The American Journal of Applied Sciences
ISSN 2689-0992 | Open Access

TYPE Original Research
PAGE NO. 57-65
DOI 10.37547/tajas/Volume07Issue05-06

Check for updates

Personalization in E-Commerce: Optimizing Recommendations for Multimodal Content

Bulycheva Mariia

Senior Applied Scientist, Zalando Germany

OPEN ACCESS

SUBMITED 25March 2025
ACCEPTED 19 April 2025
PUBLISHED 23 May 2025
VOLUME Vol.07 Issue 05 2025

CITATION
Bulycheva Mariia. (2025). Personalization in E-Commerce: Optimizing Recommendations for Multimodal Content. The American Journal of Applied Sciences, 7(05), 57–65.
https://doi.org/10.37547/tajas/Volume07Issue05-06

Abstract: This article examines modern approaches to the personalization of multimodal content in e-commerce, driven by the growing complexity of user requests and the evident need to adapt recommendations to diverse data formats and content modality. The relevance of this topic is underscored by the increasing volume of information associated with the article, including text, images, video, and audio, which necessitates the application of specialized methods for precise customization and improved personalization. The purpose of the study is to develop original proposals for optimizing recommendation algorithms based on multimodal information, enabling the consideration of both context and individual user preferences. The research reveals contradictions in the literature—while many studies focus on specific aspects of personalization, such as textual data or visual elements, integrative approaches to the analyzed content are insufficiently addressed. The author proposes solutions combining deep learning methods and behavioral model analysis to achieve more accurate results in predicting audience interests. The materials presented in this work will be useful for e-commerce professionals, developers of recommendation systems, and researchers focused on evaluating behavioral patterns.

Keywords: algorithms, deep learning, multimodal content, personalization, user behavior, recommendations, e-commerce.

Introduction: E-commerce, which has become one of the

leading forms of interaction between sellers and consumers, is currently characterized by rapid growth in data volumes and content representation formats.

As multimodal information representation technologies (texts, images, videos, audio files, etc.) evolve, numerous additional challenges arise, associated with adapting to the individual needs of users.

The relevance of this article's topic is explained by the dynamic development of digital commerce combined with the increasing consumer expectations for personalized experiences. In the context of intense competition and information overload, businesses are compelled to adapt their approaches to data processing by integrating textual, visual, and audio information to create accurate, relevant recommendations. Multimodal content enables the consideration of numerous factors influencing preferences, making the process more targeted and effective.

The research problem lies in the need to develop effective personalization methods that formulate recommendations accounting for not only user preferences but also the context of the described content's consumption.

Materials and Methods

Research dedicated to this topic can be conditionally divided into several groups, each reflecting specific aspects of personalization.

The authors of several publications focus on developing and applying methods for processing multimodal data to analyze and generate recommendations. R. Bibi and colleagues [1] propose a framework for content-oriented image search that integrates visual and textual elements, improving the accuracy of recommendations. A similar emphasis is found in the work of Yu. Liu et al. [3], where methods for analyzing multimodal content are applied to determine emotional characteristics. Yu. Lu and Y. Duan [4] presents a solution based on dynamic preferences, which enhances the relevance of recommendation steps.

Numerous studies are devoted to implementing neural network technologies for processing complex multimodal data. E.B. Boztepe and colleagues [2] describe the nuances of applying deep learning for analyzing audiovisual content, which aids in improving the perception of information. Ya. Zhu and his colleagues [10] detail a training model for analyzing videos considering affective content, which is relevant for enhancing user experience in e-commerce.

Certain researchers focus on personalizing interfaces. A. Wasilewski and G. Kolaczek [8] emphasize that a universal interface design does not suit all users, proposing a multivariate approach with a focus on adaptation to individual preferences.

Contemporary studies also address understanding the impact of multimodal content on trust. I. Syed and colleagues [6] examine the effects of facial features, voice, and text messaging on trust levels, which is valuable for fostering deeper customer engagement. P. Thangavel and R. Lourdusamy [7] describe approaches to emotional analysis using lexicon-based methods, enhancing personalization capabilities.

A number of authors concentrate on creating recommendation systems to improve user experience. S. Silvester and Sh. Kurain [5] provide an overview of the Dual-Blend Insight system, which integrates multimodal content and analytical data to improve recommendation accuracy. Z. Zhang and colleagues [9] explore the methodological foundation for synthesizing visual content based on multimodal data, ensuring more relevant and engaging recommendations.

Despite significant progress in research, a review of the literature reveals the following shortcomings:

- Insufficient attention to adaptive systems;
- Limitations in accounting for cultural and regional differences in processing analyzed content;
- A lack of studies focused on integrating multimodal approaches with machine learning systems to optimize customer experience.

To explore the topic, the following methods were employed: comparison, retrospective analysis, synthesis, systematization, and generalization.

RESULTS AND DISCUSSION

When addressing the essential characteristics, it is important to note that personalization in e-commerce is the process of adapting product/service offerings to the unique preferences and behaviors of each consumer as

well as current session context (geo location, time of the day, etc.). This involves the collection, analysis, and interpretation of user data to increase their engagement, satisfaction, and the likelihood of repeat purchases [5, 8].

Algorithms capable of processing multimodal content hold a special place, as they allow for the simultaneous consideration of visual, textual, and behavioral signals.

The history of personalization in e-commerce dates back to the 1990s, when the internet became accessible to a wide audience. Early attempts involved the use of cookies to track user behavior on websites. In the 2000s, large companies began integrating recommendation systems based on analyzing past user actions, like purchases and views, which was a significant step forward.

In the 2010s, with the growth of data volumes and the availability of powerful computing resources, a breakthrough occurred in the use of machine learning to provide more accurate recommendations. Algorithms emerged that considered a wide range of information, including demographic data, behavioral patterns, and preferences.

In recent years, the evolution of technologies—such as machine learning , deep learning, and multimodal models—has enabled the integration of data from various sources, including text, images, video, and audio. This has facilitated the shift from basic recommendations to adaptive intra-session schemes that take into account complex contexts and make it possible to predict user needs in real-time.

As for the nature of multimodal content, it integrates various formats of information presentation, creating a unique opportunity for a more accurate understanding of people's intentions (Fig. 1). For example, product images attract attention due to their aesthetic qualities, while textual descriptions add informativeness. Unlike static images or text, videos unfold over time, introducing motion, pacing, sound, and narrative flow, which can evoke stronger emotional engagement and provide richer contextual cues. A video can demonstrate fabric texture in movement, convey authenticity through a speaker’s voice, or build anticipation with a well-paced storyline—elements that are difficult to capture through still images or words alone. However, processing such inputs requires specialized technologies, including deep neural networks trained on variable sources.

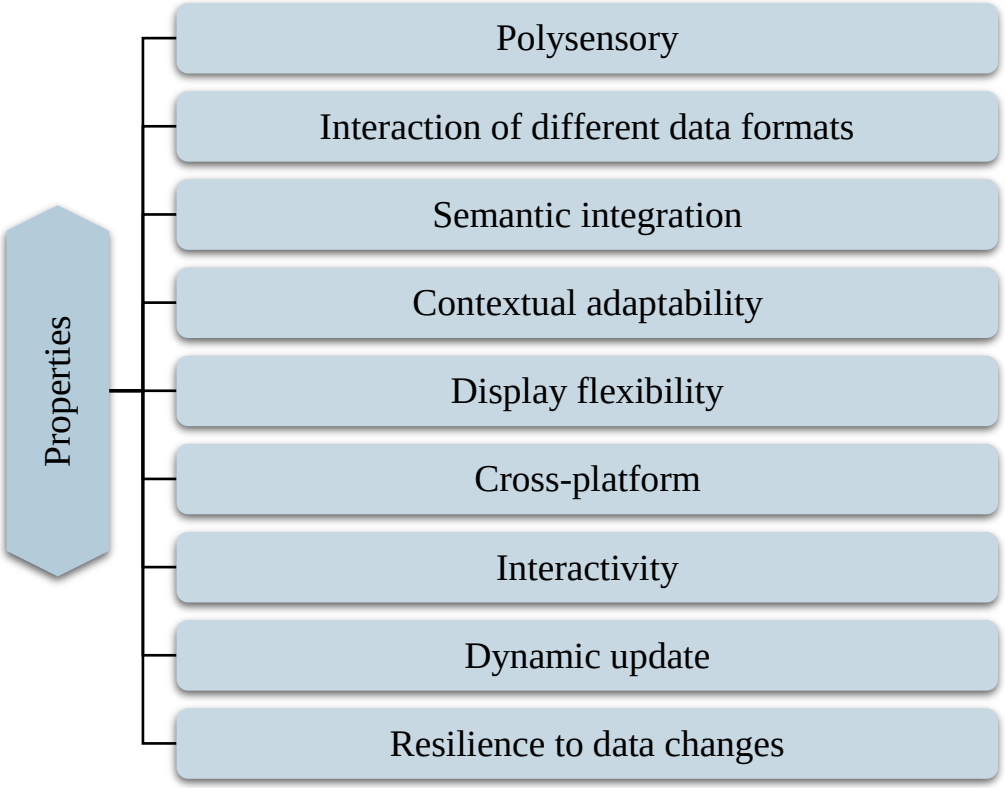

Fig. 1. Properties of multimodal content

(compiled by the author based on [1, 3, 9])

In e-commerce, the content in question consists of a combination of various data formats (text, images, video, audio, graphics) unified to create a cohesive user experience. Its key feature is the ability to account for different channels of perception, which helps adapt information to customer preferences and behavior. This enhances interaction through personalized recommendations, semantic analysis, and evaluation of user actions. It also supports cross-platform compatibility, ensuring accessibility across various devices, and fosters increased engagement by integrating interactive elements.

A powerful multimodal recommender system is built upon several key components that ensure the delivery of highly relevant and engaging content (Fig. 2). These include content understanding, user understanding, session context analysis, and the ability to mix different content modalities.

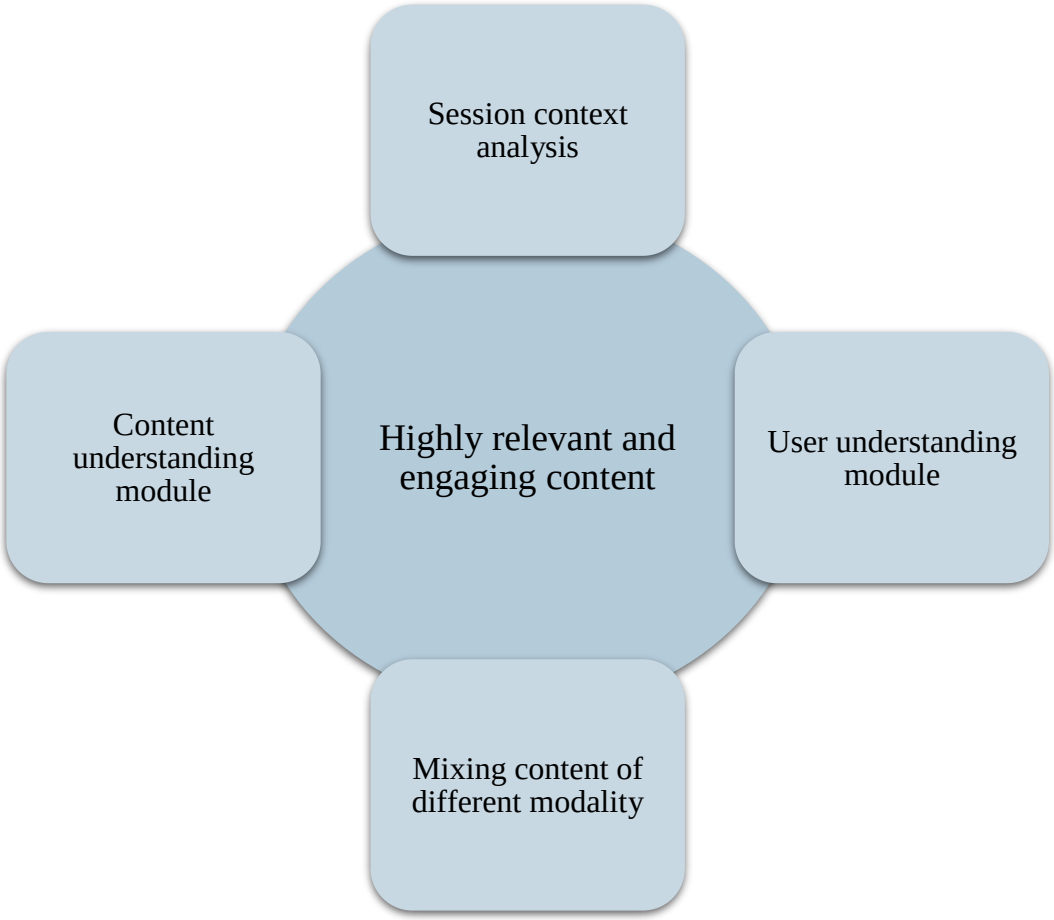

Fig. 2. Key components of a powerful multimodal recommender system

(compiled by the author based on [1, 2, 4, 7])

For the successful personalization of multimodal content in e-commerce, the involvement of integrated approaches is necessary (Fig. 3).

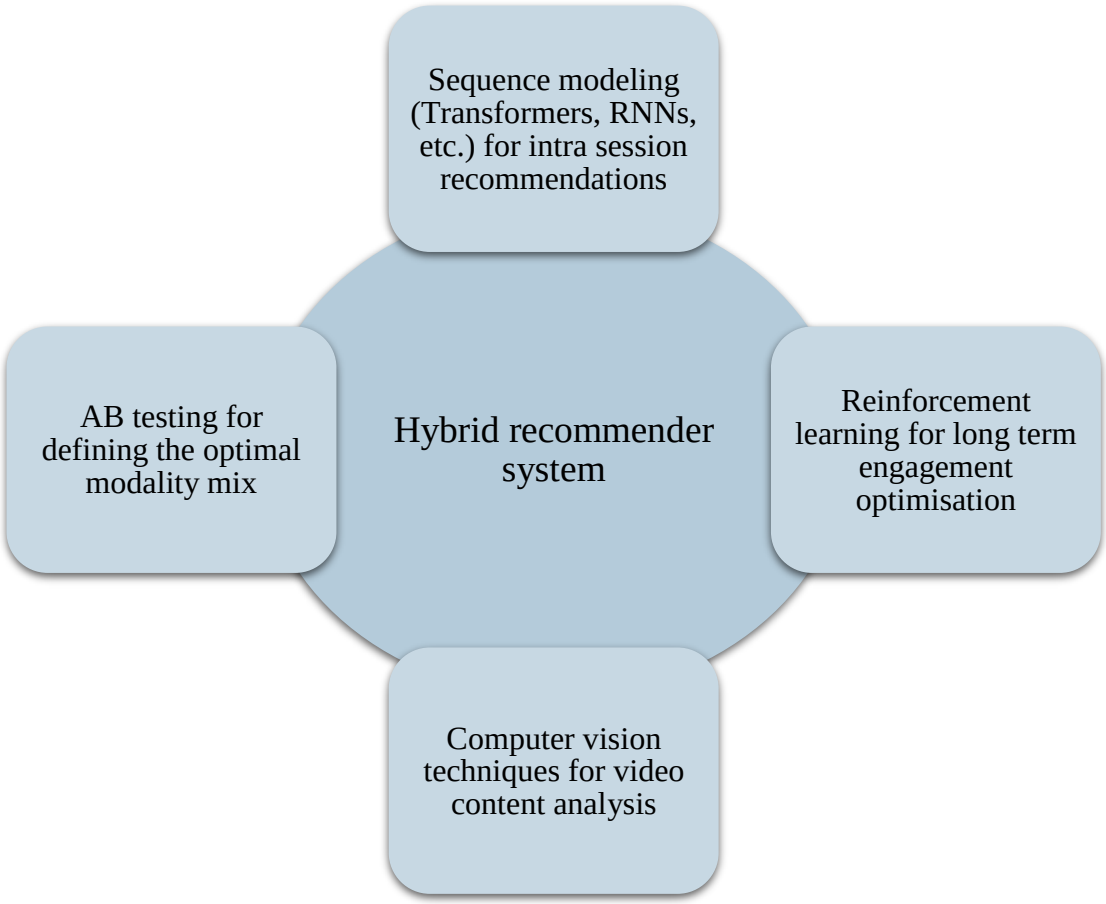

Fig. 3. Integrated approaches to personalization of multimodal content in e-commerce (compiled by the author based on [1, 2, 4, 7])

Hybrid recommendation systems combine content-oriented and collaborative schemes, enabling the consideration of both individual preferences and general patterns in customer behavior.

As videos become the dominant content format across various types of platforms, they bring a wealth of multimodal information, including visual elements, motion, sound, and narrative structure, making them more complex than static images. Computer vision techniques enable efficient analysis and understanding of this content.

- Frame-based analysis: Convolutional Neural Networks (CNNs) extract visual features from keyframes, identifying objects, colors, and branding elements.
- Motion analysis: Optical flow and action recognition models track movement patterns to detect dynamic elements like gestures or scene transitions.
- Audio-visual fusion: Models that integrate both visual and auditory cues (e.g., lip-sync analysis, speech recognition) help classify video content more accurately.
- Video summarization: Deep learning techniques, such as Temporal Segment Networks (TSN) or Transformer-based models, can summarize key moments in a video, improving indexing and retrieval for recommendation.

By leveraging these techniques, video content can be categorized more effectively, improving personalized recommendations and ensuring that users receive engaging, contextually relevant suggestions.

Reinforcement learning (RL) is a technique that helps tailor recommendations based on user interaction with the system, training algorithms to suggest the most relevant products or services. Beyond optimizing for immediate engagement, RL is also a powerful tool for maximizing long-term metrics, such as time spent, the number of content pieces viewed, or sustained user retention, by learning policies that balance short-term clicks with deeper, more meaningful engagement patterns.

User interactions during a single session often follow a sequential pattern, where earlier actions influence later choices. Sequence modeling techniques, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Transformers, are particularly effective in capturing these dependencies. Traditional RNNs and LSTMs process session data step-by-step, maintaining memory of past interactions to predict the next most relevant item. However, Transformers, with their self-attention mechanisms, can model long-range dependencies more effectively, making them ideal for handling complex user behavior patterns within a session. These models help predict what content a user is likely to engage with next, optimizing for continuous engagement and session prolongation.

Beyond traditional sequence modeling and computer vision techniques, multimodal transformers like CLIP and Flamingo play a crucial role in integrating and interpreting information across different content formats. These models align text, images, and video in a shared representation space, allowing for context-aware recommendations that go beyond single-modality analysis. By learning relationships between visual and textual elements, multimodal transformers enhance content understanding, improving search relevance, personalization, and cross-modal retrieval in recommendation systems.

And finally, with multiple content formats available (text, images, videos), determining the optimal mix for user engagement is a critical challenge. A/B testing provides a systematic approach to measuring the impact of different modalities on user behavior. By running controlled experiments, A/B testing helps validate assumptions about content effectiveness and refine recommendation strategies to balance user experience, engagement, and business goals.

Modern online platforms, such as Zalando, face the challenge of effectively understanding the complex interactions between users and multimodal content. To address this, innovative approaches are being developed that rely on the use of Graph Neural Networks (GNN) and Reinforcement Learning (RL), which open new possibilities in the field of personalization. GNNs help model intricate user-content relationships, while RL enables optimization for long-term engagement metrics, ensuring that recommendations go beyond immediate clicks to drive sustained user interaction.

Essentially, GNNs are a powerful tool for analyzing data organized as graphs, where the nodes represent users and content elements, and the edges represent their interactions, such as:

- clicks;
- views;
- likes;
- other actions along the user journey like add to cart, add to wishlist, etc.

Thanks to their ability to model nonlinear dependencies and account for context, these networks are used to train numerical representations—embeddings—that reflect implicit relationships between objects.

It is important to note that their generation in this context is focused on creating rich representations of users and content that take into account various data modalities:

- text;
- images;
- behavioral activity.

For example, graph structures allow for the consideration of both individual user preferences and general trends in their actions. These embeddings are integrated into existing recommendation models, which positively impacts the improvement of predictions for

interaction likelihoods (clicks or purchases).

One of the key advantages of using GNNs is their ability to uncover hidden patterns arising from complex interactions. For example, analyzing paths in the graph between users with similar interests helps identify new relevant connections that were previously overlooked by traditional approaches. This is particularly significant for multimodal content, where the relationships between textual, visual, and behavioral information play a crucial role.

Thus, the application of GNNs for personalizing the Zalando homepage not only enhances the relevance of recommendations but also improves the quality of the customer experience. The integration of embeddings generated through Graph Neural Networks allows the system to more accurately predict which content will interest a particular user, which in turn contributes to increased engagement and satisfaction.

Therefore, the use of GNNs in optimizing multimodal content is becoming a highly valuable tool for personalization on online platforms. This technology opens new prospects for a deeper understanding of user needs, improving content interaction, and ultimately enhancing the commercial effectiveness of the platform.

While GNNs excel at capturing intricate user-content relationships and generating meaningful embeddings, Reinforcement Learning (RL) plays a crucial role in optimizing recommendations for long-term engagement. Instead of solely predicting immediate interactions like clicks or purchases, RL enables the system to learn sequential decision-making policies, balancing short-term relevance with sustained user interaction.

By continuously adjusting recommendations based on user feedback, RL helps optimize high-impact metrics such as session duration, the number of content pieces viewed, and user retention over time. This makes RL particularly valuable in dynamic environments like large ecommerce platforms, where content selection should not only match immediate preferences but also encourage deeper exploration and prolonged engagement.

The integration of a multimodal approach into e-commerce personalization enables the following results:

- Increased conversion due to the accuracy of recommendations;
- Enhanced consumer trust through content adaptation to their expectations;
- Reduced decision-making time for purchases;
- Strengthened competitive advantages for companies through unique user experiences.

Despite the obvious advantages, the implementation of personalization for the described content is associated with several limitations. These include:

- high computational complexity;
- significant costs for processing large volumes of data;
- the risk of incorrect interpretation of preferences due to a lack of data or its one-sidedness;
- issues related to privacy and information security, which have become particularly relevant in light of tightening legislation [2, 6, 10].

Below are the author's recommendations for improving the effectiveness of multimodal content personalization in e-commerce. To begin with, it is advisable to list the proposed directions (Fig. 4).

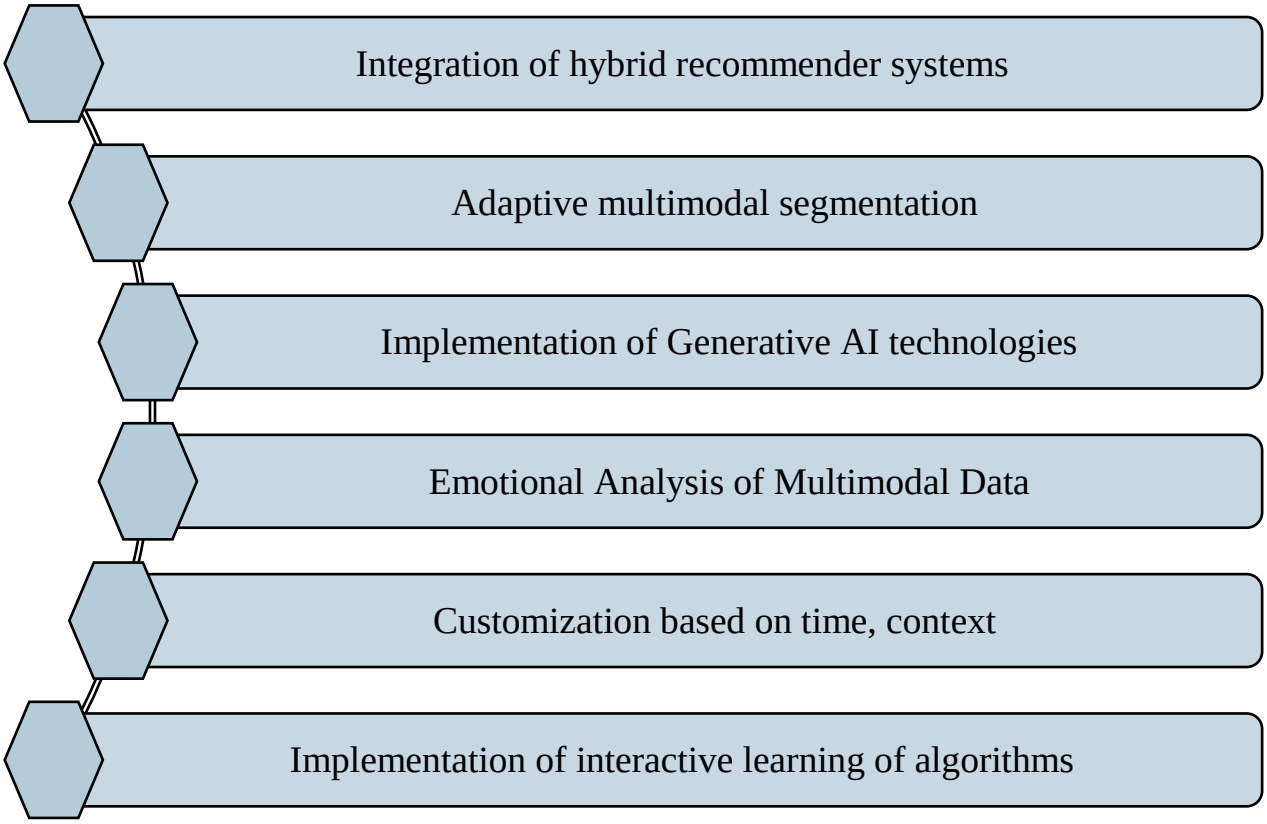

Fig. 4. Recommendations aimed at improving the effectiveness of personalization of multimodal content in the field of e-commerce

(compiled by the author)

The proposed steps are based on a combination of content-based, collaborative, and neural network methods, allowing for the consideration of both retrospective user behavior and semantic analysis of multimodal content. The use of hybrid models ensures high personalization accuracy, especially for new customers or products with minimal data.

It is recommended to create segments based on the analysis of multiple modalities (text reviews, preferences for visual content, demographic characteristics). For example, for users who prefer visual elements, recommendations should be focused on images and videos.

The use of models (such as GPT, DALL-E, etc.) for dynamically generating personalized product descriptions, titles, and images based on preferences is also suggested. This helps not only to increase content relevance but also to adapt the visual aspect to cultural and regional characteristics.

The implementation of algorithms capable of analyzing the sentiment of texts (such as reviews), expressions in images (e.g., in videos or live broadcasts) to account for reactions when generating recommendations is also advisable. The novelty lies in using emotional intelligence to predict behavioral manifestations.

Special attention should be given to customization based on time and context. This involves developing recommendations that consider chronological patterns of behavior (e.g., purchases at certain times of the day) and interaction nuances (such as holiday seasons, weather, geolocation, etc.). This approach ensures high recommendation accuracy in real-time.

Finally, a significant direction involves providing users with the opportunity to actively participate in adjustments through feedback features ("show more like this" or "hide similar items"). This improves the quality of personalization, making it highly flexible. The novelty in this case lies in ensuring continuous interaction between the customer and the system for its

adaptation.

CONCLUSIONS

Personalization of recommendations for multimodal content in e-commerce is a key factor for success in a highly competitive environment. Optimizing algorithms based on visual, textual, and behavioral data opens additional options for improving customer interaction efficiency. However, the implementation of these steps requires both a technological foundation and a careful approach to ethical and privacy concerns.

The recommendations proposed in this article offer a comprehensive approach to personalization, based on the synthesis of multimodal data, emotional analysis, and generative AI technologies. The novelty lies in combining methods that are rarely used together, which helps create personalized and adaptive user scenarios.

In conclusion, it should be emphasized that to enhance the effectiveness of personalization in the described field, it is necessary to implement technologies that can account for the multifaceted nature of data and adapt to changing user preferences. Recommendation strategies based on hybrid algorithms, emotional aspects, and interactive functions not only improve content relevance but also strengthen customer trust, contributing to long-term loyalty and increased conversion rates.

REFERENCES

Bibi R. Query-by-visual-search: multimodal framework for content-based image retrieval / R. Bibi, Z. Mehmood, R.M. Yousaf, T. Saba, M. Sardaraz, A. Rehman // Journal of Ambient Intelligence and Humanized Computing. – 2020. – Vol. 11. – No. 11. – Pp. 5629-5648.

Boztepe E.B. An approach for audio-visual content understanding of video using multimodal deep learning methodology / E.B. Boztepe, B. Karakaya, B. Karasulu, İ. Ünlü // Sakarya University Journal of Computer and Information Sciences. – 2022. – Vol. 5. – No. 2. – Pp. 181-207.

Liu Yu. Scanning, attention, and reasoning multimodal content for sentiment analysis / Yu. Liu, Zh. Li, Ke. Zhou, L. Zhang, L. Li, P. Tian, Sh. Shen // Knowledge-Based Systems. – 2023. – Vol. 268.

Lu Yu. Online content-based sequential recommendation considering multimodal contrastive representation and dynamic preferences / Yu. Lu, Y. Duan // Neural Computing & Applications. – 2024. – Vol. 36. – No. 13. – Pp. 7085-7103.

Silvester S. Dual-blend insight recommendation system for e-commerce recommendations and enhance personalization / S. Silvester, Sh. Kurain // Indonesian Journal of Electrical Engineering and Computer Science. – 2024. – Vol. 34. – No. 2. – P. 1181-1191.

Syed I. The multimodal trust effects of face, voice, and sentence content / I. Syed, M. Baart, J. Vroomen // Multisensory Research. – 2024. – Vol. 37. – No. 2. – Pp. 125-141.

Thangavel P. A lexicon-based approach for sentiment analysis of multimodal content in tweets / P. Thangavel, R. Lourdusamy // Multimedia Tools and Applications. – 2023. – Vol. 82. – No. 16. – Pp. 24203-24226.

Wasilewski A. One size does not fit all: multivariant user interface personalization in e-commerce / A. Wasilewski, G. Kolaczek // IEEE Access. – 2024. – Vol. 12. – Pp. 65570-65582.

Zhang Z. A survey on multimodal-guided visual content synthesis / Z. Zhang, Z. Li, K. Wei, S. Pan, Ch. Deng // Neurocomputing. – 2022. – Vol. 497. – Pp. 110-128.

10. Zhu Ya. Affective video content analysis via multimodal deep quality embedding network / Ya. Zhu, Zh. Chen, F. Wu // IEEE Transactions on Affective Computing. – 2022. – Vol. 13. – No. 3. – Pp. 1401-1415