Causal Localization of the English Pivot in LLaVA: Mechanistic VLM Analysis and Training-Free Multilingual Steering

Abrar Zahin Raihan, Aurchi Chowdhury
Bangladesh University of Engineering and Technology
raihanzahin95@gmail.com aurchichowdhury07@gmail.com

Abstract

Multilingual vision-language models (VLMs) consistently underperform on non-English visual queries, yet the internal mechanism behind this disparity remains unknown. As a focused case study on LLaVA-1.5-7B, we apply logit-lens analysis and causal activation patching to show that non-English visual queries are routed through an English-biased representational bottleneck in layers 5–17, extending the English-pivot phenomenon of Wendler et al. (2024) to the multimodal setting. Peak causal influence occurs at layer 8 ($\overline{\text{AIE}}$=0.49, averaged across languages), with all measurable pivot signal running through text-token positions. Without meaningful visual content (blank-image condition), language-specific representations do not emerge at any layer, showing that the pivot is image-content-dependent rather than triggered by any visual input. Building on these findings, we derive training-free language-steering vectors at the mechanistically identified pivot layers, improving Russian VQA by +6.5 percentage points (pp) and Portuguese by +4.0 pp on MMMB without any fine-tuning—the latter surpassing the English baseline. Within this case study, our results are consistent with the English pivot being a structural property of the LLM backbone that multimodal pre-training does not mitigate; extending this mechanistic methodology to other VLMs and language families remains an important direction for future work.

1 Introduction

Multilingual VLMs are deployed globally, yet a persistent gap separates their English and non-English performance on visual question answering (Fan et al., 2025; Song et al., 2025). Prior work attributes this gap to training-data imbalance or insufficient multilingual instruction tuning, and proposes remedies such as language-aware fine-tuning (Fan et al., 2025) or culturally diverse data (Nyandwi et al., 2025). Yet no prior work has asked the mechanistic question: *where* in the network does English bias arise, and what causal role does the visual modality play?

The text-only interpretability literature provides a starting hypothesis. Wendler et al. (2024) show that LLaMA-family models process multilingual input through three phases: an *input space*, a *concept space* where representations become English-biased, and an *output space* where the target language is reconstructed. Ferrando and Costa-jussà (2024) extend this to circuit-level analysis across languages. Neither work addresses VLMs, leaving open whether the visual modality interacts with or suppresses this pivot.

We fill this gap with the first causal mechanistic analysis of the English pivot in a multilingual VLM, presenting a focused case study on LLaVA-1.5-7B. We treat our findings as specific to this model and architecture; whether analogous structures arise in other VLMs is an open question we leave to future work. Our contributions are:

- We confirm that LLaVA-1.5-7B routes non-English visual queries through an English-biased concept space in **layers 5–17**, with peak causal responsibility at layer 8 (cross-language average $\overline{\text{AIE}}$=0.49; §4.3).
- We show that the pivot is image-content-dependent: language-specific representations do not emerge without semantically rich visual input (blank-image condition; §4.2).
- Under our patching design, all measurable pivot signal runs through text-token positions; isolating visual causal contributions would require paired semantically-equivalent cross-lingual images (§4.3).
- We derive training-free steering vectors at the mechanistically identified pivot layers, yielding **+6.5 pp** on Russian and **+4.0 pp** on Portuguese on the MMMB multiple-choice benchmark, without any fine-tuning (§4.4).

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 257–265

2 Related Work

Mechanistic interpretability of multilingual LLMs. Wendler et al. (2024) establish the three-phase English-pivot model for LLaMA via logit-lens analysis. Dumas et al. (2025) use activation patching to show that concept representations are language-agnostic at intermediate layers, separating the underlying concept from its surface language form. Ferrando and Costa-jussà (2024) show language-agnostic circuits are reused across languages, implying the pivot is structural. Brinkmann et al. (2025) use causal sparse-autoencoder feature analysis to show shared latent grammatical representations across typologically diverse languages; Resck et al. (2025) survey mechanistic and interpretability findings across multilingual LLMs more broadly. All of these works study text-only models; we extend causal mechanistic analysis to VLMs.

Multilingual vision-language models. Multilingual VLM architectures such as mBLIP (Geigle et al., 2024) and PALO (Maaz et al., 2024) extend English-centric models through multilingual encoders or instruction tuning, but do not examine internal mechanisms. Fan et al. (2025) propose PLAST, a language-aware approach requiring fine-tuning. Song et al. (2025) document non-English VQA failure modes without investigating internal mechanisms. Nyandwi et al. (2025) address cultural diversity at the data level. In contrast, we provide a causal mechanistic account and a training-free fix.

Mechanistic interpretability for VLMs. Golovanevsky et al. (2025) apply activation patching to LLaVA and BLIP to study how visual information is integrated under image corruption, finding that LLaVA's attention heads perform outlier suppression rather than visual grounding—a complementary analysis of the visual integration pathway rather than the language-processing pathway we study. Our steering intervention is related to representation engineering (Zou et al., 2023), but derived from mechanistically identified pivot layers rather than contrast pairs.

3 Background

LLaVA-1.5-7B. LLaVA-1.5-7B (Liu et al., 2024) connects a CLIP ViT-L/14 visual encoder to a Vicuna-7B backbone via a two-layer MLP projector. An image is encoded into 576 patch tokens (positions 1–576 in the sequence), followed by the tokenised question. The backbone has 32 Llama-2 decoder layers of hidden size 4096.

Logit lens and PMI. The logit lens (nostalgebraist, 2020) projects an intermediate hidden state $\mathbf{h}_\ell$ through the final layer norm and LM head to obtain token probabilities at any depth. We report *pointwise mutual information*: $\mathrm{PMI}_\ell[\tau] = \log p_\ell[\tau] - \log p_0[\tau]$, where p_0 is the LM-head output at a zero hidden state (the model's unconditional prior), to remove vocabulary-frequency bias.

Causal tracing and AIE. Causal tracing (Meng et al., 2022) replaces the clean hidden state at layer ℓ with the corresponding state from a corrupted run, measuring the output change. We define the *Average Indirect Effect*:

$$\mathrm{AIE}(\ell) = \mathbb{E}_e\left[p_\ell^{\text{patch}}[\tau_{\text{en}}] - p^{\text{clean}}[\tau_{\text{en}}]\right], \quad (1)$$

where the clean run uses a non-English question and the corrupted run uses the English equivalent on the same image. τ_{en} is the first token of the model's English response, determined dynamically per example (typically "The", id 450; see Appendix B). Positive AIE at layer ℓ means that layer causally carries English-pivot information.

4 Experiments and Results

4.1 Experimental Setup

Model. We use `llava-hf/llava-1.5-7b-hf` in float16, no quantisation or fine-tuning, greedy decoding.

Probing set. We build a 195-example probing set from COCO val2017 (Lin et al., 2014) across 15 object categories (person, car, bus, train, boat, orange, book, clock, knife, spoon, umbrella, bird, cat, dog, apple). For each example we ask the question in four languages: English (en), Portuguese (pt), German (de), and Russian (ru), using forced-choice prompts listing all 15 category names in the target language. We track the first sub-token of each answer word using contextual-subtraction tokenisation to match generation-time tokenisation (Appendix A). This controlled object-naming paradigm is designed to identify the earliest layers at which the model develops language-specific representations under a semantically invariant visual stimulus; the transfer of vectors derived from this set to the broader MMMB benchmark is an empirical finding.

Evaluation benchmarks. We evaluate steering on MMMB (Massive Multilingual Multimodal Benchmark; Sun et al., 2025), a multilingual multiple-choice VQA benchmark covering English, Portuguese, and Russian (200 examples each). German is included in the mechanistic analysis (Figures 1–3) and cross-lingual transfer ablation (§4.5) because its residual-stream AIE profile peaks at layer 10 with a pivot region of approximately 5–17, closely matching the range identified for all languages—making it a natural cross-lingual transfer control. MMMB does not provide a German split, so German steering is assessed indirectly via transfer.

Baselines. We compare our steered model ($\alpha>0$) against unmodified LLaVA-1.5-7B (baseline, $\alpha=0$) and against steering vectors extracted from fixed layer ranges—early (0–10), mid (11–21), and late (22–31)—to test whether mechanistic localisation is necessary.

4.2 The English Pivot in VLMs

Figure 1 shows logit-lens PMI across all 32 layers. German is included in all mechanistic figures as a cross-lingual control; it has no MMMB split, so its steering effect is assessed indirectly via cross-lingual transfer (§4.5).

Image-present condition. With a real image, both English and target-language PMI grow substantially in late layers. English PMI peaks at layers 20–23 (values 8.1–8.4) across all three languages, consistently *before* the target-language peak at layers 27–29 (values 8.6–10.1). This ordering—English representations crystallising 5–7 layers before the target language—directly instantiates Wendler et al.'s concept-space phase in the multimodal setting. The grey pivot region (layers 5–17) corresponds to where English PMI first begins to build, preceding both peaks.

Blank-image condition. Without meaningful visual content (image replaced by a blank tensor of the same spatial dimensions), PMI values are predominantly negative at every layer for every language; where marginally positive, values are substantially weaker than in the image-present condition. This shows that the pivot is image-content-dependent: it is not triggered by any visual input, but requires semantically rich visual content for language-specific representations to form.

Logit-lens vs. causal patching. These measurements are complementary: the logit lens shows where representations *crystallise* into decodable probabilities (a readout property), while causal patching (§4.3) shows where pivot information *transfers* causally—hence the pivot region (5–17) precedes the late-layer PMI peaks (20–29).

4.3 Causal Localisation via Activation Patching

Residual stream patching. Figure 2 (left column) shows strongly positive AIE in layers 5–17 and near-zero thereafter. Peak cross-language average AIE is at **layer 8** ($\overline{\text{AIE}}$=0.49); per-language peaks reach 0.52 (PT, layer 8), 0.62 (DE, layer 10), and 0.85 (RU, layer 10), reflecting Russian's substantially broader and stronger pivot. The signal remains above the 50%-of-peak threshold (0.24) through layer 17 and drops to noise floor (< 0.04) by layer 20. We define the **pivot region** as layers 5–17.

The pivot region occupies the first half of the LLM, consistent with Wendler et al.'s concept-space phase, and precedes the late-layer (20–29) representational peaks observed in §4.2.

Attention and MLP patches. Middle and right columns of Figure 2 show near-zero AIE for both attention-output and MLP-output patches across all layers (max $|\overline{\text{AIE}}|$<0.01 for attention; <0.01 for MLP except layer 0 where MLP≈0.08, reflecting sub-word tokenisation effects (Geva et al., 2022)). The English pivot operates through the integrated residual stream rather than through separable attention heads or feed-forward networks.

Design validation: visual token isolation. Figure 3 serves as a design validation rather than a mechanistic finding. Because the same image is used in both the non-English (clean) and English (corrupted) runs, the 576 CLIP visual-token activations are identical by construction and contribute zero differential signal. The identically-zero visual-token curve confirms that our patching setup correctly isolates text-token contributions—any non-zero value would indicate a confound in the implementation. What the design *does* establish is that the pivot's causal pathway does not depend on any language-specific visual signal; the blank-image ablation (§4.2) provides complementary evidence that the pivot is image-content-dependent—semantically rich visual input is required for language-specific representations to

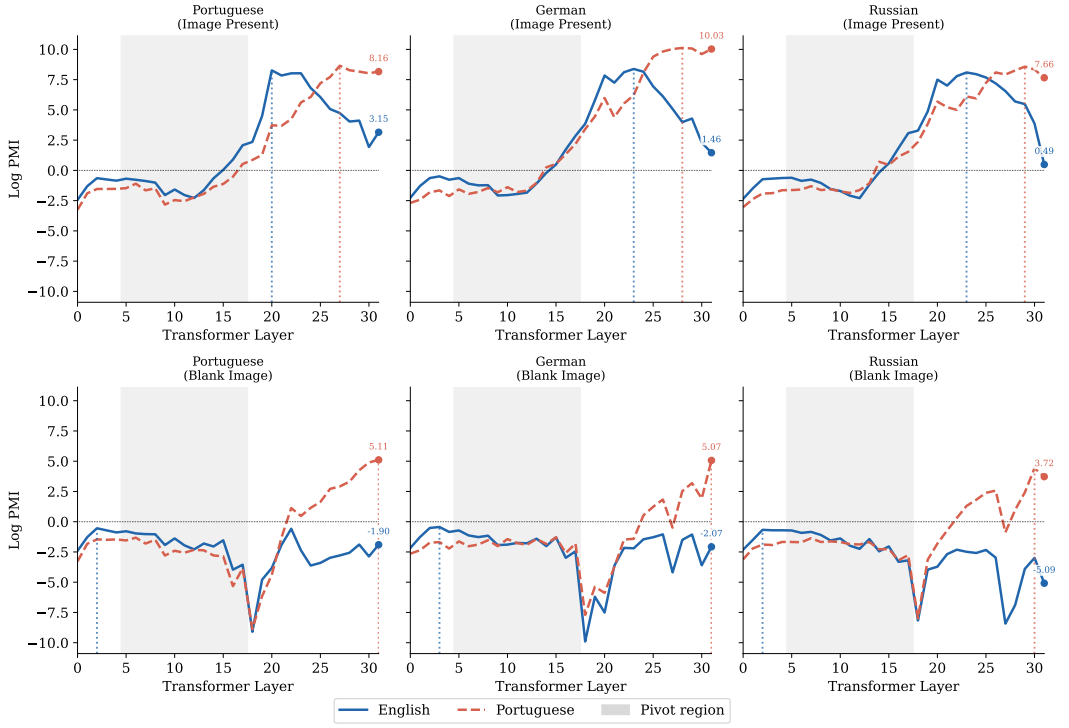

Figure 1: **Logit-lens PMI across layers.** Blue (solid): English answer token; red (dashed): target-language token. Top row: real image; bottom row: blank image. Grey band: mechanistically identified pivot region (layers 5–17). With a real image, both tokens develop strong positive PMI in late layers (20–29), with English peaking first at layers 20–23. Without meaningful visual content, PMI is predominantly negative at every layer; where marginally positive, values are substantially weaker than in the image-present condition—language-specific representations do not emerge.

form. Whether visual tokens could independently carry pivot information under a paired-image design (e.g. semantically equivalent images in different languages) remains an open question for future work.

4.4 Training-Free Steering at Pivot Layers

Steering vector extraction. For each non-English language ℓ, we extract a steering vector at each pivot layer $L \in \{5, \ldots, 17\}$:

$$\mathbf{v}_L^{(\ell)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_L^{(\ell,i)} - \mathbf{h}_L^{(\mathrm{en},i)}, \qquad (2)$$

where $\mathbf{h}_L^{(\cdot,i)}$ is the residual-stream hidden state at the last answer position for example i (N=195). Split-half cosine similarity exceeds 0.85 across all languages and pivot layers, indicating stable steering directions.

Steered inference. At inference time we add $\alpha\, \mathbf{v}_L^{(\ell)}$ to the hidden state at each pivot layer via an in-place forward hook, applied to the last active position at every decode step (KV-cache compatible; Appendix B).

Table 1: **Main results on MMMB** (4-way multiple-choice accuracy, %). "Ours": best accuracy over $\alpha \in \{0.01, 0.02, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5\}$ (optimum α=0.2 for PT and RU independently). EN: baseline only (no steering vector for English). Δ: gain over baseline.

Method	EN	PT	RU
Baseline	66.0	63.0	51.5
Ours (α=0.2)	66.0	**67.0**	**58.0**
Δ	—	+4.0	+6.5

Results (Table 1). Steering yields **+6.5 pp** on Russian MMMB (51.5% → 58.0% at α=0.2), substantially reducing the 14.5-point gap between Russian and English baselines. Portuguese gains **+4.0 pp** (63.0% → 67.0%), surpassing the English baseline of 66.0%—demonstrating that mechanistically guided steering can lift non-English performance above the monolingual baseline without any fine-tuning. The Portuguese AIE profile peaks earlier and tapers off around layer 14 (Figure 2), suggesting that language-specific pivot ranges (e.g. 5–14 for Portuguese) may yield further gains; we

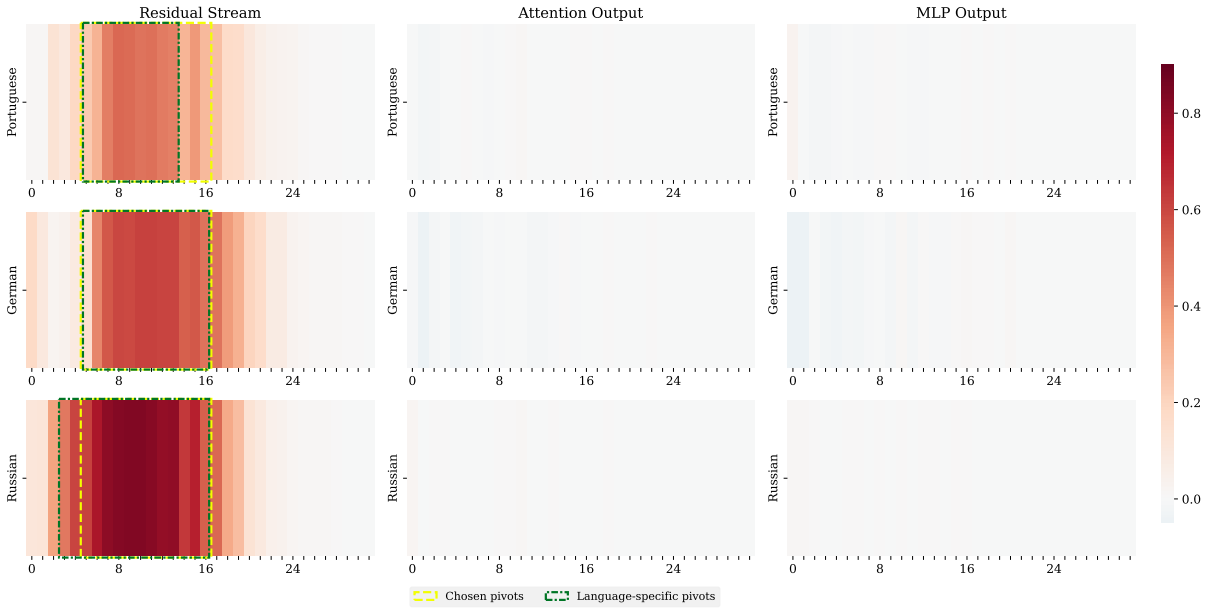

Figure 2: **Average Indirect Effect (AIE) heatmap.** Rows: Portuguese, German, Russian. Columns: residual stream / attention output / MLP output. The residual-stream pivot at layers 5–17 is pronounced across all languages; attention and MLP patches carry negligible weight. Yellow dashed box: shared pivot region (layers 5–17) used for steering. Green dashed box: language-specific pivot region inferred from each language's AIE profile, shown to highlight the per-language variation (e.g. Portuguese tapers off around layer 14 while Russian extends to 17).

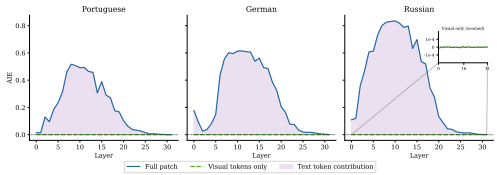

Figure 3: **Design validation: visual-token-only AIE is identically zero.** The same image appears in both passes, so the 576 CLIP patch activations are identical by construction and contribute zero differential signal. This confirms correct implementation: all measurable pivot signal under this design runs through text-token positions.

leave this optimisation as a direction for future work. Figure 4 shows a representative Portuguese example where steering shifts the model's prediction from the wrong option (C) to the correct answer (D), with the PMI curves at the pivot layers revealing the mechanism behind the correction.

4.5 Ablations

Layer range ablation (Table 2). Our mechanistic pivot region (5–17) outperforms all fixed ranges on both languages: +4.0 pp vs. +0.5 pp (early) for Portuguese, and +6.5 pp vs. +3.0 pp (mid) for Russian. Early and late ranges match the baseline for Russian; mid provides +3.0 pp but falls well short of our range's +6.5 pp. This ablation shows

Table 2: **Layer range ablation on MMMB (best α per cell, %).** Best α shown in parentheses where it differs from 0.2.

Layer range	PT	RU
Early (0–10)	63.5 (α=0.2)	51.5 (α=0.01)
Mid (11–21)	63.0 (α=0.01)	54.5 (α=0.1)
Late (22–31)	63.0 (α=0.01)	51.5 (α=0.01)
Ours (5–17)	**67.0** (α=0.2)	**58.0** (α=0.2)

Table 3: **Visual token ablation on MMMB (%).** "Steer" = best α: α=0.2 (real), α=0.02 (blank).

Condition	PT Base	PT Steer	RU Base	RU Steer
Real image	63.0	67.0	51.5	58.0
Blank image	42.0	41.5	46.5	48.5

that mechanistic localisation beats arbitrary partitioning; we note that any AIE-informed selection of similar width would likely achieve comparable gains, and the key value of the mechanistic analysis is identifying the pivot region rather than optimising layer boundaries post-hoc.

Visual token ablation (Table 3). Blank images drastically lower baseline performance (PT: -21 pp; RU: -5 pp), consistent with the image-content-dependence finding from §4.2. With blank images, steering is marginal or slightly harmful, as language-specific representations require semanti-

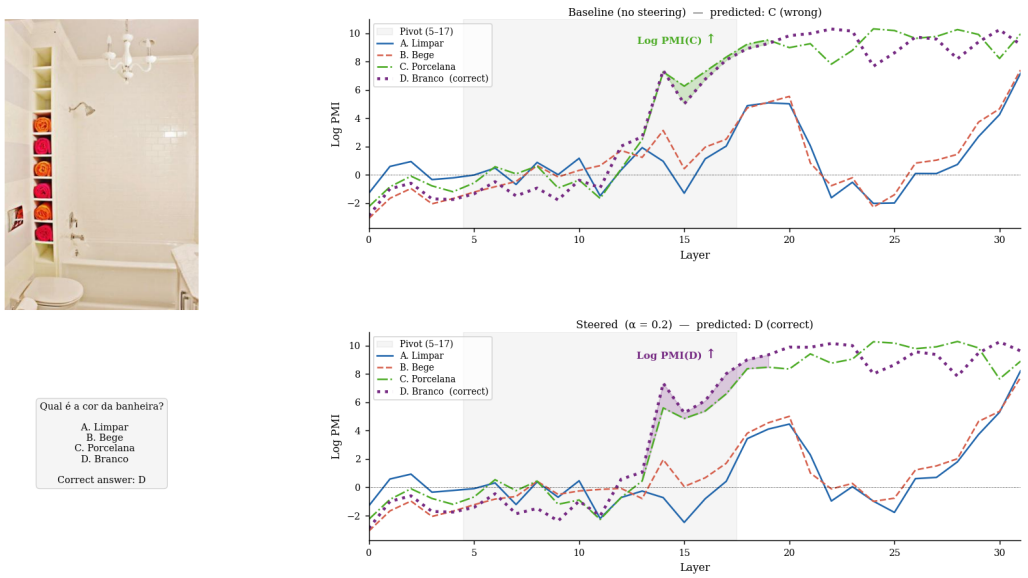

Figure 4: **Representative Portuguese MMMB case (steered wins).** *Left:* input image and question with four answer options; correct answer is D. The question asks about the colour of a bathroom fixture—a colour-attribute question, not an object-naming question, demonstrating that steering vectors transfer to qualitatively different question types beyond the probing-set distribution. *Top right (Baseline):* logit-lens PMI curves for all four options across 32 layers—the model incorrectly peaks on option C at layer 31. *Bottom right (Steered, α=0.2):* after applying the Portuguese steering vector at pivot layers 5–17, option D dominates at layer 31, recovering the correct answer. The shaded fill in the pivot region highlights where C and D compete; steering resolves this competition in favour of the correct option.

Table 4: **Cross-lingual transfer on MMMB (best α per cell, %).**

Vector source	MMMB-PT	MMMB-RU
PT (own)	—	55.5 (α=0.3)
DE	67.0 (α=0.2)	58.0 (α=0.2)
RU (own)	63.5 (α=0.3)	—

cally rich visual content to form.

Steering coefficient sensitivity (Figure 5). Performance peaks at α=0.2 for both MMMB languages (independently) and degrades sharply beyond α=0.3. Large α values corrupt generation quality (MMMB-RU drops to 0% at α=0.5, indicating degenerate outputs).

Cross-lingual transfer (Table 4). The **German steering vector transfers perfectly** to both Portuguese (67.0% = own-language) and Russian (58.0% = own-language). In contrast, the Portuguese vector applied to Russian yields 55.5% (-2.5 pp vs. own-language). The full transferability of the German vector suggests it captures a general "non-English" axis rather than a language-specific direction, consistent with the shared concept space observed in activation patching.

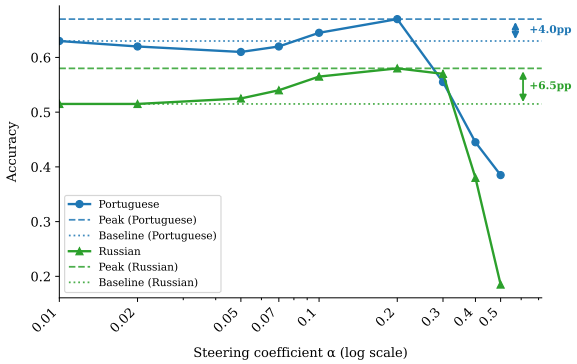

Figure 5: **Sensitivity to α on MMMB.** Performance peaks at α=0.2 and degrades sharply beyond α=0.3, consistent with prior representation-engineering results (Zou et al., 2023).

5 Discussion

Why does the pivot exist in VLMs? The pivot is a backbone property: Vicuna's English-dominant pretraining maps non-English sequences into an English-centric concept space. Because CLIP is language-agnostic, the MLP projector translates visual features into the backbone's embedding space without any language-direction component—visual tokens activate language-specific processing but cannot signal *which* language to use, explaining the

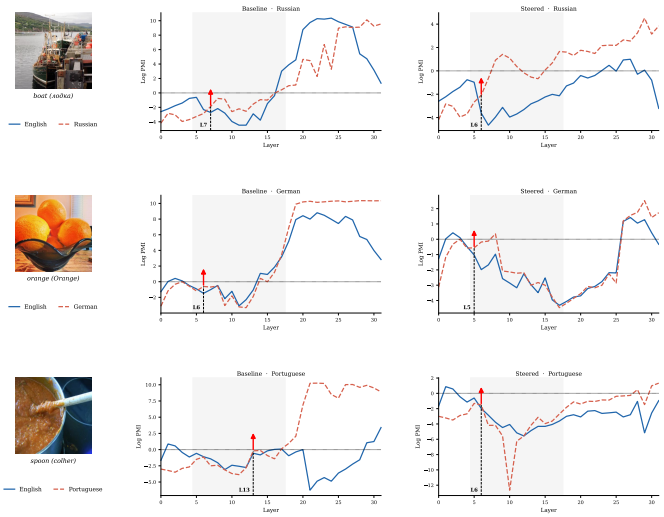

Figure 6: **Qualitative logit-lens walkthrough (probing set).** Russian (top), German (middle), Portuguese (bottom) without steering (left) and with α=0.2 (right). Red arrow: layer where target-language PMI first exceeds English. Steering shifts the crossover earlier, amplifying target-language representations at pivot layers.

zero visual-token AIE. Multimodal training optimises only the projector and leaves the backbone's language topology intact.

Steering direction geometry. The complete transferability of the German vector suggests that pt, de, and ru steering directions approximately coincide—a single "non-English" axis rather than three language-specific directions—consistent with the shared concept-space hypothesis of Wendler et al. (2024) and raising the possibility that one universal vector could suffice for multilingual correction.

Practical Implications. Our steering approach is KV-cache compatible and requires no weight updates, making it a drop-in patch for existing LLaVA-1.5 deployments serving non-English users on structured VQA tasks. The complete cross-lingual transfer of the German vector suggests that a single "non-English" vector—computed once across languages—may generalise to unseen non-English inputs, potentially reducing the per-language probing requirement; we leave this as an empirical question for future work. More broadly, the logit-lens and AIE pipeline could serve as a lightweight pre-deployment audit for new VLMs, localising whether an English pivot exists and at which layers it peaks, complementing rather than replacing full multilingual benchmarking.

Limitations. This is a single-model case study; generalisation to other VLMs and language families remains open. The steering coefficient α is selected on the MMMB test set, so reported gains are upper bounds on out-of-sample performance. The shared pivot region (5–17) may be over-inclusive for some languages (Portuguese AIE tapers at layer 14); per-language ranges are a promising direction for stronger gains. Finally, the zero visual-token AIE is a design artefact of using the same image in both passes; paired cross-lingual images would probe whether visual tokens can independently carry pivot information.

6 Conclusion

LLaVA-1.5-7B routes non-English visual queries through an English-biased representational bottleneck in layers 5–17 (peak $\overline{\text{AIE}}$=0.49 at layer 8), inherited from the Vicuna backbone and unmitigated by multimodal pre-training. The pivot is image-content-dependent—language-specific representations require semantically rich visual content to emerge—but is mediated entirely through text-token positions. Training-free steering at the mechanistically identified layers yields +6.5 pp on Russian and +4.0 pp on Portuguese (surpassing English), with German vectors transferring perfectly—establishing the pivot as structural and providing a blueprint for mechanistic analysis of other VLMs.

References

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceed-*

ings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Yuchun Fan, Yilong Wang, Yongyu Mu, Lei Huang, Bei Li, Xiaocheng Feng, Tong Xiao, and Jingbo Zhu. 2025. Language-specific layer matters: Efficient multilingual enhancement for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research*, pages 7–25.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *Preprint*, arXiv:2203.14680.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. 2025. What do VLMs NOTICE? A mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, and 1 others. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2024. PALO: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

nostalgebraist. 2020. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. What is missing in multilingual visual reasoning and how to fix it. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2654–2667.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2025. Parrot: Multilingual visual instruction tuning. In *Proceedings of the 42nd International Conference on Machine Learning*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English? On the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15209–15228.

Andy Zou, Long Phan, Sarah Chen, and 1 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

A Probing Set Details

Table 5 lists the 15 COCO categories and their Portuguese, German, and Russian answer tokens. All answer words are tracked by their first sub-token under the Vicuna tokeniser, using contextual subtraction (encoding "ASSISTANT: {word}" and stripping the prefix tokens) to match generation-time tokenisation. The leading-space token (id 29871) is filtered before selecting the first sub-token.

B Implementation Details

Our code is publicly available at https://github.com/azraihan/mul-vlm-mech-interp.

Table 5: Object categories with PT, DE, and RU answer tokens. Multi-token words tracked by first sub-token only (†). All Russian (Cyrillic) entries are multi-token under the Vicuna tokeniser.

Category	PT	DE	RU
person	pessoa	Person	человек†
car	carro	Auto	машина†
bus	ônibus†	Bus	автобус†
train	trem	Zug	поезд†
boat	barco	Boot	лодка†
orange	laranja†	Orange	апельсин†
book	livro	Buch	книга†
clock	relógio†	Uhr	часы†
knife	faca	Messer†	нож†
spoon	colher†	Löffel†	ложка†
umbrella	guarda-chuva†	Regenschirm†	зонт†
bird	pássaro†	Vogel†	птица†
cat	gato	Katze†	кот†
dog	cachorro†	Hund	собака†
apple	maçã†	Apfel†	яблоко†

Sequence alignment. Non-English questions are 2–5 tokens longer than English counterparts. We keep the clean (non-English) input at full length and truncate only the corrupted (English) input to `min_len` to preserve the correct `answer_pos`.

Pre-hook patching under transformers 5.0. `GradientCheckpointingLayer` wrappers in transformers 5.0 discard post-hook return values, so we register patching as a *pre-hook* on layer $\ell{+}1$ rather than a post-hook on layer ℓ. For layer $\ell{=}31$, the pre-hook is on `model.language_model.model.norm`.

Response-token target and steering. τ_{en} is determined dynamically via one-token greedy decoding on the English input (consistently "The", id 450). Steering applies $\mathbf{h}_L \mathrel{+}= \alpha \cdot \mathbf{v}_L^{(\ell)}$ at the last active token position at each decode step via a forward hook on `model.language_model.model.layers[L]`, covering both prefill and autoregressive steps under KV-cache.