D-Con: Domain-Partitioned Contrastive Learning Framework for Mitigating Domain Interference in Multi-Domain Text Mining

Jiarui Zhang*†(✉), Mingzhe Lu*†, Yifan Deng*†
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
Email: {zhangjiarui,lumingzhe,dengyifan}@iie.ac.cn

***Abstract*—Large-scale text mining often mixes heterogeneous domains, where semantically similar but domain-conflicting samples distort representation geometry and hurt generalization. We present D-Con, a two-stage intermediate pre-training framework with Domain-Partitioned Hard Negative Mining (D-PHNM), which mines cross-domain hard negatives via partitioned FAISS and optimizes a domain-aware contrastive objective to form disentangled subspaces. With LoRA-based adaptation and no test-time domain labels, D-Con improves in-domain performance and reduces forgetting on multi-domain machine translation and sentiment analysis in our settings.**

***Index Terms*—Representation & Reasoning, Transformer Networks, Neural Network Applications, Unsupervised Learning, Large-Scale Neural Networks**

I. Introduction

Large-scale text mining must integrate heterogeneous domains (e.g., web, medical, legal) into a unified model [1]. In this setting, semantically similar but domain-conflicting samples can confuse the model, distort representation space, and degrade cross-domain generalization. We term this *domain interference* and treat it as a central bottleneck for contrastive learning in multi-domain environments.

Contrastive learning (CL) is effective for sentence representations [2] but usually assumes a single domain-agnostic space. When it pushes apart semantically similar yet domain-conflicting samples (e.g., legal "contract breach" vs. IT "security breach"), representations can become ambiguous and averaged. D-Con addresses this at the geometry level: it separates subspaces during intermediate pre-training, so downstream routing/PEFT operates on less entangled representations.

Recent attempts at contrastive domain adaptation add domain awareness but often depend on explicit labels or task-specific designs. Standard negative sampling (random or in-batch) also overlooks cross-domain hard negatives–the main driver of interference. Systematic mining at the domain level remains underexplored, leaving domain-disentangled learning an open problem.

To address this, we propose D-Con (Domain-Partitioned Contrastive Learning), a two-stage intermediate pre-training framework centered on Domain-Partitioned Hard Negative Mining (D-PHNM). D-PHNM retrieves confusing cross-domain negatives with partitioned FAISS [3]; a tailored contrastive loss then increases inter-domain separation while preserving intra-domain coherence. In Stage 2, LoRA [4] activates these pre-learned subspaces without requiring test-time domain labels.

Contributions.

- We formalize *Domain Interference* as a key barrier to cross-domain generalization in contrastive representation learning and introduce **D-Con** to address it.
- We propose **Domain-Partitioned Hard Negative Mining (D-PHNM)**, which uses partitioned FAISS indexes to efficiently retrieve cross-domain hard negatives that encourage inter-domain separation while preserving intra-domain coherence.
- We show that D-Con improves accuracy and reduces forgetting on multi-domain MT and sentiment tasks, offering a scalable, parameter-efficient solution suitable for heterogeneous settings.

II. Related Work

A. Contrastive Learning and Hard Negative Mining

CL learns strong representations but usually assumes a single domain-agnostic space [2], which can amplify domain interference in mixed-domain data. Existing HNM mainly mines in-domain negatives [5]; cross-domain hard negatives are less studied [6]. D-PHNM targets this gap by retrieving semantically similar yet domain-mismatched negatives during pre-training.

B. Multi-Domain Learning and Domain Adaptation

PEFT methods (e.g., Adapters, LoRA) mitigate cost and forgetting [4], [7], but typically adapt after representation geometry is already formed, inheriting entanglement. D-Con is complementary: it first reshapes geometry during intermediate pre-training, then uses LoRA for lightweight downstream adaptation without test-time labels.

Routing/MoE-style PEFT focuses on architecture-level expert selection [8], while D-Con focuses on geometry-level separation before adaptation; the two are complementary. We also follow prior forgetting analyses in MT [9].

C. Representation Disentanglement and Inductive Bias Activation

Prior disentanglement often relies on adversarial or reconstruction objectives with higher complexity. D-Con uses a simpler contrastive mechanism: domain-partitioned HNM increases inter-domain separation while preserving in-domain coherence, then PEFT activates the learned bias.

Comparison with Existing Approaches. Routing/MoE PEFT is architecture-level; D-Con is geometry-level and enforces cross-domain separation during intermediate pre-training via domain-partitioned hard negatives.

III. The D-Con Framework

D-Con is a two-stage framework for mitigating domain interference in large-scale multi-domain text mining. It injects a domain-separation inductive bias during intermediate pre-training to produce a more disentangled representation space (Figure 1). Unlike DaCon [10], which uses adversarial CL for classification, D-Con emphasizes pre-training for general DM tasks.

A. Stage 1: D-Con Intermediate Pre-training

This stage learns disentangled domain geometry. D-PHNM serves as a domain-aware curriculum that exposes increasingly confusing cross-domain negatives.

Input Data and Formatting: We partition the corpus into domain subsets $D = \{D_{med}, D_{leg}, \dots\}$ and prepend a domain prefix in Stage 1 only; Stage 2 and inference use no domain tags.

Core Algorithm: D-PHNM (Domain-Partitioned Hard Negative Mining): D-PHNM dynamically identifies cross-domain hard negatives, i.e., semantically similar but domain-mismatched samples that drive interference. Mining such pairs provides stronger contrastive signals for fine-grained domain separation.

Lookup Structures and Embedding Consistency: We maintain partitioned FAISS indexes S_d [3] and use a momentum encoder pair (query $\mathcal{M}_{query}$ and key $\mathcal{M}_{key}$) [11]:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{1}$$

where m (e.g., 0.999) stabilizes key representations; indexes are periodically refreshed with $\mathcal{M}_{key}$.

D-PHNM Sampling Process: For an anchor x_i from domain d_i, we encode h_i with $\mathcal{M}_{query}$ and retrieve semantically similar embeddings from $S_{d'}$ of all other domains ($d' \neq d_i$) using cosine similarity. To handle imbalance, we allocate $K_{d'} \propto |D_{d'}|$ under a fixed total budget K_{total}, ensuring proportional cross-domain coverage. The full procedure is summarized in Algorithm 1.

Contrastive Objective ($\mathcal{L}_{\text{D-Con}}$): Positive pairs (x_i, x_i^+) are formed intra-domain via SimCSE-style augmentation [2] or parallel data (MT: parallel pairs; MDSA: dropout/augmentation variants). Negatives H_{neg} come exclusively from D-PHNM (no in-batch negatives), so training focuses on cross-domain separation:

Algorithm 1 D-PHNM Sampling Algorithm with Momentum Encoder (Cosine Similarity Metric)

Require: Query encoder $\mathcal{M}_{query}$, key encoder $\mathcal{M}_{key}$; positive pair (x_i, x_i^+) from domain d_i; partitioned lookup structures $\{S_d\}$ (populated by $\mathcal{M}_{key}$); total negative budget K_{total}

```
Encode anchor: h_i ← Encode(M_query, x_i)
Initialize negative set: H_neg ← ∅
for each domain d' ∈ D, d' ≠ d_i do
    Allocate negatives proportionally: K_d' ← ⌊K_total · |D_d'| / Σ_{d''≠d_i} |D_d''|⌋
    Retrieve cross-domain hard negatives: H_hard ← ApproxNearestNeighborSearch(h_i, S_d', K_d', metric = cosine)
    Update negative pool: H_neg ← H_neg ∪ H_hard
end for
return H_neg
```

$$\mathcal{L}_{\text{D-Con}} = -\log \frac{\exp(\text{sim}(h_i, h_i^+)/\tau)}{Z_i},$$
$$Z_i = \exp(\text{sim}(h_i, h_i^+)/\tau) + \sum_{h_j \in H_{neg}} \exp(\text{sim}(h_i, h_j)/\tau). \tag{2}$$

where $\text{sim}(u,v) = \frac{u^\top v}{\|u\|\|v\|}$ and τ is a temperature hyperparameter. This loss explicitly pushes apart semantically similar but domain-conflicting samples, enforcing disentangled representations.

Auxiliary Regularization and Overall Objective: To preserve general semantics, we optionally mix a span-denoising loss $\mathcal{L}_{\text{Aux}}$ (T5-style) with contrastive learning:

$$\mathcal{L}_{\text{Pretrain}} = \mathcal{L}_{\text{D-Con}} + \alpha\mathcal{L}_{\text{Aux}} \tag{3}$$

where α linearly decays (e.g., $1.0 \rightarrow 0.2$) to gradually shift focus from general retention to domain separation.

B. Stage 2: Downstream Task Fine-tuning

This stage transfers the learned domain separation bias to downstream tasks such as machine translation and sentiment analysis, where explicit domain identifiers are unavailable at inference time.

Fine-tuning Strategy: We remove domain prefixes and activate the pre-trained D-Con weights using LoRA [4], a parameter-efficient fine-tuning method. LoRA modules are injected into key Transformer layers, enabling controllable, low-rank adaptation.

Training Protocol: Models are initialized with D-Con pre-trained weights. To isolate the effect of D-Con, we freeze the PLM backbone and fine-tune only LoRA modules.

Mechanism Interpretation: Stage 1 constructs a domain-disentangled representation space. We hypothesize that in Stage 2, LoRA learns a lightweight domain-adaptive mapping that functions as a conditional routing mechanism, dynamically projecting input representations towards the most relevant

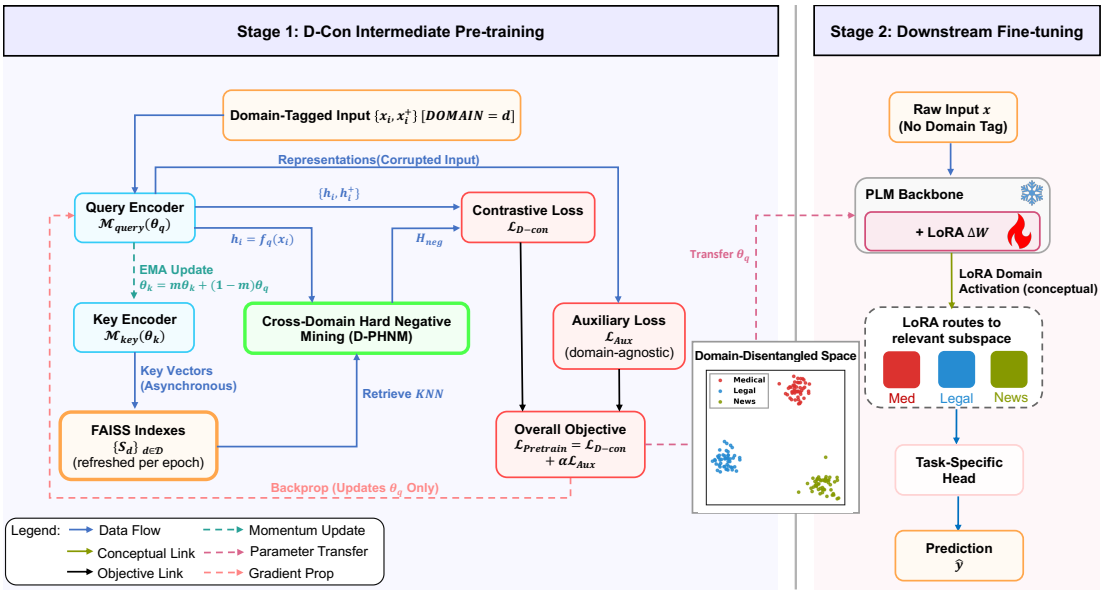

Fig. 1. Overall architecture of the D-Con framework. **Stage 1:** domain-tagged inputs flow through momentum-updated encoders (Query → Key → FAISS), enabling efficient D-PHNM to mine cross-domain hard negatives. The contrastive loss $\mathcal{L}_{\text{D-Con}}$ and auxiliary loss $\mathcal{L}_{\text{Aux}}$ jointly optimize the model, producing domain-disentangled representation spaces. **Stage 2:** Downstream fine-tuning transfers the learned disentangled knowledge to PLM parameters via LoRA (parameter-efficient adapters), enabling dynamic domain activation without explicit labels.

pre-trained domain subspace without requiring explicit domain labels. This design bridges the gap between domain-specific pre-training and domain-agnostic inference, enabling robust and scalable domain-aware learning.

Terminology. We distinguish *representation collapse* during Stage 1 from *catastrophic forgetting* during Stage 2. The former reflects over-specialization that harms generalization; the latter is loss of prior knowledge during downstream adaptation.

IV. Experiments

We evaluate whether D-Con alleviates domain interference while preserving generalization. Unless otherwise stated, we report means over three random seeds.

A. Experimental Setup

Backbone and optimization. All models use `facebook/mbart-large-cc25` [1] [27]. Stage 1 follows Section III-A with momentum query/key encoders and the joint objective $\mathcal{L}_{\text{Pretrain}}$; we train 200k steps with AdamW (3×10^{-5} peak LR, warm-up 10k, cosine decay), batch size 4096 tokens on 4×A100-80GB, gradient clipping 1.0, EMA m=0.999, and mixed precision. D-PHNM uses a simple curriculum where total cross-domain negatives K increases from 64 to 512. Stage 2 freezes the backbone and updates only LoRA and task heads (LR 5×10^{-4}, batch size 256, 30k steps) with early stopping.

Stage-1 corpus. We curate a heterogeneous DE-EN corpus covering six domains (Legal, Medical, News/General, Parliamentary, Scripted Dialogue, Spoken), standardize and filter it, and attach explicit domain tags for Stage 1. Table I summarizes statistics.

D-PHNM implementation. We deploy FAISS-GPU [3] indexes (OPQ16, IVF65536, PQ32) across four A100 (80GB) GPUs with cosine similarity, $nprobe = 64$, and $K_{\text{total}} = 512$ cross-domain negatives per anchor. Embeddings are L2-normalised in FP16. Indexes are rebuilt each epoch using momentum-key embeddings to track the evolving space.

Stage-2 fine-tuning. We remove domain prefixes, freeze the PLM backbone, and insert LoRA modules [4] into all self-/cross-attention query and value projections (rank $r = 16$, scaling $\alpha = 16$). For machine translation we keep the full encoder-decoder; for sentiment classification we use only the encoder and attach a linear classifier. During fine-tuning we update LoRA parameters and task heads only.

For the **D-Con (XLM-R)** results, we perform Stage 1 D-Con pre-training on the XLM-R encoder with the same corpus, domain tags, D-PHNM+$\mathcal{L}_{\text{Aux}}$ objective, and hyperparameters (encoder-only), and then apply the same LoRA fine-tuning protocol as above.

Tasks and datasets. We evaluate two tasks:

1) Multi-Domain Machine Translation (MDMT). We fine-tune on DE-EN domain splits (100k pairs/domain) excluded from Stage 1 to avoid leakage, with 5k/5k dev/test per domain. Forgetting is measured on a WMT News/General test mixture (2019–2022).

2) Multi-Domain Sentiment Analysis (MDSA). We use MARC [28] (EN/DE; sports/books/apparel/beauty), train on 200k reviews per language, and report official 5k dev/test results. Forgetting is measured on SST-2 translated to German (1,821 dev/test), providing a cross-task generalization check.

[1]https://huggingface.co/facebook/mbart-large-cc25

TABLE I

STATISTICS OF THE STAGE 1 INTERMEDIATE PRE-TRAINING CORPUS (COMBINING OPUS AND NON-OPUS SOURCES) AFTER PRE-PROCESSING. UNPC† DENOTES SYNTHETIC DE–EN PAIRS CREATED VIA EN→DE TRANSLATION FROM A RANDOM 2M ENGLISH-SENTENCE SAMPLE, WITH QUALITY FILTERING. MOSEL V2.0⋆ TRANSCRIPTS FILTER OUT EU PARLIAMENT [12] AND VOXPOPULI [13] IDS, ARE SUBSAMPLED TO 3,000 H, AND THEN TRANSLATED WITH NLLB-200 (3.3B) TO OBTAIN PARALLEL DATA.

Domain	Sources	Sentence pairs (M)
Legal	JRC-Acquis [14], DGT-TM [15], UNPC† [16], DCEP [17]	3.7
Medical	UFAL Medical Corpus v1.0 [18]	2.6
News / General	News-Commentary v16 [14], ParaCrawl v9 (partial) [19]	10.1
Parliamentary	Europarl v10 [20]	1.9
Scripted Dialogue	OpenSubtitles (partial) [21]	10.2
Spoken	TED2020 [22], CoVoST-2 [23], Qed v2.0a [24], MOSEL v2.0⋆ [25], [26]	2.5
Total		≈ 31.1

Baselines. To isolate the effect of D-Con and D-PHNM, we compare five groups:

1) Contrastive baselines (LoRA variants). *Vanilla mBART+LoRA* adapts directly on downstream data. *SpLoRA* performs intermediate pre-training on mixed data with only $\mathcal{L}_{\text{Aux}}$ before LoRA fine-tuning.

2) Full fine-tuning and adapters. *Vanilla mBART+Full FT* fully updates the backbone; *Adapters* [7] insert bottleneck adapters instead of LoRA.

3) Domain adaptation methods. We include PTE [29], DoSS [30], ICL [31], MoA [32], and MDAC/DDS [33], which address multi-domain NMT or continual adaptation via pruning, modular updates, MoE-style adaptation, or dynamic sampling.

We exclude methods with incompatible inputs/objectives from main tables to keep comparisons task-aligned.

4) Contrastive classification baselines (MDSA). DaCon [10] and SimCSE [2] are implemented under matched tokenization, optimization budgets, and early stopping on mBART-encoder and XLM-R backbones.

5) Alternative multilingual backbones (MDSA only). XLM-R Large [34] + LoRA, mBERT (cased/uncased) [1] + LoRA, and DistilmBERT [35] + LoRA serve as competitive baselines for sentence-level classification.

Evaluation metrics. For MDMT, we report SacreBLEU [36] and chrF [37] per domain. For MDSA, we report per-domain F1-scores on four representative categories (Sports, Books, Apparel, Beauty), each with binary sentiment labels, along with aggregate metrics over the full test set. We additionally track forgetting on a held-out general-domain benchmark (WMT News mix for MDMT; SST-2 translated to German for MDSA). Let s_{zero} denote the zero-shot score of the Stage-1 backbone on the general benchmark and s_{ft} the score after downstream fine-tuning. Forgetting is calculated as

$$\text{Forgetting}(\%) = 100 \times \max\Big(0, \frac{s_{\text{zero}} - s_{\text{ft}}}{\max(|s_{\text{zero}}|, \epsilon)}\Big), \quad (4)$$

where $\epsilon = 10^{-4}$ prevents division by zero. We use SacreBLEU for MDMT and F1-Score for MDSA when instantiating (4).

B. Main Results

Main results on MDMT and MDSA are shown in Table II and Table III.

Multi-domain Machine Translation (MDMT). D-Con achieves the best SacreBLEU/chrF across all six domains (Table II), improving average BLEU/chrF by about **+0.47/+0.59** over the strongest baselines and reducing forgetting to **1.8%** (**88-91%** relative reduction vs. LoRA/FT).

Multi-Domain Sentiment Analysis (MDSA). Under matched settings, D-Con yields the best mBART-encoder Overall F1 (**73.37**) with lowest forgetting (**0.6%**); D-Con (XLM-R) further reaches **74.47** with **0.4%** forgetting (Table III), showing consistent gains across backbones.

Summary. Across both tasks, D-Con improves in-domain performance while minimizing general-domain forgetting.

C. Analysis

We provide concise diagnostics on efficiency and robustness.

Efficiency. Compared with **SpLoRA** on four A100 (80 GB) GPUs, Stage 1 with D-PHNM shows a small throughput drop (7.6k vs 8.1k sentences/min). FAISS index rebuilds add about 1.4 GB GPU memory per device and take 11.3 min/epoch; peak usage remains within 80 GB. Stage 2 keeps only 2.36M LoRA parameters (≈0.35% of mBART-large) and adds ≈5% inference latency, far below full fine-tuning (≈18%).

Ablation and Component Analysis

We report ablations in Table IV using average SacreBLEU/chrF over six DE–EN domains and forgetting on the held-out general benchmark.

Removing D-PHNM causes the largest degradation; GHNM improves over uniform negatives but remains below partitioned mining. Removing EMA/partitioning and disabling dynamic schedules also hurts performance. Pure contrastive training (α=0) strongly increases forgetting, confirming the role of $\mathcal{L}_{\text{Aux}}$.

5) Discussion. The ablation results indicate that D-PHNM is the primary driver of gains, EMA plus partitioned indexing stabilize training, and dynamic schedules further improve the specialization–generalization trade-off.

TABLE II

MAIN RESULTS ON MULTI-DOMAIN MACHINE TRANSLATION (DE-EN). FOR EACH STAGE 1 DOMAIN WE REPORT SACREBLEU / CHRF (HIGHER IS BETTER), TOGETHER WITH FORGETTING SCORES (EQ. 4; LOWER IS BETTER) ON THE HELD-OUT GENERAL-DOMAIN MIXTURE.

Model	Med BLEU/chrF	Law BLEU/chrF	News BLEU/chrF	Parl BLEU/chrF	SD BLEU/chrF	SP BLEU/chrF	Forget (%)
LoRA	45.53 / 62.12	46.82 / 62.98	41.17 / 57.89	39.04 / 55.33	19.53 / 39.33	18.17 / 37.64	15.7
SpLoRA	46.34 / 62.94	47.35 / 63.33	41.42 / 58.18	39.26 / 55.49	19.75 / 39.59	18.36 / 37.88	15.9
FT	46.57 / 62.13	47.33 / 62.91	41.35 / 58.10	39.15 / 55.38	19.31 / 39.11	18.24 / 37.73	19.5
Adapter	45.86 / 62.23	46.94 / 63.12	41.28 / 58.01	39.12 / 55.35	19.48 / 39.21	18.43 / 37.89	14.1
PTE	47.08 / 63.11	47.86 / 63.71	42.37 / 59.33	39.94 / 56.11	19.83 / 39.69	18.57 / 38.05	3.3
DoSS	47.35 / 63.43	48.07 / 63.88	42.46 / 59.39	40.08 / 56.20	19.92 / 39.71	18.64 / 38.11	5.2
ICL	47.16 / 63.14	47.85 / 63.70	42.38 / 59.36	40.06 / 56.24	19.97 / 39.77	18.73 / 38.25	2.8
MoA	46.83 / 62.98	47.47 / 63.44	41.96 / 58.78	39.58 / 55.77	19.74 / 39.55	18.56 / 38.10	12.9
MDAC/DDS	46.95 / 63.07	47.58 / 63.53	42.18 / 59.12	39.86 / 56.02	19.82 / 39.67	18.58 / 38.09	16.0
D-Con (ours)	**47.93 / 63.78**	**48.37 / 64.26**	**42.76 / 59.91**	**40.18 / 56.86**	**20.92 / 40.31**	**19.22 / 39.12**	**1.8**

TABLE III

MAIN RESULTS ON MULTI-DOMAIN SENTIMENT ANALYSIS. UPPER BLOCK: MBART-ENCODER-BASED MODELS. LOWER BLOCK: OTHER PRE-TRAINED ENCODERS. WE REPORT FOUR-CATEGORY F1-SCORE (HIGHER IS BETTER) AND AGGREGATE METRICS ON THE FULL TEST SET, TOGETHER WITH FORGETTING SCORES (EQ. 4; LOWER IS BETTER) ON THE GENERAL SENTIMENT BENCHMARK. OUR METHOD CAN ALSO BE INSTANTIATED ON XLM-R LARGE.

Model	Sports F1	Books F1	Apparel F1	Beauty F1	Precision	Accuracy	Recall	Overall F1	Forget (%)
mBART encoder-based									
LoRA	64.80	68.31	68.04	65.69	65.34	65.82	66.21	65.77	5.8
SpLoRA	65.48	69.14	68.24	65.80	65.75	66.94	66.80	66.27	5.9
FT	65.71	68.87	68.59	66.83	66.11	67.11	67.01	66.56	6.6
Adapter	65.15	68.60	68.30	66.14	65.65	66.25	66.50	65.87	5.2
SimCSE	68.72	72.18	70.05	71.33	70.44	70.83	70.12	70.29	6.4
DaCon	70.15	74.82	71.64	72.09	71.18	71.73	71.50	71.33	2.9
D-Con	**72.79**	**77.42**	**73.20**	**74.69**	**73.41**	**74.11**	**73.34**	**73.37**	**0.6**
Other pre-trained encoders									
mBERT cased	63.13	69.32	68.82	68.54	67.27	68.15	69.95	68.63	4.9
mBERT uncased	64.21	70.21	70.06	69.91	69.44	70.03	69.89	69.98	5.1
DistilmBERT	62.96	69.07	68.39	67.74	68.55	67.82	68.32	68.43	6.9
XLM-R	72.13	76.16	73.07	74.33	73.07	73.69	73.24	73.16	3.3
SimCSE (XLM-R)	72.53	76.52	73.32	74.79	73.42	73.90	73.44	73.57	4.7
DaCon (XLM-R)	72.79	77.04	73.42	74.82	74.07	74.64	73.90	73.98	2.3
D-Con (XLM-R)	**73.02**	**77.90**	**73.98**	**75.26**	**74.43**	**74.99**	**74.51**	**74.47**	**0.4**

TABLE IV

ABLATION RESULTS ON THE OVERALL DE–EN BENCHMARK (AVERAGE BLEU / CHRF ACROSS SIX DOMAINS ↑; FORGETTING RATE ↓). ALL VARIANTS SHARE IDENTICAL TRAINING SETTINGS UNLESS SPECIFIED.

Variant	Avg. BLEU	Avg. chrF	Forget (%)
Full D-Con	**36.6**	**54.0**	**1.8**
w/o D-PHNM (uniform random negatives)	35.7	53.2	8.3
GHNM (global hard negatives)	36.2	53.6	3.2
w/o EMA (synchronous key encoder)	36.2	53.5	3.4
w/o Partition (single index)	36.3	53.7	3.0
Static K=512	36.4	53.8	2.4
Const. α=1.0	36.3	53.8	2.7
α=0 (pure contrastive)	36.1	53.5	16.5

Error Analysis and Failure Modes. D-Con helps most when domains share surface forms but differ semantically (e.g., *"breach"* in Law vs. IT). It can be less effective for genuinely hybrid inputs or noisy Stage 1 labels, and may under-sample minorities under severe imbalance; increasing K or up-weighting rare domains alleviates this. For sentiment, gains are smaller on figurative/sarcastic text. Stale negatives from infrequent FAISS refreshes also reduce hardness.

D. Limitations and Scope

(i) **Baseline scope.** We focus on MT and sentiment; we benchmark a representative subset of adaptation methods, but not a fully matched suite of recent routing/MoE PEFT variants due to architectural/training-cost differences. (ii) **Hard-negative variants.** We include uniform, GHNM, and partitioned mining ablations, but broader comparisons under larger domain sets remain future work. (iii) **Domain-count sensitivity.** We do not systematically vary the number of domains or simulate extreme imbalance. (iv) **Theoretical guarantees.** We provide empirical, not formal, guarantees of disentanglement. (v) **Language coverage.** MT experiments are DE–EN only; broader multilingual validation is still needed.

V. CONCLUSION

We presented **D-Con**, a two-stage framework for mitigating domain interference in multi-domain text mining. With **D-PHNM**, momentum alignment, and partitioned FAISS indexing, D-Con reshapes representation geometry into more disentangled subspaces during intermediate pre-training. On multi-domain MT and sentiment analysis, D-Con improves in-domain performance (up to **+0.9 BLEU**) and reduces forgetting (over **88%**) versus standard LoRA in our settings. Future work will extend language coverage and test larger-scale, potentially cross-modal settings.

VI. 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).

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 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)*, 2019, pp. 4171–4186.

[2] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910.

[3] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[5] Z. Long, G. Killick, R. McCreadie, G. A. Camarasa, and Z. Meng, "When hard negative sampling meets supervised contrastive learning," *arXiv preprint arXiv:2308.14893*, 2023.

[6] T. Luo, J. Lei, F. Lei, W. Liu, S. He, J. Zhao, and K. Liu, "Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models," *arXiv preprint arXiv:2402.12851*, 2024.

[7] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[8] S. Dou, E. Zhou, Y. Liu, S. Gao, W. Shen, L. Xiong, Y. Zhou, X. Wang, Z. Xi, X. Fan *et al.*, "Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1932–1945.

[9] D. Saunders and S. DeNeefe, "Domain adapted machine translation: What does catastrophic forgetting forget and why?" in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 12 660–12 671.

[10] Y. Dai and A. El-Roby, "Dacon: Multi-domain text classification using domain adversarial contrastive learning," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 40–52.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[12] A. Chmiel, P. Janikowski, M. Kajzer-Wietrzny, D. Koržinek, and D. Jakubowski, "EU parliament speech corpus," 2021, CLARIN-PL digital repository. [Online]. Available: http://hdl.handle.net/11321/821

[13] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," 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)*, 2021, pp. 993–1003.

[14] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *Lrec*, vol. 2012, 2012, pp. 2214–2218.

[15] R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlüter, "Dgt-tm: A freely available translation memory in 22 languages," *arXiv preprint arXiv:1309.5226*, 2013.

[16] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The united nations parallel corpus, language resources and evaluation (lrec16)," *Portorož, Slovenia*, 2016.

[17] N. Hajlaoui, D. Kolovratnik, J. Väyrynen, R. Steinberger, D. Varga *et al.*, "Dcep-digital corpus of the european parliament." in *LREC*, 2014, pp. 3164–3171.

[18] O. Bojar, B. Haddow, D. Mareek, R. Sudarikov, A. Tamchyna, and D. Vari, "Report on building translation systems for public health domain (deliverable d1. 1)," in *Technical Report*. Technical report, Health in my Language (HimL), 2017.

[19] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn *et al.*, "Paracrawl: Web-scale acquisition of parallel corpora." Association for Computational Linguistics (ACL), 2020.

[20] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *The Tenth Machine Translation Summit Proceedings of Conference*. International Association for Machine Translation, 2005, pp. 79–86.

[21] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 923–929.

[22] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525.

[23] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020.

[24] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The amara corpus: Building parallel language resources for the educational domain." in *LREC*, vol. 14, 2014, pp. 1044–1054.

[25] M. Gaido, S. Papi, L. Bentivogli, A. Brutti, M. Cettolo, R. Gretter, M. Matassoni, M. Nabih, and M. Negri, "MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, United States: Association for Computational Linguistics, Nov. 2024.

[26] N. Koluguri, M. Sekoyan, G. Zelenfroynd, S. Meister, S. Ding, S. Kostandian, H. Huang, N. Karpov, J. Balam, V. Lavrukhin, Y. Peng, S. Papi, M. Gaido, A. Brutti, and B. Ginsburg, "Granary: Speech Recognition and Translation Dataset in 25 European Languages," in *Proc. of Interspeech 2025*, Rotterdam, Netherlands, Aug. 2025.

[27] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," 2020.

[28] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The multilingual amazon reviews corpus," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4563–4568.

[29] S. Gu, Y. Feng, and W. Xie, "Pruning-then-expanding model for domain adaptation of neural machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 3942–3952.

[30] A. Hendy, M. Abdelghaffar, M. Afify, and A. Y. Tawfik, "Domain specific sub-network for multi-domain neural machine translation," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2022, pp. 351–356.

[31] Z. Man, K. Huang, Y. Zhang, Y. Chen, Y. Chen, and J. Xu, "Icl: Iterative continual learning for multi-domain neural machine translation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 7732–7743.

[32] F. Zhang, M. Tu, S. Liu, and J. Yan, "A lightweight mixture-of-experts neural machine translation model with stage-wise training strategy," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2381–2392.

[33] M.-Q. Pham, J. Crego, and F. Yvon, "Multi-domain adaptation in neural machine translation with dynamic sampling strategies," in *Conference of the European Association for Machine Translation*, 2022.

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[35] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[36] M. Post, "A call for clarity in reporting bleu scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.

[37] M. Popović, "chrf: character n-gram f-score for automatic mt evaluation," in *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.