*Journal of Digital Media & Culture Technology* Vol.6, No.1, June (2026), pp.33-44
*ISSN 2983-0664(p) 2983-0850(e)* *http://dx.doi.org/10.29056/jdmct.2026.06.04*

# Vision-Conditioned Gating for Product-Store-Level Demand Forecasting of New Fast-Fashion Products

Jeongho Kim[1], Saebom Lee[2], Chang Choi[3]*

## Abstract

Fast fashion demand forecasting at the product-store level is a critical task for improving market responsiveness, profitability, and operational stability. Nevertheless, product-store level data are often sparse and highly volatile, and forecasting becomes particularly challenging in cold-start settings where new products lack historical sales records. In response, this study proposes Vision-Conditioned Gated RNN (VIGRNN), based on the observation that the predictive usefulness of auxiliary information varies across product-store instances. VIGRNN is an adaptive forecasting model that dynamically controls attribute, release-date, and trend information on a per-sample basis, conditioned on visual information. Experimental results show that, under cold-start settings, the proposed model outperforms an existing multimodal RNN-based baseline by reducing WAPE by 3.6% and MAE by 2.47%. In addition, it reduces GFLOPs by 5.33% and the number of parameters by approximately 32%. These findings suggest that the proposed model effectively balances forecasting accuracy and computational efficiency, highlighting its potential for practical deployment in real-world industrial environments.

Keyword : Fast Fashion, Cold Start, Demand Forecasting, Multimodal Learning, Vision-Conditioned Gating

## 1. Introduction

Demand forecasting in the fast fashion industry goes beyond simple sales estimation and provides a critical basis for decision-making in the optimal allocation of limited inventory across multiple stores [1]. In particular, in an environment characterized by short product life cycles and high sensitivity to rapidly changing trends, errors in initial inventory allocation can lead to overstocking or early stockouts, thereby causing substantial business losses; therefore, precise demand forecasting is essential [2][3]. However,

1 Department of Computer Engineering, Gachon University, Seongnam, Korea [Undergraduate Student]
e-mail: kjh0409@gachon.ac.kr
2 Department of Computer Engineering, Gachon University, Seongnam, Korea [Graduate Student]
e-mail: dltoqha@gachon.ac.kr
3 Department of Computer Engineering, Gachon University, Seongnam, Korea [Professor]
e-mail: changchoi@gachon.ac.kr (Corresponding author)
* This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00559546), and this research was supported by the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (RQT-25-120226).

*Received(April 1, 2026), Review Result(1st: April 18, 2026), Accepted(June 13, 2026), Published(June 30, 2026)*

inventory allocation in practice is performed at the product-store level, where demand is highly uncertain due to substantial sales variability arising from regional characteristics, consumer preferences, and store-specific operational conditions [4]. As shown in [Fig. 1], the same product may exhibit substantial variation in both total sales volume and week-by-week sales patterns across stores, underscoring the high uncertainty of product-store-level demand forecasting. In addition, traditional time-series forecasting methods that rely on historical data are limited in estimating demand for new products with no sales history and in capturing store-level demand heterogeneity [5]. Accordingly, recent studies have explored multimodal demand forecasting approaches that incorporate external signals, such as product images, product attribute text, release date, and online interest, in order to compensate for limited sales history [6].


[Fig. 1] Variations in Sales Volume across Stores for the Same Product

However, existing multimodal demand forecasting studies have primarily been based on integrating diverse sources of information, such as product characteristics, expert knowledge, visual features, and sales information, as predictive variables [7][8]. Such approaches have limitations in improving forecasting performance because they do not sufficiently capture the complementary relationships among different information sources or the instance-specific differences in their importance. In particular, since image, text, and temporal information have different representational structures and semantic properties, simply combining them in a uniform manner makes it difficult to fully exploit the cues that are most relevant for forecasting. To address this limitation, methods have been introduced that learn the relationships among heterogeneous information sources, such as images and text, through cross-attention, and these have shown superior performance compared with simple fusion approaches [9]. However, cross-attention-based methods require repeatedly computing attention scores over multimodal fused representations at each prediction step, which increases computational cost and inference latency. These characteristics may limit their practical applicability in the fast fashion industry, where large-scale

product-store-level decision-making is required; therefore, an approach that jointly considers both forecasting performance and efficiency is needed.

Accordingly, this study proposes a Vision-Conditioned Gated RNN (VIGRNN), which adjusts the extent to which auxiliary signals are incorporated for each sample through vision-conditioned gating. VIGRNN adopts visual representations, which play a central role in consumer purchase decisions [10][11], as the primary source of information, while utilizing text and temporal information as auxiliary signals to model inter-modal interactions more flexibly. This design enables more stable demand forecasting in product-store-level cold-start settings by preserving information that is useful for prediction while suppressing irrelevant signals. Experimental results on the real-world fashion e-commerce dataset VISUELLE 2.0 [12] show that the proposed model improves WAPE and MAE by 3.6% and 2.47%, respectively, compared with an existing cross-attention-based multimodal RNN model [9]. In addition, it reduces GFLOPs by approximately 5.4% and the number of parameters by about 32%. These results demonstrate that the proposed model can simultaneously achieve forecasting accuracy and inference efficiency even under cold-start conditions with large store-level demand variation. Therefore, this study presents an effective demand forecasting approach to support more precise initial inventory allocation decisions in the highly volatile fast fashion environment.

## 2. Related Work

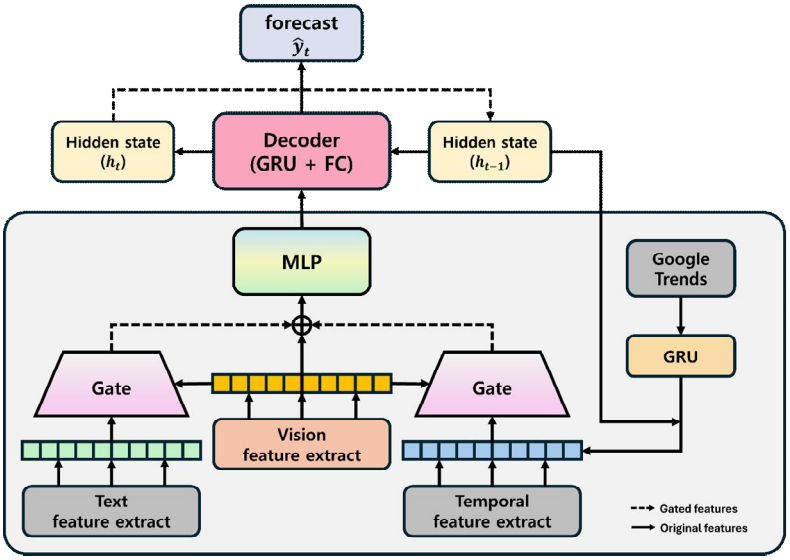
The fast fashion industry is highly sensitive to trends and exhibits substantial seasonal variability, resulting in short product life cycles; consequently, failures in demand forecasting can lead to increased inventory management costs [13]. In response, the fashion industry has introduced intelligent forecasting systems in an effort to mitigate the uncertainty of new product demand forecasting through data-driven approaches [14]. Early studies in this area were largely based on similarity-based forecasting, which refers to the demand patterns of historical products similar to a new product in order to address the lack of data in cold-start settings [15]. In addition, it has been reported that referencing the demand patterns of multiple similar products, rather than relying on a single similar product, is more effective for improving forecasting performance [16]. However, because the criteria for selecting similar historical products are not always clear and market environments change dynamically, the demand patterns of historical products often have limited ability to fully explain the demand of new products. To address these limitations, deep learning methods that learn demand patterns from large-scale data have begun to be adopted, along with multimodal approaches that incorporate unstructured data such as product images.

Recent studies on multimodal demand forecasting have proposed various methods to capture the relationships among heterogeneous information sources, such as product images, product attribute text, and sales history [9]. However, most existing multimodal demand forecasting approaches are based on integrating the features of each modality into a joint representation, which limits their ability to fully reflect the complementary relationships among different information sources and the instance-specific differences in modality importance [7][8]. To address this issue, CrossAttnRNN [9] employed cross-attention to capture complex inter-modal interactions and improved forecasting performance by incorporating additional internal variables, such as holidays and promotional events. This was a meaningful attempt to introduce diverse multimodal information and internal signals into time-series forecasting. However, because it did not incorporate external signals that reflect actual market interest, it was limited in its ability to sensitively capture consumer demand.

Subsequently, with the introduction of the VISUELLE dataset, which includes external interest signals such as Google Trends [6], research has expanded toward improving forecasting performance by compensating for the limited historical records of new products [17-19]. However, in real industrial settings, product-store-level demand forecasting for new products aimed at inventory allocation remains a challenging problem due to data sparsity and store-level demand variability [4]. In particular, in multimodal demand forecasting, the modalities that are most informative for prediction may vary across product-store instances [20], which calls for a fusion mechanism capable of flexibly adjusting the contribution of each modality on a per-sample basis. Although cross-attention-based fusion is effective for learning inter-modal interactions, it has limitations in adaptively suppressing irrelevant modality signals and selectively preserving useful ones for individual instances. In contrast, gating-based fusion methods, such as the Gated Multimodal Unit (GMU) [21], are well suited to this requirement because they can dynamically regulate the contribution of each modality according to the input sample. Therefore, this study proposes a gating-based approach that reflects sample-specific differences in modality usefulness and enables more stable demand forecasting.

## 3. Methodology

This section presents the overall architecture of the Vision-Conditioned Gated RNN (VIGRNN) for new product sales forecasting. Specifically, it describes feature extraction from image, text, and temporal modalities, the query-based gating mechanism conditioned on visual information, and the integrated forecasting process. The overall model architecture is shown in [Fig. 2].




[Fig. 2] Overall Architecture of Vision-Conditioned Gated RNN (VIGRNN)

### 3.1 Vision Feature Extraction

Product images constitute core information that directly influences consumer purchase decisions, and visual attributes such as color, pattern, material texture, and silhouette serve as important factors in shaping product demand. Accordingly, this study adopts ResNet101 [22] as the image encoder, where the early layers are frozen to preserve general low-level visual patterns, while the later layers are fine-tuned to learn high-level representations specialized for the fashion domain. Through this process, a visual embedding $E_{img} \in R^d$, which reflects the visual demand characteristics of product $p$ from its image $v_p$ is extracted. The computation is formulated as follows.

$$E_{img} = Linear_{v}(F_{ResNet101}(v_p)) \quad (1)$$

### 3.2 Text Feature Extraction

The product attribute text consists of category, color, fabric, and store ID, and can complement demand-related characteristics that are difficult to capture using time-series information alone, particularly in settings where observed sales history is limited, as in the case of new products. In particular, these attributes can effectively represent fine-grained differences that may vary according to both the intrinsic

characteristics of the product and the sales context of each store, thereby reflecting similar sales patterns within the same product group as well as store-specific distinctions. To this end, the set of textual attribute components is defined as $K=\{cat,col,fab,store\}$. Let $\alpha_k$ denote the embedding of each attribute $k \in K$. Then, the text embedding $E_{text} \in R^d$ is given as follows.

$$E_{text} = Linear_t(\alpha_{cat} + \alpha_{col} + \alpha_{fab} + \alpha_{store}) \quad (2)$$

### 3.3 Temporal Feature Extraction

Fast fashion products are highly sensitive to seasonality and trend changes; therefore, it is important to incorporate the temporal context of the release date into forecasting. To this end, this study decomposes the product release date into four components-year, month, week, and day-in order to jointly reflect long-term trends and periodic patterns that emerge at different temporal scales. Since each component has a different scale, each is normalized to the range between 0 and 1 by dividing it by its maximum possible value. Let the normalized values be denoted by $y,m,w,d$ respectively. Then, the date information embedding $E_{date} \in R^d$ is defined as follows.

$$E_{date} = Linear_y(y) + Linear_m(m) + Linear_w(w) + Linear_d(d) \quad (3)$$

In addition, Google Trends reflects external interest related to product attributes and is therefore useful for compensating for the limited historical records of new products in cold-start settings. In this study, the trend data $g_p$ for the category, color, and fabric of product $p$ are first encoded using a GRU and then combined with the previous hidden state of the decoder, $h_{t-1}$ to obtain the trend embedding $E_{trend}^{(t)} \in R^d$ at the current prediction step $t$.

$$E_{trend}^{(t)} = [h_{t-1}; F_{GRU}(g_p)] \quad (4)$$

Accordingly, the final temporal embedding is defined as follows by combining the date information and trend information.

$$E_{time}^{(t)} = E_{date} + E_{trend}^{(t)} \quad (5)$$

### 3.4 Query-Based Gating Mechanism

Since the modalities that are informative for prediction may differ across product-store instances, this study proposes an asymmetric structure in which a query embedding controls the extent to which the remaining context embeddings are incorporated through a sigmoid gate. In particular, to reflect the importance of visual information that is closely related to consumer demand, the product image embedding is designated as the query, while the product attribute text embedding and the temporal embedding, which includes date and trend information, are used as the context.

$$Q = E_{img},\ C_1 = E_{text},\ C_2 = E_{time}^{(t)} \tag{6}$$

Subsequently, each context embedding is combined with the query embedding and passed through a sigmoid gate, through which it is transformed into a representation whose degree of incorporation is adaptively modulated according to the query information.

$$C_1^{'} = \sigma(W_1 \bullet [Q; C_1] + b_1) \tag{7}$$

$$C_2^{'} = \sigma(W_2 \bullet [Q; C_2] + b_2) \tag{8}$$

Finally, the modulated context representations are integrated into the query representation to construct the final fused representation at the current prediction step as follows.

$$E_{fusion}^{(t)} = Q + C_1^{'} + C_2^{'} \tag{9}$$

### 3.5 Sales Forecasting for New Products

In this study, future sales are forecast sequentially based on the final fused representation obtained above, which reflects the product's image, text, and temporal information. To this end, the final fused representation is passed through an MLP to transform it into the decoder input representation, which is then fed into a GRU-based decoder together with the previous hidden state $h_{t-1}$. The GRU generates the output $o_t$ and hidden state $h_t$ at the current prediction step $t$. Afterwards, $o_t$ is projected through a linear layer to produce the prediction $y_t^{'}$ for the current step.

$$x_t = MLP(E_{fusion}^{(t)}) \tag{10}$$

$$o_t, h_t = F_{GRU}(x_t, h_{t-1}) \tag{11}$$

$$y_t^{'} = Linear_o(o_t) \tag{12}$$

## 4. Experiments

In this study, to validate the effectiveness of the proposed method, CrossAttnRNN [9] is adopted as the baseline, and experiments are conducted on the VISUELLE 2.0 dataset [12] for 12-week product-store-level demand forecasting in a cold-start setting. The dataset provides not only sales records for 5,355 fashion products across 110 stores, but also product images, product attribute text, and Google Trends information for attribute-related keywords. The experiments are conducted using an NVIDIA V100 32GB GPU. For model training, MSE is used as the loss function and Adafactor as the optimizer. Training is performed for 50 epochs, and the validation performance is measured at the end of each epoch; the parameters achieving the best validation performance are saved and used for final prediction. Forecasting performance is evaluated using Mean Absolute Error (MAE) and Weighted Absolute Percentage Error (WAPE). MAE, defined as the average absolute difference between the ground-truth and predicted values, is useful for measuring the intuitive magnitude of forecasting errors in sales prediction. WAPE, defined as the ratio of the total absolute error to the total actual sales, is suitable for evaluating relative forecasting performance while accounting for the demand scale of individual items. Here, $y_i$ denotes the actual sales, $y_t^{'}$ denotes the predicted sales, and $N$ denotes the total number of samples. The formulas for the evaluation metrics are given below.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i^{'}\right| \tag{13}$$

$$WAPE = \frac{\sum_{i=1}^{N}\left|y_i - y_i^{'}\right|}{\sum_{i=1}^{N}\left|y_i\right|} \times 100 \tag{14}$$

In addition, this study compares the computational cost and model size required for inference to evaluate the practical deployability of the forecasting model. The computational cost of inference is measured using Giga Floating Point Operations (GFLOPs), which quantify the number of floating-point operations incurred during inference and serve as an indicator of computational burden. Model size is compared by the number of parameters, which is closely related to physical storage requirements and

memory usage at inference time. Using these evaluation metrics, the proposed method is comprehensively assessed not only for forecasting performance but also for its practical applicability in real industrial environments.

### 4.1 Comparison of 12-Week Demand Forecasting Performance for New Products

[Table 1] presents the comparative performance of the existing methods and VIGRNN. The KNN-based approach, which selects historical products similar to a new product and utilizes their sales patterns for prediction, showed relatively lower forecasting performance than deep learning-based methods such as CrossAttnRNN [9]. Among the KNN models using a single modality as input, Attribute KNN, which relies only on text, achieved a WAPE of 91.26 and an MAE of 1.07, whereas Image KNN, which uses only image, achieved a WAPE of 90.17 and an MAE of 1.06. This suggests that visual information may be more informative than text for fashion products demand forecasting. In addition, Attribute+Image KNN, which utilizes both text and image, recorded a WAPE of 89.73 and an MAE of 1.06, indicating that the combination of multiple information sources contributes to improved forecasting performance.

[Table 1] Comparison of Forecasting Performance and Model Efficiency with Baselines

| Method | Input modality | WAPE(%) | MAE | GFLOPs | #Params |
|---|---|---|---|---|---|
| Attribute KNN | Text | 91.26 | 1.07 | - | - |
| Image KNN | Vision | 90.17 | 1.06 | - | - |
| Attribute+Image KNN | Text+Vision | 89.73 | 1.06 | | - |
| CrossAttnRNN | Text+Vision+Temporal | 83.06 | 0.97 | 30.83 | 251.29M |
| **VIGRNN** | | **80.07** | **0.94** | **29.18** | **170.90M** |

Among the comparison models, CrossAttnRNN [9] achieved the best performance, with a WAPE of 83.06 and an MAE of 0.97, showing a meaningful improvement over the KNN-based approaches. This shows that modeling complex interactions among multiple modality representations is effective for product-store-level cold-start demand forecasting. The proposed VIGRNN achieved the best forecasting performance overall, with a WAPE of 80.07 and an MAE of 0.94. Compared with CrossAttnRNN [9], it also reduced GFLOPs by 5.33% and the number of parameters by 32%. These results suggest that the gating structure, which dynamically modulates the contribution of context information based on the query modality, can adaptively incorporate useful information while reducing unnecessary computation.

### 4.2 Analysis of Forecasting Performance by Query Modality Choice

[Table 2] reports the performance variation of VIGRNN according to the choice of query modality. The results show that using Vision as the query yielded the best performance, with a WAPE of 80.07 and an MAE of 0.94. Using Text as the central modality resulted in a WAPE of 80.56 and an MAE of 0.95, and using Temporal as the central modality resulted in a WAPE of 81.09 and an MAE of 0.95. These findings indicate that the forecasting model operates most effectively when visual information is used as the query to regulate the other modality signals, suggesting that visual information can serve as an important reference signal in shaping the demand for fashion products.

[Table 2] Comparison of Forecasting Performance across Different Query Modality Choices

| Query Modality | Input modality | WAPE(%) | MAE |
|---|---|---|---|
| **Vison** | Text+Vision+Temporal | **80.07** | **0.94** |
| **Text** | | 80.56 | 0.95 |
| **Temporal** | | 81.09 | 0.95 |

## 5. Conclusion

In this study, we proposed VIGRNN for forecasting the demand of new products at the product-store level in the fast fashion industry. The proposed model fixes image representations, which are closely related to product demand, and adaptively modulates the contributions of textual and temporal information through a gating mechanism. VIGRNN achieved superior forecasting performance compared with existing methods and also demonstrated improved efficiency in terms of computational cost and model lightweight. These results suggest that the proposed approach has strong practical potential in the fast fashion industry, where rapid and accurate decision-making is essential. In future work, we plan to extend the model to more effectively incorporate external signals, such as long-term accumulated trend patterns, and to learn inter-modal interactions in a more stable manner.

## References

[1] J. Gallien, A. J. Mersereau, A. Garro, A. D. Mora, and M. N. Vidal, "Initial shipment decisions for new products at Zara," *Operations Research*, vol. 63, no. 2, pp. pp. 269-286, Feb. 2015, doi: 10.1287/opre.2014.1343.

[2] Y. Zhang, J. F. Wang, C. Lin, and G. T. M. Hult, "Assessing fast fashion overstock through time-to-peak-sales," *Journal of Retailing*, vol. 101, no. 3, pp. 431-453, Sep. 2025, doi: 10.1016/j.jretai.2025.05.001.

[3] S. Gopalakrishnan, R. Hariss, H. Moshrefi, and S. Ray, "Out-of-stock, out-of-store? Estimating cross-store fulfillment via inferring customer journeys," unpublished.

[4] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346-1364, Oct. 2022, doi: 10.1016/j.ijforecast.2021.11.013.

[5] K. Swaminathan and R. Venkitasubramony, "Demand forecasting for fashion products: A systematic review," *International Journal of Forecasting*, vol. 40, no. 1, pp. 247-267, Jan. 2024, doi: 10.1016/j.ijforecast.2023.02.005.

[6] G. Skenderi, C. Joppi, M. Denitto, and M. Cristani, "Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends," *Journal of Forecasting*, vol. 43, no. 6, pp. 1982-1997, Mar. 2024, doi: 10.1002/for.3104.

[7] A. L. Loureiro, V. L. Miguéis, and L. F. Da Silva, "Exploring the use of deep neural networks for sales forecasting in fashion retail," *Decision Support Systems*, vol. 114, pp. 81-93, Oct. 2018, doi: 10.1016/j.dss.2018.08.010.

[8] C. Giri and Y. Chen, "Deep learning for demand forecasting in the fashion and apparel retail industry," *Forecasting*, vol. 4, no. 2, pp. 565-581, Jun. 2022, doi: 10.3390/forecast4020031.

[9] V. Ekambaram, K. Manglik, S. Mukherjee, et al., "Attention based multi-modal new product sales time-series forecasting," in *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, Aug. 23-27, 2020, pp. 3110-3118, doi: 10.1145/3394486.3403362.

[10] S. Jang, H. Kim, "A Study on the Impact of Regional Characteristics in Rice Packaging Design on Consumer Purchase Behavior," *Journal of Next-generation Convergence Information Services Technology*, vol. 14, no. 1, pp. 81-94, Feb. 2025, doi: 10.29056/jncist.2025.02.08.

[11] K. Shi, H. Kim, "A Study on the Logo Design of Chinese Medical Companies," *Journal of Next-generation Convergence Information Services Technology*, vol. 14, no. 2, pp. 213-226, Apr. 2025, doi: 10.29056/jncist.2025.04.06..

[12] G. Skenderi, C. Joppi, M. Denitto, B. Scarpa, and M. Cristani, "The multi-modal universe of fast-fashion: The Visuelle 2.0 benchmark," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, New Orleans, LA, USA, Jun. 19-24, 2022, pp. 2241-2246, doi: 10.1109/CVPRW56347.2022.00245.

[13] S. Thomassey, "Sales forecasts in clothing industry: The key success factor of the supply chain management," *International Journal of Production Economics*, vol. 128, no. 2, pp. 470-483, Dec. 2010, doi: 10.1016/j.ijpe.2010.07.018.

[14] T. M. Choi, C. L. Hui, N. Liu, S. F. Ng, and Y. Yu, "Fast fashion sales forecasting with limited data and time," *Decision Support Systems*, vol. 59, pp. 84-92, Mar. 2014, doi: 10.1016/j.dss.2013.10.008.

[15] K. C. Green and J. S. Armstrong, "Structured analogies for forecasting," *International Journal of Forecasting*, vol. 23, no. 3, pp. 365-376, Jul.-Sep. 2007, doi: 10.1016/j.ijforecast.2007.05.005.

[16] P. Goodwin, K. Dyussekeneva, and S. Meeran, "The use of analogies in forecasting the annual sales of new electronics products," *IMA Journal of Management Mathematics*, vol. 24, no. 4, pp. 407-422, Oct. 2013, doi: 10.1093/imaman/dpr025.

[17] S. I. Papadopoulos, C. Koutlis, S. Papadopoulos, and I. Kompatsiaris, "Multimodal quasi-autoregression: Forecasting the visual popularity of new fashion products," *International Journal of Multimedia Information Retrieval*, vol. 11, pp. 717-729, Oct. 2022, doi: 10.1007/s13735-022-00262-5.

[18] M. Rajendran and B. Hong, "Autoregressive multimodal transformer for zero-shot sales forecasting of fashion products with exogenous data," *Applied Intelligence*, vol. 55, Art. no. 108, Dec. 2024, doi: 10.1007/s10489-024-05972-3.

[19] X. Li, J. Shen, D. Wang, W. Lu, and Y. Chen, "Multi-modal transform-based fusion model for new product sales forecasting," *Engineering Applications of Artificial Intelligence*, vol. 133, Art. no. 108606, Jul. 2024, doi: 10.1016/j.engappai.2024.108606.

[20] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Vancouver, Canada, Jun. 18-22, 2023, pp. 2575-2584, doi: 10.1109/CVPRW59228.2023.00256.

[21] J. Arevalo, T. Solorio, M. Montes-y-Gomez, and F. A. González, "Gated multimodal networks," *Neural Computing and Applications*, vol. 32, pp. 10209-10228, Jan. 2020, doi: 10.1007/s00521-019-04559-1.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 27-30, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.