction Construction

For each layer ℓ , we stack the language-specific steering vectors into a matrix $M^{(\ell)} \in \mathbb{R}^{|\mathcal{L}| \times d}$,

Question	Language	Base	Language-specific	English vector	Shared-SVD	LOO-SVD
Common	English	59.6	88.9 / 91.9	88.9 / 91.9	90.9 / 92.9	90.9 / 92.9
	Chinese	62.6	81.8 / 86.9	81.8 / 86.9	81.8 / 86.9	81.8 / 86.9
	Japanese	8.1	24.2 / 28.3	52.5 / 55.6	44.4 / 47.5	46.5 / 49.5
	Korean	16.2	35.4 / 37.4	42.4 / 43.4	39.4 / 41.4	39.4 / 41.4
	Spanish	44.4	64.6 / 73.7	70.7 / 76.8	65.7 / 73.7	64.6 / 72.7
	German	55.6	77.8 / 84.8	80.8 / 86.9	74.7 / 79.8	76.8 / 80.8
	French	33.3	60.6 / 64.6	63.6 / 70.7	56.6 / 65.7	56.6 / 65.7
	Arabic	23.2	51.5 / 52.5	55.6 / 56.6	54.5 / 55.6	54.5 / 55.6
	Hindi	29.3	37.4 / 51.5	49.5 / 60.6	40.4 / 51.5	40.4 / 51.5
	Vietnamese	36.4	50.5 / 60.6	55.6 / 64.6	54.5 / 62.6	53.5 / 61.6
	Thai	30.3	43.4 / 52.5	46.5 / 54.5	47.5 / 54.5	46.5 / 54.5
	Bengali	26.3	31.3 / 41.4	43.4 / 51.5	40.4 / 48.5	39.4 / 48.5
Grounded	English	98.0	97.0 / 98.0	97.0 / 98.0	97.0 / 98.0	97.0 / 98.0
	Chinese	91.9	77.8 / 92.9	77.8 / 92.9	80.8 / 92.9	80.8 / 92.9
	Japanese	97.0	85.9 / 97.0	83.8 / 97.0	83.8 / 97.0	83.8 / 97.0
	Korean	96.0	88.9 / 96.0	91.9 / 97.0	89.9 / 96.0	89.9 / 96.0
	Spanish	94.9	76.8 / 96.0	77.8 / 96.0	82.8 / 96.0	82.8 / 96.0
	German	87.9	83.8 / 90.9	85.9 / 92.9	87.9 / 89.9	87.9 / 88.9
	French	97.0	91.9 / 98.0	90.9 / 98.0	93.9 / 98.0	93.9 / 98.0
	Arabic	92.9	89.9 / 96.0	52.5 / 96.0	76.8 / 96.0	75.8 / 96.0
	Hindi	64.6	49.5 / 68.7	32.3 / 68.7	40.4 / 68.7	39.4 / 68.7
	Vietnamese	78.8	60.6 / 79.8	62.6 / 79.8	59.6 / 79.8	59.6 / 79.8
	Thai	84.8	81.8 / 84.8	80.8 / 84.8	81.8 / 84.8	81.8 / 84.8
	Bengali	43.4	18.2 / 47.5	15.2 / 46.5	16.2 / 46.5	16.2 / 46.5

Table 5: Per-language version of Table 1. Base is unsteered accuracy. Each steering cell reports Acc. / Rec., where Acc. steers every held-out example and Rec. is accuracy with selective recovery with original corrects preserved.

Steering source	Wrong→correct	Correct→wrong	Retention	Steer-all Acc.
Lang-specific	12/134 (9.0%)	76/361 (21.1%)	285/361 (78.9%)	60.0%
English-transfer	11/134 (8.2%)	131/361 (36.3%)	230/361 (63.7%)	48.7%
Shared-SVD	11/134 (8.2%)	100/361 (27.7%)	261/361 (72.3%)	54.9%
LOO-SVD	11/134 (8.2%)	102/361 (28.3%)	259/361 (71.7%)	54.5%

Table 6: Transition counts for lower-resource image-grounded questions under Qwen3-VL-8B. Recovery counts measure originally wrong examples corrected by steering; retention counts measure originally correct examples preserved after steering.

where each row corresponds to one language and d is the hidden dimension. We then compute an uncentered singular value decomposition of $M^{(\ell)}$. The shared steering direction is given by the top right singular vector. We sign-align this vector to the English steering direction and scale it by the mean signed projection of the language-specific vectors onto the shared axis.

To quantify how well this rank-1 direction summarizes the multilingual steering directions, we compute the fraction of vector energy explained by the first singular direction:

$$E_1^{(\ell)} = \frac{(\sigma_1^{(\ell)})^2}{\sum_j (\sigma_j^{(\ell)})^2}, \quad (1)$$

where $\sigma_j^{(\ell)}$ are the singular values of $M^{(\ell)}$. Higher values indicate that the language-specific steering vectors are well approximated by a single shared direction.

For the main resid_pre configuration over layers 26–32, the rank-1 diagnostic is high across all

selected layers (Tab. 11). The mean rank-1 energy is 85.9% with a layer-wise standard deviation of 2.4 percentage points. This diagnostic is computed over the 12 language-specific vectors used by Shared-SVD. The two stored steering directions, ck_to_ig and ig_to_ck, are sign reversals of one another, so they have identical singular values and identical explained-energy ratios.

A.10 Translation and Quality Control Details

We translate the Visual CounterFact color split into 11 target languages using HY-MT1.5-7B (Zheng et al., 2025) with the default translation prompt. The split contains 493 base examples; we use a fixed 80/20 random split with seed 3730, giving 394 training examples and 99 held-out examples. Each target language has two translated question forms per example; paired with real and counterfactual images, this yields 1,972 evaluated QA instances per language. Our use of Visual CounterFact is consistent with its research evaluation pur-

Model	Setting	Base	English-transfer	Shared-SVD	LOO-SVD
Qwen3-VL-8B	Layers 26–32, $\alpha = 0.25$	33.2 [30.5, 36.1]	58.4 [55.5, 61.3]	54.5 [51.6, 57.5]	54.5 [51.6, 57.5]
Qwen3-VL-32B	Layers 49–57, $\alpha = 0.10$	47.0 [44.1, 50.0]	59.2 [56.3, 62.1]	59.3 [56.4, 62.2]	59.4 [56.5, 62.3]
Gemma-3-4B	Layers 18–30, $\alpha = 0.20$	19.4 [17.1, 21.8]	42.8 [39.9, 45.8]	41.6 [38.7, 44.5]	41.8 [38.9, 44.7]

Table 7: Non-English counterfactual-common steer-all accuracy with 95% Wilson confidence intervals. Results are pooled over the 11 non-English target languages.

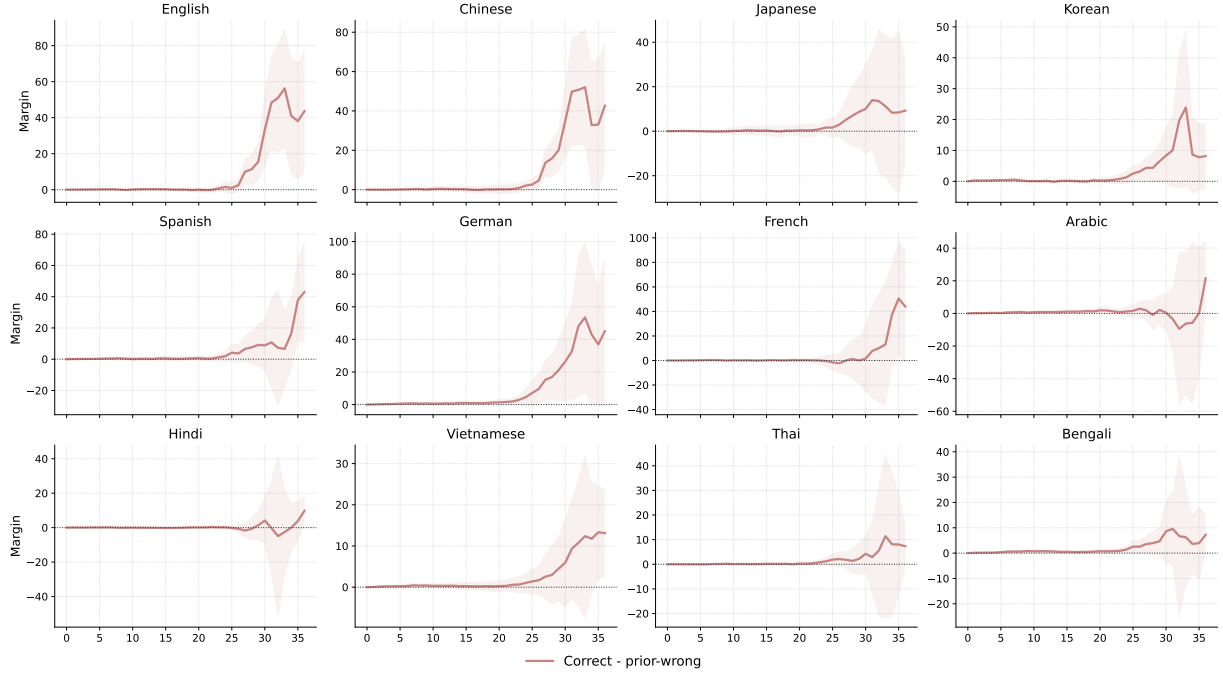


Figure 3: Layer-wise margin for counterfactual images with common questions. The margin is the correct-answer average log probability minus the prior-wrong average log probability.

Source	Base	Acc.	Δ	Rec.	Δ /	Ret.
Block output	35.44	51.94	+16.50	58.84	+23.40	77.11
Attention output	35.44	35.10	-0.34	37.63	+2.19	90.01
MLP output	35.44	39.14	+3.70	41.92	+6.48	89.52
resid_pre	35.44	53.96	+18.52	60.52	+25.08	77.52

Table 8: Activation-source ablation. Base is unsteered accuracy; Acc. is accuracy after steering all held-out examples; Rec. is recovery-only accuracy; Ret. is retention on originally correct examples. Bold marks the column maximum where values differ.

α	Base	Acc.	Δ	Rec.	Δ /	Ret.
0.25	35.44	53.96	+18.52	60.52	+25.08	77.52
0.50	35.44	47.39	+11.95	60.19	+24.75	60.50
0.75	35.44	42.42	+6.99	58.67	+23.23	51.02
1.00	35.44	31.65	-3.79	52.19	+16.75	43.14

Table 9: Steering-strength ablation. Base, Acc., Rec., and Ret. follow Tab. 8. Bold marks the column maximum where values differ.

pose, and our translated annotations are intended only for research use under the original dataset’s access conditions.

For answer labels, our translation code wraps each source color in a color context (e.g., “the color orange”) before translation to reduce word-sense ambiguity; we then convert translated labels to bare color terms. In QC, we check for color-sense mistakes such as translating the color “orange” as the fruit “orange”, object-granularity changes such

as replacing “Samoyed” with “dog”, and other clear translation errors. We use GPT-5.5 (OpenAI, 2026)-assisted consistency checks over the translated rows to flag likely issues, then manually adjudicate corrections. We also inspect translated color terms for cross-language lexicalization differences. For example, in Vietnamese, blue and green answer labels remain distinguished rather than using an underspecified blue/green label. We do not identify cases where lexical differences make the intended answer label ambiguous or inconsistent. As

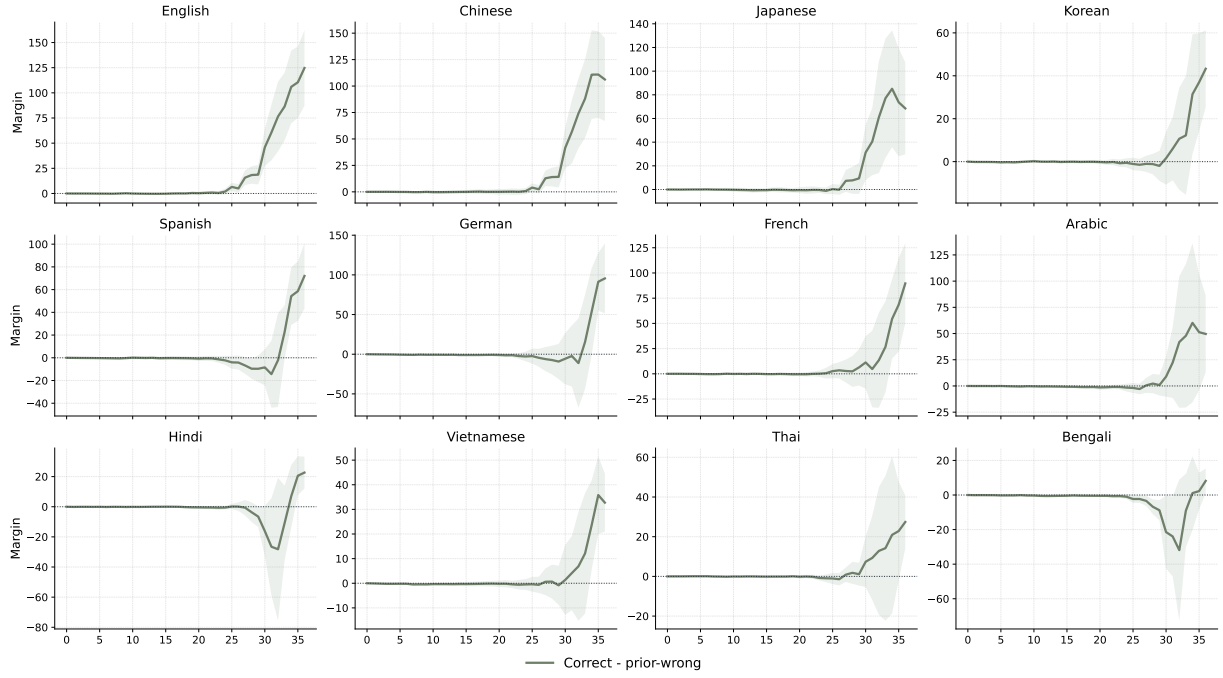


Figure 4: Layer-wise margin for counterfactual images with grounded questions. The margin is the correct-answer average log probability minus the prior-wrong average log probability.

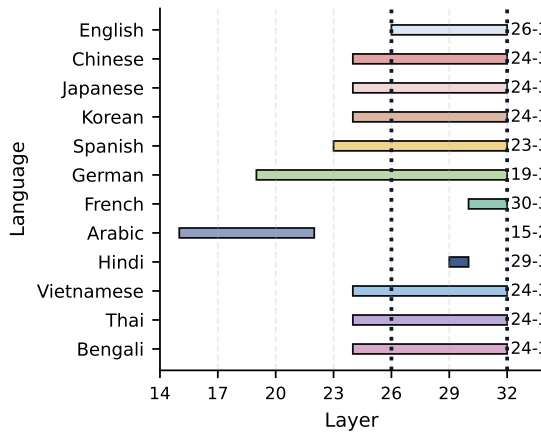


Figure 5: Language-specific intervention windows from layer-wise margin probes. Dotted lines mark the shared 26–32 window used in the main transfer experiments.

an additional native-speaker check, we inspect 100 randomly sampled Simplified Chinese examples and do not identify translation errors that change the intended question or answer. We did not conduct full native-speaker validation for every target language; therefore, residual translation artifacts may remain, as discussed in Limitations.

The row-level manual text edits are language-specific but sparse (Tab. 12). We made 32 object/question fixes across 5,423 target-language rows (0.59%). Arabic required 8 fixes; Korean,

German, and French each required 4; Spanish and Vietnamese each required 3; Japanese, Hindi, and Thai each required 2; Chinese and Bengali required none. These counts summarize object/question text fixes; the color-term inspection and Simplified Chinese spot check did not produce additional answer-changing corrections.

A.11 Use of AI Assistants

We used AI assistants for writing and editing support, as well as for monitoring experimental outputs and logs. All final claims, analyses, and corrections were manually reviewed by the authors.

Lang.	Base	S-Acc.	Δ	S-Rec.	Δ	L-Acc.	Δ	L-Rec.	Δ
English	59.60	88.89	+29.29	91.92	+32.32	88.89	+29.29	91.92	+32.32
Chinese	62.63	81.82	+19.19	86.87	+24.24	78.79	+16.16	85.86	+23.23
Japanese	8.08	24.24	+16.16	28.28	+20.20	36.36	+28.28	40.40	+32.32
Korean	16.16	35.35	+19.19	37.37	+21.21	45.45	+29.29	47.47	+31.31
Spanish	44.44	64.65	+20.20	73.74	+29.29	68.69	+24.24	78.79	+34.34
German	55.56	77.78	+22.22	84.85	+29.29	80.81	+25.25	89.90	+34.34
French	33.33	60.61	+27.27	64.65	+31.31	47.47	+14.14	50.51	+17.17
Arabic	23.23	51.52	+28.28	52.53	+29.29	31.31	+8.08	32.32	+9.09
Hindi	29.29	37.37	+8.08	51.52	+22.22	37.37	+8.08	44.44	+15.15
Vietnamese	36.36	50.51	+14.14	60.61	+24.24	55.56	+19.19	64.65	+28.28
Thai	30.30	43.43	+13.13	52.53	+22.22	48.48	+18.18	56.57	+26.26
Bengali	26.26	31.31	+5.05	41.41	+15.15	27.27	+1.01	40.40	+14.14

Table 10: Layer-window ablation. S uses the shared 26–32 window; L uses language-specific windows. Acc. and Rec. follow Table 8; Δ and Δ' are changes relative to Base. Bold marks the higher S/L value within each row.

Metric	Layer							Mean $\pm$ SD
	26	27	28	29	30	31	32	
Rank-1 energy (%)	87.9	88.1	86.9	85.7	86.2	85.5	80.8	85.9 $\pm$ 2.4
Mean cosine	0.932	0.932	0.925	0.919	0.918	0.912	0.883	0.917 $\pm$ 0.017

Table 11: Shared-SVD rank-1 diagnostic for the main resid_pre vectors on layers 26–32. Rank-1 energy is $100E_1^{(\ell)}$. Mean cosine averages the cosine between each language-specific vector and the shared direction at that layer after sign alignment.

Manual QC count	Target language											Total
	Zh	Ja	Ko	Es	De	Fr	Ar	Hi	Vi	Th	Bn	
Object/question fixes	0	2	4	3	4	4	8	2	3	2	0	32
Rows changed (%)	0.00	0.41	0.81	0.61	0.81	0.81	1.62	0.41	0.61	0.41	0.00	0.59

Table 12: Manual translation QC corrections. Each count is a language-row pair among 493 translated base examples per target language. The total percentage divides by all 5,423 target-language row pairs.