Multimodal Fusion for Effective Recommendations on a User-Anonymous Price Comparison Platform

Merve Gül Kantarcı
Research and Development Center
iLab
İstanbul, Turkey
mkantarci@ilab.com.tr

Mehmet Gönen
Department of Industrial Engineering
Koç University
İstanbul, Turkey
mehmetgonen@ku.edu.tr

***Abstract*—This study proposes a novel recommendation framework designed for a digital price comparison platform. The challenges arise from the absence of user login and gold labels in item variations, which make effective recommendations tricky. The proposed framework integrates three distinct modalities: product titles using a multilingual BERT model, product images through the CLIP model, and click data via a novel Word2Vec model named Prod2Vec. Three fusion methods were tested to obtain a single unified representation for a given product: early, intermediate, and late fusion. Offline evaluations showcased a significant performance boost when leveraging all three modalities and employing intermediate fusion. The proposed framework achieved an impressive 92% Adjusted Rand Index clustering score at the category level. Fusion with two modalities also proved to be competitively effective, yielding scores between 87% and 88%. The framework was shown to be scalable by maintaining good performance even when we increased the number of categories up to 50. For online evaluations, we selected three representative categories and deployed the best-selected fusion method on the platform through A/B testing against a click-text encoding baseline. Our framework resulted in a significant improvement by increasing the Click-Through Rate from 1.43% to 3.17% across all categories. These findings underscore the efficacy of the proposed framework in enhancing user engagement and interaction with the platform.**

***Index Terms*—Data fusion, product recommendation, representation learning.**

I. Introduction

In recent years, recommendation systems have become popular across various digital domains, including media streaming, social media, and e-commerce. This study focuses on a Turkish digital price comparison platform that is widely used by millions monthly. Given its complex ecosystem, an effective recommendation system is crucial for enhancing the user experience on the platform.

Recommendation systems come in different forms, from suggesting only the most clicked items to advanced ones offering personalized recommendations based on user preferences. The choice among systems depends on business requirements, user data, and precision needs. Our focus on recommending similar products on the platform arises from understanding its unique attributes. Due to its unconventional e-commerce structure and lack of user-generated shopping carts, personalized recommendation systems are less relevant. Additionally, the ability for users to navigate anonymously and the absence of a membership-based transaction system make personalized recommendations impractical. Our strategic shift towards delivering similar product recommendations in a single session addresses user challenges, enhances the user experience, increases site navigation, and indirectly helps advertising strategies. This aligns with our goals of increasing referral numbers and encouraging additional cross-sales.

This research introduces a novel recommendation system tailored to our platform. The framework utilizes a multilingual BERT model [1], [2] for encoding product titles, the CLIP model [3] for product images, and a Word2Vec model [4], [5] to encode click data derived from sessions on the platform. Each modality is encoded separately, and we explore three fusion strategies to merge these three representations into a single representation.

Our evaluation begins with the use of the Adjusted Rand Index (ARI) clustering score to measure the effectiveness of learned representations in separating at the category level. Additionally, qualitative analysis was conducted to understand within-category differences. Based on offline evaluations, we selected the best fusion strategy, which was integrated into the platform for three product categories. In our online evaluation, for a query product, we recommended 15 similar products pulled from the fused embedding space by minimizing the cosine distance. We compared the recommendations of our proposed framework with our existing framework through A/B testing. The existing framework uses Prod2Vec [5] trained similarly with click interactions and the same BERT model as the cold start. Our results demonstrated that adding the image modality and fusing the modalities in the proposed way significantly outperforms the baseline. The contributions of our framework along with extensive evaluations can be listed as

- With an effective fusion strategy, it is possible to encode different modalities separately and still achieve optimal performance. This approach offers the advantage of cost-effective updates and provides an intuitive solution to the cold-start problem.
- The incorporation of the user-click interaction significantly improves learned representations, further highlighting its importance in capturing data-specific features.
- By comparing with our baseline, we showed the effectiveness of adding the image modality.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00174

II. Related Work

Automated recommendation systems have been a longstanding interest within the e-commerce industry. Initial methodologies predominantly relied on a single modality and frequency-based models. However, with recent advancements, transformers [6] has emerged as the de facto standard for encoding textual, visual, and cross-modality data. Notably, CLIP [3] has attracted great attention for encoding text and image data within a unified latent space, eliminating the necessity for external fusion.

E-CLIP [7] transferred the usefulness of CLIP representations through fine-tuning to the e-commerce domain. As the authors stated such models require careful design due to their inherent high cost. Furthermore, using clean data is crucial to reduce the impact of skewed and noisy data. Usage of CLIP in a zero-shot setting is not optimal since text modality excels when it is both language- and domain-specific [8], [9]. Recognizing this, EI-CLIP [10] introduced an entity-aware approach to CLIP, mitigating the influence of commonsense and adapting both textual and visual data to the fashion domain. In this context, the joint optimization of text and image modalities plays a critical role in fine-detail retrieval within the fashion domain in the presence of high-quality metadata. Preceding CLIP, FashionBERT [11] and Kaleido-BERT [12] proposed transformer-based cross-modality vision-language models. These models, prioritizing the image module, put more emphasis on aligning text tokens and image patches to distinguish items at a finer level. In scenarios where global image features such as color, shape, and category detection suffice, the effort to optimize the image modality becomes redundant.

User-item interaction is another widely used modality in the e-commerce domain. Several works [8], [9], [13] employed click data to implicitly encode user preferences by enhancing text representations. These studies used triplet loss [9] or contrastive loss [14], aiming to minimize the distance for positive pair text representations while maximizing it for negative pairs in the latent space. Similarly, E-BERT [15] is a customized BERT model for integrating e-commerce domain knowledge. This involves training at the single product level and the product pairs level. The associated products are derived from shopping statistics, and the model learns similarity using the triplet loss and the proposed neighbor product reconstruction module. However, acquiring high-quality pairs is challenging. The poor selection of negative pairs or an imbalance in pair counts significantly impacts system performance, as noted by Chiu and Shinzato [13]. Moreover, user-item interaction is particularly prone to domain shift. Keeping a multimodal system up-to-date is challenging with this optimization scheme since it requires end-to-end training of the entire system. To address this, our goal is to disentangle all modalities, thereby covering a domain shift at a minimal cost. Several studies [5], [16], [17] applied NLP methods to obtain product representations from click sequences. A recent model, Prod2Vec-Var [5], disentangled user-item interaction from other modalities by training a Word2Vec model solely on click data. It leverages textual cues as additional information. Prod2Vec utilizes BERT representations to diversify recommendations or when there is a lack of click information, but it does not fuse text and click data. Wang *et al.* [18] further disentangled each modality to detect only visual, only textual, and common signals. Later these signals are fused by a neural network. We argued the need for such complex fusion models when single modality systems are the right fit. As proposed by Byvatov and Schneider [19], we experimented with early, intermediate, and late fusion to integrate modalities.

III. Method

In our data set, we have three modalities for a product: image, text, and user-item interactions. We separately encoded these modalities to obtain distinct representations. Subsequently, we fused these representations to obtain a unified and improved representation of products. During inference, we selected the first k similar products with respect to the cosine distances calculated from the learned representations.

Although we showed that the ideal strategy is to combine all three modalities for a product, our experiments also demonstrated that selecting two modalities remains competitive. Hence, we can still provide reliable recommendations in the absence of one of the modalities. Fig. 1 illustrates our pipeline for three modalities. We explore the technical specifications of the encoders and the fusion methods in the remainder.

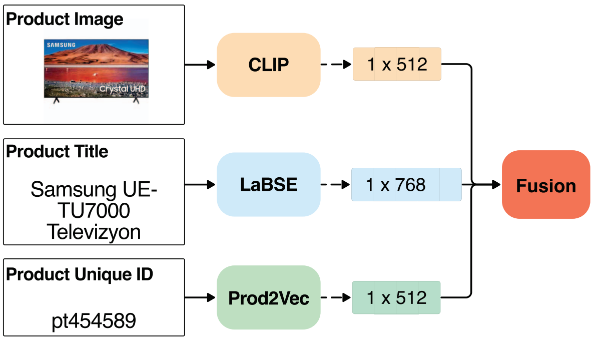

Fig. 1. Method overview with three input modalities.

A. Sentence Encoder

BERT [1], short for Bidirectional Encoder Representations from Transformers, is a natural language processing (NLP) model developed in 2018. It is a pre-trained transformer architecture, excelling in capturing context and understanding semantic relationships within language. Unlike sequential text processing in traditional models, BERT adopts a bidirectional approach, considering both the left and right context in a sentence simultaneously.

BERT comprises multiple layers of transformers with attention mechanisms, allowing the model to focus on various parts of the input sequence. During pre-training, BERT is exposed to substantial amounts of unlabeled text data. It creates a contextualized representation of words by learning to predict missing words in sentences. This pre-trained model can be fine-tuned for specific NLP tasks such as sentence encoding. For this, we employ a multilingual sentence encoding BERT model: Language-agnostic BERT Sentence Encoder (LaBSE) [2].

LaBSE distinguishes itself from other multilingual models [1], [20], [21] by directly targeting the sentence encoding task. It employs a dual-encoder with initially pre-trained BERT weights. The pre-training process involves the Masked Language Model (MLM) [1] and Translation Language Model (TLM) [20]. The training utilizes translation ranking loss with additive margin softmax loss to achieve effective results. 17 billion monolingual sentences and 6 billion translation pairs were used to train the LaBSE model.

B. Image Encoder

Contrastive Language-Image Pre-training (CLIP) [3] is a state-of-the-art model designed for cross-modal tasks. CLIP utilizes a dual-encoder architecture that simultaneously processes both images and text. During pre-training, CLIP is exposed to a vast amount of image and text pairs from the internet, learning to associate corresponding representations for images and their associated textual descriptions. The pre-training of CLIP involves a contrastive loss function, which encourages the model to bring similar image-text pairs closer in the embedding space and push dissimilar pairs farther apart. This loss function creates a semantically meaningful joint embedding space for images and text. The success of CLIP in various downstream tasks is a testament to the effectiveness of its joint embedding space such as object classification, identity detection, and optical character recognition.

C. Click Encoder

Word2Vec [4] is a popular technique in NLP used for word embedding. It basically captures semantic relationships between words by learning distributed representations of words based on their context.

Recently, Turgut *et al.* [5] proposed an adaptation of Word2Vec named Prod2Vec in the product encoding domain, derived from user-item interactions. In this model, Word2Vec is trained using sequences of product IDs. The goal is to associate frequently co-clicked items closely in the Prod2Vec embedding space. Product IDs are treated as words, and sequences are treated as sentences. During inference, the trained Prod2Vec model can provide embeddings for a query product from its vocabulary. One downside of this model is that Prod2Vec cannot easily adapt to new vocabulary since it requires re-training in that case. In their implementation, Turgut *et al.* [5] proposed using LaBSE embeddings as the cold start. In our case, we can use any or both of the other two modalities with the selected fusion method.

In our system, only this modality requires frequent updates based on changes in user behavior. This disentangled modality encoding provides us with the opportunity to cost-effectively fine-tune the click encoder while leaving others unchanged.

D. Fusion Method

After encoding three modalities using separate encoders, we explore various fusion methods at different stages: early, intermediate, and late. Fig. 2 illustrates the fusion methods.

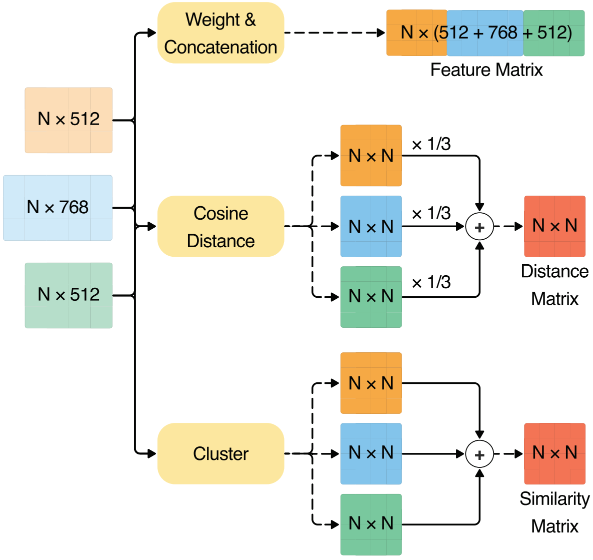

Fig. 2. Fusion methods at different stages.

1) Early Fusion: This fusion method concatenates the feature vectors and incorporates a weighting scheme to mitigate the impact of varying feature sizes. The weighted concatenation is achieved by scaling each feature vector by the reciprocal of the square root of its dimensionality.

Mathematically, for the feature vectors $\boldsymbol{v}_{\text{CLIP}}$, $\boldsymbol{v}_{\text{LaBSE}}$, and $\boldsymbol{v}_{\text{Prod2Vec}}$ with dimensions D_{CLIP}, D_{LaBSE}, and D_{Prod2Vec}, the early fusion representation $\boldsymbol{v}_{\text{early}}$ is computed as

$$\boldsymbol{v}_{\text{early}} = \frac{\boldsymbol{v}_{\text{CLIP}}}{\sqrt{D_{\text{CLIP}}}} \oplus \frac{\boldsymbol{v}_{\text{LaBSE}}}{\sqrt{D_{\text{LaBSE}}}} \oplus \frac{\boldsymbol{v}_{\text{Prod2Vec}}}{\sqrt{D_{\text{Prod2Vec}}}}$$

where $\oplus$ denotes the concatenation operation.

2) Intermediate Fusion: For intermediate fusion, cosine distance matrices are computed for each modality separately, representing the pairwise distances between all product feature vectors in the corpus.

For intermediate fusion, let $\mathbf{D}_{\text{CLIP}}$, $\mathbf{D}_{\text{LaBSE}}$, and $\mathbf{D}_{\text{Prod2Vec}}$ be the cosine distance matrices. The unified pairwise distance matrix $\mathbf{D}_{\text{intermediate}}$ is computed as the mean of the element-wise sum of these matrices:

$$\mathbf{D}_{\text{intermediate}} = \frac{1}{3}(\mathbf{D}_{\text{CLIP}} + \mathbf{D}_{\text{LaBSE}} + \mathbf{D}_{\text{Prod2Vec}})$$

3) Late Fusion: In the late fusion approach, k-means++ clustering is applied to each modality's feature vectors independently to create clusters representing different product categories. Similarity matrices for the whole corpus are then constructed for each modality, where the entry (i, j) is 1 if product i and product j belong to the same cluster, and 0

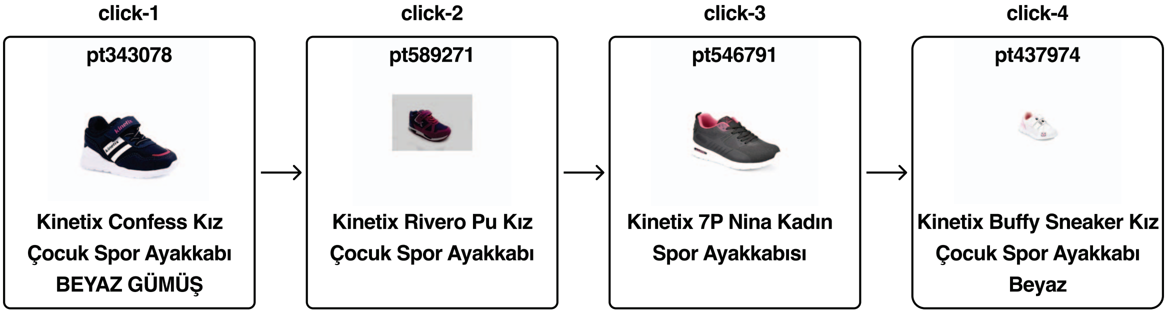
click-1
pt343078
Kinetix Confess Kız Çocuk Spor Ayakkabı BEYAZ GÜMÜŞ
click-2
pt589271
Kinetix Rivero Pu Kız Çocuk Spor Ayakkabı
click-3
pt546791
Kinetix 7P Nina Kadın Spor Ayakkabısı
click-4
pt437974
Kinetix Buffy Sneaker Kız Çocuk Spor Ayakkabı Beyaz

Fig. 3. Example of a session.

otherwise. The final fused similarity matrix $\mathbf{S}_{\text{fused}}$ is obtained by summing the individual matrices across all three modalities.

Let $\boldsymbol{c}_{\text{CLIP}}$, $\boldsymbol{c}_{\text{LaBSE}}$, and $\boldsymbol{c}_{\text{Prod2Vec}}$ be the cluster assignments obtained using k-means++. The final fused similarity matrix $\mathbf{S}_{\text{fused}}$ is computed as follows:

$$\mathbf{S}_{\text{modality}}(i,j) = \begin{cases} 1 & \text{if } \boldsymbol{c}_{\text{modality}}(i) = \boldsymbol{c}_{\text{modality}}(j) \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{S}_{\text{late}} = \mathbf{S}_{\text{CLIP}} + \mathbf{S}_{\text{LaBSE}} + \mathbf{S}_{\text{Prod2Vec}}$$

IV. Experiments

A. Data Set

We call a sequence of clicks by one user a session. Fig. 3 illustrates an example of a session with four clicks. We gathered sessions for two months from our platform and cleaned them. The key statistics of the cleaned data set are presented in Table I.

TABLE I
Data Set Characteristics

Number of unique products	915 817
Total click count	13 616 732
Number of unique users	3 065 399
Number of sessions	4 233 976
Minimum number of clicks per session	2
Maximum number of clicks per session	184
Average number of clicks per session	3.22
Minimum number of words per product title	1
Maximum number of words per product title	43
Average number of words per product title	8.24
Image size in pixels	1000 × 1000

We collected raw text descriptions to be tokenized in inference time. As the platform crawls products from various e-commerce websites, a product can have multiple images that significantly vary or are nearly identical. We used only the first image that users saw. We collected RGB images of size 1000 × 1000 pixels.

B. Evaluation Metric

1) Offline Evaluation: For offline evaluation, we initiated the process by implementing dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE) and subsequently clustering the projected feature vectors through the k-means++ algorithm. The agreement between the ground truth product category labels and the generated cluster labels is quantified using the ARI score.

The ARI score is a measure used to evaluate the agreement between two clustering solutions. It assesses how well the assignments of data points to clusters in one clustering correspond to their assignments in another clustering while adjusting for chance. In our context, this helps us assess if products from the same category are consistently assigned to the same clusters. The ARI score ranges from -1 to 1, where a score of 1 indicates perfect agreement between the two clustering solutions, 0 indicates random labeling, and -1 indicates complete disagreement.

$$\text{ARI} = \frac{\text{RI} - \text{RI}_{\text{expected}}}{\text{RI}_{\text{max}} - \text{RI}_{\text{expected}}}$$

where:

$$\begin{aligned} \text{RI} &= \text{Rand Index} \\ \text{RI}_{\text{expected}} &= \text{Expected RI under independence assumption} \\ \text{RI}_{\text{max}} &= \text{Max. possible RI given the same number of clusters} \end{aligned}$$

2) Online Evaluation: We conducted A/B testing by evenly splitting traffic between the new implementation and the baseline method. Subsequently, we calculated the Click-Through Rate (CTR) for each method. CTR is a crucial metric to evaluate the effectiveness of a recommendation system. It indicates the percentage of users who click on a recommended item after viewing it. It is calculated by dividing the number of clicks by the number of views. Higher CTR values signify greater user engagement and interest.

$$\text{CTR} = \frac{\text{Number of clicks}}{\text{Number of views}}$$

In the same manner, we calculated the Jump Rate (JR), which is defined as the ratio of the number of jumps to the

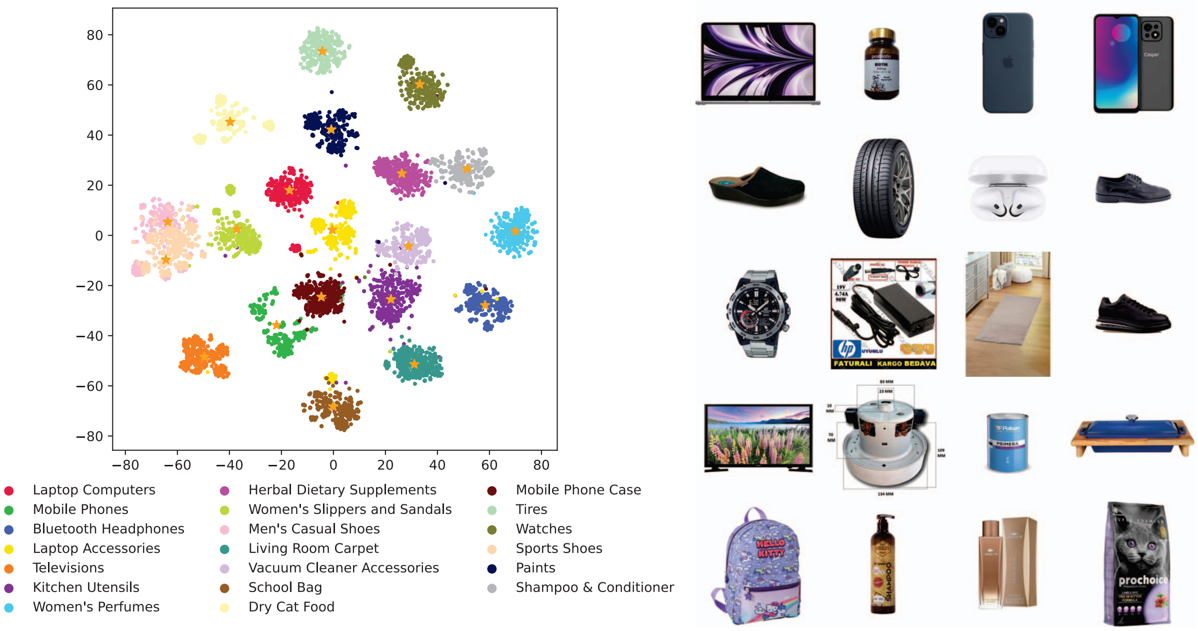

Fig. 4. t-SNE visualization of CLIP+LaBSE+Prod2Vec model with intermediate fusion of 20 categories (left). An example image from the selected categories in the same order with the legend (right).

number of views. Jump refers to a user navigating to the purchase page from the platform.

$$\text{JR} = \frac{\text{Number of jumps}}{\text{Number of views}}$$

C. Experimental Setup

We used the pre-trained versions of the CLIP[1] and LaBSE[2] models without further training and employed the default processors and tokenizers of CLIP and LaBSE. For Prod2Vec, we trained the model from scratch with an embedding size of 512 using a sliding window of 5 with skip-gram.

D. Offline Results

We first tested our framework using offline experiments where we evaluated the representations obtained by different strategies in terms of their visualization and clustering performances.

1) Effect of Modality Selection: We compared the clustering performances of representations by single modality and multiple modality solutions. Table II lists the ARI scores of seven different strategies we tried. Combining two or more modalities was clearly shown to be superior to using just a single modality. We obtained the best results with the model that integrates all three modalities using the intermediate fusion.

In Fig. 4, we visualized the two-dimensional embeddings of products obtained using t-SNE algorithm on the resulting cosine distance matrix of CLIP+LaBSE+Prod2Vec model with intermediate fusion. It is clear that our fusion framework learned representations that can separate product categories, leading to a strong potential for better recommendation performance.

TABLE II
ARI Scores Based on Different Modality Combinations (20 Categories)

	ARI Score		
Modalities	Early	Intermediate	Late
CLIP	0.7254	0.7417	–
LaBSE	0.7723	0.7909	–
Prod2Vec	0.6568	0.6852	–
CLIP+LaBSE	0.8728	0.8680	–
CLIP+Prod2Vec	0.8135	0.8726	–
LaBSE+Prod2Vec	0.8528	0.8798	–
CLIP+LaBSE+Prod2Vec	0.8840	**0.9153**	0.8270

2) Effect of Fusion Method: We also compared the fusion strategies for the CLIP+LaBSE+Prod2Vec scenario. Table II shows the ARI scores of three different fusion methods. Intermediate fusion obtained significantly better results compared to early and late fusion.

3) Effect of Category Count: We also investigated the clustering performances with varying number of product categories, namely, 3, 20, and 50 categories. Table III lists the ARI scores of these three scenarios. As expected, the clustering performance decreased with the increasing number

[1] https://github.com/openai/CLIP
[2] https://huggingface.co/setu4993/LaBSE

of product categories. However, our framework obtained very strong results even with 50 categories.

TABLE III
ARI Scores Based on Different Category Counts (Intermediate Fusion)

	ARI Score	
Category Count	**CLIP+LaBSE+Prod2Vec**	**Prod2Vec**
3	1.0000	0.9148
20	0.9153	0.6852
50	0.7889	0.5320

E. Online Results

We finally tested the online retrieval performance of the CLIP+LaBSE+Prod2Vec scenario with intermediate fusion using an A/B testing experiment, which considers three product categories, namely, herbal dietary supplements, sports shoes, and televisions. When we recommended the products to the real users based on the distance matrix learned by our method, the CTR scores were significantly improved in all three categories. If we look at the total CTR score, it was improved by more than 100% by going from 1.43% to 3.17%. Similarly, we observed that users are more likely to navigate to the purchase page with our recommendations. Overall, our model increased the JR score by approximately 35%. Specifically, in the category of herbal dietary supplements, the JR score was increased by around 7 times.

TABLE IV
Online Evaluation Results In CTR & JR

	CTR		JR	
Categories	**Ours**	**Baseline**	**Ours**	**Baseline**
Herbal Supplements	2.9077%	0.8090%	1.2974%	0.1727%
Sports Shoes	4.3963%	1.5480%	1.1173%	0.7828%
Televisions	2.6774%	1.5133%	0.9676%	0.8880%
Total	3.1741%	1.4331%	1.0569%	0.7838%

V. Conclusion

We investigated different fusion strategies for three distinct modalities of products listed in a digital price comparison platform. The integrated modalities included product titles, product images, and click streams of anonymous users. We used three deep learning models to obtain three distinct representations from three modalities for each product. Our offline and online experiments clearly showed that using a multimodal fusion approach combining two or three modalities significantly outperformed single modality approaches in terms of embedding quality and retrieval performance.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in NAACL, 2019, pp. 4171–4186.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT Sentence Embedding,” in ACL, 2022, pp. 878–891.

[3] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in ICML, 2021, vol. 139, pp. 8748–8763.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in NIPS, 2013, vol. 2, pp. 3111–3119.

[5] H. Turgut, T. D. Yetki, Ö. Bali, and T. A. Yücel, “Prod2Vec-Var: A Session Based Recommendation System with Enhanced Diversity,” in ICIKM, 2023, pp. 5253–5254.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, “Attention is All you Need,” in NIPS, 2017, pp. 6000–6010.

[7] W. Shin, J. Park, T. Woo, Y. Cho, K. Oh, and H. Song, “E-CLIP: Large-Scale Vision-Language Representation Learning in E-Commerce,” in ICIKM, 2022, pp. 3484–3494.

[8] R. Peeters and C. Bizer, “Cross-Language Learning for Product Matching,” in WWW, 2022, pp. 236–238.

[9] J. Tracz, P. I. Wójcik, K. Jasinska-Kobus, R. Belluzzo, R. Mroczkowski, and I. Gawlik, “BERT-based Similarity Learning for Product Matching,” in Proc. of Workshop on Natural Language Processing in E-Commerce, 2020, pp. 66–75.

[10] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, X. Xie, ”EI-CLIP: Entity-aware Interventional Contrastive Learning for E-commerce Cross-modal Retrieval,” in CVPR, 2022, pp. 18030–18040.

[11] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, H. Wang, “FashionBERT: Text and Image Matching with Adaptive Loss for Cross-Modal Retrieval,” in SIGIR, 2020, pp. 2251–2260.

[12] M. Zhuge, D. Gao1, D.-P. Fan, L. Jin, B. Chen, H. Zhou M. Qiu, L. Shao, “Kaleido-BERT: Vision-Language Pre-training on Fashion Domain,” in CVPR, 2021, pp. 12647–12657.

[13] J. Chiu and K. Shinzato, “Cross-Encoder Data Annotation for Bi-Encoder Based Product Matching,” in EMNLP: Industry Track, 2022, pp. 161–168.

[14] M. Guo, N. Yan, X. Cui, S.H. Wu, U. Ahsan, R. West, K.A. Jadda, “Deep Learning-based Online Alternative Product Recommendations at Scale,” in ECNLP, 2020, pp. 19–23.

[15] D. Zhang, Z. Yuan, Y. Liu, F. Zhuang, H. Chen, H. Xiong, ”E-BERT: A Phrase and Product Knowledge Enhanced Language Model for E-commerce,” 2020, arXiv:2009.02835v3.

[16] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan, D. Sharp, “E-Commerce in Your Inbox: Product Recommendations at Scale,” in SIGKDD, 2015, pp. 1809–1818.

[17] F. Bianchi, B. Yu, and J. Tagliabue, “BERT Goes Shopping: Comparing Distributional Models for Product Representations,” in ECNLP, 2021, pp. 1–12.

[18] X. Wang, H. Chen, and W. Zhu, “Multimodal Disentangled Representation for Recommendation,” in ICME, 2021, pp. 1–6.

[19] E. Byvatov and G. Schneider, “Support Vector Machine Applications in Bioinformatics,” Applied Bioinformatics, vol. 2, no. 2, pp. 67–77, 2003.

[20] A. Conneau and G. Lample, “Cross-Lingual Language Model Pretraining,” in NIPS, 2019, pp. 7059–7069.

[21] A. Conneau and D. Kiela, “SentEval: An Evaluation Toolkit for Universal Sentence Representations,” in LREC, 2018.