NRD: A Hybrid Disentanglement Framework for Mitigating Interference in Multilingual Machine Translation

Jiarui Zhang[1,2], **Yifan Deng**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
19 Shucun Road, Haidian District, Beijing 100085, P. R. China
[2]School of Cyber Security, University of Chinese Academy of Sciences
{zhangjiarui, dengyifan}@iie.ac.cn

Abstract

Negative interference from cross-lingual conflicting syntactic patterns is a primary obstacle in Multilingual Neural Machine Translation (MNMT). We trace this problem to the entanglement of transferable, universal semantics with non-transferable, language-specific syntactic structures. Existing methods, relying on disjoint training-only specialization or inference-only filtering, fail to fully resolve this fundamental entanglement. To address this, we propose NRD (Neuron Representation Disentanglement), a two-stage hybrid framework that couples training-time specialization with inference-time filtering. First, a Specialization Fine-tuning stage identifies functional neurons via a semantic-invariant activation-variance metric and reinforces intrinsic modularity through sparse updates. Second, a Dynamic Representation Filtering stage purifies semantic representations at inference by adaptively suppressing syntax-sensitive neurons, guided by each language's pre-computed gradient consistency. On the OPUS-100 benchmark, NRD outperforms strong baselines, achieving an average gain of +1.9 BLEU on supervised directions. On the WMT-10 zero-shot benchmark, it obtains a substantial +7.1 BLEU, demonstrating robust cross-lingual generalization. These results provide strong evidence that our hybrid approach effectively purifies semantic representations by mitigating syntactic interference, paving the way for more robust cross-lingual generalization.

Keywords: Multilingual NMT, Negative Interference, Neuron Disentanglement, Representation Learning, Cross-Lingual Transfer

1. Introduction

Multilingual Neural Machine Translation (MNMT) transfers knowledge across languages through shared parameters (Johnson et al., 2017; Arivazhagan et al., 2019) but still suffers from negative interference—when typologically distant languages share encoder representations, conflicting syntactic patterns degrade translation quality (Conneau et al., 2020). Even strong models such as XLM-T (Ma et al., 2020) lose several BLEU points on high-divergence pairs, revealing that multilingual sharing often entangles transferable semantics with language-specific syntax (Aharoni et al., 2019).

Recent neuron-level studies (Tan et al., 2024; Huo et al., 2024; Kim et al., 2024) treat cross-lingual interference as a representation-level entanglement rather than an architectural limitation. Two strategies have emerged: (1) Training-time specialization, which restricts updates to language-relevant neurons; and (2) Inference-time filtering, which prunes syntax-sensitive activations during decoding. However, these are disjoint—training-only methods shape structure without maintaining it, while inference-only methods clean activations without grounding. Each mitigates interference partially but cannot sustain semantic purity beyond training.

We posit that effective multilingual generalization requires training–inference coupling: disentanglement must be encoded during training and enforced during inference through a consistent neuron partition.

To this end, we propose NRD (Neuron Representation Disentanglement), a hybrid framework linking specialization and filtering via shared functional neurons. In Stage 1 (Specialization Fine-tuning), neurons are identified by paraphrase-level activation variance, separating semantic-stable from syntax-sensitive units. In Stage 2 (Dynamic Representation Filtering), the same partition guides inference-time suppression using gradient consistency as a language-specific controller. This coupling aligns parameter updates with activation behavior, maintaining semantic clarity across languages. **Our contributions are threefold**:

- We propose the first coupled neuron-centric framework unifying specialization and filtering through shared neuron partitions.
- We show that activation variance and gradient consistency provide complementary, interpretable signals for disentanglement.
- NRD achieves a +1.9 BLEU average gain on OPUS-100, and a substantial +7.1 BLEU gain on the WMT-10 zero-shot benchmark. This result provides strong evidence that our hybrid

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8577–8586
11–16 May 2026.

co-design succeeds in mitigating syntactic interference, not just improving general generalization.

NRD demonstrates that maintaining representational purity requires coordinated design across optimization and inference—a training–inference co-design paradigm for future multilingual systems.

2. Related Work

Our work advances multilingual parameter sharing by adopting the Neuron-Centric Paradigm. We shift the focus from macro-level parameter allocation to the micro-level behavior of individual neurons, contrasting with prior approaches that either statically partition parameters or quantify interference without intervention.

2.1. Parameter Sharing vs. Isolation

Early MNMT systems used full parameter sharing (Johnson et al., 2017; Aharoni et al., 2019), which led to negative interference. In response, methods like Adapters (Pfeiffer et al., 2020) introduced lightweight, language-specific modules to isolate parameters. While effective, these approaches rely on heuristic capacity allocation (Xie et al., 2021) and impose rigid language partitions that limit transfer. Later work such as CaPA (Huo et al., 2024) measures gradient similarity to adapt the amount of capacity granted to each language, yet the intervention still happens only during fine-tuning. Other studies quantify interference without prescribing how to intervene in representations (Wang et al., 2020a). NRD differs by regulating both *how* multilingual representations are formed during specialization and *how* they are filtered at inference: the same neuron partition guides sparse updates and the subsequent activation control, eliminating the gap between training-time budgeting and inference-time behaviour.

2.2. Neuron-Centric Paradigm

Recent neuron-centric methods have focused on either training-time specialization or inference-time pruning, but not both.

1). Static Specialization vs. Dynamic Filtering: (Tan et al., 2024) identify sparse, language-specific sub-networks via activation frequency and update them during training, but the selected neurons remain fixed afterwards. CaPA (Huo et al., 2024) similarly uses gradient consistency to grant extra capacity to cooperative languages, yet it still lacks a mechanism to adapt activations once decoding begins. NRD instead keeps the specialization signal alive at inference by reusing gradient consistency as a control knob for activation filtering.

2). Heuristic Importance vs. Functional Identification: Frequency- or magnitude-based criteria (Tan et al., 2024; Kim et al., 2024) deem neurons ”important” when they fire often or with large amplitudes, regardless of whether they track syntax or semantics. Our variance-based metric distinguishes neurons by their functional role—stable semantics versus syntax-sensitive fluctuations—providing the necessary partition for coordinated specialization and filtering.

3). Universal vs. Language-Specific Pruning: (Kim et al., 2024) prune a universal set of ”non-translation” neurons, assuming a single set of important features for all languages. This ignores that interference severity varies across languages and even across sentences. NRD conditions the masking strength on language-specific gradient consistency, yielding per-language control without retraining.

NRD reinterprets these techniques within a unified training-and-inference framework. Unlike static methods, we repurpose gradient consistency as an inference-time control knob that dynamically modulates syntax-sensitive neurons identified via a semantics-grounded variance metric. This co-design keeps disentanglement both structurally embedded and contextually enforced, rather than relying on isolated interventions.

3. The NRD Framework

To address the entanglement of semantic and syntactic representations in Multilingual Neural Machine Translation (MNMT), we propose NRD (Neuron Representation Disentanglement), a two-stage hybrid framework that synergizes training-time specialization with inference-time adaptation. In simple terms, this means we first train the model to better separate meaning from structure, and then refine its outputs during use to reduce language-specific noise. This section outlines the framework’s design, detailing its two core stages: Specialization Fine-tuning and Dynamic Representation Filtering. We start with an overview, followed by a step-by-step explanation of each stage, with formal definitions and algorithms for clarity.

3.1. Overview

NRD maintains disentangled multilingual representations through a two-stage interaction between training-time neuron specialization and inference-time adaptive filtering. In this work, we focus on the feed-forward networks (FFNs) within the Transformer’s encoder layers. Here, the term ”neuron” refers to a single hidden unit within the

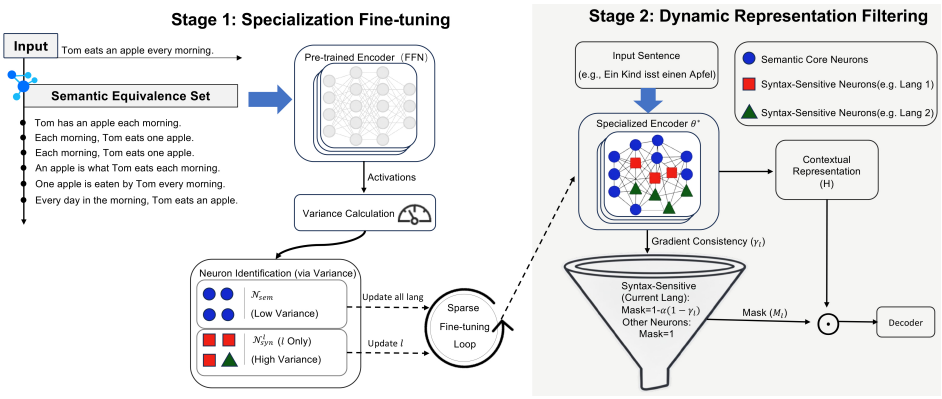

Figure 1: Overview of the NRD framework. The Specialization Fine-tuning stage (left) identifies semantic and syntax-sensitive neurons via activation variance and updates them sparsely. The Dynamic Representation Filtering stage (right) applies a language-specific mask to purify representations at inference, guided by gradient consistency.

FFN's intermediate activation layer—essentially, these are the building blocks that process information in the model. Our method operates on the activations (outputs) of these individual units across the input sequence. The NRD framework aims to disentangle transferable, universal semantics (core meaning that works across languages) from non-transferable, language-specific syntactic structures (like word order or grammar rules unique to one language). As illustrated in Figure 1, NRD operates in two complementary stages:

1). Specialization Fine-tuning: This offline stage reinforces the model's natural modularity by identifying and strengthening neuron groups handling semantics versus syntax, using a variance-based metric to spot stability or sensitivity.

2). Dynamic Representation Filtering: This online stage cleans up the model's output by adaptively suppressing syntax-related noise, based on a pre-calculated measure of how much a language "interferes" with others.

Our framework relies on two complementary metrics precisely because the two stages answer questions at different granularities:

1). Stage 1 (Identification): Requires a neuron-level metric. We use activation variance to answer: "What is this neuron's function?" (i.e., distinguishing semantic core from syntax-sensitive).

2). Stage 2 (Control): Requires a language-level metric. We use gradient consistency to answer: "How much interference does this language cause?" (i.e., determining the filtering strength).

In short, our hybrid framework first uses variance to precisely identify what to filter and gradient consistency to dynamically decide how much to filter during training—while also building disentanglement into model parameters then, and actively maintaining this separation during inference to ease negative interference.

3.2. Stage 1: Specialization Fine-tuning

The first stage enhances the model's built-in modularity by embedding disentanglement directly into its parameters through selective, sparse updates. Think of this as "specializing" parts of the model: we identify key neuron groups and update them in a targeted way to promote clear separation between shared meaning and language-specific details.

3.2.1. Identifying Functional Neurons via Activation Variance

We identify functional neurons by measuring activation variance across paraphrases that preserve meaning but alter syntax. Semantic invariance implies that meaning-bearing neurons should remain stable, whereas syntax-tracking neurons will fluctuate. Paraphrase sets are generated once per language via controlled back-translation and reused during training without additional cost. This ensures syntactic variety (e.g., different word orders) while keeping the meaning intact. For example, take the English sentence "The child eats an apple", paraphrases might include:

1). "An apple is eaten by the child" (passive voice).

2). "The kid is consuming the fruit" (synonym substitution and word order shift).

For a multilingual model's FFN layers, let $a_j(s_i^l)$ denote the activation of neuron j for paraphrase s_i^l in the equivalence set of size k for language l. We measure variance to capture stability:

$$\text{Var}_j(l) = \frac{1}{k}\sum_{i=1}^{k}(a_j(s_l^i) - \bar{a}_j(l))^2, \quad \text{where} \quad \bar{a}_j(l) = \frac{1}{k}\sum_{i=1}^{k} a_j(s_l^i). \tag{1}$$

In plain terms, this variance tells us how much a neuron's output "wiggles" across equivalent sentences. Low variance means stability (semantic core), high variance means sensitivity to changes (syntax-related).

Compared with frequency-based heuristics that merely count how often a neuron fires, the variance criterion explicitly checks whether a neuron reacts to paraphrase-level perturbations that preserve meaning. A unit that triggers frequently because of a dominant word order or morphological suffix will exhibit large variance when that surface cue is altered, and is therefore routed to the syntax-sensitive pool. Semantic neurons, in contrast, must remain stable across these controlled transformations to receive a low-variance score. The sensitivity of our method to paraphrase quality in Table 4 mirrors this behaviour: once semantic equivalence is broken, the variance signal collapses, whereas frequency counts would stay unchanged.

Neurons in the bottom p_{sem} percentile (e.g., 20%) are classified as semantic core neurons (N_{sem}), stable across variations. Those in the top p_{syn} percentile (e.g., 20%) are syntax-sensitive for language l (N^l_{syn}). This metric is more grounded in linguistics than frequency-based methods (Tan et al., 2024), directly linking to real language functions for better disentanglement. Specific implementation details are provided in Section 4.4.1.

Variance is a necessary condition for semantic neurons. Let $\mathcal{P}(x)$ denote a semantic equivalence set around source sentence x, containing paraphrases that differ only through syntactic operators (voice, agreement, reordering). If neuron j encodes purely semantic content, its activation can be written as $a_j(s) = f_j(\mathcal{M}(s))$, where $\mathcal{M}(s)$ maps a sentence to its meaning representation (constant within $\mathcal{P}(x)$). It follows that $a_j(s)$ is constant for all $s \in \mathcal{P}(x)$ and therefore $\text{Var}_j(l) = 0$. Conversely, suppose $\text{Var}_j(l) = 0$ on a set that includes a syntactic transformation T with non-zero Jacobian $\partial T/\partial s$. By the mean value theorem, $a_j(T(s)) - a_j(s) = \nabla a_j(\xi)^\top (T(s) - s) = 0$, implying $\nabla a_j(\xi)$ is orthogonal to every syntactic direction. Hence the neuron's gradient with respect to those transformations vanishes, and the neuron cannot encode syntax-sensitive signals. Variance thus acts as a necessary certificate: only neurons with near-zero paraphrase variance can represent transferable semantics, while high-variance neurons must respond to syntactic cues.

3.2.2. Sparse Specialization Training

Having identified $\mathcal{N}_{\text{sem}}$ and $\mathcal{N}^l_{\text{syn}}$ for each language, we fine-tune the model to strengthen this modularity. For a training batch from language l:

1). Semantic core neurons ($\mathcal{N}_{\text{sem}}$): Update parameters linked to these for all languages, fostering shared semantic knowledge that transfers across them.

2). Syntax-sensitive neurons($\mathcal{N}^l_{\text{syn}}$): Update parameters only for batches in language l, freezing them for others to avoid cross-language conflicts.

We implement this by masking gradients during standard multilingual training. After calculating gradients for a batch from language l, we apply a binary mask: 1 for parameters in $\theta_{sem} \cup \theta^l_{syn}$ (update) and 0 otherwise (freeze). This masked gradient is then used in the optimizer step, creating sparse, targeted updates.

Importantly, this strategy preserves and boosts cross-lingual transfer. All shared knowledge flows through the semantic core neurons (N_{sem}), updated universally, building a language-neutral meaning space. Syntax-sensitive neurons (N^l_{syn}) capture unique patterns without spillover, allowing the decoder to generate accurate outputs while leveraging the clean semantics.

3.3. Stage 2: Dynamic Representation Filtering

The second stage keeps representations disentangled during inference by dynamically filtering out syntax-sensitive elements, tailored to the source language's interference level. This acts like a "clean-up" filter, adjusting based on how conflicting a language is.

3.3.1. Quantifying Interference with Gradient Consistency

We introduce gradient consistency as a measure of how well the training signals from a particular language align with the overall multilingual objective. For each language l, compute the gradient of the loss $\mathcal{L}_l$ with respect to the model's encoder parameters θ_{enc} on a validation set $\mathcal{V}_l$. The gradient consistency γ_l is defined as the cosine similarity between its gradient g_l and the average gradient across all languages $\bar{g}$:

$$\gamma_l = \cos(g_l, \bar{g}) = \frac{g_l \cdot \bar{g}}{\|g_l\| \|\bar{g}\|}. \tag{2}$$

Cosine similarity measures angle between vectors: close to 1 means alignment (low interference), close to -1 means opposition (high interference). This γ_l captures a language's inherent properties, making it stable and robust to small data shifts between validation and testing.

3.3.2. Adaptive Filtering at Inference

At inference, for an input in language l, the encoder outputs a representation H. Before decoding, we apply a language-specific mask M_l to dampen syntax-sensitive neurons:

$$M_l[j] = \begin{cases} 1, & \text{if } j \in \mathcal{N}_{\text{sem}}, \\ 1 - \alpha \cdot (1 - \gamma_l), & \text{if } j \in \mathcal{N}^l_{\text{syn}}, \\ 1, & \text{otherwise.} \end{cases} \quad (3)$$

The filtered output is $H_{filtered} = H \odot M_l$ (element-wise multiplication).

In essence, this mask leaves semantic core activations untouched (value 1). For syntax-sensitive ones, it scales down based on interference: low γ_l (high conflict) means stronger suppression (closer to $1 - \alpha$). The "otherwise" preserves neurons sensitive in other languages, avoiding over-pruning shared traits. We focus suppression on the current source's syntax to purify the signal. The full process is outlined in Algorithms 1 (offline training and profiling) and 2 (online inference).

Algorithm 1 NRD: Offline Procedure for Training and Profile Generation

Input: Pre-trained model θ, dataset $\{\mathcal{D}_l\}_{l \in \mathcal{L}}$, validation sets $\{\mathcal{V}_l\}_{l \in \mathcal{L}}$, percentile thresholds $p_{\text{sem}}, p_{\text{syn}}$.
Output: Fine-tuned model θ^*, identified neuron sets $\mathcal{N}_{\text{sem}}$, $\{\mathcal{N}^l_{\text{syn}}\}_{l \in \mathcal{L}}$, interference tendencies $\{\gamma_l\}_{l \in \mathcal{L}}$.
Stage 1: Specialization Fine-tuning
Compute variance $\text{Var}_j(l)$ for each neuron j and language l using semantic equivalence sets.
Identify $\mathcal{N}_{\text{sem}}$ and $\{\mathcal{N}^l_{\text{syn}}\}_{l \in \mathcal{L}}$ based on variance percentiles.
while not converged **do**
    Sample a batch from a language $l \in \mathcal{L}$.
    Compute loss and gradients.
    Update only parameters in $\theta_{\text{sem}} \cup \theta^l_{\text{syn}}$ by masking the gradients.
end while
Stage 2: Pre-computing Interference Profiles
for each language $l \in \mathcal{L}$ **do**
    Compute gradient g_l on $\mathcal{V}_l$ and interference tendency γ_l using Eq. 2.
end for
return $\theta^*, \mathcal{N}_{\text{sem}}, \{\mathcal{N}^l_{\text{syn}}\}, \{\gamma_l\}$

Algorithm 2 NRD: Online Inference Procedure

Input: Input sentence x_l, fine-tuned model θ^*, neuron sets $\mathcal{N}_{\text{sem}}$, $\{\mathcal{N}^l_{\text{syn}}\}_{l \in \mathcal{L}}$, interference tendencies $\{\gamma_l\}_{l \in \mathcal{L}}$, hyperparameter α.
Output: Filtered representation H_{filtered}.
Compute representation $H = f_{\text{enc}}(x_l; \theta^*)$.
Construct mask M_l based on γ_l and α using Eq. 3.
return $H_{\text{filtered}} = H \odot M_l$ for decoding.

4. Experimental Setup

4.1. Datasets and Tasks

To ensure fair and direct comparison with established baselines, our experimental setup closely follows the methodology of Ma et al. (2020). We conduct comprehensive evaluations on two large-scale multilingual benchmarks: OPUS-100 and WMT-10.

4.1.1. Primary Benchmark: OPUS-100

Zhang et al. (2020) introduced OPUS-100 as a standard benchmark for **massively multilingual** machine translation. We adopt this English-centric dataset as our primary training and evaluation corpus, as it provides ideal conditions for analyzing cross-lingual transfer across diverse linguistic families.

Following Ma et al. (2020), we evaluate on 94 language pairs with approximately 55M training sentences and 2,000 test sentences per language. Languages are categorized following Zhang et al. (2020) into high-resource (≥0.9M pairs, 45 languages), medium-resource (0.1M-0.9M, 28), and low-resource (<0.1M, 21).

4.1.2. Complementary Benchmark: WMT-10

We complement our evaluation with the WMT-10 dataset (Wang et al., 2020b) to specifically assess NRD's effectiveness in mitigating syntactic interference across typologically diverse languages. This benchmark includes ten languages (Fr, Cs, De, Fi, Lv, Et, Ro, Hi, Tr, Gu) with substantial syntactic variation relative to English, particularly between SVO and SOV structures. We evaluate the performance of zero-shot translation on this dataset (X $\rightarrow$ X, excluding English, 90 directions).

4.2. Baseline

All experiments build upon the same strong foundation model XLM-T (Ma et al., 2020) to ensure fair comparison and demonstrate NRD's ability to enhance pre-trained multilingual systems. To com-

Model	OPUS-100 (X → En)				OPUS-100 (En → X)			
	High	Medium	Low	Avg_{94}	High	Medium	Low	Avg_{94}
MNMT (from scratch)	31.1	32.8	32.3	31.9	23.9	28.1	28.7	26.2
XLM-T (Foundation Model)	32.4	35.9	36.4	34.3	26.1	30.9	31.0	28.6
+ Adapters	32.6	36.4	36.9	34.7	26.4	31.4	31.9	29.1
+ Neuron Specialization	33.6	37.0	37.2	35.4	27.8	31.9	32.2	30.0
+ CaPA	33.4	37.3	37.1	35.3	28.3	31.9	32.0	30.1
+ NRD (Ours)	**33.9**	**37.7**	**37.9**	**35.9**	**28.9**	**32.2**	**32.8**	**30.8**

Table 1: Main translation performance (BLEU scores) on the OPUS-100 test sets. NRD significantly outperforms all baselines, demonstrating enhanced cross-lingual transfer. Best results are in **bold**.

Model	BLEU Score
MNMT (from scratch)	10.1
XLM-T (Foundation Model)	15.7
+ Adapters	16.6
+ Neuron Specialization	19.1
+ NRD (Ours)	**22.8**

Table 2: Zero-shot translation performance on the WMT-10 dataset (The average across 90 non-English directions).

prehensively evaluate the effectiveness of NRD, we compare it against the following models:

Multilingual Transformer (from scratch): A standard Transformer trained from scratch on OPUS-100. This baseline quantifies the gains from pre-training and provides a task difficulty reference.

XLM-T + Adapters (Bapna and Firat, 2019): We insert and train only language-specific Adapter modules into the frozen XLM-T backbone, comparing against a leading parameter-efficient method for mitigating interference.

XLM-T + Neuron Specialization (Tan et al., 2024): We apply this sparse fine-tuning method to XLM-T, providing a direct comparison against another state-of-the-art, training-phase, neuron-centric approach.

XLM-T + CaPA (Huo et al., 2024): We apply this gradient consistency-based parameter allocation method to XLM-T, comparing against a modern adaptive parameter allocation strategy.

4.3. Implementation Details

All models are implemented using the fairseq toolkit [1] (Ott et al., 2019). Our foundation model architecture is identical to the one used for the OPUS-100 experiments in (Ma et al., 2020) to ensure a fair comparison. We use Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$) with 4,000 warm-up steps and learning rate of $5e^{-4}$. All models employ label smoothing (0.1) and standard dropout (0.1). For OPUS-100's imbalanced data, we use temperature-based sampling ($T = 5.0$) with batch size of 2,048 tokens and gradient accumulation over 32 steps. We use beam search (size=5, length penalty=1.0), average last 5 checkpoints, and report case-sensitive BLEU using sacreBLEU [2] (Post, 2018).

4.4. NRD-Specific Settings

4.4.1. Stage 1: Specialization Fine-tuning

Semantic Equivalence Sets: Generate paraphrases for 100k sentences per language via XLM-T back-translation (nucleus sampling, $p = 0.9$)

Neuron Identification: Set $p_{\text{sem}} = 20\%$, $p_{\text{syn}} = 20\%$ based on grid search

4.4.2. Stage 2: Dynamic Representation Filtering

Gradient Consistency: Pre-compute γ_l on validation sets, averaged over 16 batches for stability

Filtering Strength: Set $\alpha = 0.5$ to balance interference mitigation and information preservation

5. Results and Analysis

In this section, we present the empirical results of our NRD framework, comparing it against the baselines described in Section 4. We first report the main translation performance on both supervised and zero-shot tasks, followed by a series of in-depth analyses to validate our claims regarding disentanglement, interference mitigation, and knowledge transfer.

[1] https://github.com/facebookresearch/fairseq

[2] BLEU+case.mixed+lang.{src}-{tgt}+numrefs.1+smooth.exp+tok.13a+version.1.4.14

5.1. Main Results

Table 1 presents the main results on the supervised OPUS-100 test sets, covering both X→En and En→X directions across all resource tiers. The findings clearly demonstrate the consistent and significant advantages of our NRD framework. As expected, the pre-trained XLM-T foundation model vastly outperforms the model trained from scratch, confirming the critical role of pre-training. While state-of-the-art neuron-centric methods like Neuron Specialization and CaPA build upon this strong baseline, our NRD framework achieves the highest BLEU scores in every single category. Overall, NRD obtains an average gain of +1.6 BLEU for X→En and +2.2 BLEU for En→X over the powerful XLM-T baseline (Average +1.9). Notably, NRD delivers substantial improvements for low-resource languages (+1.5 BLEU for X→En and +1.8 BLEU for En→X), suggesting our two-stage approach successfully mitigates negative interference and allows for more effective knowledge transfer to resource-poor languages.

To further challenge our model's ability to disentangle linguistic representations, we evaluate it on the WMT-10 zero-shot benchmark, which is specifically designed to test performance on typologically diverse and syntactically challenging language pairs. The results, presented in Table 2, are even more striking in this demanding setting. While the XLM-T baseline already shows strong performance, our NRD framework delivers a dramatic improvement, achieving a score of 22.8 BLEU. This represents a massive +7.1 BLEU gain over the XLM-T foundation model and a substantial +3.7 BLEU gain over the strongest competitor. This significant leap in performance on language pairs with severe syntactic conflicts provides the most compelling evidence that NRD successfully mitigates syntactic interference by filtering out conflicting neurons, leading to cleaner and more language-specific representations.

5.2. Ablation Study

To prove that both stages of our NRD framework are essential for achieving effective disentanglement, to validate the necessity of dynamic adaptation, and crucially, to demonstrate the effectiveness of our variance-based neuron identification strategy, we conduct a detailed ablation study. We analyze how each component contributes to the final performance, with results presented in Table 3.

NRD-FT (Specialization Fine-tuning only): Applying only Specialization Fine-tuning brings limited improvement. This indicates that parameter-level specialization lays the foundation for disentanglement, but the lack of explicit constraints during inference hinders the purity of representations.

Model Variant	OPUS-100 X→En	WMT-10 Zero-Shot
XLM-T	34.3	15.7
+ NRD-FT (only)	35.1	18.9
+ Static Masking	35.4	19.3
+ NRD-DRF (only)	34.9	18.6
+ NRD (w/ Frequency)	35.5	21.0
+ NRD (Full)	**35.9**	**22.8**

Table 3: Ablation study results (BLEU scores). Both Specialization Fine-tuning (NRD-FT) and Dynamic Representation Filtering (NRD-DRF) contribute significantly, with their synergistic combination (Full NRD) achieving the best performance.

NRD-FT + Static Masking: In this comparison, we use NRD-FT to detect syntax-sensitive neurons and then permanently mask their activations during inference—without any gradient-aware adaptation. This provides a strong baseline for static post-specialization filtering. Although it outperforms NRD-FT alone, it falls short of full NRD, demonstrating that global, fixed suppression of syntax-sensitive neurons is less effective than adaptive filtering.

NRD-DRF (Dynamic Representation Filtering only): Applying Dynamic Representation Filtering directly to the original XLM-T model also yields some gains in zero-shot scenarios, but the overall performance remains limited. This suggests that while filtering is effective, its potential is constrained by the underlying entangled representations.

NRD (w/ Frequency-based Neuron ID): This variant replaces our variance-based neuron identification with a frequency-based method (Tan et al., 2024) to select the N_{sem} neurons. Although this approach shows modest benefits, it underperforms the full NRD model. This confirms that while activation frequency correlates with modularity, our variance-based method more precisely targets neurons that encode syntax-invariant semantics, which is critical for effective disentanglement and cross-lingual transfer.

The results clearly demonstrate that neither training-time specialization nor inference-time filtering alone is sufficient. Moreover, merely identifying and statically filtering specialized neurons, or relying solely on activation frequency for identification, is suboptimal. Our synergistic and dynamic framework, anchored by a principled variance-based neuron identification strategy, is critical for producing a truly language-agnostic semantic representation, which is the key to the substantial gains observed in the zero-shot setting.

5.3. Impact of Semantic Equivalence Sets

The effectiveness of our neuron identification strategy hinges on the quality of the Semantic Equivalence Sets. We ablate this component by varying the sentence sets used to compute activation variance in Stage 1, with results in Table 4.

NRD (w/ Random Sets): Using semantically unrelated sentences causes performance to drop below the XLM-T baseline, confirming that semantic consistency is essential for meaningful variance signals.

NRD (w/ Identical Sentences): Using identical sentence copies leads to significant degradation, as near-zero variance prevents distinguishing semantic and syntactic neurons.

These results validate that NRD's success relies not merely on applying a variance metric, but on the principled construction of high-quality Semantic Equivalence Sets that provide both syntactic diversity and semantic consistency.

Var Calculation	OPUS-100	WMT-10
w/ Random Sets	33.5	14.2
w/ Identical Sentences	34.1	15.1
NRD (Full Method)	**35.9**	**22.8**

Table 4: Ablation study on the quality of sentence sets used for variance calculation.

5.4. Direct Measurement of Negative Interference

To directly measure negative interference, we conduct a focused analysis on the high-resource En→De and En→Fr pairs from WMT-10, using strong bilingual models as interference-free upper bounds. The architecture of these bilingual models is identical to that of MNMT baselines and the XLM-T model. As shown in Figure 2, both the standard MNMT and the powerful XLM-T baselines underperform their bilingual counterparts, confirming that negative interference degrades performance even on strong pre-trained models.

Remarkably, our NRD framework not only recovers this performance loss but significantly surpasses the specialist bilingual models on both language pairs, achieving a gain of +1.8 BLEU on En→De and +2.8 BLEU on En→Fr over the XLM-T baseline. This indicates that NRD's disentanglement process does more than just mitigate interference; it creates a cleaner semantic representation that allows the model to leverage shared knowledge from the entire multilingual training pool more effectively than a model exposed only to bilingual data. This provides direct evidence that NRD unlocks a higher performance ceiling for high-resource languages within a multilingual system.

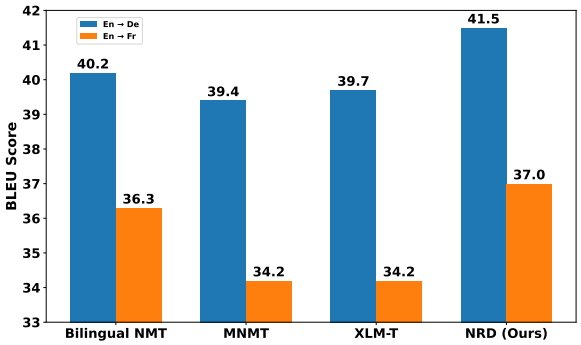

Figure 2: Direct measurement of negative interference on the WMT-10 EN-DE/FR test set (BLEU scores). NRD not only closes the interference gap but also surpasses the strong bilingual model.

6. Conclusion

In this paper, we addressed the fundamental problem of negative interference in Multilingual Neural Machine Translation, which we attribute to the entanglement of semantic and syntactic representations. We introduced NRD, a novel two-stage framework that synergizes training-time neuron specialization with inference-time dynamic filtering to effectively disentangle these representations.

Our empirical results validate the framework's effectiveness. NRD achieves a consistent average improvement of +1.9 BLEU on the large-scale OPUS-100 benchmark. More importantly, on the challenging WMT-10 dataset, which features significant syntactic conflicts, NRD demonstrates remarkable gains of up to +2.8 BLEU over the XLM-T baseline. This highlights our method's strength in mitigating interference. The most compelling evidence comes from the zero-shot setting, where NRD achieves a substantial +7.1 BLEU improvement, demonstrating that it produces more language-agnostic representations.

The success of NRD highlights a promising new paradigm for designing multilingual models: actively managing information flow at inference time. This approach paves the way for more robust and equitable systems, especially for low-resource languages. For future work, we plan to extend NRD to large language models (LLMs) pre-training, explore extending NRD to other cross-lingual tasks, and automate the neuron identification process.

7. Potential Ethical Statement

This work does not involve human subjects or personal data collection. We acknowledge that multilingual translation systems may still inherit or amplify societal biases present in training corpora, especially for low-resource or underrepresented language communities. To mitigate potential harm, future deployments should include bias auditing across language pairs, transparent reporting of failure cases, and human oversight in high-stakes applications.

8. Acknowledgments

We extend our sincere gratitude to the anonymous reviewers for their invaluable and insightful feedback. This research was supported by the National Natural Science Foundation of China (Grant No.U21B2009).

9. Bibliographical References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440. Association for Computational Linguistics.

Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Hui Wang, and Bing Qin. 2024. Gradient consistency-based parameter allocation for multilingual neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7901–7912.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Hwichan Kim, Jun Suzuki, Tosho Hirasawa, and Mamoru Komachi. 2024. Pruning multilingual large language models for multilingual inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9921–9942.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, et al. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020a. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1022–1034, Online. Association for Computational Linguistics.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.