applied sciences

MDPI

Article

From Popularity Signals to Semantic Descriptions: Dual-Stream Driven Fashion Trend Forecasting

Shenguo Fang [1,2], Jin Cui [3,4], Xiaofen Ji [1,*], Jing Zhang [3], Shijing Shen [3], Tuocheng Zeng [3,4], Gang Chen [5], Houyong Yu [1] and Xiaohua Pan [3,4]

1 College of Textile Science and Engineering (International Institute of Silk), Zhejiang Sci-Tech University, Hangzhou 310018, China; 202010201006@mails.zstu.edu.cn (S.F.)
2 Zhejiang Institute of Industry and Information Technology, Hangzhou 310023, China
3 Information Technology and Emerging Service Innovation Research Center, Binjiang Institute of Zhejiang University, Hangzhou 310053, China
4 College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
5 School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
* Correspondence: xiaofenji@zstu.edu.cn

Abstract

Fashion trend forecasting is crucial for proactive supply chain management, sustainable production, and personalized marketing. However, accurately predicting fine-grained fashion trends across diverse user groups remains challenging, as conventional numerical models rely solely on historical popularity signals and cannot explicitly exploit the high-level semantic context underlying trend evolution. To address this limitation, we propose a Dual-Stream Driven Fashion Trend Forecasting (DDFTF) framework that integrates numerical forecasting with textual semantic reasoning. The numerical stream employs a multi-scale patch Transformer with metadata-aware feature fusion to capture temporal dynamics, while the semantic stream converts historical trends into structured textual descriptions, extracting interpretable semantic representations of global trends and turning points. A residual fusion module combines both streams by treating semantic signals as complementary corrections to numerical predictions. Extensive experiments on the FIT and GeoStyle benchmarks demonstrate that DDFTF consistently outperforms state-of-the-art methods, achieving up to a 14.6% relative MAE reduction in long-term forecasting. Ablation and qualitative analyses further show that the performance gains arise from meaningful semantic information rather than increased model capacity, while also improving the interpretability of fashion trend prediction. Our code is publicly available.

Keywords: fashion trend forecasting; time series forecasting; multimodal learning; consumer behavior analysis

Check for updates

Academic Editor: Paolo Renna

Received: 24 July 2026
Revised: 20 August 2026
Accepted: 25 August 2026
Published: 11 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The fashion industry [1–3] is a multi-trillion-dollar global market where accurate trend forecasting is critical for supply chain management [4], product design [5], and marketing [6]. At the same time, increasing concerns over textile waste have highlighted the need for more sustainable production practices [7,8], where precise demand forecasting enables on-demand manufacturing and reduces overproduction.

Despite its importance, fashion trend forecasting has become increasingly challenging due to the rapid growth of e-commerce and the rise of personalized consumer preferences,

leading to fragmented and highly dynamic markets [3,9]. These challenges call for more advanced data-driven approaches capturing complex and evolving trend dynamics.

To address these challenges, prior work has largely focused on numerical modeling. Early approaches relied on classical statistical methods to capture basic trend patterns. With the rise of deep learning, neural networks [10] and recurrent models such as LSTMs [11] enabled more effective modeling of non-linear temporal dynamics. More recently, advanced architectures, including N-BEATS [12] and Transformer-based models [13], have achieved strong performance by capturing long-range dependencies and complex statistical structures. Despite these advances, existing approaches remain predominantly numerical, lacking the ability to model high-level semantic factors that are crucial in shaping fashion trends.

Building on these advances, trend forecasting systems [14–16] have been developed to support proactive decision-making in fashion design and retail. For example, Mall et al. [17] uncover localized trends from large-scale street imagery, while Ma et al. [18] formalize fine-grained fashion trend forecasting for modeling the evolution of fashion elements across heterogeneous user groups using both sequential data and external knowledge.

Despite these advancements, a critical gap remains: existing methods predominantly rely on historical popularity signals, while fashion trends are also described through rich semantic contexts [19]. As a result, numerical-stream models may overlook high-level cues associated with trend formation and turning points. As illustrated by the "turtleneck" trend in Figure 1, abrupt changes can be difficult to distinguish from ordinary fluctuations using numerical extrapolation alone. In this work, we use the observed historical sequence and metadata to construct semantic trend priors; external events and causal signals are outside the scope of the present study.

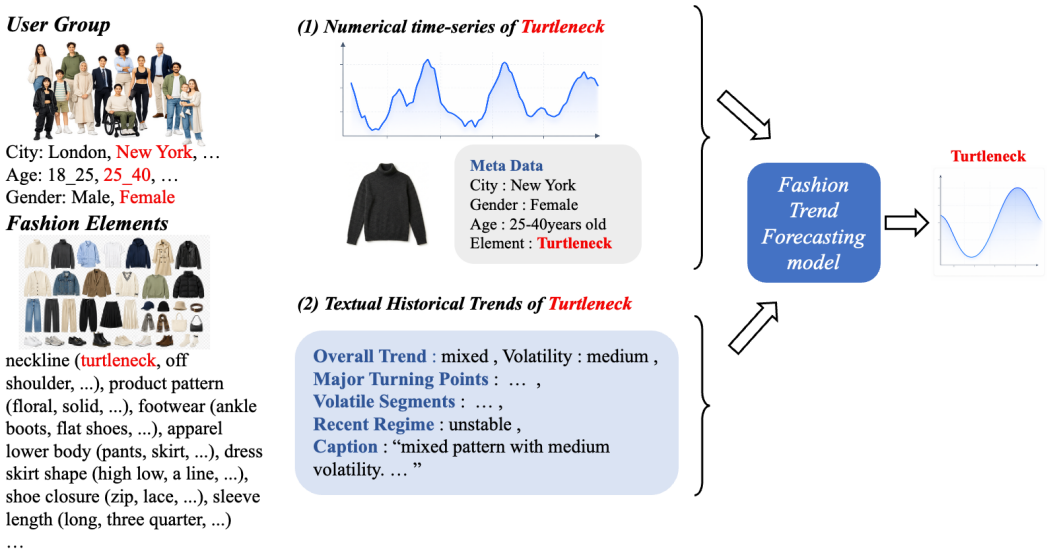


Figure 1. Fashion trend forecasting predicts future popularity of fashion elements, and our method enhances it by integrating textual semantics with numerical signals.

Recent advances in Transformer architectures [20] and Pre-trained Language Models (PLMs), such as BERT [21], have demonstrated strong capabilities in semantic understanding and reasoning. More recently, Large Language Models (LLMs) [22] have shown promising potential in modeling temporal patterns and generalizing across domains. However, directly applying LLMs to time series forecasting remains challenging due to their limited numerical precision and the difficulty of aligning heterogeneous modalities.

To address this gap, we propose a dual-stream framework that combines data-driven forecasting with semantic trend representations. The framework predicts future popularity while providing high-level descriptions of salient historical patterns, without requiring external event data. DDFTF incorporates two functionally complementary components: a numerical forecasting stream and a textual semantic stream.

The numerical stream captures fine-grained temporal dynamics through multi-scale temporal segmentation and metadata-aware feature fusion. In parallel, the textual stream converts historical numerical trends into semantically enriched textual descriptions using prompt-based generation and pretrained language encoding, injecting interpretable high-level knowledge into forecasting. A dedicated fusion module then jointly integrates both streams to predict future popularity trends.

The main contributions of this work are as follows:

(1) We introduce a metadata-aware patch-based numerical stream with adaptive importance weighting for fine-grained temporal modeling under diverse user group distributions.
(2) We propose a textual semantic stream that distills global trend semantics and critical turning-point information from historical popularity sequences.
(3) We present a flexible dual-stream fusion framework that uses residual refinement as the default strategy and evaluates weighted and attention-based variants as ablation references.

2. Related Work

2.1. Fashion Trend Forecasting and Analysis

Research in the fashion domain has shifted from static recognition to dynamic trend analysis. Early efforts primarily focused on descriptive analytics, such as detecting clothing attributes from runway shows [23,24] or mapping the spatio-temporal distribution of street styles [17,25]. While effective at quantifying current aesthetics, these methods lack predictive foresight. To address this, recent studies have transitioned toward proactive modeling; notably, GeoStyle [17] employs periodic functions to forecast attribute popularity. To enhance robustness, knowledge-aware frameworks like KERN [18] and REAR [26] incorporate structural domain knowledge (e.g., category hierarchies) to stabilize predictions across heterogeneous user groups. Despite these strides, most fashion-specific models remain tethered to numerical signals. They often fail to capture the underlying semantic drivers, resulting in "lagging" predictions that struggle to anticipate abrupt trend inflections. Most closely related, ref. [27] introduces a dual-stream caption-guided framework that we extend with metadata-aware numerical modeling and flexible fusion; refs. [18,24] provide earlier knowledge-enhanced and visual-style foundations.

2.2. Time Series Forecasting Model

General time series forecasting has evolved from traditional statistical methods to sophisticated deep learning architectures [28,29]. While classical models like ARIMA [30] perform well in stationary environments, they often falter when faced with the inherent non-stationarity and volatility of fashion data. Deep learning approaches, particularly LSTMs [31] and Transformers [32,33], have set new benchmarks by capturing long-range dependencies. However, these "black-box" models frequently suffer from a lack of interpretability and an inability to assimilate high-level conceptual knowledge. To bridge this gap, emerging research has pivoted toward semantic-aware forecasting and time-series captioning [34,35]. Frameworks such as TESSA [36] and DualSG [27] suggest that natural language descriptions can provide structural priors for numerical models. Nevertheless, these methods are predominantly general-purpose and remain agnostic to the metadata-

rich, fine-grained nuances of the fashion industry. Consequently, a significant opportunity remains to integrate numerical precision with semantic reasoning, enabling a more holistic and explanatory approach to fashion trend evolution. Recent long-horizon forecasters, including state-space models such as TimeMachine and Transformer-based models such as PatchTST [37,38], further extend long-range forecasting, while our framework complements these efforts by injecting domain-specific semantic priors.

3. Problem Formulation

Consistent with the framework established by Ma et al. [18], we define the task of fine-grained fashion forecasting as predicting the temporal evolution of style elements across diverse user segments. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$ denote a set of M fine-grained fashion elements (e.g., white, high-neck), and $\mathcal{G} = \{g_1, g_2, \ldots, g_K\}$ represent a set of K specific user groups defined by demographic attributes such as age, gender, and city. For a given fashion element $f \in \mathcal{F}$ and a user group $g \in \mathcal{G}$, the historical popularity is represented as a time series $y_t = N_t^{g,f} / N_t^g$, where $N_t^{g,f}$ denotes the number of occurrences of fashion element f in user group g at time t, and N_t^g represents the total number of fashion items (e.g., apparel, bags, footwear) observed in group g at time t. Given a historical time range $[1, T]$, our objective is to predict the future values over a forecasting horizon $[T+1, T+H]$, where T denotes the length of the historical sequence, H denotes the prediction window (forecasting horizon), and $T \gg H$. In our implementation, both datasets are organized as preprocessed instances, each consisting of a historical popularity window, a future forecasting horizon, and associated metadata. All experiments are conducted at weekly granularity. For FIT, we use a 48-step historical window and consider two forecasting settings with $H = 12$ and $H = 24$, corresponding to half-year and one-year prediction, respectively. For GeoStyle, each instance uses a 52-step historical window and a 26-step forecasting horizon.

4. Methodology

4.1. Overall Framework

As shown in Figure 2, our proposed DDFTF comprises a numerical forecasting stream, a textual semantic stream, and a dual-stream fusion module, which jointly model historical trends and generate fine-grained fashion trend predictions.

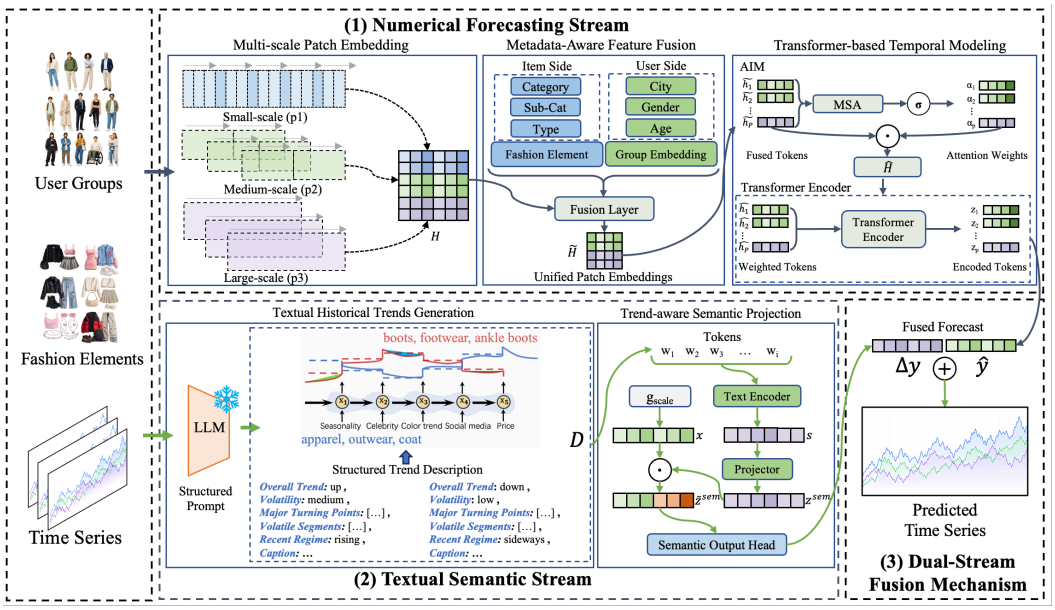


Figure 2. Overall framework of our proposed fashion trend prediction model.

4.2. Numerical Forecasting Stream

The numerical forecasting stream is designed as a lightweight yet expressive module dedicated to capturing fine-grained temporal variations in fashion popularity. Departing from traditional point-wise modeling [39], this stream adopts a patch-based strategy that integrates multi-scale temporal representations with metadata-aware feature fusion. Concretely, the numerical stream is instantiated as a patch-based Transformer encoder rather than a recurrent backbone. The historical popularity sequence is segmented into temporal patches, encoded as patch tokens, reweighted by the adaptive importance mask, and then processed by self-attention layers to capture inter-patch temporal dependencies before forecasting the future horizon.

Multi-Scale Patch Representation. To overcome the rigid resolution of fixed-length patches, we segment the historical sequence at multiple temporal scales. This generates non-uniform temporal tokens, allowing the model to capture both high-resolution fluctuations and low-resolution structural trends.

Let the historical popularity sequence be $\mathbf{x} = [x_1, x_2, \ldots, x_T]^\top$. We extract block-wise representations at three scales: Small (p_1), Medium (p_2), and Large (p_3). For the k-th scale, the sequence is partitioned into patches of length l_k and stride s_k:

$$\mathbf{P}^{(k)} = \mathcal{P}(\mathbf{x}, l_k, s_k), \quad \mathbf{P}^{(k)} \in \mathbb{R}^{N^k \times l_k}, \quad k \in \{1, 2, 3\}, \tag{1}$$

where N^k is the number of patches at scale k. We use $(l_k, s_k, p_k) = (8, 4, 2)$, $(16, 8, 4)$, and $(32, 16, 8)$ for the Small, Medium, and Large scales, respectively, where p_k denotes right-side replication padding. Thus, $N^k = \lfloor (T + p_k - l_k)/s_k \rfloor + 1$, giving $(11, 5, 2)$ available patches for FIT ($T = 48$) and $(12, 6, 2)$ for GeoStyle ($T = 52$). No historical observation is truncated before patch extraction. Each patch is projected into a shared d-dimensional embedding space:

$$\mathbf{H}^{(k)} = \mathbf{P}^{(k)} \mathbf{W}_k, \tag{2}$$

where $\mathbf{W}_k \in \mathbb{R}^{l_k \times d}$ is a scale-specific learnable projection. The resulting embeddings are concatenated along the patch dimension in a fixed Small-to-Large order to form a unified sequence $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_P] \in \mathbb{R}^{P \times d}$, where $P = \sum N^k$. The current design uses scale-specific projections and ordered token positions rather than an additional scale-ID embedding; an explicit hierarchical scale encoding is left as a future extension.

Metadata-Aware Feature Fusion. To account for the heterogeneity across different demographics and items, we incorporate structured metadata as context. Let $\mathbf{e}^{\text{city}}$, $\mathbf{e}^{\text{gender}}$, $\mathbf{e}^{\text{age}}$, and $\mathbf{e}^{\text{elem}}$ denote the d-dimensional embeddings for city, gender, age, and fashion element, respectively. In practice, these metadata are encoded by separate learnable embedding layers. The city, gender, and age embeddings are concatenated and transformed by an MLP to form a group-level representation, while the fashion element is retained as an independent element embedding.

Demographic attributes are first fused into a group-level representation:

$$\mathbf{e}_{\text{group}} = f_{\text{group}}([\mathbf{e}^{\text{city}}; \mathbf{e}^{\text{gender}}; \mathbf{e}^{\text{age}}]), \tag{3}$$

where $f_{\text{group}}(\cdot)$ is a multi-layer perceptron (MLP). This group prior and the element embedding $\mathbf{e}^{\text{elem}}$ are then broadcast and fused with each temporal patch $\mathbf{h}_p$:

$$\tilde{\mathbf{h}}_p = f_{\text{fuse}}([\mathbf{h}_p; \mathbf{e}_{\text{group}}; \mathbf{e}^{\text{elem}}]), \tag{4}$$

where $f_{\text{fuse}}(\cdot)$ maps the concatenated temporal and semantic features into a unified latent space, yielding the metadata-enriched sequence $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_P] \in \mathbb{R}^{P \times d}$. This design adopts concatenation-based fusion rather than cross-attention: for each temporal patch,

the patch token, the group embedding, and the element embedding are concatenated and projected back to the shared latent space, after which the enriched tokens are fed into the adaptive importance masking and Transformer encoding stages.

Transformer-based Temporal Modeling. Recognizing that temporal segments contribute unequally to future trends, we introduce an Adaptive Importance Mask (AIM) to assign an importance weight $\alpha_p \in [0,1]$ to every patch token before the main Transformer aggregation. The weights are obtained by applying a shared linear projection to a multi-head self-attention (MHA) output:

$$\boldsymbol{\alpha} = \sigma\big(\mathrm{MHA}(\tilde{\mathbf{H}})\mathbf{W}_\alpha\big), \quad \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_P] \in [0,1]^P, \tag{5}$$

where $\mathbf{W}_\alpha \in \mathbb{R}^{d\times 1}$ maps each patch token to a scalar. The gated sequence is then obtained by element-wise modulation, with the scalar α_p broadcast across the feature dimension and multiplied onto every feature of the corresponding token:

$$\hat{\mathbf{H}}[p,:] = \alpha_p \, \tilde{\mathbf{H}}[p,:], \quad \hat{\mathbf{H}} = \boldsymbol{\alpha} \odot \tilde{\mathbf{H}}. \tag{6}$$

In words, the importance weight is shared across the feature dimension of a patch but varies across patches. Placing AIM before the main Transformer encoder serves a two-fold purpose. (i) It reweights the patch representations before global temporal interaction, allowing the encoder to emphasize segments that are more relevant to the forecasting objective. (ii) The resulting $\boldsymbol{\alpha}$ is directly interpretable, and inspecting α_p along the look-back window reveals which historical periods the model considers most predictive of the future horizon.

The gated sequence $\hat{\mathbf{H}}$ is processed by a Transformer encoder that aggregates information across patches via self-attention:

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_P] = \mathrm{Transformer}(\hat{\mathbf{H}}), \quad \mathbf{Z} \in \mathbb{R}^{P\times d}, \tag{7}$$

yielding context-enhanced patch representations $\mathbf{z}_i$ that provide a robust foundation for prediction.

Prediction Head. Given the patch-level representations, the numerical stream generates future predictions by a flat linear projection from a fixed-order vectorization of patch tokens onto the target forecasting horizon.

With $\mathbf{Z} \in \mathbb{R}^{P\times d}$ from above, we stack all entries in a fixed order to obtain the global feature vector $\mathbf{u} = \mathrm{flatten}(\mathbf{Z}) \in \mathbb{R}^{d\cdot P}$. The forecast is then produced by

$$\hat{\mathbf{y}} = \mathbf{W}_o\mathbf{u} + \mathbf{b}_o, \quad \mathbf{W}_o \in \mathbb{R}^{H\times(d\cdot P)}, \ \mathbf{b}_o \in \mathbb{R}^H, \tag{8}$$

with $\hat{\mathbf{y}} \in \mathbb{R}^H$ predicting the future horizon H. The head is intentionally kept as a flat linear map without per-step recurrence or probabilistic output; richer horizon-aware decoders and uncertainty-aware heads are interesting orthogonal directions that we leave to future work. Through this architecture, the numerical stream completes the transition from historical sequence modeling to future trend estimation.

4.3. Textual Semantic Stream

We further introduce a textual semantic stream to capture high-level semantic information in trend evolution that is difficult to model from numerical signals alone. In contrast to the numerical stream, which focuses on local fluctuations, temporal dependencies, and structured metadata, the textual stream models global trend semantics and provides complementary guidance for forecasting. Specifically, it constructs descriptive trend narratives from historical time series and encodes them into continuous semantic representations us-

ing a pre-trained language model. These representations serve as trend-level priors and are further transformed via trainable projection and fusion modules to support final prediction.

Textual Historical Trends Generation. The textual semantic stream abstracts the numerical trends within a historical observation window into high-level, interpretable representations. Let the time series be denoted as $\mathbf{x}_{1:T} = \{x_1, x_2, ..., x_T\}$, where T signifies the look-back window. To bridge the gap between raw data and semantic understanding—analogous to the classic textile challenge of mapping subjective fabric handle to objective parameters [40], we employ a prompt-based generation strategy. Specifically, an LLM distills the sequence into key trend dimensions, including Overall Trend, Volatility, Major Turning Points, Volatile Segments, and Recent Regime, along with a summarizing Caption. Such structured descriptions provide a compact and interpretable representation of trend dynamics, enabling the model to capture latent factors and irregular events (e.g., seasonality, celebrity-driven shifts) beyond purely numerical patterns.

This structured description avoids raw popularity values and instead emphasizes high-level evolutionary patterns. During annotation generation, we enforce a structural consistency check requiring valid JSON format and one-to-one sample alignment within each batch; invalid outputs are discarded and regenerated. The structured JSON representation is used for the main DDFTF experiments, while free-form or long-text descriptions are reserved for comparison settings. Finally, $\mathcal{D}$ is encoded into a latent semantic space via a pretrained text encoder, serving as a semantic prior for the numerical forecasting stream.

Subsequently, the structured profile $\mathcal{D}$ is mapped into a continuous semantic space: $\mathbf{s} = f_{\text{text}}(\mathcal{D})$, where $\mathbf{s} \in \mathbb{R}^{768}$ is obtained by mean-pooling the final hidden states of a pre-trained GPT-2 base encoder. The encoder is used offline with frozen parameters, and the cached vector is then mapped by a trainable projection to $\mathbf{z}^{sem} \in \mathbb{R}^{128}$ before fusion. This design provides a fixed semantic interface while keeping language-model inference outside the forecasting optimization loop.

Quality Control and validation. To ensure that LLM-generated semantic descriptions are reliable rather than hallucinated, we adopt a two-stage validation pipeline. (1) Structural validation. All outputs are validated using rule-based checks to ensure valid JSON format with the required six fields. Invalid descriptions are automatically regenerated using the same prompt. As shown in Table 1, the first-pass validation rate reaches 96.1%, while the final pass rate exceeds 99.8% after the regeneration procedure. (2) Semantic consistency verification. We further validate the faithfulness of generated descriptions using the LLM-as-judge paradigm with GPT-4o. GPT-4o receives both the historical numerical sequence and its generated description, and evaluates whether the description faithfully reflects the underlying trend without fabrication or omission. The evaluation covers three aspects: trend direction, volatility level, and turning-point identification, each rated as correct, partially correct, or incorrect. Descriptions rated as either correct or partially correct are considered acceptable. Partially correct descriptions are included because they usually capture the main trend pattern, although they may differ in granularity or omit minor details. This automatic validation serves as a quality-control step rather than a substitute for human annotation, and future work will consider human evaluation and stricter filtering of incorrect descriptions. As shown in Table 2, the overall acceptance rate reaches 91.7%, indicating that the generated semantic descriptions reliably capture the temporal characteristics of the underlying numerical sequences. These results also suggest that the structured prompt and fixed JSON schema improve the stability of the semantic stream. The fixed semantic fields and text encoder help reduce sensitivity to variations in language model outputs.

Table 1. Structural validation pass rate of LLM-generated descriptions.

Dataset	Total Instances	First-Pass Rate (%)	Final Pass Rate (%)
GeoStyle	2024	97.2	100.0
FIT	8808	95.8	99.7
Overall	10,832	96.1	99.8

Table 2. Semantic consistency verification via LLM-as-judge on 500 randomly sampled descriptions. A stronger LLM (GPT-4o) rates each dimension as correct (C), partially correct (PC), or incorrect (I).

Dimension	Correct (%)	Partially Correct (%)	Incorrect (%)
Trend direction	82.4	11.8	5.8
Volatility level	76.2	15.4	8.4
Turning-point identification	71.6	17.8	10.6
Overall (C+PC)	—	91.7	—

Trend-aware Semantic Projection. To integrate the semantic representation with the numerical forecasting stream, we first map the text embedding into a latent space compatible with numerical features. Let $\mathbf{s}$ denote the semantic vector encoded from the structured description. We employ a learnable projection layer to obtain an aligned representation.

$$\mathbf{z}^{sem} = f_{proj}(\mathbf{s}), \tag{9}$$

where $f_{proj}(\cdot)$ denotes the trainable semantic projection layer, and $\mathbf{z}^{sem} \in \mathbb{R}^{128}$ represents the aligned semantic embedding.

Semantic Control Representation. After obtaining the projected semantic representation $\mathbf{z}^{sem}$, the textual stream provides horizon-level control signals for the residual refinement module. Rather than producing an independent forecast, the semantic branch supplies coefficients that modulate numerical residual bases, allowing high-level trend information to complement the numerical stream while preserving its temporal structure.

4.4. Dual-Stream Fusion Mechanism

Once the numerical forecast and semantic control representation are obtained, we produce the final prediction through a dual-stream fusion module. The numerical stream models fine-grained temporal dynamics, while the textual stream provides high-level semantic signals capturing global trend evolution. We adopt *residual refinement* as the default fusion strategy, since the numerical stream already forms a strong forecasting backbone and the textual stream mainly contributes complementary global trend corrections; two alternative strategies, *weighted interpolation* and *attention-based fusion*, are explored as ablation references and detailed below.

The final prediction $\mathbf{y}_f \in \mathbb{R}^H$ is formulated as:

$$\Delta\mathbf{y} = \eta\, \mathbf{r}(\mathbf{z}^{sem}, \mathbf{h}^{hist}) \odot \tanh(\mathbf{B}(\mathbf{h}^{num})\, \mathbf{c}(\mathbf{z}^{sem})), \quad \mathbf{y}_f = \hat{\mathbf{y}} + \Delta\mathbf{y}, \tag{10}$$

where $\mathbf{B}(\mathbf{h}^{num})$ denotes numerical residual bases, $\mathbf{c}(\mathbf{z}^{sem})$ provides semantic coefficients, and $\mathbf{r}(\mathbf{z}^{sem}, \mathbf{h}^{hist})$ provides a context-conditioned per-step amplitude. The saturation function bounds the correction, while η is a global residual scale; together these factors stabilize the semantic refinement while keeping the numerical backbone as the primary source of fine-grained temporal structure.

Because the semantic vector $\mathbf{s}$ summarizes the observation window globally, the semantic branch provides a bounded horizon-level correction, while the numerical stream preserves fine-grained temporal structure.

Ablation-only variants. The weighted interpolation variant produces a learnable per-step gate $\mathbf{g} = c\,\sigma(f_g(\mathbf{z}^{sem}, \mathbf{h}^{num}))$ and interpolates between the two predictions: $\mathbf{y}_f = (1-\mathbf{g}) \odot \hat{\mathbf{y}} + \mathbf{g} \odot \mathbf{y}_{\text{text}}$. The attention-based variant injects the semantic representation into numerical patch tokens via cross-attention, $\tilde{\mathbf{H}} = \text{CrossAttn}(\mathbf{Q} = \hat{\mathbf{H}}, \mathbf{K} = \mathbf{V} = \mathbf{W}_c\mathbf{z}^{sem})$, and then feeds $\tilde{\mathbf{H}}$ into the same output head. Both variants use the fusion hidden dimension of 128 and dropout of 0.1 reported in Section 5.3. They are evaluated only as ablation references in Tables 3–5, with residual refinement retained as the default strategy.

Table 3. Ablation results on FIT under the half-year setting.

Group	Variant	MAE	MAPE (%)
Num.	Num. (w/o Meta)	0.1165	43.60
Num.	Num. (w/Meta)	0.0841	30.26
Fus.	Res (Random)	0.0859	30.23
Fus.	Res (Ours, Str.)	0.0713	25.23
Fus.	Direct (Weight, Str.)	0.0715	25.42
Fus.	Cross (Cross, Str.)	0.0722	26.25
Imp. (%)	Fus. vs. Num. (w/Meta)	15.22	16.62

Table 4. Ablation results on FIT under the one-year setting.

Group	Variant	MAE	MAPE (%)
Num.	Num. (w/o Meta)	0.1102	41.97
Num.	Num. (w/Meta)	0.0844	31.22
Fus.	Res (Random)	0.0840	30.96
Fus.	Ours (Res., Str.)	0.0704	25.38
Fus.	Direct (Weight, Str.)	0.0716	26.05
Fus.	Cross (Cross, Str.)	0.0724	26.36
Imp. (%)	Fus. vs. Num. (w/Meta)	16.59	18.71

Table 5. Ablation results on GeoStyle under the half-year setting.

Group	Variant	MAE	MAPE (%)
Num.	Num. (w/o Meta)	0.0143	16.53
Num.	Num. (w/Meta)	0.0120	15.30
Fus.	Res (Random)	0.0125	14.69
Fus.	Ours (Res., Str.)	0.0108	13.04
Fus.	Direct (Weight, Str.)	0.0119	13.41
Fus.	Direct (Cross, Str.)	0.0116	14.02
Imp. (%)	Fus. vs. Num. (w/Meta)	10.00	14.77

5. Experiments

This section presents the experimental results of the proposed DDFTF framework on the fashion trend forecasting task. Experiments are conducted on the FIT [18] and GeoStyle [17] datasets, where we report overall performance comparisons, ablation studies on key components, performance under different forecasting settings, and qualitative visualizations of trend prediction cases.

5.1. Experimental Setting

Following prior work [18], we evaluate our method on two public datasets. Both model fashion element popularity across user groups, but GeoStyle contains limited city-level

attributes with strong periodicity, while FIT includes richer user attributes (city, gender, age) and more diverse categories, leading to more complex trend dynamics.

We use the provided preprocessed files and split them into train/validation/test sets (7:1:2) using contiguous indices without shuffling, preserving the ordering of the preprocessed instances. Each (group, element) instance is treated as an independent prediction unit with no temporal causal dependency across instances; the historical-to-future ordering is strictly preserved within each instance. Instance normalization further removes absolute magnitude discrepancies across instances, so the contiguous index split is equivalent to a random instance-level assignment in this setting. All experiments use a fixed random seed of 2026 to ensure reproducibility.

For evaluation, we adopt **MAE** and **MAPE** as the primary metrics to measure forecasting performance. The evaluation pipeline also computes MSE, RMSE, and WAPE; we retain MAE and MAPE in the main tables to remain directly comparable with prior fashion forecasting studies. Directional, peak-timing, and group-stratified analyses are left for future evaluation. Table 6 reports the statistical comparison of the datasets used in experiments with the proposed method.

Table 6. Statistical Comparison Between FIT and GeoStyle Datasets. **#FE** denotes the number of fashion elements, and **#Seq** denotes the number of sequences.

Dataset	City	Gender	Age	#FE	Time	#Seq
GeoStyle	44	N/A	N/A	46	3 years	2024
FIT	14	2	4	173	5 years	8808

5.2. Baselines

To comprehensively evaluate our proposed method on the fashion trend forecasting task, the experiments consider a diverse set of representative baseline approaches:

- **Mean** and **Last**: statistical forecasting methods include simple history-based baselines.
- **AR** and **VAR**: time-series models based on linear autoregressive principles, which are used to capture local temporal dependencies.
- **Linear** and **Cyclic** [17]: linear and cyclic models, which fit historical sequences by explicitly modeling trend and periodic components.
- **GeoStyle** [17]: combines linear trends with cyclic components and is a commonly used parametric forecasting model on the GeoStyle dataset.
- **KERN** [18]: a knowledge-enhanced RNN model for time-series forecasting.
- **REAR** [26]: extends KERN with a relation-enhanced attention recurrent network modeling interactions among fashion elements and user groups for cross-trend dependencies.
- **DKERNQL** [41]: extends KERN with dynamic knowledge representations and quantile-loss optimization for fashion trend forecasting.
- **MAFT** [42] employs a Multivariable Fusion Attention mechanism to capture interactions among heterogeneous variables and effectively integrate multivariate features.
- **TS-BiLSTM** [43]: a BiLSTM-based model integrating attribute extraction, user-aware aggregation, and co-occurrence modeling.
- **DualSG** [27]: leverages LLMs as semantic guides by introducing explicit textual descriptions and a caption-guided fusion module for forecasting.

The baseline set emphasizes representative statistical, recurrent, knowledge-enhanced, and caption-guided fashion forecasters. Modern general-purpose forecasters are discussed in Related Work, while their full reimplementation under the same fashion-specific protocol is left for future study.

To ensure a consistent comparison, the proposed approach and the reported baselines use the same dataset definitions and forecasting horizons. Baseline values remain identified by their published or plot-extracted provenance in Table 7; we do not claim a complete equal-budget reimplementation of every external baseline.

Table 7. Overall forecasting performance of different methods on the GeoStyle and FIT datasets. Unless otherwise noted, baseline values are taken from the corresponding published reports under their stated settings. "*" indicates values extracted from performance plots reported in the original TS-BiLSTM paper, while "—" indicates that the corresponding value was not available in the cited source.

Method	GeoStyle (Half-Year)		FIT (Half-Year)		FIT (One-Year)	
	MAE	MAPE(%)	MAE	MAPE (%)	MAE	MAPE (%)
Mean	0.029	25.79	0.132	65.31	0.135	63.21
Last	0.023	21.04	0.125	46.45	0.147	54.04
AR	0.021	20.69	0.114	54.36	0.119	51.96
VAR	0.015	17.95	0.157	62.97	0.126	47.35
Linear	0.037	24.40	0.112	43.30	0.133	45.89
Cyclic	0.017	16.64	0.129	49.92	0.143	51.66
GeoStyle	0.015	16.03	0.136	52.40	0.149	53.14
KERN	0.013	14.24	0.083	30.02	0.094	33.45
REAR	0.013	14.36	0.086	29.45	0.095	32.26
DKERNQL	<u>0.012</u>	14.78	0.082	29.87	0.095	32.92
MAFT	0.013	<u>13.97</u>	—	—	<u>0.082</u>	<u>29.23</u>
TS-BiLSTM	0.013	14.68	<u>0.079</u> *	<u>28.66</u> *	0.084 *	29.40 *
DualSG	0.014	15.66	0.082	30.00	<u>0.082</u>	30.33
DDFTF	**0.011**	**13.04**	**0.071**	**25.23**	**0.070**	**25.38**
Impro. (%)	8.33	6.65	10.13	13.59	14.63	15.17

Note: Bold values indicate the best result in each column. Underline indicates the second-best result in the corresponding column.

5.3. Implementation Details

All models are implemented in a unified framework using PyTorch 2.2.2. For FIT experiments, we use a batch size of 200 and an initial learning rate of 1×10^{-3}. For FIT fusion, we further adopt split optimization with $\mathrm{lr}_{num} = 1 \times 10^{-4}$ and $\mathrm{lr}_{text} = 5 \times 10^{-4}$, together with early stopping based on validation performance. The GeoStyle dataset follows the same optimizer, early stopping strategy, and architecture hyperparameters as FIT. For the numerical stream, the hidden dimension is set to 64, the feed-forward dimension to 256, the number of attention heads to 4, and the encoder depth to 2. For the fusion module, we set the text hidden dimension to 128, num_feat_dim to 64, fusion_hidden to 128, and fusion_dropout to 0.1. Training uses MAE loss by default (MSE is also supported as an alternative), with the numerical and textual streams optimized jointly under the same forecasting objective. The main DDFTF configuration uses joint optimization; comparison runs may initialize the numerical branch from its numerical-only counterpart, but the fusion training objective remains joint. Caption embeddings are produced offline by a frozen pre-trained GPT-2 base encoder: mean pooling yields a 768-dimensional vector, which is cached and mapped to a 128-dimensional semantic representation by the trainable projection layer. Thus, the language model itself is not invoked or fine-tuned during forecasting optimization. All experiments use random seed 2026. Fusion experiments use the same batch-size and validation-based early-stopping protocol as the numerical baselines; split optimization assigns separate learning rates to the numerical and semantic/fusion parameter groups. Frozen language-model parameters are excluded from trainable-parameter accounting, and caption generation is a one-time offline preprocessing step rather than a per-epoch cost.

Caption descriptions are generated by the standard version of DeepSeek-R1 [44], and all experiments are conducted on an Nvidia GeForce RTX 4090 GPU with 24GB memory.

5.4. Overall Performance

Table 7 presents the overall forecasting results of the proposed method and the compared baselines. MAE and MAPE are used as main evaluation metrics, with lower values indicating better predictive accuracy. Several observations can be drawn from the results:

(1) Our method demonstrates strong and competitive performance across the evaluated datasets and forecasting settings. On GeoStyle (half-year), it attains 0.011 MAE and 13.04 MAPE, representing an 8.33% and 6.65% relative improvement over the second-best results. The gains are more pronounced on FIT, particularly in the one-year setting, where MAE and MAPE are reduced by 14.63% and 15.17%, respectively. We also observe improvements over DualSG, supporting the effectiveness of metadata-aware modeling and structured semantic guidance within the evaluated settings.

(2) On the FIT dataset, knowledge-based methods consistently outperform traditional statistical and parametric models. In particular, KERN-based approaches (KERN, REAR, DKERNQL), along with MAFT and TS-BiLSTM, achieve substantially lower errors. Among them, TS-BiLSTM is the strongest baseline under the half-year setting, while MAFT performs best in the one-year setting. Despite these strong baselines, DDFTF achieves the best performance across both FIT horizons and remains competitive on GeoStyle. These results indicate that DDFTF provides more effective modeling of complex trend dynamics than existing knowledge-enhanced and recurrent approaches.

(3) On GeoStyle, DDFTF achieves an 8.33% reduction in MAE compared to the strongest baseline, while its MAPE remains nearly identical to the best reported result. The marginal difference in MAPE may be attributed to the characteristics of the GeoStyle dataset, which features relatively sparse metadata and pronounced periodic patterns. In such settings, classical trend components and existing baselines already exhibit competitive performance, leaving limited room for further improvement.

5.5. Ablation Analysis

We conduct ablation studies on the FIT and GeoStyle datasets to validate the key components of our model and training strategies. Under consistent experimental settings, we systematically examine the effects of input features, textual representations, and fusion mechanisms.

In Tables 3–5, “Num.” denotes the with numerical stream only, while “Fus.” refers to models integrating numerical and textual inputs. “Res.”, “Weight”, and “Cross” correspond to residual, weighted, and attention-based fusion, respectively. “Random” indicates randomly generated text, whereas “Str.” denotes our structured LLM-generated semantic descriptions defined in Appendix A.

On the FIT dataset (Table 3), introducing metadata significantly improves numerical modeling performance, highlighting the importance of contextual information. Relative to the metadata-aware numerical backbone, structured residual fusion further reduces MAE by 15.22% in the half-year setting. Replacing structured text with random inputs leads to higher error, supporting the contribution of informative semantic guidance beyond the numerical backbone. Differences among fusion strategies are relatively small, with residual fusion showing a slight but consistent advantage.

Under the more challenging one-year setting (Table 4), numerical models exhibit increased errors even with metadata, reflecting the difficulty of long-term forecasting. Relative to the metadata-aware numerical backbone, structured residual fusion reduces MAE

by 16.59%. Notably, random text yields performance comparable to the metadata-enhanced numerical model, suggesting that the effectiveness of fusion depends on semantic quality.

On the GeoStyle dataset (Table 5), metadata provides relatively modest gains, reflecting the more localized nature of the task. Relative to the metadata-aware numerical backbone, structured residual fusion reduces MAE by 10.00% and MAPE by 14.77%. While random text offers only marginal benefits, the residual fusion design maintains competitive performance under the evaluated setting.

Overall, the ablation studies demonstrate that (1) metadata fusion improves numerical modeling, (2) structured semantic descriptions outperform random text within the same fusion architecture, and (3) residual fusion provides the most robust integration mechanism. These results support the role of structured semantic content in the observed improvements. A fully factorized comparison of individual text fields, encoder choices, and capacity-matched controls is outside the scope of the present study.

5.6. Case Study

Figure 3 shows representative fashion element trend forecasts on the FIT dataset for two user groups, each with two selected fashion elements. The blue curves represent the ground-truth trends, while the purple curves indicate the model predictions. For each subplot, the top and bottom rows correspond to the half-year and one-year forecasting settings, respectively. Below each subplot, the corresponding textual description is provided, which serves as part of the model's textual input stream to provide semantic guidance for trend prediction. These examples are illustrative rather than an exhaustive failure analysis or user study.

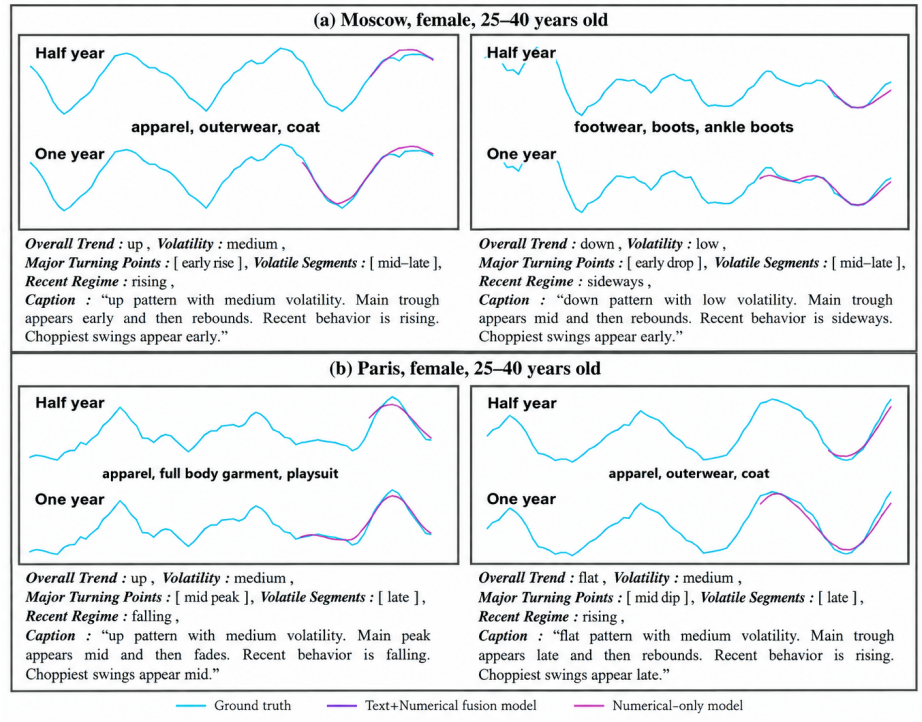


Figure 3. Examples of popularity trend forecasting for different user groups and fashion elements under half-year and one-year settings. The blue curves denote the ground-truth values, while the purple curves represent the predicted results. The text below each subplot is the structured semantic description used by the textual stream.

Overall, the model closely tracks the general trajectory of the ground-truth trends under both forecasting horizons and visually aligns with many peaks and turning points. Although prediction errors increase in the one-year setting, the discrepancies mainly appear

as amplitude shifts, while the overall trend directions remain largely consistent with the ground truth. Moreover, different user groups exhibit distinct fluctuation patterns and temporal rhythms even along similar attribute dimensions. By integrating user attributes with textual semantic information, the model generates predictions that align with the characteristics of each user group, demonstrating its ability to capture conditional trend variations across user populations.

5.7. Fashion Trend Analysis

Figure 4 compares trend forecasting results for two representative fashion elements across user groups, highlighting the gains from textual semantic fusion. Examples are selected from the held-out test set and are not used for training or hyperparameter tuning. Blue denotes ground truth, red denotes the numerical-only model, and purple denotes the fusion model. Columns: Sydney, Paris, and London females aged 25–40.

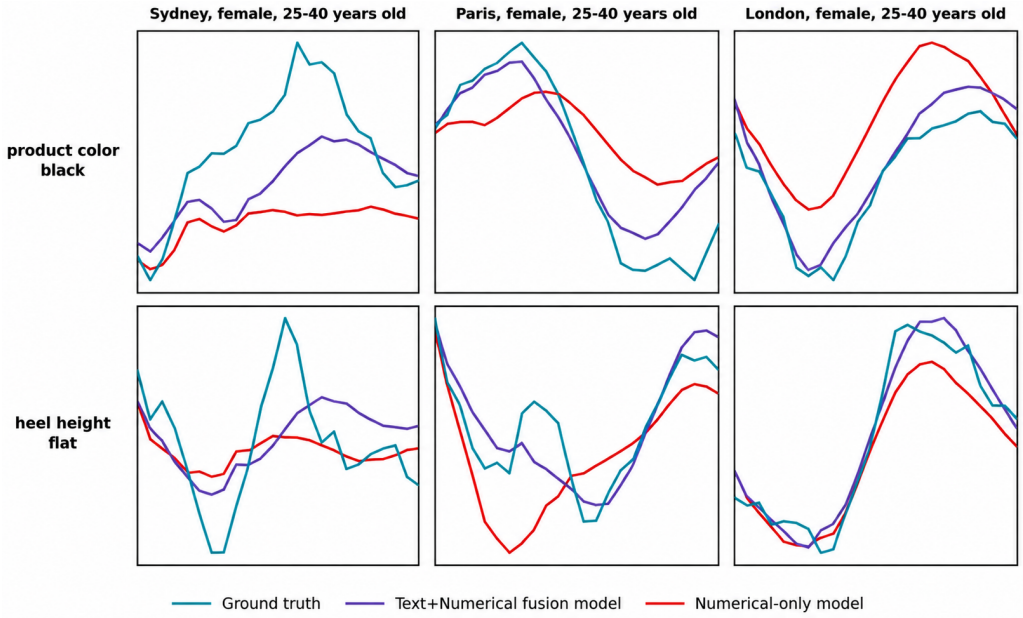


Figure 4. Trendforecast comparison between the numerical-only model and the fusion model. Blue denotes ground truth, red denotes the numerical-only prediction, and purple denotes the prediction with textual semantic fusion.

It can be observed that the pure numerical model produces relatively smoother predictions and tends to underreact to pronounced fluctuations and turning points in the ground-truth trends. For example, it fails to adequately capture the stage-wise rise followed by a decline in the "black" element in Paris, and the abrupt peak toward the end of the sequence in London. In contrast, the fusion model achieves better alignment with key peaks, valleys, and overall trend directions in most subplots, exhibiting slopes and local fluctuation amplitudes that are closer to the ground truth. These observations indicate that incorporating textual semantic information enhances the model's ability to capture fine-grained, group-specific trend variations of fashion elements.

Based on batch forecasting results for multiple types of fashion elements (including categories, colors, patterns, styles, and fine-grained attributes), trend reports tailored to specific user groups can be further generated. These reports summarize key rising elements, stage-wise peaks, and emerging style combinations over a future time horizon, providing practical support for assortment planning, design decisions, and marketing strategies. As illustrated in Figure 5, we present a representative visualization report for reference. Systematic error stratification by volatility and broader user evaluation are left for future work.

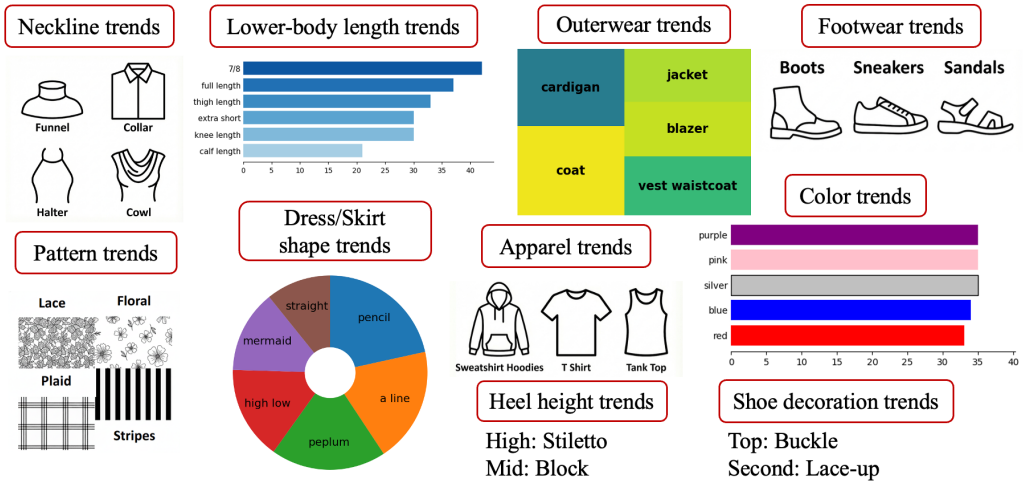


Figure 5. Example of group-specific fashion trend reports generated from predictions across multiple fashion elements. The report summarizes rising items, trend peaks, and style combinations to support fashion design, assortment planning, and marketing decisions.

6. Conclusions

In this paper, we propose a dual-stream framework for fashion trend forecasting that integrates numerical temporal modeling with textual semantic understanding. By jointly modeling fine-grained popularity dynamics and high-level semantic representations of historical trends, the proposed approach captures both statistical patterns and trend-level semantic cues. Experiments on two real-world fashion datasets show competitive performance across the evaluated settings, with the clearest gains observed on the FIT long-horizon tasks. Qualitative analyses further illustrate how the semantic stream complements numerical forecasting by highlighting trend-level patterns.

While the proposed framework achieves strong performance, there is still room for improvement. The current semantic stream relies on LLM-generated textual representations, which introduce additional computational overhead compared with purely numerical forecasting models. The additional cost mainly comes from semantic description generation and text encoding. In practical deployment, these descriptions can be generated offline, processed in batches, and cached when the historical window and metadata remain unchanged, thereby reducing repeated LLM inference cost. For large-scale deployment, further optimization of training time, memory usage, and inference efficiency is still needed.

Although evaluated on fashion datasets, the proposed dual-stream design may be adapted to other forecasting tasks that combine temporal signals with semantic information. Its robustness under domain shifts, noisy data, non-stationary distributions, and extreme events remains an important direction for future validation.

Building upon these observations, future work will focus on lightweight semantic modeling, multimodal trend forecasting, and adaptive prompt generation. We will also explore transfer learning, domain adaptation, uncertainty estimation, and hybrid forecasting architectures to improve robustness and deployability.

Author Contributions: Conceptualization, S.F. and J.C.; methodology, S.F. and J.C.; software, S.F. and J.Z.; validation, S.F., J.C. and J.Z.; formal analysis, S.F.; investigation, T.Z.; resources, X.J.; data curation, S.F.; writing—original draft preparation, S.F. and J.C.; writing—review and editing, X.J. and G.C.; visualization, J.Z.; supervision, X.J. and H.Y.; project administration, X.P.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Science and Technology Program of Zhejiang Province (2025C01087), the Major Scientific and Technological Projects of CNTC (110202401031(SZ-05)), and the Zhejiang Key Laboratory Project (2024E10001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in GeoStyle [17] and FIT [18] at https://drive.google.com/file/d/1OtwOoHYMuLKy_Yjk-_rgjJL5uMWWhPn8/view?pli=1 (accessed on 24 August 2026).

Acknowledgments: The authors acknowledge and appreciate the support received from the Zhejiang Sci-Tech University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DDFTF	Dual-stream Driven Fashion Trend Forecasting
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
DualSG	Dual-Stream Explicit Semantic-Guided Framework
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

Appendix A. Prompt Template for Semantic Description Generation

For reproducibility, we provide the prompt template used to generate the default structured semantic descriptions in DDFTF. The prompt only uses the observed historical sequence and the associated metadata, without access to the future forecasting horizon. The generated output is constrained to a structured JSON object that summarizes overall trend, volatility, major turning points, volatile segments, recent regime, and a concise caption.

System Prompt

You are a fashion trend analyst. Your task is to summarize the historical popularity evolution of a fashion element within a specific user group. Use only the provided historical sequence and metadata. Do not predict future values or use external knowledge. Return a faithful structured description of the observed historical trend.

User Prompt

You are given a historical fashion popularity sequence and its metadata.
Metadata.

- `fashion_element`: {element}
- `city`: {city}
- `gender`: {gender}
- `age_group`: {age}
- `historical_window_length`: {T}
- `time_granularity`: weekly

Historical sequence.

$$\{x_1, x_2, \ldots, x_T\}$$

Analyze only the historical window above and return a valid JSON object with the following fields:

- `overall_trend`: dominant long-range pattern of the historical sequence.
- `volatility`: overall fluctuation intensity.
- `major_turning_points`: at most three salient turning points in the history.
- `volatile_segments`: at most three intervals with relatively strong fluctuations.
- `recent_regime`: summary of the final part of the historical sequence.
- `caption`: a concise natural-language summary in 2–4 sentences.

Generation constraints.

- Use only the provided historical sequence and metadata.
- Do not forecast the future horizon.
- Do not copy large amounts of raw numerical values.
- If some metadata are unavailable, use "unknown" rather than inventing information.
- Return valid JSON only.

Expected output schema.

The returned JSON object contains the six fields above, where `major_turning_points` and `volatile_segments` are list-valued fields, and the remaining fields are textual summaries.

Example Output

- `overall_trend`: Gradual increase over the first half, followed by a plateau with mild fluctuations in the second half.
- `volatility`: Moderate overall, with concentrated spikes around Weeks 20–25.
- `major_turning_points`:
  - Week 18: Starts accelerating after a relatively flat period.
  - Week 26: Reaches a peak and begins a declining phase.
  - Week 40: Stabilizes at a new baseline.
- `volatile_segments`:
  - Weeks 18–28: Rapid rise followed by an initial correction.
  - Weeks 35–42: Second volatility cluster.
- `recent_regime`: Stable with a slight upward drift during the last eight weeks, remaining around 0.22–0.25.
- `caption`: The popularity initially increases and peaks around the middle of the historical period, followed by stabilization at a moderately higher level. The recent regime indicates sustained interest with relatively low volatility.

Appendix A.1. Caption Generation Decoding Settings

For reproducibility, all caption descriptions are generated by DeepSeek-R1 with the following decoding configuration: temperature 0.7, top-p 0.9, maximum output tokens 2048, and a fixed random seed of 2026. Batches that fail structural validation are regenerated when necessary, with a maximum of three attempts. The resulting captions are then encoded by a pre-trained text encoder (see Section 5.3) and reused throughout training, so the language model itself is not invoked during optimization.

References

1. Choi, T.M.; Hui, C.L.; Liu, N.; Ng, S.F.; Yu, Y. Fast fashion sales forecasting with limited data and time. *Decis. Support Syst.* **2014**, *59*, 84–92. [CrossRef]
2. Jin, B.E.; Shin, D.C. The power of 4th industrial revolution in the fashion industry: What, why, and how has the industry changed? *Fash. Text.* **2021**, *8*, 31. [CrossRef]
3. Imtiaz, A.; Pathirana, N.; Saheel, S.; Karunanayaka, K.; Trenado, C. A Review on the Influence of Deep Learning and Generative AI in the Fashion Industry. *J. Future Artif. Intell. Technol.* **2024**, *1*, 201–216. [CrossRef]
4. Choi, T.M.; Cheng, T.E. *Sustainable Fashion Supply Chain Management*; Springer: Cham, Switzerland, 2015.
5. Gornostaeva, G. The development of digital commerce in the fashion industry: The typology of emerging designers in London. *Technol. Forecast. Soc. Change* **2023**, *186*, 122122. [CrossRef]
6. Zahran, A.M. The impact of marketing strategies on the success of the fast fashion industry: A systematic review. *JUMDER J. Bisnis Digit. Dan Ekon. Kreat.* **2025**, *1*, 1–15.
7. Shi, M.; Chussid, C.; Yang, P.; Jia, M.; Dyk Lewis, V.; Cao, W. The exploration of artificial intelligence application in fashion trend forecasting. *Text. Res. J.* **2021**, *91*, 2357–2386. [CrossRef]
8. Abrishami, S.; Shirali, A.; Sharples, N.; Kartal, G.E.; Macintyre, L.; Doustdar, O. Textile recycling and recovery: An eco-friendly perspective on textile and garment industries challenges. *Text. Res. J.* **2024**, *94*, 2815–2834. [CrossRef]
9. Avogaro, A.; Capogrosso, L.; Toaiari, A.; Fummi, F.; Cristani, M. New Fashion Products Performance Forecasting: A Survey on Evolutions, Models and Emerging Trends. *arXiv* **2025**, arXiv:2501.10324.
10. Loureiro, A.L.D.; Miguéis, V.L.; da Silva, L.F.M. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decis. Support Syst.* **2018**, *114*, 81–93. [CrossRef]
11. Singh, P.; Dwivedi, P. Fashion retail sales forecasting using linear and non-linear models. *Int. J. Retail. Distrib. Manag.* **2019**, *47*, 1092–1111.
12. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv* **2019**, arXiv:1905.10437.
13. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
14. Xu, J.; Zhou, Y.; Zhang, L.; Wang, J.; Lefloch, D. Sportswear retailing forecast model based on the combination of multi-layer perceptron and convolutional neural network. *Text. Res. J.* **2021**, *91*, 2980–2994. [CrossRef]
15. Ding, Y.; Lai, Z.; Mok, P.; Chua, T.S. Computational technologies for fashion recommendation: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–45. [CrossRef]
16. Xu, Y.; Wang, W.; Feng, F.; Ma, Y.; Zhang, J.; He, X. Diffusion models for generative outfit recommendation. In Proceedings of the 47th international ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 1350–1359.
17. Mall, U.; Matzen, K.; Hariharan, B.; Snavely, N.; Bala, K. Geostyle: Discovering fashion trends and events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 411–420.
18. Ma, Y.; Ding, Y.; Yang, X.; Liao, L.; Wong, W.K.; Chua, T.S. Knowledge enhanced neural fashion trend forecasting. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 82–90.
19. Cabeza-Ramírez, L.J.; Sánchez-Cañizares, S.M. Exploring the effectiveness of fashion recommendations made by social media influencers in the centennial generation. *Text. Res. J.* **2023**, *93*, 3017–3034. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 2–7 June 2019; pp. 4171–4186.
22. Zhou, T.; Niu, P.; Wang, L.; Liang, J.; Zhang, J. One Fits All: Power General Time Series Analysis by Pretrained LM. In Proceedings of the Advances in Neural Information Processing Systems, Louisiana, LA, USA, 10–16 December 2023; Volume 36.
23. Vittayakorn, S.; Yamaguchi, K.; Berg, A.C.; Berg, T.L. Runway to realway: Visual analysis of fashion. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2015; pp. 951–958.
24. Al-Halah, Z.; Stiefelhagen, R.; Grauman, K. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 388–397.
25. Matzen, K.; Bala, K.; Snavely, N. Streetstyle: Exploring world-wide clothing styles from millions of photos. *arXiv* **2017**, arXiv:1706.01869.
26. Ding, Y.; Ma, Y.; Liao, L.; Wong, W.K.; Chua, T.S. Leveraging multiple relations for fashion trend forecasting based on social media. *IEEE Trans. Multimed.* **2021**, *24*, 2287–2299. [CrossRef]

27. Ding, K.; Fan, F.; Wang, Y.; Jian, R.; Wang, X.; Gong, L.; Jiang, Y.; Luo, C.; Zhan, J. Dualsg: A dual-stream explicit semantic-guided multivariate time series forecasting framework. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; pp. 508–517.
28. Lea, C.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks: A unified approach to action segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 47–54.
29. Wang, W.; Xu, Y.; Shen, J.; Zhu, S.C. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 4271–4280.
30. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
31. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [CrossRef]
32. Wang, Y.; Zhu, J.; Kang, R. Destformer: A transformer based on explicit seasonal–trend decomposition for long-term series forecasting. *Appl. Sci.* **2023**, *13*, 10505. [CrossRef]
33. Su, L.; Zuo, X.; Li, R.; Wang, X.; Zhao, H.; Huang, B. A systematic review for transformer-based long-term series forecasting. *Artif. Intell. Rev.* **2025**, *58*, 80. [CrossRef]
34. Jhamtani, H.; Berg-Kirkpatrick, T. Truth-conditional captions for time series data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 719–733.
35. Trabelsi, M.; Boyd, A.; Cao, J.; Uzunalioglu, H. Time series language model for descriptive caption generation. *arXiv* **2025**, arXiv:2501.01832.
36. Lin, M.; Chen, Z.; Liu, Y.; Zhao, X.; Wu, Z.; Wang, J.; Zhang, X.; Wang, S.; Chen, H. Decoding Time Series with LLMs: A Multi-Agent Framework for Cross-Domain Annotation. In *Findings of the Association for Computational Linguistics: EACL 2026*; Association for Computational Linguistics: Rabat, Morocco, 2026; pp. 6244–6281. [CrossRef]
37. Ahamed, M.A.; Cheng, Q. TimeMachine: A Time Series is Worth 4 Mambas for Long-term Forecasting. In Proceedings of the European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 19–24 October 2024.
38. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
39. Miller, J.A.; Aldosari, M.; Saeed, F.; Barna, N.H.; Rana, S.; Arpinar, I.B.; Liu, N. A survey of deep learning and foundation models for time series forecasting. *axrXiv* **2024**, arXiv:2401.13912. [CrossRef]
40. Ismail, M.; Wang, H. AI for predictive analytics in textile quality assurance: A comprehensive review. V *Text. Res. J.* **2020**, *90*, 1183–1195.
41. Rooholamini, F.; Azmi, R.; Khademhossein, M.; Zarvani, M. Dynamic Knowledge Enhanced Neural Fashion Trend Forecasting with Quantile Loss. In *Proceedings of the 2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*; IEEE: Piscataway, NJ, USA, 2023; pp. 601–606.
42. Chen, J.; Zhao, Y.; Zhong, S.; Hu, X. Fashion Trend Forecasting Based on Multivariate Attention Fusion. In *Proceedings of the International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2023, pp. 68–81.
43. Huang, J.; Zhang, X. Application of Fashion Element Trend Prediction Model Integrating AM and EfficientNet-b7 Models in Art Design. *Scalable Comput. Pract. Exp.* **2024**, *25*, 2756–2769. [CrossRef]
44. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.