MIND: Multimodal Shopping Intention Distillation from Large Vision-language Models for E-commerce Purchase Understanding

Anonymous ACL submission

Abstract

Improving user experience and providing personalized search results in E-commerce platforms heavily rely on understanding purchase intention. However, existing methods for acquiring large-scale intentions bank on distilling large language models with human annotation for verification. Such an approach tends to generate product-centric intentions, overlook valuable visual information from product images, and incurs high costs for scalability. To address these issues, we introduce MIND, a multimodal framework that allows Large Vision-Language Models (LVLMs) to infer purchase intentions from multimodal product metadata and prioritize human-centric ones. Using Amazon Review data, we apply MIND and create a multimodal intention knowledge base, which contains 1,264,441 intentions derived from 126,142 co-buy shopping records across 107,215 products. Extensive human evaluations demonstrate the high plausibility and typicality of our obtained intentions and validate the effectiveness of our distillation framework and filtering mechanism. Further experiments reveal the positive downstream benefits that MIND brings to intention comprehension tasks and highlight the importance of multimodal generation and role-aware filtering. Additionally, MIND shows robustness to different prompts and superior generation quality compared to previous methods.

1 Introduction

Understanding customers' intentions behind their purchase behaviors remains crucial in E-commerce as it potentially benefits several downstream tasks, such as product recommendation (Grbovic et al., 2015; Zhao et al., 2014; Li et al., 2020) and search query answering (Zhao et al., 2019; Hirsch et al., 2020). Unlike traditional factual knowledge related to products, intentions are implicit mental states of customers, which typically require commonsense

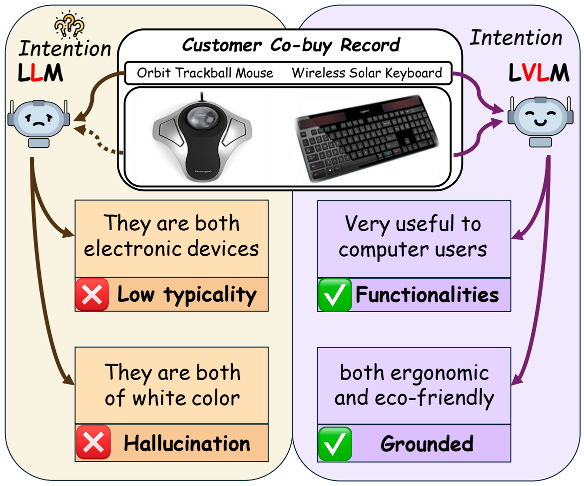

Figure 1: Examples showing the process of distilling purchase intentions from large language models and large vision-language models. Without product images, large language models tend to generate intentions with low typicality and hallucinated facts, while leveraging large vision-language models resolve such issue.

knowledge to understand and reason upon (Bratman, 1984). For example, in Figure 1, the intentions of purchasing a mouse and a keyboard can be *they are very useful to computer users*, which is not mentioned either in the customer's query or products' metadata. Thus, due to such implicitness, it is infeasible to perform large-scale automatic extraction from text to obtain them.

To combat this, Yu et al. (2023) proposed to distill purchase intentions from large language models, such as OPT (Zhang et al., 2022), by prompting them with real purchasing records and relevant product metadata. Human-in-the-loop annotations are also carried out to verify the plausibility and typicality of the generated intentions and train a discriminator for large-scale critic filtering. Yu et al. (2024) further entangled human annotations with instruction tuning to align the distilled intentions with a human-centric perspective. While these works provide a straightforward approach to intention acquisition, several limitations still persist.

First, previous works on E-commerce intention knowledge base construction have solely focused on the text modality, thereby sacrificing significant supervision signals from visual modalities, such as product images. This oversight hinders the model from obtaining a more comprehensive understanding of the product, consequently compromising the quality of the generated intentions, as demonstrated in the left lower part of Figure 1. Furthermore, recent work has shown that intentions derived using current distillation methods exhibit bias towards product-centric aspects, excessively emphasizing product properties and metadata (Zhou et al., 2024). Consequently, interactions between the products and customers, including potential use cases and features of interest to customers, are absent from the derived intentions, despite being fundamental in facilitating customers' shopping experience. Finally, human annotations are heavily deployed in current intention collection methods, which serve as a critical step in controlling the quality of the generated results. This poses a challenge towards constructing scalable yet diverse intention knowledge bases with minimum human supervision cost.

To address these issues, we propose MIND, a **M**ultimodal Shopping **I**ntentio**N** **D**istillation framework. MIND instructs Large Vision-Language Models (LVLMs) to generate purchase intentions in a three-step manner, based on real user co-buy records and product metadata. Specifically, we select LLaVa (Liu et al., 2023a) as a representative LVLM and incorporate both visual information from the product images and text information from the product name into the generation process. To better align the generated raw intentions with human preferences and alleviate human annotation costs for further quality control, we propose a human-centric role-aware mechanism. This mechanism first instructs LLaVa to discover similar features between the products and then imitates a customer agent to decide whether the products would be bought together under previously generated intentions.

By applying MIND to a subset of the Amazon Review Dataset (Ni et al., 2019), we construct a multimodal intention knowledge base. It features 1.26 million of intentions over 126,142 co-buy shopping records across 107,215 products. Human evaluations further confirm: (1) the exceptional quality of our generated intentions, which have higher plausibility and typicality than previous generation methods, and (2) the effectiveness of our proposed human-centric role-aware mechanism. Furthermore, we apply our generated intentions to two downstream tasks in the IntentionQA benchmark (Ding et al., 2024), which evaluates a language model's abilities to discriminate and utilize purchase intentions. Extensive experiments show that distilling our generated intentions into large language models' provide substantial benefits on both tasks via fine-tuning. Further ablation studies reveal the importance of incorporating visual cues of products in MIND and the necessity of integrating our proposed role-aware filter mechanism. Moreover, analyses demonstrate the remarkable diversity of MIND's intentions and the exceptional robustness of MIND when dealing with different prompts. Our code, data, and models will be released upon acceptance.

2 Related Works

2.1 Shopping Intention in E-commerce

Shopping intention is an implicit mental state that motivates purchase-related behaviors from the customer's perspective (Koo and Ju, 2010). Various studies have been conducted to examine the impact of consumer shopping intentions on downstream applications (Dai et al., 2006; Zhang et al., 2016; Hao et al., 2022). Recently, Ni et al. (2019) suggested using customer reviews to investigate the underlying purchase intentions in consumer purchase behavior and created a large-scale review dataset based on Amazon. Building upon this, Yu et al. (2023) proposed FolkScope, which aims to guide LLMs in generating user co-buy intentions for different product pairs by grounding them in ConceptNet relations (Speer et al., 2017). While human evaluations confirmed its effectiveness, Zhou et al. (2024) argued that it not only remains expensive to scale up but also fails to align the resulting shopping intentions with human preferences, which encompass a wide range of factors beyond product properties and similarities. To tackle these issues, in our work, we propose MIND, a framework that undermines online co-buy intentions and aligns better with human perceptions.

2.2 Multimodal Knowledge Distillation

Since VLMs have yield significant advance recently (Liu et al., 2023d; Li et al., 2023; Zhu et al., 2023), distilling domain-specific knowledge from them has become an effective yet cost-friendly

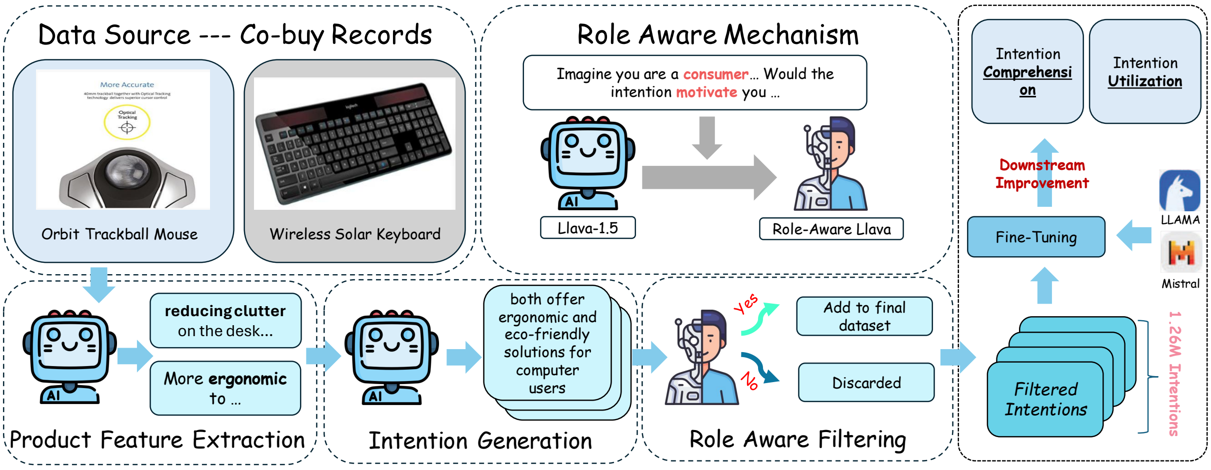

Figure 2: An overview of MIND. We first extract features from products in real-world co-buy records, generate intentions multimodally, and apply a human-centric role-aware filter for quality optimization.

trend in multimodal studies (Liu et al., 2023c; Lu et al., 2024; Jin et al., 2021). Liu et al. (2023c) proposed a framework that applies self-distillation to stimulate the pre-train process of BERT to improve its performance in E-commerce product understanding tasks. Lu et al. (2024) similarly instructed MiniGPT4 (Zhu et al., 2023) to generate user intention form social media posts text and its associated images. Jin et al. (2021) also designed a framework to instruct the student model to imitate teacher model's behavior, which successfully preserved the teacher model's capabilities with fewer parameters. In our work, we share the same aspiration and leverage distillation as a tool for data collection that provides downstream benefits in the E-commerce domain. Specifically, we designed a framework to distill E-commerce intentions from LLaVa (Liu et al., 2023a) and construct a comprehensive intention knowledge base based on the resulted generations.

3 The MIND Framework

3.1 Overview of MIND

Following Yu et al. (2023), the objective of MIND is formulated as a text generation task. Given a record that shows a customer's co-buy (purchasing together) of two products, along with the detailed metadata of both products, MIND aims to generate the intentions behind such purchase behaviors that best align with the customer's mental state during the purchase, which includes their beliefs, desires, and intents (Georgeff et al., 1999).

Formally, for a given co-buy record, we define the two products as p^1 and p^2, along with their associated images $\{p_i^1, p_i^2\}$, and features and attributes $\{p_f^1, p_f^2\}$. MIND aims to leverage a LVLM F to generate the intentions $I(p^1, p^2)$ of purchasing both products based on a pre-defined commonsense relation r, denoted as $I(p^1, p^2, r) = F(p_f^1, p_f^2, p_i^1, p_i^2, r)$. In this paper, we follow Yu et al. (2023, 2024) and use relations from ConceptNet (Speer et al., 2017) to model the intentions. LLaVa-1.5-13b (Liu et al., 2023a) is used as the LVLM F.

To achieve this objective, we design three sequentially connected steps within MIND, which are shown in Figure 2. These steps are termed as: (1) product feature extraction; (2) co-buy intention generation; (3) human-centric role-aware filtering. Together, they form a collective pipeline for systematic intention acquisition without the need for human supervision and quality filtering.

3.2 Source Data Collection

We utilize the Amazon Review Data released by Ni et al. (2019), which contains millions of products from 18 domains. Each product is accompanied by detailed reviews, co-buy records, and metadata, including its product title, features, attributes, and images provided by the retailer. Following Yu et al. (2023), we select products from the *Electronics* and *Clothing, Shoes and Jewelry* domains as representative products to demonstrate the effectiveness of MIND. To fit our framework, we filter out products without accessible images that may have been removed from the Amazon platform.

3.3 Product Feature Extraction

We begin processing the collected products by first extracting key features with the aid of LVLMs. This is motivated by our observations that product descriptions and attributes, inputed by retailers, tend to be noisy and unorganized, probably for promotion and style organization purposes. Thus, we explicitly instructs LVLMs to augment source product metadata by extracting implicit features from each product's image and title by leveraging a zero-shot prompt:

> **Prompt Template for Product Feature Extraction**
>
> **Visual Input:** p_i
>
> ---
>
> **Textual Input:** `<Instruction>`. Given the product shown in the image: p_f, generate additional features by focusing on the product's attribute, design, and quality.

where `<Instruction>` is a detailed task instruction, and p_i, p_f are the respective image and details (title, descriptions, etc.) of the product. This enables LVLM to comprehend the product from both visual and textual modalities, thereby providing us with a richer set of features that complements those provided by the retailers.

3.4 Co-buy Intention Generation

Then, for each co-buy pair of products (p^1, p^2), we provide LVLM with the acquired features together with all details of both products, and instructs it again to reason the intentions for purchasing them simultaneously. Specifically, we follow Yu et al. (2023) and leverage 20 commonsense relations from ConceptNet (Speer et al., 2017) as waymarks to lead LVLM in generating purchase intentions with controllable commonsense groundings. Similar to the previous step, a zero-shot prompt is used:

> **Prompt Template for Intention Generation**
>
> **Visual Input:** p_i^1, p_i^2
>
> ---
>
> **Textual Input:** `<Instruction>`. A customer purchased a pair of products, as shown in the images. They are: p_f^1, p_f^2. Act as the customer and infer a potential intention behind such purchase. Start the intention with `<Relation>`.

Where `<Instruction>` is a detailed task instruction and `<Relation>` is the corresponding text template of a commonsense relation from ConceptNet. For every relation, we generate only one intention per pair of products due to the large amount of products and co-buy records. However, this is not restricted and can easily scale up.

3.5 Human-centric Role-aware Filtering

To effectively manage a large amount of purchase intentions, quality control measures have become imperative. While previous works relied on human annotaions for this purpose, recent works (Zhou et al., 2024) show that co-buy intentions generated by LLMs, despite undergoing human filtering, still fail in capturing the customers' mental states but rather focus on factual similarities of the products, as demonstrated in Figure 1. This phenomenon, refered to as "product-centric," restricts the potential downstream applications of the generated intentions. To address both issues, inspired by recent works on theory-of-mind (Kosinski, 2023), we propose to incorporate a filtering module, powered by a LVLM, after the generation process. We instruct the LVLM to assume the role of an E-commerce customer and provide it with a generated intention as the objective in the customer's mental state. Based on this intention, we present the LVLM with a pair of products and ask it to first determine whether the intention successfully motivates the purchase behavior and then generate a rationale to support its decision. This process simulates a real-world scenario where the LVLM functions as a customer, making purchase decisions. By filtering intentions that result in a positive response for purchasing, we obtain intentions that are "human-centric" in the sense that they satisfy the mental state of an agent that is aware of its role as a customer. We term this approach as *human-centric role-aware filtering*, which serves as an automatic filter to replace manual annotations. We apply this module to all the intentions we collected in previous steps and select the product-intention pairs that are accepted by the module as the final outcomes of our framework. Detailed prompts are provided in Appendix A.

4 Intrinsic Evaluations

By applying MIND to products we collected from Amazon Reviews (Ni et al., 2019), we construct a multimodal intention knowledge base, with statistics shown in Table 1. In total, 1.26 million intentions are preserved after applying our proposed filtering module, spanning across 20 relations. Therefore, in this section, we first evaluate MIND intrinsically by examining the quality of the generated intentions and the effectiveness of our proposed filter module through human annotation.

4.1 Annotation Setup

We hire human annotators from the Amazon Mechanical Turk platform to evaluate the generated intentions. For a generated intention, we task each worker to evaluate four aspects:

• **Plausibility** refers to the degree to which an intention of a co-buy purchase appears correct and reasonable given both products.

• **Typicality** evaluates how well the intention reflects a specific feature that causes the user behavior, which emphasizes on *informativeness* and *causality* (Yu et al., 2023).

• **Human-centric** evaluates the extent to which the intention considers and aligns with the mental state and preferences of a human customer.

• **Filter rationale** evaluates the correctness of the reasoning or justification provided by the filtering module for accepting or rejecting a product-intention pair.

For each aspect, we ask the annotators to rate them as a binary classification task. A random sample of 5,000 generated intentions are annotated, and the final vote is determined by the majority vote from three annotators. The requirement for the annotators could be found in Appendix B.

4.2 Annotation Results

The results of the annotations are presented in Table 1. The annotators achieved a pairwise agreement of 73.1% and a Fleiss's κ (Fleiss, 1971) of 0.56, indicating satisfactory internal agreement. The results reveal that MIND effectively generates purchase intentions that are both highly plausible (**94%** on average) and typical (**90%** on average) across all relations. This indicates the strong product understanding and intention reasoning capabilities of MIND. Additionally, our proposed human-centric role-aware filter correctly identifies 82% of intentions on average, with 80% of them having appropriate justifications for filtering. These high percentages further validate the effectiveness of our proposed method, which serves as a cost-efficient and highly reliable quality control measure, replacing the need for human annotations. More details and analysis regarding the filtered out intentions are attached in Appendix C

5 Experiments and Analyses

In this section, we first study the downstream benefits brought by intentions generated by MIND. Then, we conduct in-depth analyses to demonstrate the advantages of multimodal generation in MIND compared to generating only with textual information, the superior capability of the human-centric role-aware filter in comparison to other filtering measures, knowledge diversity in MIND generations, and its robustness when generating with different prompts.

Relation	#Int.	Pla.	Typ.	Fil.	Rat.
Effect	97,047	0.90	0.83	0.73	0.70
MannerOf	50,563	0.93	0.89	0.83	0.82
isA	62,069	0.94	0.88	0.82	0.80
Other	545	0.94	0.90	0.79	0.75
MadeOf	40,593	0.95	0.92	0.85	0.82
SimilarTo	63,558	0.94	0.87	0.83	0.80
UsedFor	52,383	0.94	0.88	0.81	0.79
Can	90,392	0.95	0.91	0.82	0.78
CauseDesire	95,097	0.94	0.90	0.82	0.80
RelatedTo	64,152	0.93	0.89	0.81	0.79
PartOf	81,230	0.92	0.87	0.79	0.77
Open	122,296	0.93	0.89	0.83	0.82
CreatedBy	35,723	0.94	0.88	0.78	0.76
DeriveFrom	60,347	0.95	0.89	0.80	0.77
DefinedAs	51,680	0.96	0.92	0.84	0.84
PropertyOf	57,947	0.97	0.90	0.83	0.82
CapableOf	86,772	0.95	0.90	0.82	0.82
Cause	61,860	0.95	0.92	0.83	0.82
SymbolOf	64,477	0.95	0.92	0.84	0.82
DistinctFrom	27,710	0.94	0.89	0.84	0.83
Total	1,264,441	0.94	0.90	0.82	0.80

Table 1: Statistics of the intention knowledge base constructed via MIND and human annotation results.

5.1 Evaluation Setup

We explore the effectiveness of MIND on the IntentionQA benchmark (Ding et al., 2024), a comprehensive multiple-choice question answering dataset comprising two challenging subtasks that require language models to comprehend and utilize intentions in E-commerce scenarios accurately. The first task assesses LLMs' capability in accurately inferring the intention given a co-buy product pair together with 3 distractors sampled from other product pairs, while the second task evaluates LLMs' capability in utilizing purchase intention to make reasonable product recommendation by selecting the product that best aligns with the user's intention from four choices.

While existing results show that language models struggle with both tasks, we aim to examine whether MIND can enhance LLMs' intention understanding capabilities through fine-tuning. Specifically, from all intentions generated by MIND, we transform them into instruction-following format via natural language templates following Zhou et al. (2023). Then, we fine-tune LLAMA2-7B-

Methods	Backbone	IntentionUnderstanding				IntentionUtilization			
		Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
Random	-	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Majority Vote	-	26.37	25.24	26.27	26.15	25.97	28.57	28.57	26.60
PTLM	RoBERTa-Large *214M*	41.46	41.98	38.98	41.43	54.95	35.06	30.08	49.84
	DeBERTa-v3-Large *435M*	36.40	38.72	37.62	36.90	26.52	29.35	32.33	27.39
	T5-v1.1-xxl *11B*	24.84	25.47	25.42	24.99	26.71	26.23	25.56	26.55
	Flan-T5-xxl *11B*	75.98	73.58	63.56	74.88	79.26	81.82	81.95	79.89
	T0-pp *11B*	71.70	68.87	64.41	70.78	77.11	76.10	78.20	76.99
Commonsense	HyKAS *435M*	71.81	67.17	46.69	69.61	47.02	45.97	48.12	46.90
	CAR *435M*	73.69	71.46	54.38	72.20	36.18	43.12	44.36	37.94
	CANDLE *435M*	74.34	70.75	52.54	72.52	35.94	43.90	43.61	37.84
	VERA *11B*	69.82	70.52	61.02	69.49	59.20	58.18	64.66	59.36
	VERA-CANDLE *11B*	70.59	71.33	63.41	70.02	62.18	60.13	66.13	61.81
Open LLM	LLAMA2-7B-chat	64.98	66.54	53.85	64.61	59.90	54.86	47.37	58.04
	LLAMA2-13B-chat	69.63	63.96	60.78	68.06	45.53	41.95	39.71	44.52
	Gemma-2B-instruct	48.77	47.23	48.21	48.45	39.45	39.15	38.17	39.32
	Gemma-7B-instruct	65.55	64.31	52.04	64.61	33.18	36.01	41.51	34.20
	Mistral-7B-Instruct-v0.2	76.57	74.53	63.56	75.50	59.78	62.60	65.41	60.64
	Falcon-7B-instruct	24.54	22.17	28.26	24.25	26.15	28.05	26.32	26.50
	Vicuna-7B-v1.5	57.13	57.08	55.43	57.05	27.88	30.13	23.31	28.00
MIND Distilled	LLAMA2-7B-chat	65.78	64.61	55.75	66.15	59.43	57.13	60.03	59.04
	Mistral-7B-Instruct-v0.2	78.57	74.31	80.89	76.97	61.14	65.42	62.16	62.02
LLM API	ChatGPT	75.06	73.76	68.64	74.48	80.74	76.62	68.42	79.23
	ChatGPT (CoT)	76.07	74.53	63.56	75.12	78.89	75.32	78.20	78.21
	ChatGPT (CoT-SC)	76.51	73.82	63.56	75.32	85.72	77.14	82.71	83.99
	GPT 4	78.12	75.41	66.10	76.97	86.03	82.34	84.96	85.30
	GPT 4 (CoT)	78.12	75.41	66.10	76.97	86.03	82.34	84.96	85.30
	GPT 4 (CoT-SC)	78.80	72.88	65.25	76.97	84.00	80.78	84.96	83.48
Human	-	89.96	90.00	80.96	89.33	95.50	85.19	100.0	94.00

Table 2: Evaluation results (Accuracy%) of various language models on both tasks of the IntentionQA benchmark.

chat (Jiang et al., 2023) and Mistral-7B-Instruct-v0.2 (Touvron et al., 2023) on the retrieved data as a type of knowledge injection. They are then evaluated in a zero-shot manner by being prompted to select the most plausible choice for every QA pair in IntentionQA. Accuracy is used as the evaluation metric. More information about the baselines can be found in Appendix D

5.2 Results

The results are presented in Table 2, demonstrating significant improvements in both tasks when LLMs are fine-tuned on intentions generated by MIND. For instance, LLAMA2 achieves accuracy gains of 1.54% and 1.00% for both tasks, respectively. Notably, Mistral yields a remarkable performance gain that even becomes comparable to GPT-4, despite having a significantly lower number of parameters. However, for the intention utilization task, while both fine-tuned LLMs show performance improvements, they still fall behind GPT-4. One potential reason for this gap could be the misalignment between the fine-tuning objective and the evaluated ability of the task, which involves generating intentions for a pair of products and selecting a product based on a given intention. Nevertheless, these results underscore the effectiveness and efficiency of MIND in enhancing LLMs' capabilities in E-commerce intention comprehension and utilization.

5.3 Analyses

In this section, we study the superiority of MIND by examining three aspects. First, we demonstrate the positive impact of acquiring intentions in a multimodal manner instead of relying solely on textual hints. Next, we show that our proposed human-centric filtering leads to better downstream results and is more effective than traditional critic filtering based on a supervised scoring discriminator. Finally, we illustrate the robustness of MIND when using different prompts and its superior quality compared to FolkScope.

5.3.1 Multimodal vs. Unimodal Generation

We first study the ablation of incorporating visual information in MIND by comparing the downstream benefits of intentions generated in mul-

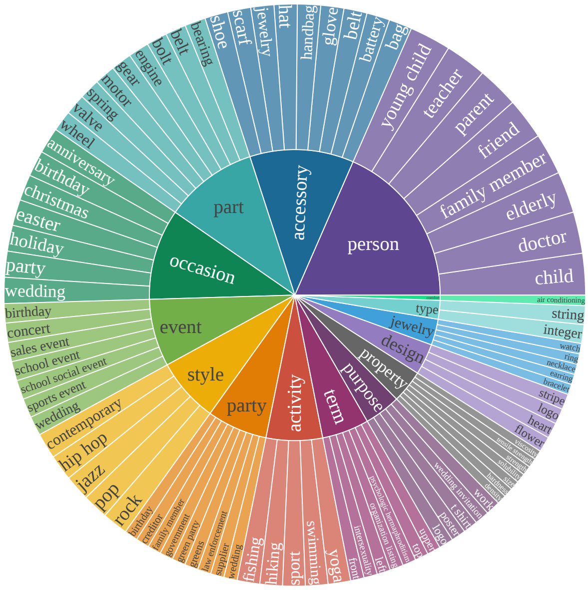

Figure 3: Distribution of hypernyms sourced from Probase in MIND with top frequencies.

timodal versus unimodal (text-only) paradigms. For a fair comparison, we exclude visual input in LLaVa when generating in a unimodal manner and instruct it to generate intentions for the same purchasing records as in MIND with prompts that are as identical as possible. We then fine-tune LLMs on the collected intentions, evaluate the resulting models on the IntentionQA benchmark, and compare the performance of the two types of distilled models. The results are shown in Table 3. We observe that fine-tuning LLMs on intentions generated with textual information can only merely improve their performances on downstream tasks, certifying the need of additional visual signals.

5.3.2 Diversity of MIND

Moreover, the semantic diversity of the generated intentions are another flag of the quality as featuring intentions that cover a diverse collection of topics, events, and even mental states makes it more possible to model purchase intentions comprehensively. Thus, following Wang et al. (2024), we sample 30,000 intentions from MIND, extract the nouns in them via dependency parsing, and plot the distribution of the hypernyms of these nouns, matched against Probase (Wu et al., 2012), according to their number of occurences. The resulting plot is shown in Figure 3. Remarkably, the intentions generated by MIND display elevated levels of diversity, which signifies the broad semantic coverage of purchasing different products in the generated intentions. We posit that such high semantic diversity may provide implicit benefits to downstream tasks when employing MIND in E-commerce applications.

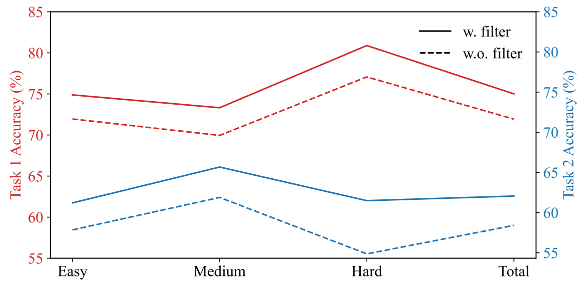

Figure 4: Ablation results on IntentionQA tasks by Mistral-7B distilled on intentions generated by MIND with/without filtering.

5.3.3 Impact of Role-Aware Filter

Ablation Study on IntentionQA. We then study the ablation of MIND by focusing on the role of our proposed human-centric role-aware filter mechanism in its impact toward quality control of the generated intentions. Specifically, we leverage the IntentionQA (Ding et al., 2024) as the evaluation benchmark and separately train two models on (1) intentions that are filtered by our proposed mechanism (*w. filter*) and (2) intentions without filtering (*w.o. filter*). All setups follow the same as described in Section 5.1, and we use Mistral-instruct-7B-v0.2 as the backbone and train it using a unified hyper-parameter setting. The results are plotted in Figure 4. We observe that, without filtering, performances on both tasks across all difficulty levels drop significantly, which is possibly due to the inclusion of more noisy intentions in the training data. This shows that our proposed filtering module is indeed functioning well in controlling high-quality intentions and is beneficial to downstream tasks.

Critic Filter v.s. Role-Aware Filter. Afterward, we validate the effectiveness of our role-aware filter by comparing it against a traditional critic filter provided by Yu et al. (2023). The critic filter is obtained by training a language model with a regression objective to predict the typicality of generated intentions in the range of 0 to 1. We adopt the released critic scorer, pre-trained on annotated intentions in FolkScope, and use it to score intentions in MIND under identical settings. By setting a critic threshold to 0.8 and discarding intentions below this threshold, we obtain a sibling subset of MIND. LLMs are then fine-tuned on this sibling subset and evaluated on the testing sets of IntentionQA. The results are presented in Table 3. It can be observed that LLMs exhibit inferior performance when using the critic filtering mechanism.

Backbones	Training Recipe	IntentionUnderstanding				IntentionUtilization			
		Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
LLAMA-7B-chat	Zero-shot	64.98	**66.54**	53.85	64.61	**59.90**	54.86	47.37	58.04
	w. Unimodal	65.03	65.49	**56.71**	64.99	59.08	54.71	45.59	57.34
	w. Critic Filter	61.88	64.56	51.22	61.67	59.27	54.13	46.89	57.88
	MIND Distilled	**65.78**	64.61	55.75	**66.15**	59.43	**57.13**	**60.03**	**59.04**
Mistral-7B-Instruct-v0.2	Zero-shot	76.57	**74.53**	63.56	75.50	59.78	62.60	**65.41**	60.64
	w. Unimodal	75.02	72.33	62.17	73.72	58.35	61.48	62.81	58.51
	w. Critic Filter	74.78	71.23	62.87	72.29	58.32	61.09	58.98	57.63
	MIND Distilled	**78.57**	74.31	**80.89**	**76.97**	**61.14**	**65.42**	62.16	**62.02**

Table 3: Ablation experiment results (Accuracy%) on IntentionQA benchmark.

One possible reason is that the pre-trained critic filter only captures the pattern of intentions at different levels of typicality without considering their relation to the products. This further verifies the need for a role-aware filtering mechanism.

5.3.4 Robustness of MIND

According to Chang et al. (2024), generations by LLMs can be significantly impacted by even slight changes in the prompts. This warrants a potential weakness of MIND which heavily relies on prompting in collecting intentions. Hence, we aim to overcome this by proving that intentions generated with modified prompts are generally semantically consistent at high quality. Specifically, we exclude the prompts which explicitly instructing the LVLMs to rely on visual cues from the product images and only retain the prompts that require the LVLMs to generate intentions. Then, 100 product pairs are randomly sampled from MIND to generate intentions utilizing the modified prompts. Finally, the sentence embedding are calculated using SentenceBERT (Reimers and Gurevych, 2019), and the cosine similarity between each modified intention and its corresponding original intention generated by MIND is derived. The results revealed an average cosine similarity of 0.85 between the intentions generated with modified prompts and those generated by MIND. This high similarity indicates the robustness of intention generation process. Inter-relation intention comparison examples are provide in Appendix E

5.3.5 Comparisons Against FolkScope

We then compare MIND against FolkScope, the previous state-of-the-art method for large-scale intention acquisition, by analyzing the typicality distribution of intentions across all relations. Specifically, we adopt the same annotation protocols designed by Yu et al. (2023); Wang and Song (2024); Wang et al. (2023b) and transfer our annotation results into a four-point Likert scale (Joshi et al., 2015). Then, for each relation, we compute the average typicality scores among all intentions and plot them for comparison, as shown in Figure 5. From the plot, we observe that intentions generated by MIND exhibit higher typicality scores across nearly all relations compared to those generated by FolkScope, which further demonstrates the superiority of MIND. More exemplar-based case comparisons are presented in Appendix F and a relation-wise filter result analysis is attached in Appendix G.

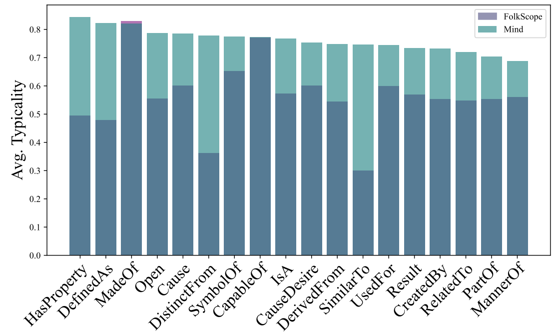

Figure 5: Relation-wise comparison of typicality scores across all relations between MIND and FolkScope.

6 Conclusions

In this work, we present MIND, a multimodal distillation framework for enhancing E-commerce purchase understanding by automating the pipeline of intention generation and quality filtering via multiple-step instructions over LVLMs. By applying MIND to real-world E-commerce data, we construct the very first multimodal purchase intention knowledge base featuring over 1.2 million intentions. These intentions have been proven to be invaluable in distilling student models that exhibit improved performance in E-commerce intention comprehension and utilization tasks. Further analyses reveal the effectiveness of MIND by validating the proposed filtering mechanism and highlighting the strengths of MIND in comparison to FolkScope. Our work sheds light on improving large-scale E-commerce intention acquisition and application.

Limitations

First, MIND generates intention by leveraging several zero-shot prompts without additional exemplars. This decision is made as we observe that few-shot prompts may "guide" LVLM to generate intentions that tend to be similar to the provided exemplars, which harms diversity. However, it remains an open question whether more advanced prompting methods (Song et al., 2023; Parnami and Lee, 2022) would help in the generation process. It's also worth noting that the LVLM used in our work may be outdated as new products show up on E-commerce platforms. However, switching LLaVa to more up-to-date LVLMs, preferably pre-trained on E-commerce data, can address this concern. Finally, MIND utilizes an automatically functioning filter as quality control. While we have shown its effectiveness, it remains challenging to effectively regulate the filter mechanism to be either lenient or strict. Further investigation is required to provide insights into the alignment between the values of VLMs and the real world, enhancing our understanding of them.

Ethics Statement

To avoid generating harmful intentions and toxic filter rationales in MIND, we recruit 4 expert annotators who are graduate students specializing in multilmodality and natural language processing to evaluate the generated intentions and rationales. We ask all experts to go through 200 sampled data and no harmful contents are reported. The crowd-sourced annotators are paid a wage that complies with the local law. The expert annotators involved in this research are knowledgeable about the annotation protocol and the intended utilization of their annotations. They are willingly to contribute without expecting any compensation. The training and evaluation datasets utilized in this study are publicly available, anonymized, and shared under open-access licenses for research purposes, adhering to their intended usage. Thus, we believe this paper does not yield any ethical issue.

References

Michael Bratman. 1984. Two faces of intention. *The Philosophical Review*, 93(3):375–405.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *CoRR*, abs/2404.01077.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Honghua (Kathy) Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li. 2006. Detecting online commercial intention (OCI). In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 829–837. ACM.

Wenxuan Ding, Weiqi Wang, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Junxian He, and Yangqiu Song. 2024. Intentionqa: A benchmark for evaluating purchase intention comprehension abilities of language models in e-commerce.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer.

Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1809–1818. ACM.

Zhenyun Hao, Jianing Hao, Zhaohui Peng, Senzhang Wang, Philip S. Yu, Xue Wang, and Jian Wang. 2022. Dy-hien: Dynamic evolution based deep hierarchical intention network for membership prediction. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 363–371. ACM.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query reformulation in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1319–1328. ACM.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren, and Hamed Firooz. 2021. MSD: saliency-aware knowledge distillation for multimodal understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3557–3569. Association for Computational Linguistics.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Dong-Mo Koo and Seon-Hee Ju. 2010. The interactional effects of atmospherics and perceptual curiosity on emotions and online shopping intention. *Computers in human behavior*, 26(3):377–388.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023b. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023c. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023*, pages 1039–1047. ACM.

Xin Liu, Zheng Li, Yifan Gao, Jingfeng Yang, Tianyu Cao, Zhengyang Wang, Bing Yin, and Yangqiu Song. 2023d. Enhancing user intent capture in session-based recommendation with attribute patterns. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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. *CoRR*, abs/1907.11692.

Feihong Lu, Weiqi Wang, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, and Jianxin Li. 2024. MIKO: multimodal intention knowledge distillation from large language models for social-media commonsense discovery. *CoRR*, abs/2402.18169.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 188–197. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Archit Parnami and Minwoo Lee. 2022. Learning from few examples: A summary of approaches to few-shot learning. *CoRR*, abs/2203.04291.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.*, 55(13s):271:1–271:40.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. MARS: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference* on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Changlong Yu, Xin Liu, Jefferson Maia, Tianyu Cao, Laurence Yang Li, Yifan Gao, Yangqiu Song, Rahul Goutam, Haiyang Zhang, Bing Yin, et al. 2024. Cosmo: A large-scale e-commerce common sense knowledge generation and serving system at amazon. In *Proceedings of the 2024 International Conference on Management of Data, SIGMOD 2024*.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu. 2016. Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 1373–1384. ACM.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Jiashu Zhao, Hongshen Chen, and Dawei Yin. 2019. A dynamic product-aware learning model for e-commerce query intent understanding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1843–1852. ACM.

Wayne Xin Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1935–1944. ACM.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

Wendi Zhou, Tianyi Li, Pavlos Vougiouklis, Mark Steedman, and Jeff Z. Pan. 2024. A usage-centric take on intent understanding in e-commerce. *CoRR*, abs/2402.14901.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

Appendices

A Prompts

In this section we show the instructions used in feature extraction, intention generation, and human-centric role-aware filtering stages. The prompts are shown in Table 4.

B Annotator Requirement

For strict quality control, we only invite workers satisfying the following requirements: 1) at least 1K HITs approved, and 2) at least 95% approval rate. Then, we conduct two rounds of qualification rounds using a qualification question set crafted by authors of this paper, which includes both straightforward and tricky questions. Over 600 workers participated and only 90 (15%) of them are deemed qualified by achieving over 87% accuracy.

C Error Analysis of Filtered Intentions

While human annotation results in Section 4.2 show that, after filtering, most of the remaining intentions are highly plausible and typical, we observe that only 46.7% generations passed our proposed filtering module as the last step of MIND. Thus, in this section, we first study the role of such human-centric filtering by looking into the causes of why the intentions get discarded, and further seek insights to resolve such a high filtering loss. To achieve this, we randomly sample 200 intentions that are abandoned by MIND during the last step and manually annotate the reasons behind based on the rationale provided by the LVLM. Three types of errors are observed and they are categorized as:

Task	Prompt
Feature Extraction	The $[IMAGE\_1, IMAGE\_2]$ contains a product and name of it is $[PROD\_NAME]$. Please analyze the product image, together with the product name, provide a detailed description focusing on the product's features, design, and apparent quality. Highlight any unique characteristics or visible elements that distinguish this product from similar items. Additionally, speculate on the potential uses and benefits of this product for a consumer, based on its appearance or any information in the image and the name.
Intention Generation	The two $[Image\_1, Image\_2]$ are two different products. The product name of the upper image is $[Prod\_A\_Name]$. The product detail and the potential purchase intention is $Prod\_A\_Desc$. The product name of the lower image is $Prod\_B\_Name$. The product detail and the potential purchase intention is $Prod\_B\_Desc$. Based on information provided, together with the product images, what could be the potential intention for people buying these two products in one purchase simultaneously based on the relation of $[Relation\_Prompt[Relation]$, take the image features into consideration, limit your word count within 120 words. Start with the potential co-buy intention could be $Relation\_Prompt[Relation]$
Human-centric Role-aware Filtering	The two images $[Image\_1, Image\_2$ are two different products. The product name of the upper image is $[Prod\_A\_Name$. The product detail and the potential purchase intention is $[Prod\_A\_Desc]$. The product name of the lower image is $[Prod\_B\_Name$. The product detail and the potential purchase intention is $[Prod\_B\_Name]$. Under the relation of $[Relation\_Prompt[Relation]$, the potential co-buy intention would be $[Intention]$. If you are a consumer who are eager to buy product a or product b, would this intention encourage you to buy the two products simultaneously? be critical on your choice, output yes or no together with the reason for your answer. For example, the output should be Yes, ... or No, ...

Table 4: Prompts used for evaluating LLM baselines across various tasks in a zero-shot scenario.

- 81.0% of the filtered intentions, while plausible, do not provide strong enough evidence to motivate a LVLM agent to execute the purchase behavior for two products. For example, the intention "*they both are related to home audio systems*" for purchasing a pair of audio adapters lacks customer interaction and solely focuses on the products themselves. A more appropriate intention, for example, "*they both are able to help in connecting audio devices*," would retain a stronger bond between the products and customers by aligning with their functionalities.

- 13.0% of the intentions result from misjudgment by the LVLM, where the agent fails to make the correct decision despite the intention being sufficiently plausible and typical. This highlights the need for future improvements, including a more refined filter to enhance our framework.

- 6.0% of the intentions are discarded due to being implausible or containing factual errors that do not align with the products.

Overall, 87% of intentions are being properly discarded, which is considerably high for an automatic filter without human supervision.

D Baseline Backbone

For both tasks, we first incorporate random and majority voting to reflect the characteristics of the benchmark. Five Pre-Trained Language Models (PTLMs) are included: RoBERTa (Liu et al., 2019) DeBERTa-v3 (He et al., 2023), T0 (Sanh et al., 2022), T5 (Raffel et al., 2020), and Flan-T5 (Chung et al., 2022). Then, performances by five commonsense-injected PTLMs are also reported, including HyKAS (Ma et al., 2021), CAR (Wang et al., 2023a), VERA (Liu et al., 2023b), CANDLE (Wang et al., 2024), and VERA-CANDLE. We also report the performances of several LLMs, such as LLaMA2 (Touvron et al., 2023), Gemma (Mesnard et al., 2024), Mistral (Jiang et al., 2023), ChatGPT (OpenAI, 2022), and GPT-4 (OpenAI, 2023). For the latter two, we also adopt Chain-of-Thought (COT; Wei et al., 2022) and CoT with Self-Consistency (COT-SC; Wang et al., 2023c) prompting.

Item 1	Item 2	Relation	Intention
Girls Prewalker Toddler Cute Flower Bowtie Antiskid Shoes Sneaker 	Fisher-Price Brilliant Basics Rock-a-Stack 	symbolOf	They both represent the early stages of a child's development.
		can	They both help to develop explore and develop children's skills.
		capableOf	They both provide young children with a safe and engaging environment.
		cause	The person wants to purchase both products as gifts for a young child.
		isA	They both cater to the needs of young children.
Rubies 18th Century Colonial Man Wig Adult One Size 	Pirate Boot Toppers - Fun Costume Accessory 	usedFor	They are both used for costume or theatrical performances.
		symbolOf	They both symbolize a pirate or colonial theme.
		isA	They are both costume accessories for a pirate-themed outfit.
		cause	The person wants to create a complete and authentic pirate costume.

Figure 6: MIND co-buy intentions generated under different relations.

E MIND Inter-relation Case Study

In this section, we showcase various co-buy intentions for the same product pairs generated under different relations. The examples are provided in Figure 7.

It is evident from the table that the intentions consistently capture the key aspect of the co-buy intention. i.e., for young kids, for costume, pirate. Though for certain relations the intention doesn't follow the instruction strictly in terms of format, the quality of the intention remains reasonable and informative. The content of these intentions is still aligned with the intended purpose of the designed relation.

F MIND Against FolkScope Case Study

Aside from empirical analyses, we also show the advantages of MIND over FolkScope through additional case studies to highlight key benefits of MIND. To this end, we randomly selected 7 pairs of co-buy products and compared the intentions generated by both frameworks, as shown in Table 5. Our findings from the table indicate that MIND-generated intentions exhibit a stronger focus on the usage and functionalities that potentially fulfill customers' needs and intentions when purchasing these products. Conversely, intentions generated by FolkScope tend to be biased towards properties and features that can be easily inferred from the product titles, which are of lesser interest to customers' shopping intentions. Take the second row in Table 5 as an example. The intentions *both are "Women's Shoes" and "Women's Handbags"* generated by FolkScope merely represent an aggregation of the product categories for the two items. In contrast, MIND produces intentions such as *looking for stylish and functional combination for daily activities* , which better captures a customer's intention when shopping for both products. This example further reinforces our previous conclusions that MIND can generate intentions that are more human-centric and better reflect the customers' intentions as mental states.

G Relation-wise Filter Analysis

In this section, we present the Relation-wise Filter Preserve Rate (RFP Rate) of MIND, which represents the proportion of intentions that are retained among all intentions for every relation. We report our result in Figure 7.

Our observations indicate that the open relation has the lowest RFP Rate at 0.17 yet other relations demonstrate RFP Rates ranging from 0.2 to 0.4.

We hypothesize that the under-performance of open relation generation could be attributed to its less specific instruction adopted in generation process. The lack of specific information in the instruction may hinder the LVLM's ability to generate persuasive and informative intentions for the

Item1	Item2	Intentions
Samsung SmartCam HD Pro	Samsung SmartThings Smart Home Hub	They are designed to **work together in a smart home system** They are derived **from the same category.**
Clarks Women's Ankle Bootie	The Sak Kendra Hobo Shoulder Bag	The consumer is looking for **a stylish and functional combination** for their daily activities. They both are **a manner of 'Women's Shoes' and 'Women's Handbags'** respectively.
Western Party Mustaches	Forum Novelties Adult Cowboy Costume Vest	They are both part of **a costume or a themed party.** They both are a **part of the 'Adult Costume' category.**
Columbia Women's Loveland Shorty Omni-Heat Snow Boot	Columbia Sportswear Women's Thermarator Glove	They are designed to **keep the wearer warm and comfortable during cold weather conditions** They both are a **part of the Columbia brand.**
Banded Arc Welded Waterproof Backpack Polyester	Banded Deluxe UFS Fleece Face Mask	They are both used for **outdoor activities and protection from harsh weather conditions.** They are both used for **outdoor activities.**
Novelty Video Games Sarcastic Mens Very Funny T Shirt	Super Mario All Over Character 100 Score Print Socks	They both **appeal to gamers and video game enthusiasts.** They both are of **high quality and are really cool.**
Women's Exaggerative Vintage Leafs Shape Necklace Sets	Plated Oil Drip Rhinestone Flower Necklace Earring Sets	They both have a **vintage-inspired design and feature colorful flowers and beads** They both have a property of **'High Quality'.**
Xbox 360 4gb Kinect Bundle	Controller Charger - Xbox 360	They both **cater to the needs of Xbox 360 gamers.** They both are **made of plastic.**

Table 5: Case studies of purchase intentions generated by MIND and FolkScope. Intentions generated by MIND are highlighted in **blue** and those generated by FolkScope are marked in **green**.

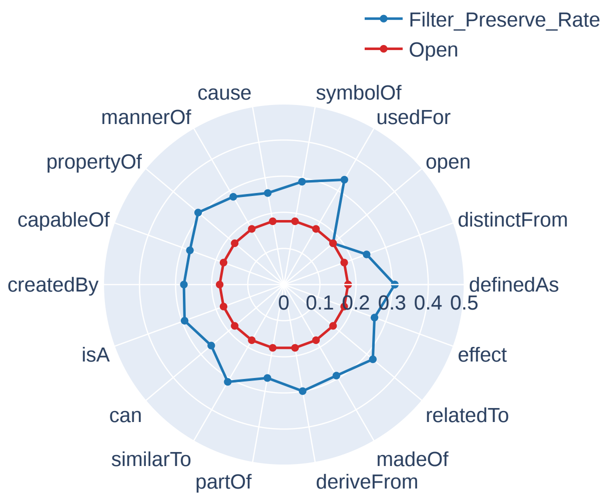

Figure 7: The rate of preserved intentions after filtering under different relations.

filter LVLM, resulting in the low preserve rate.

This finding emphasizes the importance of future intention mining research. It suggests that solely relying on the expressive power of LVLMs to undermine potential intentions is not feasible. Instead, a meticulous instruction constraint aligns with research purpose is required. Specifically, incorporating detailed relation information during intention mining is indispensable in E-commerce co-buy behavior understanding domain. This could improve the intention mining process, leading to a better construction of a credible and comprehensive intention knowledge base.