Image-Seeking Intent Prediction for Cross-Device Product Search

Mariya Hendriksen*
University of Oxford
United Kingdom

Svitlana Vakulenko
Vienna University of Economics and
Business
Austria

Jordan Massiah
Amazon
United Kingdom

Gabriella Kazai
Amazon
United Kingdom

Emine Yilmaz
Amazon
United Kingdom

Abstract

Large Language Models (LLMs) are transforming personalized search, recommendations, and customer interaction in e-commerce. Customers increasingly shop across multiple devices, from voice-only assistants to multimodal displays, each offering different input and output capabilities. A proactive suggestion to switch devices can greatly improve the user experience, but it must be offered with high precision to avoid unnecessary friction. We address the challenge of predicting when a query requires visual augmentation and a cross-device switch to improve product discovery. We introduce Image-Seeking Intent Prediction, a novel task for LLM-driven e-commerce assistants that anticipates when a spoken product query should proactively trigger a visual on a screen-enabled device. Using large-scale production data from a multi-device retail assistant, including 900K voice queries, associated product retrievals, and behavioral signals such as image carousel engagement, we train IRP (Image Request Predictor), a model that leverages user input query and corresponding retrieved product metadata to anticipate visual intent. Our experiments show that combining query semantics with product data, particularly when improved through lightweight summarization, consistently improves prediction accuracy. Incorporating a differentiable precision-oriented loss further reduces false positives. These results highlight the potential of LLMs to power intelligent, cross-device shopping assistants that anticipate and adapt to user needs, enabling more seamless and personalized e-commerce experiences.

Keywords

image-seeking intent, virtual assistants, generative models, cross-device search

---

*Work done while interning at Amazon.

1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art performance across a variety of tasks, including text generation, reasoning, and question answering [5, 7, 21]. Their integration into virtual assistants (VAs) offers new opportunities for intelligent, context-aware interaction [9, 25, 27]. In the e-commerce domain, these assistants operate within a diverse ecosystem of devices, ranging from voice-only smart speakers to screen-enabled smartphones and tablets [2, 26, 34]. Each device supports distinct input and output modalities, and product catalogs increasingly contain multimodal representations that include visual, textual, and structured data. These assistants must retrieve and present such information in modality-appropriate ways. As customer expectations for personalized and visually rich recommendations continue to grow, the ability of LLM-driven assistants to adapt content presentation to device capabilities is becoming a critical requirement.

A growing challenge in this setting is *proactive device switching*, which involves determining when a query issued to a screenless assistant requires a visual response that would be better served on a screen-enabled device. Prior research in cross-device search has primarily focused on predicting the next device a user will access. In contrast, our objective is to determine when a device switch should be suggested to better serve the user's intent. This distinction is particularly important when the user's current device lacks the modalities necessary to meet the request, such as the ability to display product images or other forms of visually enhanced recommendations.

We address this challenge through Image-Seeking Intent Prediction, a novel task for LLM-driven e-commerce assistants. The goal is to predict whether a user query, issued via speech on a screenless device, will require a visual modality to fulfill the user's need. This problem arises when a voice-only assistant must decide if presenting images or other visual content would improve the experience. Once visual intent is detected, the assistant can proactively suggest switching to a screen-enabled device or adapt the interface accordingly. We cast this task as a binary, highly imbalanced classification problem, where a positive label indicates that the current voice-only interaction should be escalated to a screen-based device. The task is motivated by two key use cases: (i) enhancing multi-device user experiences by recommending a switch to a more appropriate modality [10, 19]; and (ii) supporting adaptive user interfaces that respond intelligently to predicted modality needs, for example, by automatically showing images when necessary [20]. The problem presents several challenges: (i) Minimizing false positives — unnecessary device switches disrupt the user experience, making high

precision essential; (ii) Label imbalance — device-switch events are rare, and thus switch-worthy queries form only a small fraction of the data; and (iii) Multimodal context — utterances are spoken, while retrieved products contain a mixture of textual, structured, and visual content.

Dataset. Direct supervision for device-switching behavior is rarely available, and manually annotating visual intent at scale is prohibitively expensive. We address this by defining a proxy task that leverages behavioral signals on screen-enabled devices. Specifically, we use the action of tapping to open the image carousel as a weak but scalable indicator of visual intent. This interaction enables large-scale supervision from real-world data without manual labeling. The model is trained to predict, given an utterance and retrieved products, whether the user would have activated the image carousel, signaling a transition to a visual modality. We construct a large-scale dataset of 2.2M user interactions collected over six months from a production-grade multi-device assistant system. The dataset includes over 900,000 spoken utterances paired with the top-10 retrieved products per query, along with behavioral labels from downstream interaction signals.

Image Request Predictor. To address Image-Seeking Intent Prediction, we propose IRP, a model with a generative backbone that predicts image-seeking intent from the user's utterance and associated retrieved product information. IRP jointly encodes the utterance and product summaries to make predictions about visual needs. We compare IRP against strong baselines using only utterance transcriptions, intent classifications, or raw product titles. We find that combining utterance data with structured, summarized product information yields significant performance gains.
We answer the following research questions:

(RQ1) How does the integration of utterance-level and product-level features influence predictive performance in Image-Seeking Intent Prediction?

(RQ2) What is the effect of representing product information through lightweight summarization compared to using raw product titles or attributes?

(RQ3) To what extent does optimizing the loss function explicitly for precision influence performance in this high-precision, imbalanced classification setting?

The principal contributions of our research are the following:

(i) We define Image-Seeking Intent Prediction, a novel task for language model-based cross-device assistants in e-commerce, and propose the Image Request Predictor (IRP), a model for the task.

(ii) We conduct extensive evaluation on a large-scale proprietary dataset of real user interactions in a multi-device e-commerce assistant system.

(iii) We discover that combining query semantics with summarized product data yields consistent accuracy gains, while a differentiable precision-oriented loss further reduces false positives.

2 Methodology

We formalize the Image-Seeking Intent Prediction task and present IRP, an architecture with three modules: an utterance processing
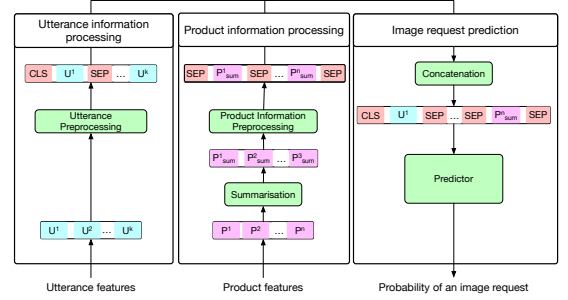
Figure 1: IRP architecture. The query and product information are processed through respective pipelines, fused into a joint representation, and passed to a prediction head that estimates the likelihood of the image-seeking intent.

pipeline, a product information processing pipeline, and an image request prediction module. Together, these components learn to determine whether a spoken query requires switching from a voice-only to a screen-enabled device.

2.1 Task definition

Following prior work [4, 28, 35], we represent each user interaction as a pair $(U, \mathcal{P})$, where U is the user query and $\mathcal{P} = \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)}$ is the set of the top-k products retrieved in response to U. Each query U consists of a transcription u_q and an associated intent label u_{int}. Each retrieved product $\mathbf{p}^{(j)} \in \mathcal{P}$ is described by a structured set of attributes:

$$\mathbf{p}^{(j)} = \{p_{title}^{(j)},\ p_{brand}^{(j)},\ p_{size}^{(j)},\ p_{color}^{(j)},\ p_{reviews}^{(j)},$$
$$p_{price}^{(j)},\ p_{style}^{(j)},\ p_{group}^{(j)},\ p_{type}^{(j)}\} \tag{1}$$

Due to the lack of labeled supervision for device switching events and the high cost of manual annotation, we adopt a scalable proxy task. Specifically, we use the user's interaction with an image carousel (i.e., whether they tap to open it) as an implicit signal of visual intent. This assumption enables the training of models at scale using real behavioral data.

2.2 Image Request Predictor

Figure 1 illustrates the architecture of IRP. The model processes utterance and product information through separate pipelines, summarizes retrieved product representations, and fuses them to predict whether the user's request implies a need for visual content.

Utterance Processing Pipeline. The user query U is transcribed into text and passed to a preprocessing function f_{up}, which converts it into a tokenized sequence:

$$seq_u = f_{up}(U) = ([CLS], u^1, [SEP], \ldots, u^K, [SEP]), \tag{2}$$

where [CLS] and [SEP] are standard special tokens.

Product Information Processing Pipeline. The retrieved products $\mathcal{P}$ are summarized using a function f_{sum}, which produces an aggregated vector:

$$\mathbf{p}_{sum} = f_{sum}(\mathcal{P}) = \frac{1}{k} \sum_{j=1}^{k} \mathbf{p}^{(j)}. \tag{3}$$

This summary is then tokenized via a product preprocessing function f_{pp}:

$$\text{seq}_p = f_{pp}(\mathbf{p}_{sum}) = ([SEP], p^1, [SEP], \ldots, p^N, [SEP]). \quad (4)$$

Fusion and Prediction. The utterance and product sequences are integrated by a fusion module f_{fusion}, parametrized to operate either as direct sequence concatenation or as concatenation over a reduced set of product tokens obtained via Maximal Marginal Relevance (MMR) selection [6]. Formally:

$$X = f_{fusion}(\text{seq}_u, \text{seq}_p) \quad (5)$$
$$= ([CLS], u^1, \ldots, u^K, [SEP], p^1, \ldots, p^N, [SEP]), \quad (6)$$

where $(p^1, \ldots, p^N)$ denotes either all product tokens or the MMR-selected subset. The fused representation X is then passed through a classification head f_{pred} to produce the probability of image-seeking intent:

$$p = f_{pred}(X), \quad 0 \leq p \leq 1. \quad (7)$$

2.3 Training Objectives

We model Image-Seeking Intent Prediction as binary classification and evaluate three loss formulations.

Binary Cross-Entropy (BCE). The BCE loss minimizes cross-entropy between predicted probability p and label $y \in \{0, 1\}$:

$$\mathcal{L}_{BCE} = -y \log(p) - (1 - y) \log(1 - p). \quad (8)$$

Precision Loss. To penalize false positives explicitly, we use a differentiable approximation of precision [3, 22]:

$$\mathcal{L}_{Precision} = 1 - \frac{TP}{TP + FP}, \quad (9)$$

where $TP = \sum \hat{y} \cdot y$ and $FP = \sum \hat{y} \cdot (1 - y)$, with $\hat{y}$ as the model's sigmoid output.

Combined Loss. To balance general accuracy with false-positive reduction, we define:

$$\mathcal{L}_{Sum} = \alpha \mathcal{L}_{BCE} + \beta \mathcal{L}_{Precision}, \quad (10)$$

where α and β are hyperparameters controlling the trade-off between optimizing overall classification accuracy and prioritizing high precision.

3 Experiments

We empirically evaluate the Image-Seeking Intent Prediction formulation and IRP on large-scale real-world interaction data from a production conversational e-commerce assistant. We describe the dataset, model backbones, and evaluation metrics, then report results for feature contributions, product summarization, precision-oriented loss functions, and backbone comparisons, followed by implementation details.

3.1 Dataset

We construct a large-scale dataset from de-identified interaction logs collected between August and December 2022 from a production conversational e-commerce assistant. The system supports both screenless devices (e.g., smart speakers) and screen-enabled devices (e.g., mobile phones, tablets), making it a natural setting for studying cross-modality behavior.

Data Sampling and Filtering. We start with 2,200,000 user interactions sampled over five months, containing both utterance data and downstream behavioral signals such as taps. Approximately 10% of these interactions include an image gallery opening action. To focus on voice-initiated interactions, we filter for utterances explicitly triggered by the user. This yields a high-quality subset of 900,000 utterance-based interactions, in which 20% are followed by an image gallery opening. This proportion provides a sufficiently large and meaningful set of positive examples for our proxy task.

Dataset Characteristics. The resulting dataset reflects realistic user behavior in a high-traffic, multi-device e-commerce assistant. Each instance in $\mathcal{D}$ is represented as a tuple $(U^i, \mathcal{P}^i, y^i)$, where U^i is a spoken utterance with both transcription and intent; $\mathcal{P}^i = \mathbf{p}^{(i,1)}, \ldots, \mathbf{p}^{(i,k)}$ is the set of top-k retrieved products for U^i; and $y^i \in 0, 1$ is a binary label indicating whether the user opened the image carousel, serving as a proxy for visual intent. Each product $\mathbf{p}^{(i,j)}$ contains structured metadata including title, brand, size, color, price, reviews, style, and product category. The dataset thus captures: (i) natural language voice queries across diverse product search contexts, (ii) real-world retrieval results from a production search system, and (iii) implicit behavioral signals indicative of multimodal information needs.

This configuration enables training and evaluation for the Image-Seeking Intent Prediction task at scale, using weak supervision derived from real-world behavioral data in place of costly manual annotation.

3.2 Model Backbones

In all experiments, the Image Request Predictor (IRP) uses transformer-based language model backbones. We evaluate three pretrained architectures:

- **DistilBERT** [24]: A compressed variant of BERT that reduces parameters by 40% and achieves up to 60% faster inference while retaining over 95% of BERT's performance. It is trained via knowledge distillation with a composite loss combining language modeling, distillation, and cosine embedding objectives.
- **RoBERTa** [17]: An optimized variant of BERT that removes Next Sentence Prediction, uses larger mini-batches and learning rates, trains on substantially more data, and applies dynamic masking, resulting in consistent performance improvements over BERT.
- **XLNet** [32]: A generalized autoregressive model that captures bidirectional context without input corruption by maximizing expected log-likelihood over all factorization permutations. It incorporates Transformer-XL features such as segment recurrence and relative positional encoding to better handle long-range dependencies.

These backbones are integrated into IRP for encoding the concatenated utterance and product sequences (see Section 2), and we use them consistently throughout training and evaluation. Backbone comparisons are reported in an ablation study later in this section.

Table 1: Feature Contribution Analysis.

	Precision	Recall	F_{05}
Utterance	77.83	64.92	73.41
Intent	76.12	40.53	56.32
Title	76.59	56.76	68.90
Utterance, intent	77.93	65.11	73.55
Intent, title	77.34	57.81	69.77
Utterance, title	77.91	63.75	72.94
Utterance, intent, title	78.11	64.19	73.23

3.3 Metrics

In Image-Seeking Intent Prediction, the high cost of unnecessary cross-device switches requires prioritizing *precision*, the proportion of predicted positives that are correct:

$$P = \frac{TP}{TP + FP}, \tag{11}$$

where TP and FP denote true and false positives, respectively. We also report *recall*, the proportion of actual positives correctly identified:

$$R = \frac{TP}{TP + FN}, \tag{12}$$

where FN denotes false negatives. To jointly assess both metrics, we use the F_β score:

$$F_\beta = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}, \tag{13}$$

where β controls the relative weight of recall. We set $\beta = 0.5$, placing greater emphasis on precision while retaining sensitivity to recall.

3.4 Experimental Setup and Results

We conduct three experiments to answer the research questions outlined in Section 1. Each experiment investigates a specific aspect of the model architecture or training strategy. For each, we present the setup, results, and key insights.

3.4.1 Feature Contribution Analysis. We begin by evaluating the contribution of different feature types to model performance. We test models trained with: (i) utterance transcription, (ii) utterance intent, (iii) product title (top-1 product), (iv) combinations of the above. Utterance features U are tokenized via f_{up} to produce seq_u, and product titles P are processed via f_{pp} to produce seq_p. Combined features are fused via f_{concat} and passed to the image request predictor.

Results. Table 1 shows that utterance features alone achieve the best performance across all metrics. However, combining utterance with intent and product title yields even better Precision (78.11) and a strong overall performance. Notably, utterance + intent gives the highest Recall (65.11) and $F_{0.5}$ (73.55).

Upshot. Utterance features are the strongest individual signal, but combining them with product metadata enhances performance. Feature fusion significantly improves Precision, which is critical for avoiding unnecessary modality switches.

Table 2: Retrieved Product Summarization.

	Precision	Recall	F_{05}
Agg. by mean (f_{sum})	78.47	60.42	71.61
MMR [6]	78.27	61.94	72.27

3.4.2 Product Summarization and Representation. We expand the product representation by including all selected product attributes (title, brand, size, color, reviews, price, style, group, and type). Due to input length constraints, we summarize the top-k retrieved products using: (i) Mean pooling (f_{sum}), (ii) Maximal Marginal Relevance (MMR) [6]. The summarized vector $\mathbf{P}_{sum}$ is then tokenized and concatenated with utterance features as in Experiment 1.

Results. As shown in Table 2, aggregation by mean improves Precision (78.47), while MMR yields the highest Recall (61.94) and $F_{0.5}$ (72.27). Both summarization techniques lead to better results than using product titles alone.

Upshot. Incorporating a richer product representation improves model performance, especially when combined with effective summarization. Mean pooling favors precision, while MMR offers a better balance between precision and recall.

3.5 Ablation Studies

We conduct a series of ablation experiments to isolate the effects of main design choices in IRP.

3.5.1 Ablation on Loss Functions. We compare three training objectives: (i) $\mathcal{L}_{BCE}$, the standard binary cross-entropy loss; (ii) $\mathcal{L}_{Precision}$, a differentiable surrogate for precision; and (iii) $\mathcal{L}_{Sum}$, a weighted combination of the two. Each loss function is evaluated across multiple input configurations, including single-feature models (e.g., utterance, intent, or title), feature pairs, and the full input setting using all features.

Results. Table 3 summarizes the outcomes. Models trained with $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Sum}$ achieve the best performance in eleven configurations each, while $\mathcal{L}_{Precision}$ leads in only three. In most settings, $\mathcal{L}_{Sum}$ is competitive or superior to the others in $F_{0.5}$.

Upshot. While BCE remains a strong baseline, incorporating precision into the loss yields consistent improvements in high-precision use cases. The combined loss strikes a favorable balance, especially for rare-event detection where false positives are costly.

3.5.2 Ablation on Predictor Backbone. To assess the trade-off between performance and model efficiency, we conduct an ablation study evaluating different transformer backbones for the image request predictor module in IRP. Specifically, we compare three widely used sequence classification architectures: distilBERT [24], RoBERTa [17], and XLNet [32].

As shown in Table 4, XLNet achieves the best overall performance across all metrics. However, distilBERT performs competitively while using significantly fewer parameters – just one-fifth the size of XLNet – highlighting its suitability for deployment in resource-constrained environments.

Table 3: Ablation on Loss Functions.

	Loss	Precision	Recall	F_{05}
Utterance	$\mathcal{L}_{\text{BCE}}$	77.83	64.92	73.41
	$\mathcal{L}_{\text{Precision}}$	77.59	65.54	73.55
	$\mathcal{L}_{\text{Sum}}$	77.89	64.02	73.05
Intent	$\mathcal{L}_{\text{BCE}}$	76.12	40.53	56.32
	$\mathcal{L}_{\text{Precision}}$	76.32	31.03	42.43
	$\mathcal{L}_{\text{Sum}}$	76.11	40.53	56.33
Title	$\mathcal{L}_{\text{BCE}}$	76.59	56.76	68.90
	$\mathcal{L}_{\text{Precision}}$	76.36	53.39	66.87
	$\mathcal{L}_{\text{Sum}}$	76.56	56.63	68.82
Utterance, intent	$\mathcal{L}_{\text{BCE}}$	77.93	65.11	73.55
	$\mathcal{L}_{\text{Precision}}$	77.48	65.63	73.53
	$\mathcal{L}_{\text{Sum}}$	77.91	63.67	72.90
Intent, title	$\mathcal{L}_{\text{BCE}}$	77.34	57.80	69.77
	$\mathcal{L}_{\text{Precision}}$	77.23	52.38	66.52
	$\mathcal{L}_{\text{Sum}}$	77.09	58.52	70.04
Utterance, title	$\mathcal{L}_{\text{BCE}}$	77.91	63.75	72.93
	$\mathcal{L}_{\text{Precision}}$	77.64	63.64	72.75
	$\mathcal{L}_{\text{Sum}}$	77.99	64.88	73.48
Utterance, intent, title	$\mathcal{L}_{\text{BCE}}$	78.11	64.19	73.23
	$\mathcal{L}_{\text{Precision}}$	77.86	62.61	72.39
	$\mathcal{L}_{\text{Sum}}$	78.20	63.60	73.01
IRP	$\mathcal{L}_{\text{BCE}}$	78.47	60.42	71.61
	$\mathcal{L}_{\text{Precision}}$	77.99	62.87	72.58
	$\mathcal{L}_{\text{Sum}}$	78.22	63.27	72.87

Table 4: Ablation on Predictor Backbone.

Architecture	Precision	Recall	$F_{0.5}$	#Params
DistilBERT [24]	78.47	60.42	71.61	66M
RoBERTa [17]	78.78	59.28	71.14	125M
XLNet [32]	78.79	60.64	71.85	340M

3.6 Implementation Details

All models are trained for 20 epochs using a batch size of $\beta = 128$, a learning rate of $\eta = 2e-5$ and a weight decay of $\lambda = 1e-2$. Optimization is performed using the AdamW optimizer [18]. We initialize all models using publicly available pretrained checkpoints [30].

4 Related Work

Multimodal Retrieval in E-Commerce. mariyaE-commerce retrieval systems increasingly leverage multimodal inputs such as text, images, and structured product metadata for ranking and recommendation [11, 12, 26, 35]. Vision–language models have been used to improve product understanding and retrieval quality [15, 31], while virtual assistants are evolving toward richer multimodal interactions [20]. However, most prior work assumes a fixed modality, focusing on improving retrieval accuracy or dialogue generation. In contrast, we address the problem of deciding when a query is best served via a visual modality. Rather than fusing modalities for ranking, we use retrieved product metadata to infer visual content needs, enabling proactive e-commerce assistant behavior.

User Intent Prediction. Inferring user intent from implicit behavioral signals is a long-standing topic in information retrieval and recommendation [1, 16]. Clicks, hovers, and scrolls have been widely used in learning-to-rank and personalized search. More recent work has modeled multiple latent intents from multimodal data and implicit feedback [31], or inferred visual intent from user actions in conversational settings [33]. Our approach shares the use of implicit feedback but differs in focus and supervision: (i) we predict modality preference rather than item relevance or task intent, and (ii) we employ a specific behavioral signal (image carousel taps) as a scalable, weakly supervised proxy for visual information need, enabling high-precision intent models without manual labeling.

Cross-Device Search. Cross-device behavior has been extensively studied in web and mobile search [10, 19, 29], often modeling device transitions within multi-session tasks. Recent work explores proactive adaptation in assistants, where systems adjust to situational context such as attention, environment, or device constraints [8, 13, 23]. We differ in scope and formulation: instead of predicting the next device, we predict whether a voice-issued query on a screenless device requires a visual response. This prediction can proactively trigger a device-switch suggestion. By combining utterance understanding, product metadata, and weak supervision, our method enables assistants to anticipate and act on latent modality needs in real time.

5 Conclusions

We introduced Image-Seeking Intent Prediction, the task of predicting when a query to a screenless assistant should trigger a proactive switch to a screen-enabled device. This setting is important for cross-device e-commerce assistants, where some interactions require visual augmentation. Given the high cost of unnecessary switches, we emphasized precision and optimized for both Precision and $F_{0.5}$. We proposed IRP, which integrates features from user utterances with product data, and improved training with a combined binary cross-entropy and differentiable precision loss, reducing false positives while maintaining strong coverage.

Using a large-scale proprietary dataset of over 900,000 utterance–product pairs, we observed three main findings. First, while utterance features alone are strong predictors, augmenting them with structured product attributes yields consistent gains. Lightweight summarization methods such as mean pooling and Maximal Marginal Relevance (MMR) improve accuracy without substantial overhead. Second, a precision-oriented auxiliary loss further reduces incorrect visual triggers, which is essential for maintaining a seamless user experience. Third, in backbone comparisons, XLNet achieved the highest predictive performance, whereas distilBERT offered a favorable trade-off between accuracy and efficiency.

These results underline the value of multimodal fusion and task-aligned optimization for cross-device virtual assistants. They also highlight the promise of lightweight generative models, adapted for image-seeking intent prediction, as a practical path toward scalable and responsive production systems. Compact generative models, when paired with effective summarization and loss design, can approach the accuracy of larger architectures while remaining suitable for latency- and resource-constrained environments.

Limitations. Our approach is trained using a proxy signal—image carousel interactions—that may not capture all cases of visual intent,

leading to potential label noise. The method also depends on structured, high-quality product metadata, which may be unavailable or noisy in other domains. Moreover, while we evaluate multiple transformer backbones, we do not explore recent vision–language or multimodal pretraining approaches that could further improve performance. Finally, our evaluation is offline; live deployment studies are required to assess real-world effectiveness, user satisfaction, and unintended consequences.

Future Work. Future directions include incorporating cross-modal attention within the product encoder, exploring multimodal transformers to improve fusion, and collecting explicit human annotations for visual intent to enable more accurate benchmarking [14]. Live A/B testing could provide deeper insights into user experience impacts and system-level trade-offs.

References

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 19–26.

[2] Yang Bai, Nakul Garg, Harshvardhan Takawale, Anupam Das, and Nirupam Roy. 2023. Natural Voice Interface for the Next Generation of Smart Spaces. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 140–140.

[3] Gabriel Bénédict, Hendrik Vincent Koops, Daan Odijk, and Maarten de Rijke. [n. d.]. sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification. *Transactions on Machine Learning Research* ([n. d.]).

[4] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. 2020. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 677–694.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020).

[6] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[8] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[9] Sashank Gondala, Lyan Verwimp, Ernest Pusateri, Manos Tsagkias, and Christophe Van Gysel. 2021. Error-driven pruning of language models for virtual assistants. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7413–7417.

[10] Shuguang Han, Zhen Yue, and Daqing He. 2015. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 1–34.

[11] Mariya Hendriksen. 2022. Multimodal retrieval in e-commerce: From categories to images, text, and back. In *European Conference on Information Retrieval*. Springer, 505–512.

[12] Mariya Hendriksen, Maurits Bleeker, Svitlana Vakulenko, Nanne Van Noord, Ernst Kuiper, and Maarten De Rijke. 2022. Extending CLIP for Category-to-image Retrieval in E-commerce. In *European Conference on Information Retrieval*. Springer, 289–303.

[13] Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke. 2020. Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers. In *eCOM 2020: The 2020 SIGIIR Workshop on eCommerce*. ACM.

[14] Mariya Hendriksen, Shuo Zhang, Ridho Reinanda, Mohamed Yahya, Edgar Meij, and Maarten de Rijke. 2025. Benchmark Granularity and Model Robustness for Image-Text Retrieval: A Reproducibility Study. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3183–3193.

[15] Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, and Ana Peleteiro Ramallo. 2024. Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation. *arXiv preprint arXiv:2409.11860* (2024).

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

[19] George D Montanez, Ryen W White, and Xiao Huang. 2014. Cross-device search. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 1669–1678.

[20] Bruno Oliveira and Carla Teixeira Lopes. 2023. The Evolution of Web Search User Interfaces-An Archaeological Analysis of Google Search Engine Result Pages. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 55–68.

[21] OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4.

[22] Yash Patel, Giorgos Tolias, and Jiří Matas. 2022. Recall@ k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7502–7511.

[23] Chris Samarinas and Hamed Zamani. 2024. ProCIS: A benchmark for proactive retrieval in conversations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 830–840.

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[25] Sonal Sannigrahi, Thiago Fraga-Silva, Youssef Oualil, and Christophe Van Gysel. 2024. Synthetic query generation using large language models for virtual assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2837–2841.

[26] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2020. Challenges and research opportunities in eCommerce search and recommendations. *SIGIR Forum* 54, 1 (2020), 2:1–2:23.

[27] Christophe Van Gysel, Manos Tsagkias, Ernest Pusateri, and Ilya Oparin. 2020. Predicting entity popularity to improve spoken entity recognition by virtual assistants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1613–1616.

[28] Ali Varamesh, Ali Diba, Tinne Tuytelaars, and Luc Van Gool. 2020. Self-supervised ranking for representation learning. *arXiv preprint arXiv:2010.07258* (2020).

[29] Yu Wang, Xiao Huang, and Ryen W White. 2013. Characterizing and supporting cross-device search tasks. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 707–716.

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[31] Wei Yang and Qingchen Yang. 2024. Multimodal-aware Multi-intention Learning for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5663–5672.

[32] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[33] Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. 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)*. 6142–6152.

[34] Youyuan Zhang, Sashank Gondala, Thiago Fraga-Silva, and Christophe Van Gysel. 2024. Server-side rescoring of spoken entity-centric knowledge queries for virtual assistants. *International Journal of Speech Technology* 27, 2 (2024), 367–375.

[35] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*. PMLR, 2–25.