Cross-Attention Eigenvalues Know the Word Order in Multilingual Machine Translation

Anonymous ACL submission

Abstract

Multilingual machine translation models must adapt their generation strategy to the grammatical structure of the target language, yet how structural information shapes decoder computation remains unclear. We show that cross-attention geometry in NLLB-200 is organized around the target language: eigenvalues of per-layer-head Gram matrices carry a strong, distributed signal about target-language word order (SVO vs. SOV) when English is the source, but this signal collapses to near chance at the layer-head level when English is the target. This asymmetry suggests that cross-attention geometry reflects the demands of target-side generation rather than source-side structure. Leave-one-language-out evaluation confirms that probe failures track corpus-level word-order distributions rather than typological labels, indicating that the signal is grounded in actual generation patterns.

1 Introduction

Multilingual neural machine translation models must implicitly represent the structural properties of dozens or hundreds of languages simultaneously. Among the most fundamental of these properties is word order: whether a language places the verb before or after its object (SVO vs. SOV) has pervasive consequences for how source and target sequences align during generation. Locating this typological information within the model is essential for the interpretability of multilingual NLP models.

Prior work on probing transformer representations has focused predominantly on encoder-side models (Conneau et al., 2018; Belinkov, 2022) or on self-attention in encoder-decoder architectures (Raganato and Tiedemann, 2018). The cross-attention mechanism of the decoder, which mediates the interface between source representations and target generation, has received comparatively little attention as a locus of typological information. It is also unclear whether word-order signals, if present, are carried by specific layer-head pairs or distributed across the network.

We probe the cross-attention mechanism of NLLB-200 (NLLB Team et al., 2022) across three model sizes and 22 typologically diverse languages to investigate whether cross-attention geometry encodes word-order typology, and if so, in which layers and heads, and under which translation conditions.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes the experimental setup. Section 4 presents the results. Section 5 concludes the paper. Section 6 and Section 7 discuss limitations and ethical considerations respectively. Our implementation is available online.[1]

2 Related Work

Probing transformer representations. The probing classifier framework (Conneau et al., 2018; Belinkov, 2022) has been widely used to study what linguistic information is encoded in neural model representations. Applied to NMT, Belinkov et al. (2017) showed that encoder representations capture morphological information, while decoder representations are weaker on this task. Raganato and Tiedemann (2018) probed encoder representations in transformer-based MT and found that higher layers encode more semantic information. Xu et al. (2021) showed that word translation occurs progressively in encoder layers, with cross-attention in the decoder serving primarily as a sequential alignment mechanism. Our work extends this line of research to the cross-attention geometry of a massively multilingual model, targeting typological rather than lexical information.

Attention analysis. Voita et al. (2019) showed that a small subset of self-attention heads in NMT

[1] https://anonymous.4open.science/r/Word-Order-ARR

encoders performs most of the useful work, with the remainder prunable without performance loss. Zhang et al. (2023) analyzed functional head specialization in a multilingual MT transformer and found that different language pairs share the same heads for the same functions, with deep cross-attention heads cooperating for word reordering. The question of whether attention weights constitute meaningful explanations has been debated: Jain and Wallace (2019) argued that attention distributions are largely uncorrelated with feature importance, while Wiegreffe and Pinter (2019) contested this claim. In contrast to the latter line of work, which analyzes attention weights directly, we use the eigenvalues of the Gram matrix of attention as features, capturing the geometric structure of the attention distribution rather than its raw values.

Word order in multilingual models. Nikolaev and Pado (2022) demonstrated that mBERT's zero-shot performance on subordinate-clause detection is dominated by SVO/SOV typology. Otmakhova et al. (2022) showed that typologically different languages encode morphosyntactic features differently across BERT layers. Ji et al. (2023) showed that typologically similar languages share similar positional representations in a multilingual dependency parser, with SVO/SOV typology correlating with the structure of the positional space. Brinkmann et al. (2025) found that LLMs develop cross-lingual abstractions of morphosyntactic concepts shared across typologically diverse languages. Our work differs from all of the above in that we target the decoder's cross-attention in a generative MT model rather than encoder representations, and we explicitly control for the direction of translation to isolate source-side from target-side effects.

3 Experimental Setup

3.1 Model

We experimented with three sizes of NLLB-200 (NLLB Team et al., 2022): 600M, 1.3B, and 3.3B parameters. All models share the same encoder-decoder Transformer architecture and vocabulary, differing only in depth and size of hidden dimension.

3.2 Data

We used the FLORES-200 development set (NLLB Team et al., 2022), which provides professionally translated sentences across all 200 NLLB languages. We selected 22 languages paired with English: 11 SVO (French, German, Thai, Indonesian, Chinese, Russian, Vietnamese, Portuguese, Swahili, Finnish, Swedish) and 11 SOV (Japanese, Korean, Turkish, Persian, Hindi, Georgian, Tamil, Azerbaijani, Mongolian, Urdu, Bengali), giving 997 sentences per language. Word-order labels were taken from WALS (Dryer and Haspelmath, 2013).

3.3 Translation Modes

We defined two experimental modes to isolate the contribution of source versus target language to cross-attention geometry. In the **forward direction** (eng→X), English was the fixed source and the target language varied across the 22 languages. In the **reverse direction** (X→eng), English was the fixed target and the source language varied.

3.4 Feature Extraction

At each decoding step t, the cross-attention module produced a tensor of shape $[L, H, T, S]$, where L is the number of layers, H the number of heads, T the number of generated tokens, and S the number of source tokens. For each layer l and head h, we extracted the full attention matrix $\mathbf{A} \in \mathbb{R}^{T \times S}$ and formed the Gram matrix $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{T \times T}$. We computed the top-$k$ eigenvalues of this matrix, normalized them by their sum, and concatenated the resulting vectors across all $L \times H$ layer-head pairs to form the sentence-level feature vector. We set $k = 12$ throughout.

3.5 Probing Classifier

We trained a logistic regression probe (Belinkov, 2022) with standardized features (`StandardScaler`) to predict a binary label: $y = 1$ if source and target language share word order (both SVO), $y = 0$ otherwise. A linear probe is deliberately chosen to ensure that any recovered signal reflects structure present in the features rather than the capacity of the classifier.

We used two evaluation protocols: **cross-validation** with StratifiedGroupKFold (groups = sentence ID) with 5 folds averaged over 3 seeds, measuring generalization to unseen sentences from the same languages; and **leave-one-language-out** (LOO), where the probe is trained on 21 languages and evaluated on the held-out one, testing generalization to unseen languages and ruling out language identity as a confound.

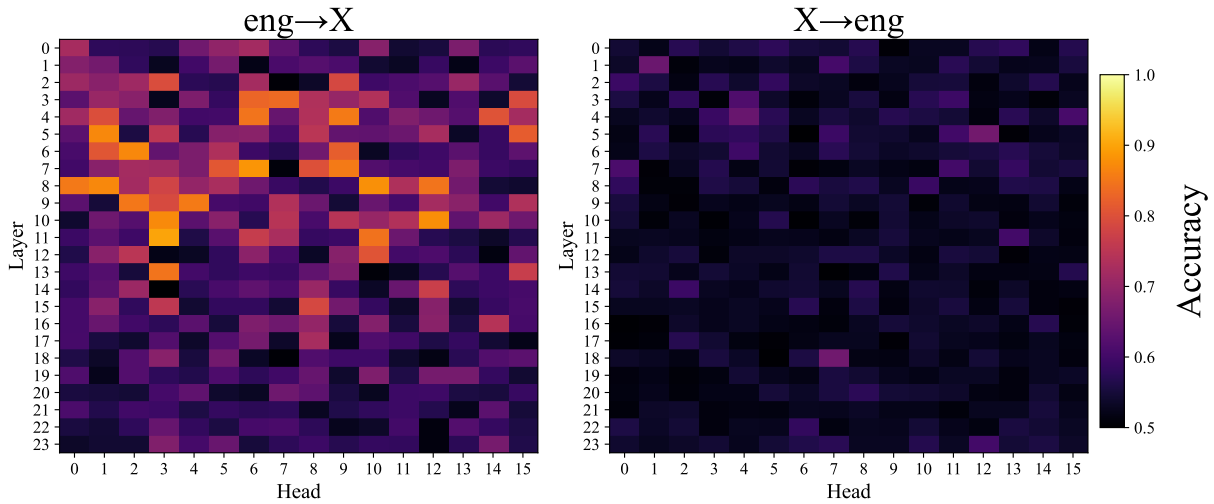

Figure 1: Probe accuracy per (layer, head) averaged over 5 folds and 3 seeds for the 3.3B model. The eng→X map (left) shows a sparse but distinct signal, while the signal in the X→eng map (right) is virtually non-existent.

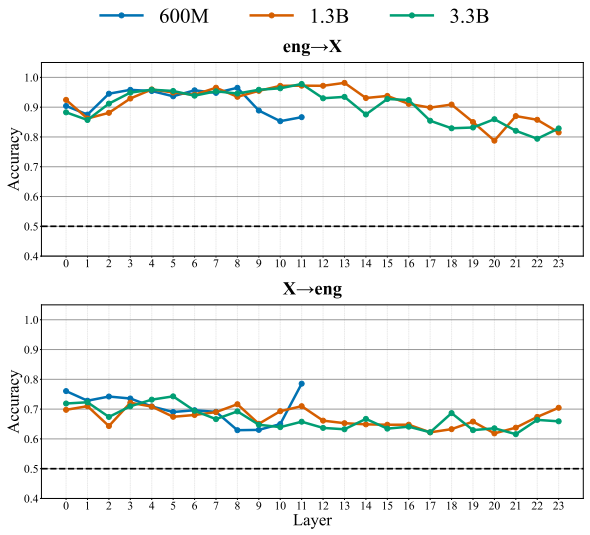

Figure 2: Layer-wise probe accuracy averaged over 5 folds and 3 seeds for eng→X (top) and X→eng (bottom) translation across three model sizes.

4 Results and Discussion

4.1 Cross-Validation

Cross-validation accuracy exceeded 0.999 across all three model sizes in the forward direction, indicating that the eigenvalue features carry a reliable signal; generalization to unseen languages is assessed via LOO below. In the reverse direction, accuracy remained substantial when features were aggregated across all layer-head pairs: 0.904 (600M), 0.906 (1.3B), and 0.919 (3.3B).

Per-layer profiles. In the forward direction, the per-layer accuracy profiles showed a clear depth-dependent structure (Figure 2). In the 600M model, accuracy remained high across all layers with a modest decline toward the final layers. In the 1.3B and 3.3B models, accuracy rose to a broad plateau in the middle layers, reaching a peak around layer 13, then degraded in the final layers, suggesting that word-order structure is most legible in the middle of the network. In the reverse direction, no pronounced mid-layer peak was observed: accuracy remained below 0.8 across all layers and declined gradually with depth.

Per-head profiles. The per-head accuracy profiles were flat across all three model sizes in the forward direction: every head achieved high accuracy with only minor variation, and no head consistently outperformed or fell substantially below the others. The 3.3B profile was both higher and flatter than in the smaller models. In the reverse direction, individual heads fluctuated between roughly 0.62 and 0.76 with no head standing out as a consistent carrier of the signal.

Per-(layer, head) structure. The joint accuracy heatmaps showed no consistent structure: high-accuracy layer-head pairs were scattered across the grid without forming coherent horizontal or vertical bands (Figure 1). This indicates that word-order information is not concentrated in a dedicated subset of layer-head pairs but is distributed across the cross-attention mechanism as a whole.

4.2 Leave-One-Language-Out Analysis

The signal generalizes well to unseen languages: in the forward direction, the majority of languages achieved accuracy at or above 0.99 across all three

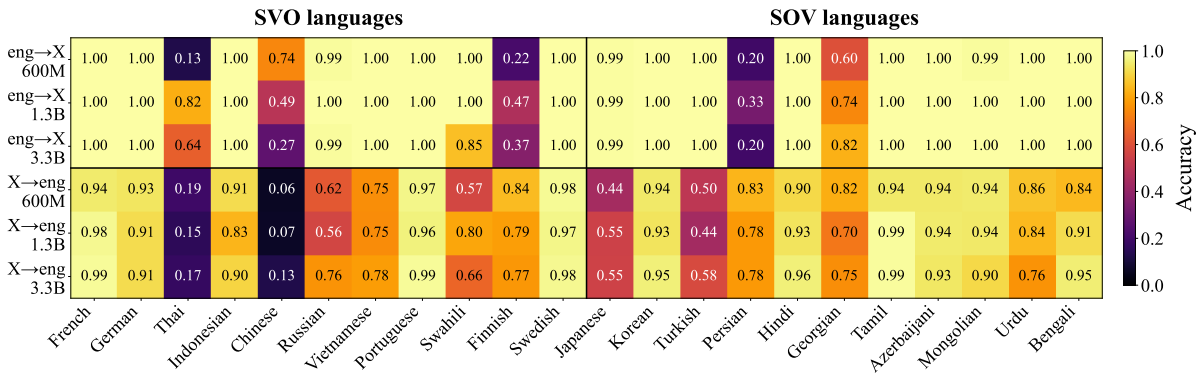

Figure 3: Leave-one-language-out accuracy for all six experimental conditions (three model sizes × two translation directions). Each cell shows the probe accuracy when the corresponding language is held out from training.

model sizes (Figure 3). In the reverse direction, accuracy was lower but remained high for most languages, typically between 0.84 and 0.99. The outliers are interpretable in terms of known linguistic properties and collectively suggest that the probe tracks corpus-level word-order distributions rather than typological labels assigned by WALS.

Chinese and Thai were outliers in both translation directions. In the forward direction, Chinese accuracy degraded monotonically with model size, while Thai was unstable across models. In the reverse direction, both languages showed consistently low accuracy, consistent with their status as highly analytic isolating languages in which individual tokens are not constrained to a single grammatical role.

In the forward direction, Finnish (SVO) was misclassified as SOV across all model sizes, and Persian (SOV) was misclassified as SVO, because their surface order in FLORES-200 diverges from their WALS labels: Finnish frequently uses SOV order for pragmatic reasons, and written Persian is predominantly SVO in news-domain text. Georgian improved monotonically with model size in the forward direction, suggesting that its complex morphosyntax requires greater model capacity to produce a geometrically distinctive signal. Japanese and Turkish achieved near-perfect accuracy in the forward direction but fell near chance in the reverse direction, consistent with the overall asymmetry.

Finally, writing direction is not a confound. Persian (RTL) is a LOO outlier, but Urdu — also RTL and typologically similar — generalizes well. This confirms that the misclassification of Persian is due to its atypical corpus-level word order rather than any artifact of right-to-left script processing.

4.3 Geometric Properties of the Signal

The signal is strongly asymmetric across translation directions: it is strong and layer-structured in the forward direction, but degrades at every layer-head pair in the reverse direction. Since the same features and probe are used in both settings, and the effect persists under leave-one-language-out evaluation, this asymmetry is consistent with the hypothesis that cross-attention geometry is shaped by target-side generation: producing a language with a particular word order induces structured attention geometry, whereas this target-side influence is largely uniform when English is the target.

That the signal is recoverable from the spectrum of $\mathbf{A}\mathbf{A}^\top$ alone—without token identities or raw attention weights—suggests that word-order information is encoded in the global shape of the attention distribution rather than in specific alignments. Qualitatively, SVO targets exhibit a dominant diagonal trend, while SOV generation results in a more dispersed, noisy distribution.

5 Conclusion

We investigated whether the geometry of cross-attention in NLLB-200 encodes target-language word order, using eigenvalues of per-layer-head Gram matrices as features for a linear probe. We found that the signal is strong, distributed across the network, and driven by the target language: it is highly accurate when English is the source and collapses at the layer-head level when English is the target. Leave-one-language-out evaluation showed that generalization failures are linguistically interpretable, tracking corpus word-order distributions rather than typological labels.

6 Limitations

Due to space constraints, we focus on a fixed feature set consisting of the top-k eigenvalues with $k = 12$. While we did not perform an exhaustive sensitivity analysis, preliminary experiments with $k \in \{4, 8, 12\}$ showed that classification accuracy remains high and stable, suggesting that our findings are robust to the specific choice of k. For the same reason, we omit comparisons with simpler feature baselines (e.g., raw attention statistics); however, eigenvalue spectra are specifically chosen as they capture global geometric structures that are not directly accessible from token-level summaries. Finally, our analysis is limited to 22 languages paired with English. Whether these findings generalize to non-English-centric translation pairs remains an open question for future work.

7 Ethics Statement

This work uses publicly available models (NLLB-200 licensed by CC-BY-NC 4.0) and datasets (FLORES-200 licensed by CC-BY-SA 4.0). No human subjects were involved. We foresee no ethical concerns.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Ji, Yuanbin Wu, and Xiaoling Wang. 2023. Typology guided multilingual position representations: Case on dependency parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13524–13541, Toronto, Canada. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Pado. 2022. Word-order typology in multilingual BERT: A case study in subordinate-clause detection. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 11–21, Seattle, Washington. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yulia Otmakhova, Karin Verspoor, and Jey Han Lau. 2022. Cross-linguistic comparison of linguistic feature encoding in BERT models for typologically different languages. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 27–35, Seattle, Washington. Association for Computational Linguistics.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2021. Probing word translations in the transformer and trading decoder for encoder layers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–85, Online. Association for Computational Linguistics.

Jingyi Zhang, Gerard de Melo, Hongfei Xu, and Kehai Chen. 2023. A closer look at transformer attention for multilingual translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 496–506, Singapore. Association for Computational Linguistics.