

# Large Language Models Meet Extreme Multi-label Classification: Scaling and Multi-modal Framework

Diego Ortego<sup>1</sup>, Marlon Rodríguez<sup>2</sup>, Mario Almagro<sup>1</sup>, Kunal Dahiya<sup>3</sup>,  
David Jiménez<sup>1</sup>, Juan C. SanMiguel<sup>2</sup>

<sup>1</sup>NielsenIQ

<sup>2</sup>Universidad Autónoma de Madrid (UAM)

<sup>3</sup>IIT Delhi

## Abstract

Foundation models have revolutionized artificial intelligence across numerous domains, yet their transformative potential remains largely untapped in Extreme Multi-label Classification (XMC). Queries in XMC are associated with relevant labels from extremely large label spaces, where it is critical to strike a balance between efficiency and performance. Therefore, many recent approaches efficiently pose XMC as a maximum inner product search between embeddings learned from small encoder-only transformer architectures. In this paper, we address two important aspects in XMC: how to effectively harness larger decoder-only models, and how to exploit visual information while maintaining computational efficiency. We demonstrate that both play a critical role in XMC separately and can be combined for improved performance. We show that a few billion-size decoder can deliver substantial improvements while keeping computational overhead manageable. Furthermore, our Vision-enhanced eXtreme Multi-label Learning framework (ViXML) efficiently integrates foundation vision models by pooling a single embedding per image. This limits computational growth while unlocking multi-modal capabilities. Remarkably, ViXML with small encoders outperforms text-only decoder in most cases, showing that an image is worth billions of parameters. Finally, we present an extension of existing text-only datasets to exploit visual metadata and make them available for future benchmarking. Comprehensive experiments across four public text-only datasets and their corresponding image enhanced versions validate our proposals' effectiveness, surpassing previous state-of-the-art by up to +8.21% in P@1 on the largest dataset.

**Code** — <https://github.com/DiegoOrtego/vixml>

## 1 Introduction

Scaling neural network architectures has proven to be an effective strategy for improving the performance of baseline models, as evidenced by powerful foundation models across natural language processing (Brown et al. 2020; Dubey et al. 2024; Abdin et al. 2024; Yang et al. 2025) and computer vision (Radford et al. 2021; Zhai et al. 2023; Tschannen et al. 2025; Bolya et al. 2025). Despite the success of this scaling

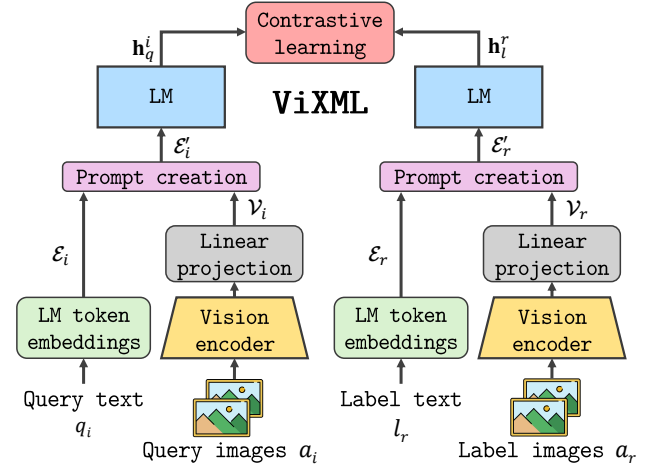


Figure 1: Overview of ViXML multi-modal framework, which supports both encoder language models (LMs) and decoder LLMs. ViXML efficiently incorporates visual metadata in queries ( $a_i$ ) and labels ( $a_r$ ) while freezing the vision encoder for efficiency. Prompts ( $\mathcal{E}'_i$  and  $\mathcal{E}'_r$ ) combine text and projected image embeddings ( $\mathcal{E}$  and  $\mathcal{V}$ ). Sentence embeddings ( $\mathbf{h}_q^i$  and  $\mathbf{h}_r^i$ ) are learned via contrastive learning.

paradigm, efficiency and computational constraints pose significant challenges when leveraging Large Language Models (LLMs) for complex real-world tasks. In particular, this paper tackles this challenge for eXtreme Multi-label Classification (XMC) (Babbar and Schölkopf 2017; Zhang et al. 2021; Dahiya et al. 2021b; Jain et al. 2023; Gupta et al. 2024; Prabhu et al. 2025), where the task is to predict the most relevant subset of labels for a given query from an extremely large label space, often containing millions of labels (Bhatia et al. 2016), *e.g.*, in product recommendation, sponsored search and document tagging applications.

Balancing high-quality predictions with computational efficiency is paramount for XMC tasks. Despite the success of decoder-only LLMs (Lee et al. 2025; Zhang et al. 2025) in extracting embeddings, their adoption in XMC is still an open challenge. While this scaling paradigm represents one potential avenue for performance enhancement, it introduces substantial computational overhead. Therefore, we

also study efficient strategies to boost XMC performance. In particular, we devote our attention to item metadata, which has shown improvements in recent work (Mohan et al. 2024; Prabhu et al. 2025) by exploiting category or hyperlink information. However, limited attention has been directed towards exploiting visual metadata, with MUFIN (Mittal et al. 2022) being the only notable example to date.

To address these challenges, this paper embraces dual-encoder learning (Dahiya et al. 2021a; Gupta et al. 2024) for XMC, also known as Siamese-style training. This approach trains transformer encoders using contrastive learning, addressing XMC via maximum inner product search of the extracted embeddings. Building upon this learning paradigm, we explore scalability and efficiency in two ways: (1) we propose a simple yet effective strategy for the adoption of decoder-only transformers, scaling them up to 7B parameters, which strikes a balance between capacity and efficiency for XMC; (2) we explore images as a key metadata and design an efficient approach to inject visual information using foundation vision models. For decoder-only models, we embed texts within structured prompt templates that provide the LLM with contextual information about the input, and extract sentence embeddings to perform dual-decoder learning. Additionally, we introduce Vision-enhanced eXtreme Multi-label Learning (ViXML), a general multi-modal framework that can be paired with any Siamese-style method for efficiently incorporating visual metadata into XMC. ViXML utilizes a single image embedding per image and concatenates it with input text token embeddings. For decoder models, we modify the structured prompt template to explicitly acknowledge the availability of image information. Crucially, ViXML preserves efficiency by eliminating the need to train visual encoders and avoiding substantial increases in sequence length. In Figure 1, we present a general diagram of ViXML. Our main contributions are as follows:

- We propose a dual-decoder learning approach to effectively adapt decoder-only architectures for XMC.
- We introduce ViXML, a novel multi-modal framework that incorporates images as highly effective metadata for Siamese-style XMC.
- We extend three text-only datasets with visual metadata from Amazon Reviews (Hou et al. 2024) and make them publicly available to foster multi-modal XMC research.
- We systematically examine how scaling transformer backbones affects performance, achieving state-of-the-art results with the biggest decoder-only models.
- We comprehensively evaluate ViXML in several datasets against state-of-the-art methods, reaching remarkable improvements that range from +5.07% to +8.21% points in P@1. Notably, ViXML with a 66M parameter encoder outperforms text-only billion-parameter models, demonstrating the effectiveness of visual metadata.

## 2 Related Work

**Extreme Multi-label Classification (XMC).** Siamese-style XMC using contrastive learning has received significant attention in recent years (Kharbanda et al. 2024; Gupta

et al. 2024; Mohan et al. 2024; Dahiya, Ortego, and Jiménez 2025; Prabhu et al. 2025) due to their efficiency in handling a large number of labels. Early contrastive learning approaches (Dahiya et al. 2021a) did not use deep neural networks and their subsequent adoption marked a significant advancement in model performance (Dahiya et al. 2023b; Gupta et al. 2024; Dahiya, Ortego, and Jiménez 2025). For example, (Dahiya et al. 2023a) reduced training complexity with negative sampling, while (Gupta et al. 2024) demonstrated state-of-the-art results using all negative labels batch-wise. To overcome short-text ambiguity (Dahiya et al. 2023b; Dahiya, Ortego, and Jiménez 2025) proposed enriched label representations. Alternatively, embedding-based predictions can be improved by learning extreme classifiers (Dahiya et al. 2023a) or jointly training the embeddings and classifiers (Kharbanda et al. 2024). Other works pose XMC as meta-classifier learning (Jiang et al. 2021; Mittal et al. 2021; Chien et al. 2023; Zhang et al. 2021) or brute-force extreme classifier learning (Jain et al. 2023).

**Metadata in XMC.** Short-text ambiguity in queries and labels limits XMC performance. In (Mittal et al. 2022) they demonstrated that visual cues can boost performance, while authors in (Mohan et al. 2024) exploited textual metadata (e.g. category information) available only at training time and inferred it at test time. More recently, (Prabhu et al. 2025) exploited metadata to learn a teacher model that is later leveraged during metadata-free training. In this paper, we extend existing XMC Amazon datasets with images and exploit them as metadata for training and inference.

**Text embeddings with decoder-only LLMs.** The text-embedding community has successfully adopted LLMs. For example, authors in (Jiang et al. 2024) demonstrated that prompting an LLM with “*This sentence: [text] means in one word:*” provides a robust text representation in the last token. More recently, authors in (Thirukovalluru and Dhingra 2025) extended this idea by ensembling representations for the same text with different perturbations. Nevertheless, finetuning improves results (Jiang et al. 2024; BehnamGhader et al. 2024) and numerous works followed this path studying: uni-directional vs bi-directional attention (BehnamGhader et al. 2024; Lee et al. 2025; Zhao et al. 2025), in-context-learning (Li et al. 2025) or keeping generative capabilities (Muennighoff et al. 2025). Task adaptation is usually conducted with query instructions (Lee et al. 2025; Zhang et al. 2025). Recent embedding-based XMC (Dahiya et al. 2023b; Gupta et al. 2024) predominantly use deep encoders. However, to the best of our knowledge, only two works exploited LLMs with no clear gains: QUEST (Zhou et al. 2024) and MOGIC (Prabhu et al. 2025). QUEST results with Llama-7B significantly underperformed state-of-the-art encoder models, while MOGIC’s LLM based oracles failed to surpass encoder-only ones. Therefore, effectively leveraging decoder-only LLMs for Siamese-style XMC is an open challenge that we address in this work.

**Multi-modal embeddings with Vision-Language Models (VLMs).** Following LLMs for text embeddings, authors in (Jiang et al. 2025) recently proposed a contrastive learning framework to convert VLMs into an embedding models and address multiple embedding tasks using instructions.

Other works embraced the same spirit: (Liu et al. 2025) employed a language-only pretraining and multi-modal instruction tuning to progressively enhance retrieval performance; (Chen et al. 2025) employed bi-directional attention to enhance context-aware reasoning during continual pre-training, and later performed contrastive finetuning on diverse tasks; (Gu et al. 2025) performed knowledge distillation from a powerful LLM and instruction tuning with hard negatives. In XMC, finetuning these popular VLMs would require adding hundreds of visual tokens, dramatically increasing computational requirements. We, therefore, explore how to efficiently meet multi-modality in this work.

### 3 Method

#### 3.1 Problem Formulation

In XMC, we consider the availability of a dataset  $\mathcal{D} = \{x_i, \mathcal{P}_i\}_{i=1}^M$  with  $M$  data points or queries. Here, the query  $x_i = \{q_i, a_i\}$  is endowed with the textual description  $q_i$  and visual metadata  $a_i$ .  $\mathcal{P}_i$  and  $\mathcal{N}_i$  are the set of positive and negative labels for the  $i$ -th query, respectively. Note that  $\mathcal{P}_i \cup \mathcal{N}_i = \mathcal{Y}$ , where  $\mathcal{Y}$  is the set of L labels. Additionally, each label  $y$  is endowed with a textual description ( $l_r$ ) and visual metadata ( $a_r$ ) (Dahiya et al. 2021a; Kharbanda et al. 2024; Gupta et al. 2024; Mittal et al. 2022). We define the visual metadata sets as  $\mathcal{A}_q = \{a_i\}_{i=1}^M$  and  $\mathcal{A}_l = \{a_r\}_{r=1}^L$  for queries and labels, respectively. In practice,  $a_i$  and  $a_r$  can contain one or multiple images and, occasionally, be empty when no visual information is available.

Predicting the positive labels for every query  $x_i$  can be achieved by posing XMC as a maximum inner product search between the query and label embeddings. The training process involves learning a function  $f_\theta : (\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}^d$ , where  $\theta$  denotes the parameters of a neural network that separately encodes query and labels into the  $d$ -dimensional sentence embeddings  $\mathbf{h}_q^i$  and  $\mathbf{h}_l^r$ . Note that  $\mathbf{h}_l^r$  can be defined as  $\mathbf{h}_p^r$  and  $\mathbf{h}_n^r$  to denote positive and negative label embeddings, respectively. The text-only setup ignores visual metadata and is a special case of our proposed formulation, which is widely used in the existing literature (Dahiya et al. 2021a; Gupta et al. 2024; Mohan et al. 2024).

#### 3.2 Dual-decoder Learning

Siamese-style XMC (Dahiya et al. 2023b; Kharbanda et al. 2024; Gupta et al. 2024) predominantly uses deep encoders as backbone neural networks, *i.e.*, dual-encoder learning, and demonstrate that bigger backbones bring consistent improvements in performance. From this observation, the natural step is to explore bigger decoder-only architectures and exploit them for XMC. However, as discussed in Section 2, no Siamese-style method has effectively leveraged decoder-only LLMs in this field. These results contrast decoder LLMs’ dominance in text embeddings benchmarks (Wang et al. 2024; Muennighoff et al. 2025; Lee et al. 2025). Motivated by this disparity, we investigate effective strategies to leveraging the strong generalization capabilities of LLMs and propose a text-only dual-decoder learning methodology that boosts encoder-only performance while maintaining manageable computational requirements.

**Prompting.** We define a sequence of input embeddings  $\mathcal{E}'_i$  for each query  $q_i$ , so queries are embedded within a structured prompt template. We do so by concatenating to the sequence of text token embeddings  $\mathcal{E}_i = \{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_E\}$ , a prefix, and end-of-sequence (EOS) token as

$$\mathcal{E}'_i = \mathcal{T} \oplus \mathcal{E}_i \oplus \mathbf{e}_{EOS}, \quad (1)$$

where  $\oplus$  denotes the concatenation operator and  $\mathcal{T}$  is the sequence of token embeddings for the prefix. In practice, we use the text prefix “*This product text*” and the end-of-sequence “ $\langle \text{endof} \text{text} \rangle$ ” with token embedding  $\mathbf{e}_{EOS}$ . The same structured prompt is applied to each label  $l_r$  to build  $\mathcal{E}'_r$  from its sequence of text token embeddings  $\mathcal{E}_r$ . This concise template enables input contextualization and keeps a low sequence length to constrain the memory overhead. Alternative prefix and EOS choices are analyzed in Section 4.4.

**Embedding extraction.** During LLMs forward pass we keep uni-directional attention as we found it to work well and aligns with pre-training. However, we leave for future work the open question of whether bi-directional (BehnamGhader et al. 2024; Lee et al. 2025) or uni-directional (Zhao et al. 2025) attention works best.

**Optimization.** The proposed dual-decoder learning framework can exploit any contrastive learning strategy. In this paper we adopt the triplet loss as proposed in (Dahiya et al. 2023a) for either encoder or decoder architectures:

$$\mathcal{L} = \sum_{i=1}^B \sum_{\substack{j \in \mathcal{P}_i \\ k \in \mathcal{N}_i}} [\mathbf{h}_q^i \cdot \mathbf{h}_n^k - \mathbf{h}_q^i \cdot \mathbf{h}_p^j + m]_+, \quad (2)$$

where  $B$  denotes the number of batch queries,  $m$  is a margin and  $\mathbf{h}$  refers to L2-normalized embeddings. For efficiency, we use NGAME (Dahiya et al. 2023a) hard negative mining strategy, thus introducing abuse of notation for  $\mathcal{N}_i$  when using it to denote the subset of negatives selected rather than all negative labels. This vanilla optimization is used in NGAME (Dahiya et al. 2023a), however there are better ways of learning embeddings for XMC (Kharbanda et al. 2024; Mohan et al. 2024; Dahiya, Ortego, and Jiménez 2025). In the first half of Table 5 (Section 4.3) we compare the performance of three approaches (NGAME, DEXA (Dahiya et al. 2023b) and PRIME (Dahiya, Ortego, and Jiménez 2025)). In the remaining of the paper and unless otherwise stated, we adopt PRIME as our base method. In particular, PRIME introduces a shallow label prototype network to aggregate text embeddings, learnable auxiliary vectors and centroids for building enriched label representations. Note that PRIME extends the optimization in Eq. 2 with additional terms and we refer the reader to (Dahiya, Ortego, and Jiménez 2025) for a detailed explanation.

Fine-tuning LLMs is resource-intensive, and inference often suffers from high latency compared to encoder-only transformers. To mitigate these challenges, we adopt relatively small LLMs (up to 7B parameters) and leverage their sample efficiency capabilities to reduce training times. While inference throughput will always be impacted due to its increased parameter size, specialized hardware or kernel-based implementations like vLLM, can make small LLMs

practical for XMC. The results of text-only dual-decoder learning are presented in Section 4.2, showing that scaling to decoder-only LLMs clearly surpass encoder architectures.

### 3.3 The ViXML Framework

The adoption of item metadata (Mittal et al. 2022; Mohan et al. 2024; Prabhu et al. 2025) poses a promising path towards more robust XMC. Following this trend, we propose Vision-enhanced eXtreme Multi-label Learning (ViXML), a multi-modal framework that extends Siamese-style XMC to leverage visual metadata. ViXML is architecture-agnostic, supporting both encoder and decoder models, and operates by integrating image embeddings  $\mathbf{v}$  together with text token embeddings  $\mathbf{e}$ . In particular, we use a pre-trained foundation vision model  $g_\phi : \mathcal{A} \rightarrow \mathbb{R}^m$  with frozen parameters  $\phi$  to map images into  $m$ -dimensional image embeddings. We follow works like (Liu et al. 2023) and learn a linear layer  $w_\psi : \mathbb{R}^m \rightarrow \mathbb{R}^d$  with parameters  $\psi$  for adaptation of visual information to text inputs. We leave the study of more complex adaptation choices for future work. ViXML is, therefore, computationally efficient as: (i) utilizing a single embedding per image preserves low sequence lengths, and (ii) employing a frozen vision encoder enables storing its embeddings as a feature bank, thereby minimizing the memory overhead during training. We detail below ViXML for both encoder and decoder architectures.

**ViXML with encoder models** Our framework can be seamlessly integrated into encoder models with minimal computational overhead. We do so by concatenating the sequence of image embeddings  $\mathcal{V}_i = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{V_i}\}$  for the  $i^{th}$  query with its corresponding sequence of text token embeddings  $\mathcal{E}_i$ . Then, the final sequence of input embeddings for query  $q_i$  is

$$\mathcal{E}'_i = \mathcal{V}_i \oplus \mathcal{E}_i. \quad (3)$$

Similar concatenation builds the label sequence  $\mathcal{E}'_r$ . During forward propagation through the network, text and visual representations are enriched with each other through attention mechanisms, leading to multi-modal sentence embeddings  $\mathbf{h}_q^i$  and  $\mathbf{h}_r^i$  for queries and labels, respectively.

**ViXML with decoder models** As presented for text-only dual-decoder learning, we embed images and text within a structured prompt template, which can be defined as

$$\mathcal{E}'_i = \mathcal{T} \oplus \mathcal{E}_i \oplus \mathcal{I} \oplus \mathcal{V}_i \oplus \mathbf{e}_{EOS}, \quad (4)$$

where  $\mathcal{T}$  and  $\mathcal{I}$  are the sequences of token embeddings for the text and image prefixes, respectively. We build multi-modal sentence embeddings  $\mathbf{h}_q^i$  and  $\mathbf{h}_r^i$  by mean pooling token representations. We initially mimicked the encoder mechanism to build the input sequence  $\mathcal{E}'_i$  and fell into two main problems: (i) a drop in performance compared to using a structured prompt template, and (ii) degenerated performance when concatenating image embeddings as the first token in the sequence. Prompting and placing image information after the text, both helped in achieving robust performance (see Section 4.4). Note that the latter decision implies contextualizing image information with text given the uni-directional attention of decoder LLMs.

## 4 Experimental Work

### 4.1 Framework

**Details.** We experiment with four publicly available XMC text-only datasets (Bhatia et al. 2016) for product-to-product recommendation; only MM-AmazonTitles-300K has visual metadata available. Therefore, for multi-modal evaluation we extend LF-AmazonTitles-131K, LF-Amazon-131K and LF-AmazonTitles-1.3M, for which images can be extracted from Amazon Reviews data (Hou et al. 2024). In particular, we selected image URLs from 2023 version and moved towards older versions if no images were available. Note that these datasets vary the amount of labels and text length, thus supporting a robust evaluation in diverse contexts. We present dataset details in the supplementary material. For evaluation, we adopt standard XMC metrics defined at (Bhatia et al. 2016), *e.g.* precision, propensity-scored precision or recall ( $P@k$ ,  $PSP@k$  and  $R@k$ ).

**Baseline methods.** We compare against the following text-only methods (extended comparison in the supplementary material): DEXA (Dahiya et al. 2023b), PINA (Chien et al. 2023), Renée (Jain et al. 2023), DEXML (Gupta et al. 2024) UniDEC (Kharbanda et al. 2024), OAK (Mohan et al. 2024), PRIME (Dahiya, Ortego, and Jiménez 2025) and MOGIC (Prabhu et al. 2025). We further compare with MUFIN (Mittal et al. 2022) on MM-AmazonTitles-300K.

**Implementation details.** We train encoder models of different sizes: MiniLM-L3, DistilBERT and BERT (Reimers and Gurevych 2019; Devlin et al. 2019). For decoder models, we focus on Qwen2.5-Instruct (Yang et al. 2025), but experiment with other LLMs in Section 4.2 to demonstrate generalization. Unless otherwise stated, we adopt SigLIPv2 (Tschannen et al. 2025) as vision encoder and aggregate tokens with model’s multi-head attention pooling, use 3 images in ViXML and perform mean pooling for sentence embedding extraction. We use low-rank adaptation (LoRA) during finetuning of decoder models and run all experiments in a single 80GB GPU as the budget constraint. We always report performance in the last training epoch, *i.e.*, we do not conduct any custom checkpoint selection based on validation metrics. We provide further implementation details, inference latencies, a study on the impact of image count and an example of random seed variability (standard deviation of 0.103 in  $P@1$ ) in the supplementary material.

### 4.2 Scaling Text-only Architectures for XMC

Related literature mostly operates with DistilBERT 66M parameters model which keeps XMC extremely efficient, but hinders understanding the scaling potential. We, therefore, scale in this work encoder sizes and, more importantly, propose a dual-decoder learning strategy to leverage decoder-only LLMs. We adopt Qwen2.5-Instruct with varying sizes, reaching 3B parameters for all datasets and 7B in LF-AmazonTitles-131K. Table 1 presents our results (we discard visual metadata for MM-AmazonTitles-300K), demonstrating that foundation LLMs can deliver better performance for XMC. Additionally, we keep computational overhead manageable as (i) all our experiments fit on a single

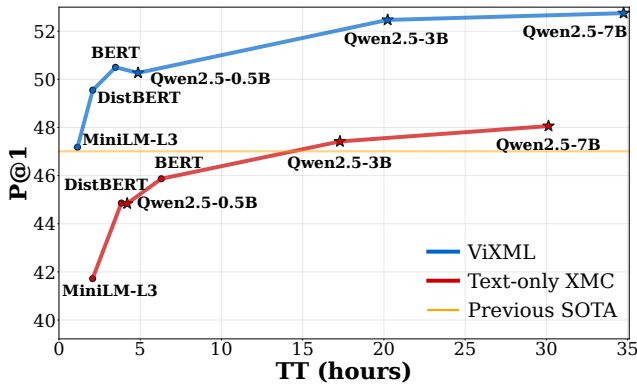


Figure 2: Performance (P@1) and Training Time (TT) for dual-encoder (dots) and dual-decoder (stars) learning in LF-AmazonTitles-131K. The ViXML multi-modal framework (blue) improves text-only alternatives (red), while decoder models boost encoder performance in both setups. Previous state-of-the-art (SOTA) is represented by MOGIC method.

GPU, and (ii) we bound the ineludible training time growth when scaling to decoder-only models by significantly reducing the number of training epochs. Red line in Figure 2 shows the performance and training times across backbone architectures in LF-AmazonTitles-131K. We demonstrate that scaling the size of the models from encoder to decoder-only architectures surpasses previous state-of-the-art results using text-only inputs (orange line). Despite requiring longer training times, decoder-only models are more sample efficient; encoders are trained for 300 epochs and decoders only for 30, which effectively reduces training times by lowering the number of training epochs one order of magnitude. Note that 0.5B decoder model and 66M distilBERT train in comparable time in our setup.

**Different families of LLM.** We demonstrate in the second half of Table 2 that text-only dual-decoder learning with Llama-3.2, Gemma-3 and Qwen variants achieve robust results in LF-AmazonTitles-131K, surpassing encoder models reported in Table 1. While we focus on task-specific XMC embeddings, strong general-purpose embeddings might appear adequate. However, Table 2 shows that the off-the-shelf Qwen3-Embedding-4B model performs significantly worse than all trained alternatives. Additionally, we tested Qwen3-Emb as recommended in (Zhang et al. 2025) (\*), *i.e.*, following a prompt template for the queries (“*Instruct: Retrieve a semantically related product.* \n*Query: [text] <|end-of-text|>*”) and not for the labels, while extracting embeddings with last token pooling. As reported, this strategy did not lead to improvements.

### 4.3 Exploiting Visual Meta-data

This section presents the benefits of ViXML multi-modal framework to leverage visual metadata in XMC. First, Table 3 presents ViXML’s performance when paired with the same backbones of Table 1. Notably, a small DistilBERT model with 66M parameters paired with ViXML surpasses

Backbone	P@1	P@5	PSP@1	R@100
<b>LF-AmazonTitles-131K</b>				
MiniLM-L3	41.72	20.16	36.17	64.68
DistilBERT	44.86	21.45	39.62	68.05
BERT	45.87	21.88	40.54	69.67
Qwen2.5-0.5B-I	44.84	21.66	39.64	70.30
Qwen2.5-3B-I	47.42	22.89	41.88	74.34
Qwen2.5-7B-I	<b>48.06</b>	<b>23.07</b>	<b>42.43</b>	<b>75.24</b>
<b>MM-AmazonTitles-300K</b>				
MiniLM-L3	49.54	31.93	33.83	69.30
DistilBERT	52.40	34.29	36.36	72.59
BERT	52.95	34.57	<b>36.85</b>	73.34
Qwen2.5-0.5B-I	52.30	34.23	36.22	74.44
Qwen2.5-3B-I	<b>54.71</b>	<b>35.93</b>	36.29	<b>77.22</b>
<b>LF-AmazonTitles-1.3M</b>				
MiniLM-L3	51.59	39.28	29.50	58.44
DistilBERT	58.49	45.27	33.28	63.76
BERT	59.10	45.62	<b>34.10</b>	64.30
Qwen2.5-0.5B-I	58.02	45.11	31.68	63.96
Qwen2.5-3B-I	<b>60.74</b>	<b>47.39</b>	34.01	<b>66.88</b>
<b>LF-Amazon-131K</b>				
MiniLM-L3	41.99	20.42	35.44	68.37
DistilBERT	47.98	23.21	40.99	76.22
BERT	49.15	23.71	42.18	78.94
Qwen2.5-0.5B-I	49.23	23.99	42.40	81.24
Qwen2.5-3B-I	<b>53.17</b>	<b>25.72</b>	<b>45.74</b>	<b>85.61</b>

Table 1: Impact of backbone size on text-only datasets.

Backbone	#p	P@1	PSP@1	R@100
MoCa*	3.75B	<b>25.67</b>	<b>26.31</b>	<b>54.48</b>
Qwen3-Emb*	4.02B	18.33	18.32	42.67
Qwen3-Emb	4.02B	22.33	22.49	46.69
Qwen2.5-I	3.09B	47.42	41.88	<b>74.34</b>
Llama-3.2	3.21B	<b>48.19</b>	<b>42.60</b>	73.95
Qwen3	4.02B	47.24	41.79	73.42
Qwen3-Emb	4.02B	47.51	42.11	73.62
Gemma-3-I	4.3B	47.74	42.34	73.70

Table 2: Pre-trained general-purpose embeddings () vs dual-decoder learning with different LLM families. #p is the number of parameters and (\*) denotes using the authors’ embedding extraction methodology. MoCa is a vision-language embedding model, *i.e.*, it uses visual metadata.

the best performance achieved with billion-size text-only models in most cases. These results underscore that an image is worth billions of parameters for achieving robust and efficient XMC. There is only one case where encoder-based ViXML approaches do not surpass text-only 3B model (LF-Amazon-131K). In contrast to the rest of the datasets, LF-

Backbone	P@1	P@5	PSP@1	R@100
<b>LF-AmazonTitles-131K</b>				
MiniLM-L3	47.18	22.69	41.01	73.81
DistilBERT	49.55	23.73	43.78	76.73
BERT	50.50	24.08	44.67	78.43
Qwen2.5-0.5B-I	50.27	24.22	44.31	79.41
Qwen2.5-3B-I	52.47	25.26	46.46	82.21
Qwen2.5-7B-I	<b>52.75</b>	<b>25.37</b>	<b>46.70</b>	<b>83.27</b>
<b>MM-AmazonTitles-300K</b>				
MiniLM-L3	52.97	34.13	35.51	73.08
DistilBERT	55.03	35.91	37.52	76.07
BERT	55.59	36.10	37.83	76.67
Qwen2.5-0.5B-I	55.28	36.09	36.48	77.61
Qwen2.5-3B-I	<b>56.55</b>	<b>37.04</b>	<b>37.60</b>	<b>79.38</b>
<b>LF-AmazonTitles-1.3M</b>				
MiniLM-L3	60.36	45.78	34.88	67.35
DistilBERT	64.17	49.43	36.58	70.65
BERT	64.59	49.66	37.35	71.08
Qwen2.5-0.5B-I	63.57	49.08	35.21	70.35
Qwen2.5-3B-I	<b>66.01</b>	<b>51.21</b>	<b>37.71</b>	<b>73.13</b>
<b>LF-Amazon-131K</b>				
MiniLM-L3	46.88	22.63	39.97	75.80
DistilBERT	51.30	24.69	44.09	81.40
BERT	52.28	25.12	45.08	83.40
Qwen2.5-0.5B-I	51.76	25.05	45.08	84.52
Qwen2.5-3B-I	<b>55.11</b>	<b>26.46</b>	<b>47.73</b>	<b>87.73</b>

Table 3: Impact of backbone size on ViXML. Note that we use visual metadata across all backbones and datasets.

Amazon-131K contains product titles and descriptions. We argue that decoder-only models benefit from longer textual inputs by exploiting their parametric knowledge, struggling in those cases where only product titles are presented.

Regarding training times, Figure 2 presents the impact of ViXML (blue line) over text-only training (red line) in LF-AmazonTitles-131K. Remarkably, ViXML with encoder models significantly reduces training time due to a faster convergence to top performance that enables halving the training epochs (300 to 150). On the other hand, for decoder models we only train for 30 epochs for both text-only and ViXML models, showing that ViXML introduces an overhead of around 15%-17% which mostly comes from longer inputs due to prompts and visual metadata.

**Early-fusion or late-fusion.** ViXML framework conducts early-fusion of visual metadata, while MUFIN employs late-fusion via self-attention to integrate text and image embeddings. To demonstrate the benefits of our early-fusion strategy, we compare ViXML and MUFIN in Table 4. Using MUFIN’s ViT-32 encoder, ViXML achieves +1% in P@1 (53.30 vs 52.30) despite requiring only an encoder training (MUFIN trains extreme classifiers and prediction combina-

MM	XMC	V	P@1	P@5	R@10
MUFIN	MUFIN	ViT-32	52.30	34.76	50.63
MUFIN	PRIME	ViT-32 SigLIP2	52.62 53.44	34.35 34.79	49.28 50.18
ViXML	PRIME	ViT-32 SigLIP2	53.30 <b>55.03</b>	34.80 <b>35.91</b>	50.19 <b>51.98</b>

Table 4: Comparison of multi-modal (MM) strategies with different vision encoders (V). First row shows original MUFIN (trains encoder, classifier and fusion). Then, we present MUFIN (late-fusion) and ViXML (early-fusion) on top of PRIME method for fair comparison.

MM	XMC	P@1	P@5	PSP@1	R@100
-	NGAME	42.47	20.44	37.95	65.43
	DEXA	44.38	20.96	38.64	65.37
	PRIME	44.86	21.45	39.62	68.05
ViXML	NGAME	47.81	22.79	43.04	74.09
	DEXA	49.48	23.19	43.21	73.57
	PRIME	<b>49.55</b>	<b>23.73</b>	<b>43.78</b>	<b>76.73</b>

Table 5: Effect of ViXML for different XMC methods.

tion). However, this comparison has limitations due to differences in training stages, image counts, and Siamese-style methods. For a fair comparison, we pair both MUFIN and ViXML with the same XMC method to isolate the impact of visual metadata injection. This evaluation reveals that the proposed early fusion consistently outperforms MUFIN over 1.5% in P@1 across both ViT-32 and SigLIP2 encoders. Additionally, these results highlight the importance of leveraging large foundation vision models, as performance significantly improves when scaling from 86M ViT-32 to 1.14B SigLIP2 in both MUFIN and ViXML.

**ViXML across methods.** We use PRIME (Dahiya, Ortego, and Jiménez 2025) as the base method due to its efficiency and robustness. However, ViXML is a general multi-modal framework that can be used with any Siamese-style method. We support this claim by pairing ViXML with NGAME (Dahiya et al. 2023a), DEXA (Dahiya et al. 2023b) and PRIME (Dahiya, Ortego, and Jiménez 2025) methods in LF-AmazonTitles-131K dataset (see Table 5). The results demonstrate that ViXML delivers an important boost across all metrics for all methods.

**Pre-trained VLMs.** Section 2 discussed VLM-based embeddings and their challenging adoption in XMC due to using hundreds of visual tokens. Conversely, the proposed ViXML stands as an efficient alternative. However, a reasonable question to address is whether off-the-shelf VLMs can generalize well to XMC. Table 2 demonstrates that recent off-the-shelf embedding VLM MoCa (Chen et al. 2025) yields weak results as compared to finetuned alternatives.

Prompt templates	P@1
$\mathcal{E}$	43.14
$\mathcal{T}_1 \oplus \mathcal{E}$	44.15
$\mathcal{T}_1 \oplus \mathcal{E} \oplus \mathbf{e}_{EOS}$	<b>44.66</b>
$\mathcal{V} \oplus \mathcal{E} \oplus \mathbf{e}_{EOS}$	46.12
$\mathcal{E} \oplus \mathcal{V} \oplus \mathbf{e}_{EOS}$	49.45
$\mathcal{I}_1 \oplus \mathcal{V} \oplus \mathcal{T}_2 \oplus \mathcal{E} \oplus \mathbf{e}_{EOS}$	49.29
$\mathcal{T}_1 \oplus \mathcal{E}_i \oplus \mathcal{I}_2 \oplus \mathcal{V}_i \oplus \mathbf{e}_{EOS}$	<b>49.67</b>

Table 6: Prompting effect for dual-decoder learning. We consider sequences of embeddings for text prefixes  $\mathcal{T}_1$  (“*This product text*”) and  $\mathcal{T}_2$  (“*and its text*”), image prefixes  $\mathcal{I}_1$  (“*This product image*”) and  $\mathcal{I}_2$  (“*and its image*”), text  $\mathcal{E}$  and images  $\mathcal{V}$ . EOS stands for end-of-sequence token.

#### 4.4 Prompting Effect

Table 6 compares the performance of different strategies to prompting decoder-only models with and without images (Qwen2.5-0.5B-Instruct trained for 20 epochs on LF-AmazonTitles-131K dataset). For text-only inputs, adding prefix  $\mathcal{T}_1$  (“*This product text*”) and end-of-sequence token  $\mathbf{e}_{EOS}$  (“*<lendoftext>*”) both improve performance over not using them. With visual metadata, prefixes  $\mathcal{T}_1$  or  $\mathcal{T}_2$  (“*and its text*”) for text and  $\mathcal{I}_1$  (“*This product image*”) or  $\mathcal{I}_2$  (“*and its image*”) also benefit performance. These results, together with further prompt variations reported in the supplementary material, suggest that performance gains stem from structural cues that help leveraging pretrained LLMs as no additional information is injected with these prompts. We also investigate the order of visual tokens in the input sequence for decoder models, thus we need to account for the pretraining dynamics of these models. As demonstrated in (Barbero et al. 2025), pretraining LLMs generates attention sinks on the first token to avoid representational collapse. We observe that phenomena when breaking the pretraining dynamics by placing image tokens first without prefixes ( $\mathcal{V} \oplus \mathcal{E} \oplus \mathbf{e}_{EOS}$ ), significantly dropping performance. Conversely, placing images at the end and using prefixes improves performance, thus we adopt this robust configuration across the experiments.

#### 4.5 Comparative Evaluation

We conclude our experimentation in Table 7, which shows large improvements of the ViXML framework (with dual-decoder learning and PRIME) over the best previous results across all metrics and datasets, ranging from +5.07% to +8.21% points in P@1. Note that MOGIC does not balance PSP@1 and P@1, reporting +0.88% in PSP@1 with respect to ViXML in LF-AmazonTitles-1.3M, while dropping 16.88% in P@1. As many other works, we incorporate inference enhancements for ViXML. In particular, we extend the search space using training data samples to refine label predictions (Wang et al. 2025; Dahiya et al. 2021b). We refer the reader to the supplementary material for further details.

Method	E	M	P@1	P@5	PSP@1
<b>LF-AmazonTitles-131K</b>					
<b>DEXML</b>	✗	✗	42.52	20.64	-
<b>UniDEC</b>	✓	✗	44.35	21.03	39.23
<b>PRIME</b>	✓	✗	45.26	21.48	39.29
<b>Renée</b>	✓	✗	46.05	22.04	39.08
<b>DEXA</b>	✓	✗	46.42	21.59	39.11
<b>OAK</b>	✓	✓	46.42	21.88	39.76
<b>MOGIC</b>	✓	✓	47.01	22.40	40.62
<b>ViXML (Ours)</b>	✓	✓	<b>53.08</b>	<b>25.74</b>	<b>46.22</b>
<b>MM-AmazonTitles-300K</b>					
<b>MUFIN</b>	✓	✓	52.30	34.76	-
<b>ViXML (Ours)</b>	✓	✓	<b>57.37</b>	<b>38.08</b>	<b>36.12</b>
<b>LF-AmazonTitles-1.3M</b>					
<b>OAK</b>	✓	✓	49.46	38.61	34.92
<b>MOGIC</b>	✓	✓	50.95	39.95	<b>36.28</b>
<b>Renée</b>	✓	✗	56.04	45.32	28.54
<b>DEXA</b>	✓	✗	56.63	43.90	29.12
<b>UniDEC</b>	✓	✗	57.41	45.89	30.10
<b>DEXML</b>	✗	✗	58.40	45.46	31.36
<b>PRIME</b>	✓	✗	59.62	46.75	31.20
<b>ViXML (Ours)</b>	✓	✓	<b>67.83</b>	<b>53.72</b>	35.40
<b>LF-Amazon-131K</b>					
<b>PINA</b>	✓	✗	46.76	23.20	-
<b>DEXA</b>	✓	✗	47.16	22.42	38.70
<b>UniDEC</b>	✓	✗	47.80	23.35	40.28
<b>Renée</b>	✓	✗	48.05	23.26	40.11
<b>PRIME</b>	✓	✗	48.20	23.28	40.16
<b>OAK</b>	✓	✓	48.36	22.20	-
<b>MOGIC</b>	✓	✓	50.05	23.72	-
<b>ViXML (Ours)</b>	✓	✓	<b>55.57</b>	<b>26.84</b>	<b>47.41</b>

Table 7: Comparison of ViXML against related work regardless of inference enhancements (E) or meta-data (M).

## 5 Conclusion

In this paper we demonstrate the impact of the scaling laws in XMC and present for the first time an efficient strategy to enable massively pretrained decoder-only transformers for XMC. We show that increasing the backbones’ size improves performance, getting state-of-the-art results. In addition, we present ViXML, an effective approach to exploit visual information in XMC. By using an early-fusion strategy and tailored prompting templates for decoder models, we bring another significant leap in XMC performance. We also open to the community three new extensions of existing datasets where we incorporate visual metadata information. Extensive experimentation across multiple datasets and methods demonstrate all our claims, advancing state-of-the-art performance in XMC by a significant margin. In the supplementary material we discuss the main limitations of our work and suggest potential lines for future work.

## Acknowledgements

The authors want to thank Javier Martínez Cebrián for the fruitful discussions during the development of this work.

## References

- Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; and Eldan, R. e. a. 2024. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.
- Babbar, R.; and Schölkopf, B. 2017. DiSMEC: Distributed Sparse Machines for Extreme Multi-label Classification. In *ACM International Conference on Web Search and Data Mining (WSDM)*.
- Barbero, F.; Arroyo, A.; Gu, X.; Perivolaropoulos, C.; Bronstein, M.; Veličković, P.; and Pascanu, R. 2025. Why do LLMs attend to the first token? *arXiv preprint arXiv:2504.02732*.
- BehnamGhader, P.; Adlakha, V.; Mosbach, M.; Bahdanau, D.; Chapados, N.; and Reddy, S. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *Conference on Language Modeling (COLM)*.
- Bhatia, K.; Dahiya, K.; Jain, H.; Kar, P.; Mittal, A.; Prabhu, Y.; and Varma, M. 2016. The Extreme Classification Repository: Multi-label Datasets & Code.
- Bolya, D.; Huang, P.; Sun, P.; Cho, J.; Madotto, A.; Wei, C.; Ma, T.; Zhi, J.; Rajasegaran, J.; Rasheed, H.; Wang, J.; Monteiro, M.; Xu, H.; Dong, S.; Ravi, N.; Li, D.; Dollár, P.; and Feichtenhofer, C. 2025. Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; and Kaplan, J. e. a. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chen, H.; Liu, H.; Luo, Y.; Wang, L.; Yang, N.; Wei, F.; and Dou, Z. 2025. MoCa: Modality-aware Continual Pre-training Makes Better Bidirectional Multimodal Embeddings. *arXiv preprint arXiv:2506.23115*.
- Chien, E.; Zhang, C.-J., J. Hsieh; Jiang, J.-Y.; Chang, W.-C.; Milenkovic, O.; and Yu, H.-F. 2023. PINA: Leveraging Side Information in eXtreme Multi-label Classification via Predicted Instance Neighborhood Aggregation. In *International Conference on Machine Learning (ICML)*.
- Dahiya, K.; Agarwal, A.; Saini, D.; Gururaj, K.; Jiao, J.; Singh, A.; Agarwal, S.; Kar, P.; and Varma, M. 2021a. SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels. In *International Conference on Machine Learning (ICML)*.
- Dahiya, K.; Gupta, N.; Saini, D.; Soni, A.; Wang, Y.; Dave, K.; Jiao, J.; Gururaj, K.; Dey, P.; Singh, A.; Hada, D.; Jain, V.; Paliwal, B.; Mittal, A.; Mehta, S.; Ramjee, R.; Agarwal, S.; Kar, P.; and Varma, M. 2023a. NGAME: Negative mining-aware mini-batching for extreme classification. In *ACM International Conference on Web Search and Data Mining (WSDM)*.
- Dahiya, K.; Ortego, D.; and Jiménez, D. 2025. Prototypical Extreme Multi-label Classification with a Dynamic Margin Loss. In *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.
- Dahiya, K.; Saini, D.; Mittal, A.; Shaw, A.; Dave, K.; Soni, A.; Jain, H.; Agarwal, S.; and Varma, M. 2021b. DeepXML: A Deep Extreme Multi-Label Learning Framework Applied to Short Text Documents. In *ACM International Conference on Web Search and Data Mining (WSDM)*.
- Dahiya, K.; Yadav, S.; Sondhi, S.; Saini, D.; Mehta, S.; Jiao, J.; Agarwal, S.; Kar, P.; and Varma, M. 2023b. Deep encoders with auxiliary parameters for extreme classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- Devlin, J.; Chang, M. W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; and Al-Dahle, A. e. a. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432*.
- Gupta, N.; Khatri, D.; Rawat, A.; Bhojanapalli, S.; Jain, P.; and Dhillon, I. 2024. Dual-Encoders for Extreme Multi-Label Classification. In *International Conference on Learning Representations (ICLR)*.
- Hou, Y.; Li, J.; He, Z.; Yan, A.; Chen, X.; and McAuley, J. 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952*.
- Jain, V.; Prakash, J.; Saini, D.; Jiao, J.; Ramjee, R.; and Varma, M. 2023. Renée: End-to-end Training of Extreme Classification Models. In *Conference on Machine Learning and Systems (MLSys)*.
- Jiang, T.; Huang, S.; Luan, Z.; Wang, D.; and Zhuang, F. 2024. Scaling Sentence Embeddings with Large Language Models. In *Conference on Empirical Methods in Natural Language Processing, Findings Track (EMNLP-F)*.
- Jiang, T.; Wang, D.; Sun, L.; Yang, H.; Zhao, Z.; and Zhuang, F. 2021. LightXML: Transformer with Dynamic Negative Sampling for High-Performance Extreme Multi-label Text Classification. In *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)*.
- Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2025. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *International Conference on Learning Representations (ICLR)*.
- Kharbanda, S.; Gupta, D.; K., G.; Malhotra, P.; Hsieh, C.-J.; and Babbar, R. 2024. UniDEC : Unified Dual Encoder and Classifier Training for Extreme Multi-Label Classification. *arXiv preprint arXiv:2405.03714*.

- Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. In *International Conference on Learning Representations (ICLR)*.
- Li, C.; Qin, M.; Xiao, S.; Chen, J.; Luo, K.; Lian, D.; Shao, Y.; and Liu, Z. 2025. Making Text Embedders Few-Shot Learners. In *International Conference on Learning Representations (ICLR)*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Liu, Y.; Zhang, Y.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2025. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.
- Mittal, A.; Dahiya, K.; Malani, S.; Ramaswamy, J.; Kuruvilla, S.; Ajmera, J.; Chang, K.; Agrawal, S.; Kar, P.; and Varma, M. 2022. Multimodal Extreme Classification. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.
- Mittal, A.; Sachdeva, N.; Agrawal, S.; Agarwal, S.; Kar, P.; and Varma, M. 2021. ECLARE: Extreme Classification with Label Graph Correlations. In *International Conference on World Wide Web (WWW)*.
- Mohan, S.; Saini, D.; Mittal, A.; Chowdhury, S.; Paliwal, B.; Jiao, J.; Gupta, M.; and Varma, M. 2024. OAK: Enriching Document Representations using Auxiliary Knowledge for Extreme Classification. In *International Conference on Machine Learning (ICML)*.
- Muennighoff, N.; SU, H.; Wang, L.; Yang, N.; Wei, F.; Yu, T.; Singh, A.; and Kiela, D. 2025. Generative Representational Instruction Tuning. In *International Conference on Learning Representations (ICLR)*.
- Prabhu, S.; Singh, B.; Mittal, A.; Asokan, S.; Mohan, S.; Saini, D.; Prabhu, Y.; Kumar, L.; Jiao, J.; Singh, A.; Tandon, N.; Gupta, M.; Agarwal, S.; and Varma, M. 2025. MOGIC: Metadata-infused Oracle Guidance for Improved Extreme Classification. In *International Conference on Machine Learning (ICML)*.
- Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*.
- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Thirukovalluru, R.; and Dhingra, B. 2025. GenEOL: Harnessing the Generative Power of LLMs for Training-Free Sentence Embeddings. In *Findings of the Association for Computational Linguistics (NAACL-F)*.
- Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; Hénaff, O.; Harmsen, J.; Steiner, A.; and Zhai, X. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*.
- Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024. Improving Text Embeddings with Large Language Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wang, Y.-S.; Chang, W.-C.; Jiang, J.-Y.; Zhang, J.; Yu, H.-F.; and Vishwanathan, S. 2025. Retrieval-augmented Encoders for Extreme Multi-label Text Classification. *arXiv preprint arXiv:2502.10615*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; and Zheng, B. e. a. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Zhang, D.; Li, S.-W.; Xiao, W.; Zhu, H.; Nallapati, R.; Arnold, A.; and Xiang, B. 2021. Pairwise Supervised Contrastive Learning of Sentence Representations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.
- Zhao, K.; Wu, Q.; Miao, Z.; and Tsuruoka, Y. 2025. Prompt Tuning Can Simply Adapt Large Language Models to Text Encoders. In *Workshop at Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL-W)*.
- Zhou, C.; Dong, J.; Huang, X.; Liu, Z.; Zhou, K.; and Xu, Z. 2024. QUEST: Efficient Extreme Multi-Label Text Classification with Large Language Models on Commodity Hardware. In *Conference on Empirical Methods in Natural Language Processing, Findings Track (EMNLP-F)*.