Information Sciences 753 (2026) 123630

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

GeCo: Towards effective GAN-based fashion compatibility modeling and retrieval

Matteo Attimonelli [a, b,*], Claudio Pomo [a], Dietmar Jannach [c], Tommaso Di Noia [a]

[a] Politecnico Di Bari, Via Edoardo Orabona, 4, Bari, 70125, Italy
[b] Sapienza University of Rome, Via Ariosto, 25, Rome, 00185, Italy
[c] University of Klagenfurt, Universitätsstraße, 65–67, Klagenfurt, 9020, Austria

HIGHLIGHTS

- **GeCo**: generative fashion compatibility model and complementary item retrieval.
- GeCo uses a two-stage GAN under realistic hardware limits.
- Human studies find its garments realistic and stylistically coherent.
- FashionTaobaoTB and benchmarking codebase have been released.

ARTICLE INFO

Keywords:
Generative compatibility modeling
Fashion image retrieval
Fashion image synthesis

ABSTRACT

Visual compatibility modeling is central to modern fashion recommendation systems. A key task is *complementary item retrieval*, where the goal is to identify a garment that harmonizes with a reference item, such as retrieving a compatible bottom for a given top. Recent generative approaches synthesize candidate garments to guide retrieval, but they either treat generation as an auxiliary signal or rely on computationally demanding architectures, limiting their practicality in large-scale deployments.

In this work, we introduce GeCo, a generative–compositional framework that couples image synthesis and retrieval within an effective, two-stage design. In the first stage, a conditional GAN generates visually coherent bottom templates from top images; in the second stage, the generated template and top form a composed visual query for compatibility-based retrieval. This decoupled design avoids the heavy optimization pipelines used in prior generative approaches, resulting in stable training and low computational cost, while allowing the generated images to guide compatibility modeling.

Experiments on three benchmarks, including the new *FashionTaobaoTB* dataset released with this work, show that GeCo offers competitive retrieval accuracy with a low memory footprint. Human evaluations further indicate that its generated items are perceived as realistic and stylistically compatible, supporting its suitability for practical, resource-constrained fashion recommendation scenarios.

* Corresponding author at: Politecnico Di Bari, Via Edoardo Orabona, 4, Bari, 70125, Italy.
Email address: matteo.attimonelli@poliba.it (M. Attimonelli).

https://doi.org/10.1016/j.ins.2026.123630
Received 4 December 2025; Received in revised form 11 May 2026; Accepted 11 May 2026
Available online 15 May 2026

1. Introduction

The rapid growth of e-commerce and the demand for personalized shopping experiences have driven major advances in visual fashion retrieval systems [1]. A central task in this domain is *complementary item retrieval*, where the goal is to identify an item that completes an outfit given a reference garment [2]. Within this paradigm, *pairwise compatibility learning* remains the foundational problem: a model learns to predict a complementary item from a single seed piece, and this capability can then be extended to full-outfit composition. The most widely studied instance is *top–bottom retrieval*, which aims to retrieve a bottom garment that visually harmonizes with a given top [3]. This task is crucial for incremental outfit construction and serves as a canonical benchmark for studying visual compatibility under practical constraints.

Despite extensive research efforts over the past years [1], complementary retrieval remains challenging, particularly in realistic deployment scenarios: compatibility judgments are subjective and context-dependent; fashion catalogs are large and diverse; and textual metadata is often noisy, incomplete, or multilingual. As a result, many real-world systems must operate using visual signals alone, without relying on clean attribute annotations or behavioral histories [4]. This setting imposes strict practical constraints: models must reliably assess compatibility using only images, while meeting tight computational budgets and low-latency requirements needed in production-level scenarios.

Generative compatibility models such as MGCM and its variants [5] accomplish these tasks by synthesizing hypothetical complementary garments to guide retrieval, drawing inspiration from *image-to-image (I2I) translation* [6]. In principle, a generative formulation allows the model to learn richer relationships than direct embedding-based ranking [5]. However, unlike classical I2I tasks where source and target domains share geometry (e.g., horses → zebras [7]), tops and bottoms belong to heterogeneous domains, which makes direct mappings unstable and prone to mode collapse [8]. Moreover, early generative models often used the synthesized images only as auxiliary regularizers rather than as core drivers of compatibility reasoning, which limited their impact on compatibility modeling and retrieval performance. Subsequent works attempted to address these issues in two directions. Attribute-conditioned methods such as Attribute-GAN [9] inject semantic or textual signals to ease cross-domain translation, but such metadata is rarely reliable or consistently available in large-scale catalogs. Alternatively, recent methods including COutfitGAN [10], FCBoost-Net [11], and BC-GAN [12], retrofit StyleGAN backbones with GAN inversion to facilitate image conditioning without relying on I2I architectures, and introduce multiple compatibility-driven losses to guide synthesis. While these approaches enhance visual fidelity, they also introduce notable computational overhead and large memory requirements. Additionally, the complex multi-objective optimization pipelines are challenging to align with the practical constraints mentioned above, which limits the deployability of these methods in real-world scenarios. Compounding these difficulties, the literature exhibits limited reproducibility: many works rely on non-public preprocessing pipelines or data-subset selection strategies and do not release code, making comparison across methods difficult and slowing progress in the field.

In this work, we revisit the generative paradigm for fashion compatibility from a deployment-oriented perspective. We focus on the realistic scenario in which only visual data are available, metadata are unreliable, and hardware budgets are limited. We note that existing generative approaches either (i) use synthetic items only as auxiliary signals, or (ii) rely on computationally heavy GAN inversion and multi-loss adversarial training. We instead hypothesize that placing the generated item *at the core* of the retrieval process and improving its visual quality can directly enhance compatibility modeling while remaining scalable.

To this end, we propose **GeCo**, an effective, visual-only *generate–then–retrieve* architecture that couples I2I garment synthesis with compatibility-aware retrieval. GeCo employs an enhanced GAN module to produce diverse, stylistically coherent bottom images from a given top, referred to as templates. These generated templates are then combined with the top to form a composed visual query used to rank candidate items. The overall architecture is shown in Fig. 1. This decoupling yields stable training, reduced memory usage (4 GB vs. 9 GB for StyleGAN-based models), and flexible component replacement. While diffusion [13] and autoregressive [14] models offer state-of-the-art generative fidelity, they remain impractical for interactive recommendation due to the *generative trilemma*: the inherent trade-off between quality, diversity, and efficiency [15]. Even accelerated or distilled variants require multiple denoising steps and large amounts of VRAM, making them unsuitable for industrial scenarios.

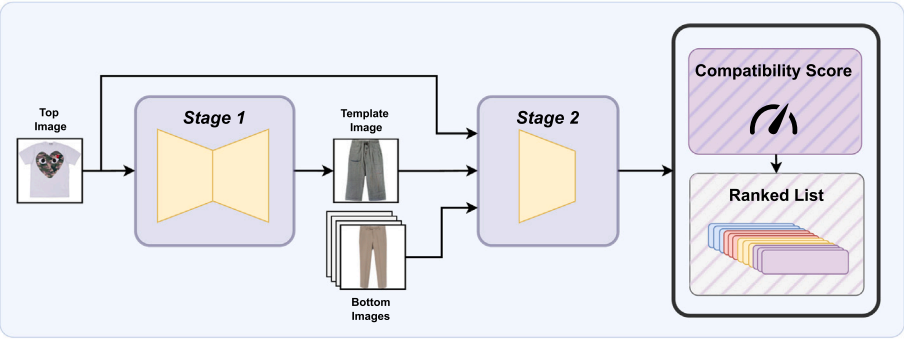

Fig. 1. Overview of the proposed two-stage architecture. GeCo first generates a bottom template image from a given top and then evaluates compatibility between the top, generated template, and candidate bottoms.

We show that this simple yet principled design (i) achieves competitive compatibility modeling and retrieval under modest hardware budgets, (ii) produces items that users perceive as realistic and stylistically coherent, and (iii) provides a reproducible visual-only pipeline that is well aligned with practical deployment scenarios. To support transparent evaluation, we also release a complete codebase and **FashionTaobaoTB**, a benchmark of over 80 000 top–bottom pairs with unified splits and processing scripts.[1] The remainder of this paper reviews related work, presents the GeCo architecture and methodology, reports experimental results, and concludes with discussion and future directions.

2. Related work

Fashion recommendation has gained increasing attention for its impact on e-commerce, driving research on visual retrieval and compatibility learning [1]. This section reviews the most relevant studies in two complementary areas: (*i*) fashion item retrieval and recommendation, and (*ii*) generative models for compatibility modeling and retrieval.

Fashion item retrieval and recommendation. The fashion recommendation landscape encompasses several related tasks, including *fashion item recommendation*, *outfit recommendation*, and *complementary item retrieval* [1]. Item-level recommendation [16] focuses on predicting individual garments from user behavior or contextual cues, while outfit recommendation [17] extends this to constructing coherent ensembles of multiple items. A widely studied instance is *top–bottom retrieval* [3], where a system retrieves a bottom garment compatible with a given top.

Early discriminative methods such as BPR–DAE [3] combine visual and textual features using the Bayesian Personalized Ranking (BPR) loss, while GP–BPR [18] adds a personalization signal, and NOR [19] introduces recurrent reasoning for sequential outfit prediction. These approaches are effective but are inherently limited to ranking existing catalog items and cannot generate unseen combinations or model latent compatibility structures beyond observed data.

Generative models for compatibility modeling and retrieval. Generative compatibility models aim to overcome the limitations of purely discriminative ranking by synthesizing complementary garments as part of the learning process. Early works such as DVBPR [20], FARM [21], and MGCM [5] generate hypothetical items to guide retrieval. MGCM introduces a template mechanism that evaluates both item–item and item–template compatibilities, while [22] adopts a CycleGAN [7] formulation for bidirectional style translation. Although these models capture richer relationships than simple embedding-based methods, they generally treat generated items as auxiliary signals within tightly coupled end-to-end formulations, and still face instability due to the heterogeneous mapping between tops and bottoms [23]. In contrast, our approach incorporates the generated item directly into the query and adopts a decoupled two-stage design, allowing generation and retrieval to be optimized independently and enabling the generated template to explicitly guide retrieval. Later approaches pursue two strategies to mitigate these issues. Attribute-conditioned models such as Attribute–GAN [9] inject textual or semantic features to bridge domain gaps, but they depend on reliable metadata, which is rarely available in practice. StyleGAN-based architectures including BC–GAN [12], COutfitGAN [10], and FCBoost–Net [11] employ latent inversion and auxiliary compatibility losses to integrate generated images into retrieval. While these methods improve fidelity, they require extensive GPU resources and multi-objective optimization, limiting scalability and reproducibility.

Recently, diffusion [13] and autoregressive [14] models have achieved unprecedented realism and diversity in image generation, yet their high computational costs and long inference times hinder practical use in large-scale retrieval pipelines [24].

Position of our work. Unlike previous approaches, GeCo combines: (i) a conditional GAN capable of synthesizing high-quality garments in a single forward pass under a modest memory budget, and (ii) a decoupled retrieval stage in which the generated item directly participates in query formation rather than acting as an auxiliary regularizer. By avoiding GAN inversion, operating without textual metadata, and remaining effective under strict resource constraints, GeCo provides a practical and fully reproducible framework for visual compatibility modeling.

3. Methodology

This section presents the proposed GeCo framework in detail. Operating in a visual-only setting, GeCo relies exclusively on visual features extracted from item images, without requiring textual metadata or multimodal inputs. The overall workflow follows a two-stage design that separates generative synthesis from compatibility modeling, as summarized below.

Stage 1 **Template Generation:** Given an input top image, GeCo generates a visual template of a compatible bottom garment via image-to-image translation.

Stage 2 **Compatibility Scoring and Retrieval:** The generated template is combined with the input top to form a composed query, which is used to retrieve real compatible bottoms from the catalog via a learned similarity metric.

Unlike previous approaches that entangle generation and retrieval within a single adversarial pipeline, GeCo explicitly decouples these objectives. This design simplifies the optimization process, allowing each component to be tuned independently. Empirically, we observe that this separation improves both visual fidelity and retrieval consistency (see Section 4).

[1] Code and data are available at: https://github.com/sisinflab/GeCo.

GECO employs a conditional generative model based on the cGAN framework [25], adapted to address the instability of traditional GANs [23]. In the first stage, the generator is trained to capture compatibility relationships by synthesizing candidate bottoms aligned with a given top. In the second stage, these synthetic exemplars serve as structured inputs to the retrieval module, which learns to rank real items that best match the composed query. The following subsections motivate this generative choice and describe the architecture and training strategy of each stage.

3.1. Preliminaries

As discussed in the previous sections, our framework builds on GAN-based I2I translation owing to its favorable trade-off between realism, speed, and computational efficiency. We briefly outline the rationale for this design and its relation to alternative generative paradigms.

Choice of generative framework. Recent diffusion [26] and transformer-based [13] models have achieved impressive fidelity and diversity but at high computational cost, reflecting the *generative trilemma*. Diffusion models typically require hundreds of sequential denoising steps during inference (e.g., Stable Diffusion 1.5 demands approximately 10–12 GB of memory and 10–20 s per image), whereas GANs can generate samples in a single forward pass with sub-second latency and a fraction of the memory [27]. Given our goal of scalable and deployable compatibility modeling, GECO adopts a GAN-based formulation that balances fidelity and efficiency.

Improving GAN stability and diversity. Despite their efficiency, GANs are susceptible to training instability [23] and mode collapse [28], issues that become more pronounced in cross-domain generation tasks where source and target domains differ structurally and semantically. To mitigate these issues, GECO extends the cGAN framework [6] by injecting stochastic noise at the bottleneck of the generator. This introduces controlled variation across samples while preserving the paired-translation structure, reducing the tendency of standard I2I models to collapse toward overly deterministic mappings in cross-domain fashion generation and helping the model learn across geometrically different domains. As shown in Section 4, this yields more stable training and improves the realism of the generated templates without increasing computational cost. The overall pipeline remains memory efficient, requiring only 4 GB of GPU memory for all experiments.

3.2. Problem formulation

GECO aims to model the joint distribution of tops $\mathcal{T}$ and bottoms $\mathcal{B}$ so that visual compatibility can be learned and exploited for retrieval. Unlike prior works that treat generation as auxiliary regularization, GECO places the generative process at the center of compatibility reasoning, enabling interpretable and controllable retrieval.

Our approach draws inspiration from the paradigm of *Composed Image Retrieval* (CoIR), in which a query is formed by combining multiple modalities, typically an image and a textual modification, and retrieval is guided by both components [29]. In our setting, we replace the textual signal with a generated bottom template, forming a composed query that emphasizes joint visual semantics rather than isolated features. This perspective generalizes beyond fashion and can apply to any task requiring the retrieval of compatible or complementary visual entities.

Given a dataset of top–bottom pairs, we define the set of positive pairs $\mathcal{P}$ as the observed compatible tuples (t_i, b_j), with $t_i \in \mathcal{T}$ and $b_j \in \mathcal{B}$. In contrast, the set of negative pairs $\mathcal{N}$ consists of unobserved or randomly sampled tuples (t_i, b_k) assumed to be incompatible. Formally, given empirical distributions $P_{\text{data}}(\mathcal{T})$ and $P_{\text{data}}(\mathcal{B})$, the goal is to approximate a conditional distribution $P_{\text{GECO}}(\mathcal{B}, | \mathcal{T})$ that generates semantically consistent bottoms for a given top $t \in \mathcal{T}$. The generator G samples a template $\hat{b}_i = G(t_i)$ from this conditional distribution as in Eq. (1),

$$G : \mathcal{T} \to \mathcal{B}, \quad \hat{b}_i = G(t_i). \tag{1}$$

which is subsequently used in the retrieval phase to form a composed query:

$$f : \mathcal{T} \times \hat{\mathcal{B}} \to \mathbb{R}^d, \quad \mathbf{q}_i = f(t_i, \hat{b}_i), \tag{2}$$

where $\mathbf{q}_i$ represents the joint embedding of the real top and generated bottom in a shared latent space. Each candidate bottom b_j is projected as $\mathbf{v}_j$, and compatibility is evaluated via a similarity function:

$$m : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad m_{ij} = \text{sim}(\mathbf{q}_i, \mathbf{v}_j). \tag{3}$$

This formulation explicitly decouples synthesis from retrieval, reducing adversarial–retrieval interference while providing a controlled mechanism to study how generative representations affect compatibility learning.

3.3. First stage – template generation

The first stage of GECO estimates a conditional generative distribution $P_{\text{GECO}}(\mathcal{B} mid \mathcal{T})$ that approximates the true distribution of compatible bottoms given a top, as formalized in Eq. (1). In essence, this module learns a visual translation function between domains $\mathcal{T}$ and $\mathcal{B}$, enabling the generation of plausible complementary items directly conditioned on a reference image. Unlike prior approaches that rely on GAN inversion or multi-objective optimization, GECO adopts a paired image-to-image translation architecture that removes the need for costly latent-space optimization. This design reduces GPU memory consumption during training, improves training speed, and enhances convergence stability.

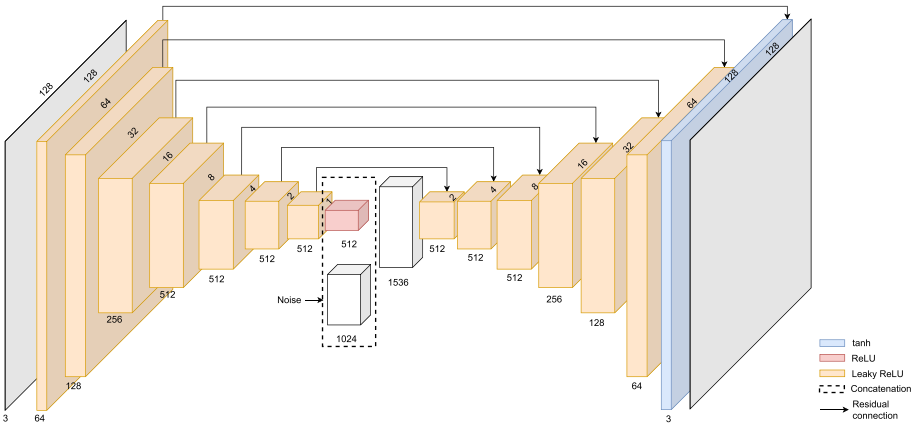

Fig. 2. Architecture of the conditional GAN used in the first stage of GECO. A UNet generator with noise injection at the bottleneck produces bottom templates conditioned on top images, guided by a PatchGAN discriminator.

The generative module extends the Pix2Pix framework [6] and comprises two main components: a U-Net generator and a PatchGAN discriminator. The U-Net autoencoder employs skip connections to retain fine-grained details and alleviate vanishing gradients [30]. PatchGAN is adopted because it evaluates realism at the level of local $N \times N$ patches rather than a single global score, which has been shown to better capture texture, edges [6], and fabric-level details that are crucial in fashion imagery. By focusing on local statistics, PatchGAN provides stronger and more stable gradients than full-image discriminators, enabling the generator to focus on patterns that directly influence perceived realism.

To mitigate mode collapse [23] and encourage diversity, we inject stochastic noise at the bottleneck of the generator. Given the encoder bottleneck feature $h_i^{\text{enc}} \in \mathbb{R}^{d_h}$, we sample a noise vector $z \sim \mathcal{N}(0, I_{d_z})$ and form:

$$\tilde{h}_i = [h_i^{\text{enc}} \parallel z] \in \mathbb{R}^{d_h + d_z},$$

where d_h is the dimension of the UNet bottleneck and d_z is the dimension of the noise vector. The resulting representation $\tilde{h}_i$ is then fed to the decoder. This controlled perturbation introduces variation across generated samples while preserving the efficiency of paired translation. Skip connections propagate low-level structural information to maintain geometric consistency between source and target domains. An overview of the architecture is shown in Fig. 2.

Training optimizes two complementary objectives: (*i*) an adversarial loss, derived from the Jensen–Shannon divergence, that drives the discriminator to distinguish real from synthetic pairs; and (*ii*) an ℓ_1 reconstruction loss enforcing proximity between the generated template and its ground-truth counterpart. Formally, the discriminator $D : \mathcal{T} times \mathcal{B}! \to \{0,1\}^{c \times c}$ outputs patch-level probabilities, with c denoting the patch dimension. Real (compatible) pairs are expected to produce values close to one, whereas generated pairs should produce values close to zero:

$$D(t_i, b_j) = 1^{c \times c}, \qquad D(t_i, \hat{b}_i) = 0^{c \times c}, \quad \forall (t_i, b_j) \in \mathcal{P}. \tag{4}$$

The overall adversarial objective is:

$$\min_G max_D mathcal L_{\text{GAN}} = \mathbb{E}_{t_i, b_j}[\log D(t_i, b_j)] + \mathbb{E}_{t_i, z}[\log(1 - D(t_i, G(t_i, z)))]. \tag{5}$$

The generator and discriminator are updated via the following loss functions:

$$\mathcal{L}_G = -\mathbb{E}_{t_i, z}[\log D(t_i, G(t_i, z))] + \lambda \| G(t_i, z) - b_j \|_1, \tag{6}$$

$$\mathcal{L}_D = -\mathbb{E}_{t_i, b_j}[\log D(t_i, b_j)] - \mathbb{E}_{t_i, z}[\log(1 - D(t_i, G(t_i, z)))]. \tag{7}$$

Here, λ balances fidelity and regularization. After convergence, only the generator G is retained within GECO to synthesize bottom templates; the discriminator serves exclusively as a training stabilizer. Algorithm 1 summarizes this first-stage training procedure.

3.4. Second stage – compatibility scoring and retrieval

The second stage of GECO operationalizes the generated templates for retrieval. Unlike prior generative compatibility models [5], which use generated items only as auxiliary regularizers, GECO explicitly incorporates them into the query representation, with the generator trained in the previous stage and its parameters fixed. This formulation aligns with the CoIR paradigm, where multiple inputs jointly define the retrieval intent. Consequently, the model can exploit the fine-grained visual cues embedded in the generated templates while relying solely on images, without textual supervision.

The objective of this stage is to align the composed query and candidate bottom embeddings so that compatible items exhibit maximal similarity. Given the set $\mathcal{P}$ of positive (compatible) pairs, GECO extracts visual features using a pre-trained visual model. ResNet-18 [31] was selected as the backbone for its efficiency and robust performance; alternative architectures are evaluated in the

Algorithm 1 Stage One Training.

Input: A set of top-bottom pairs $(t_i, b_j) \in \mathcal{P}$, the parameters of the cGAN within GECO $\{\theta_G, \theta_D\}$, learning rate η_1, regularization parameter λ.
Output: optimized $\{\theta_G, \theta_D\}$.
Initialize the parameters $\{\theta_G, \theta_D\}$.
repeat
  Sample a batch of top-bottom pairs $(t_i, b_j) \in \mathcal{P}$.
  for each parameter θ in $\{\theta_G, \theta_D\}$ **do**
    Generate a template image $\hat{b}_i$ following Eq. (1).
    Compute the output patch following Eq. (4).
    Compute $\mathcal{L}_D$ following Eq. (7).
    Update $\theta_D \leftarrow \theta_D - \eta_1 \nabla_{\theta_D} mathcal L_D$.
    Compute $\mathcal{L}_G$ following Eq. (6).
    Update $\theta_G \leftarrow \theta_G - \eta_1 \nabla_{\theta_G} mathcal L_G$.
  end for
until convergence

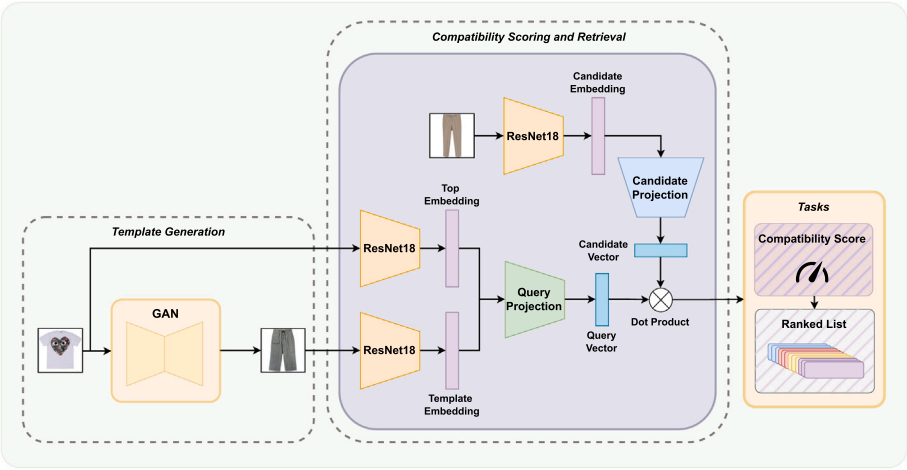

Fig. 3. Overview of the complete GECO pipeline. Stage 1 generates complementary bottoms; Stage 2 composes the query and ranks candidate items.

ablation study (Section 4). As defined in Eq. (2), the top t_i and the corresponding generated template $\hat{b}_i$ are first encoded independently. Their feature vectors are then *concatenated* into a 1024-dimensional representation and projected through a linear layer to obtain the composed query embedding $\mathbf{q}_i \in \mathbb{R}^{128}$. Candidate bottoms b_j are encoded using the same backbone and mapped to the compatibility space through a separate linear projection from $\mathbb{R}^{512}$ to $\mathbb{R}^{128}$, yielding embeddings $\mathbf{v}_j$. Both projection heads consist of a single fully connected layer without additional hidden layers, dropout, or normalization layers. Similarity is then computed using the dot product between $\mathbf{q}_i$ and $\mathbf{v}_j$, as in Eq. (3). Fig. 3 shows the overall framework.

To learn compatibility relations, the loss integrates Bayesian Personalized Ranking (BPR) [32] and Noise-Contrastive Estimation (NCE) [33], which contribute complementary supervision signals. BPR optimizes the relative ordering between a positive and a single negative sample. This directly targets the pairwise ranking objective used in top–bottom retrieval but provides only a local constraint per triplet. In contrast, the NCE component uses in-batch negatives: for each query q_i, all bottoms in the mini-batch except the positive one are treated as negatives. The temperature-scaled softmax encourages the positive similarity m_{ij} to dominate over all in-batch negatives simultaneously. Within the InfoNCE term, embeddings are ℓ_2-normalized before temperature scaling, improving training stability by removing the influence of vector magnitude on similarity. This yields a stronger, more global signal that structures the embedding space by pulling compatible items together while pushing incompatible items apart in aggregate. Combining BPR and NCE therefore enforces both (i) correct local ranking for each triplet and (ii) global discriminability across the batch. As confirmed in our ablation study (Section 4.7), using both losses leads to higher performance than either objective alone.

Formally, given $Qcoloneqq\{(i,j,k) \mid (t_i, b_j) \in \mathcal{P},\ b_k \in \mathcal{B} setminus b_j\}$, the pairwise scores are:

$$m_{ij} = \mathrm{sim}(\mathbf{q}_i, \mathbf{v}_j), \qquad m_{ik} = \mathrm{sim}(\mathbf{q}_i, \mathbf{v}_k), \tag{8}$$

where m_{ij} and m_{ik} denote the scores of the positive and negative pairs, respectively. The overall objective combines both losses and a regularization term:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\mathrm{bpr}} + \beta\, \mathcal{L}_{\mathrm{nce}} + \gamma \|\theta_{\mathrm{GECO}}|_2, \tag{9}$$

Algorithm 2 Stage Two Training.

Input: A set of triples $(t_i, b_j, b_k) \in \mathcal{Q}$, GECO's generator parametrized by θ_G, the trainable GECO parameters with the generator freezed $\theta_{\text{GECO}setminusG} := \theta_{\text{GECO}}setminus\theta_G$, learning rate η_2, hyperparameters $\alpha, \beta, \gamma, \tau$.
Output: optimized $\theta_{\text{GECO}setminusG}$.
Initialize the parameters $\theta_{\text{GECO}setminusG}$.
repeat
  Sample a batch of triples $(t_i, b_j, b_k) \in \mathcal{Q}$.
  for each parameter θ in $\theta_{\text{GECO}setminusG}$ **do**
    Compute $\hat{b}_i$ following Eq. (1).
    Extract query and candidate embeddings following Eq. (2).
    Compute the compatibility scores between the pairs (t_i, b_j) and (t_i, b_k) following Eq. (8).
    Compute $\mathcal{L}_{\text{GECO}}$ following Eq. (9).
    Update $\theta_{\text{GECO}setminusG} \leftarrow \theta_{\text{GECO}setminusG} - \eta_2 \nabla_{\theta_{\text{GECO}setminusG}} mathcal L_{\text{GECO}}$.
  end for
until convergence

with

$$\mathcal{L}_{\text{bpr}} = -\log \sigma(m_{ij} - m_{ik}), \qquad \mathcal{L}_{\text{nce}} = -\frac{1}{\tau} \log \frac{\exp\left(\frac{1}{\tau} m_{ij}\right)}{\sum_{k=1}^{K} exp\left(\frac{1}{\tau} m_{ik}\right)}. \tag{10}$$

Here, α, β, and γ regulate the influence of BPR, NCE, and the regularization term, respectively, and τ denotes the temperature parameter. During inference, the trained generator G (parameterized by $\theta_G \subset \theta_{\text{GECO}}$) synthesizes bottom templates that condition the composed queries for retrieval. The overall second-stage training procedure is shown in Algorithm 2.

4. Experiments

This section presents a comprehensive evaluation of the proposed generative compatibility framework. We benchmark the model against representative baselines across three datasets of varying scales and diversity. The evaluation is structured as follows. We first introduce the datasets and how they were prepared, then describe the adopted evaluation metrics and implementation settings. Finally, we report qualitative and quantitative results, including ablation analyses and memory-requirement analysis, to assess the effects of each design choice on performance and scalability.

4.1. Datasets

We evaluate GECO on three public and curated datasets representative of different levels of scale and visual diversity: **FashionVC** [3], **ExpFashion** [19], and the newly prepared **FashionTaobaoTB**. All datasets consist of top–bottom combinations annotated as compatible. As the notion of compatibility is inherently subjective, we follow prior works in assuming that pairs curated by fashion platforms or user communities correspond to visually coherent and stylistically consistent combinations. Nevertheless, we acknowledge that multiple valid complementary pairs may exist beyond those explicitly observed.

FashionVC. This dataset includes 18,640 outfits, containing 13,716 unique tops and 12,257 unique bottoms, collected from the Polyvore platform. We adopt the official training and testing splits defined in [5] to ensure comparability with the existing literature.

ExpFashion. This dataset comprises 853,991 outfits with 168,682 tops and 117,668 bottoms, also sourced from Polyvore. Following [5], we use both the full dataset and a reduced subset, denoted **ExpReduced**, which contains 17,140 top–bottom pairs sampled randomly to match the size of FashionVC. This smaller version facilitates controlled comparisons under balanced data conditions.

FashionTaobaoTB. To better represent large-scale and visually diverse fashion scenarios, we constructed the **FashionTaobaoTB** dataset from the large-scale Taobao collection introduced by [34], which originally contained over 1.01 M outfits and 583 K items spanning 80 categories (e.g., sweaters, coats, t-shirts, skirts, boots). Since the source data encoded categories using alphanumeric identifiers with no public mapping to item types, direct extraction of garment-level subsets was infeasible. To address this, we systematically inspected items associated with each category identifier to determine the garment type encoded by that category. We retained only identifiers that could be consistently associated with top or bottom garments, while categories that belonged to other item groups were discarded. All selected images were standardized to a uniform format (JPEG, RGB, 512×512 resolution) with a white background. Incomplete, corrupted, or ambiguous entries were removed to ensure data integrity. Particular care was taken to prevent data leakage, an issue frequently observed in fashion datasets where identical pairs appear across splits. Thus, we enforced mutually exclusive training, validation, and test partitions and released them together with the complete preprocessing scripts to enable full reproducibility. The final curated dataset comprises **88,326** verified top–bottom pairs, including approximately **35,639 unique tops** and **28,037 unique bottoms**. To the best of our knowledge, this represents the first adaptation of Taobao data for controlled fashion compatibility and retrieval studies. Table 1 summarizes the main statistics of the three datasets.

Table 1
Summary of dataset characteristics: each of the columns shows the number of instances of tops, bottoms, and positive pairs, respectively.

Dataset	Tops	Bottoms	Positives
FashionVC	13,716	12,257	18,640
ExpReduced	12,454	10,538	18,640
FashionTaobaoTB *(ours)*	35,639	28,037	88,326

4.2. Evaluation metrics

Performance is assessed along two dimensions: (i) **generation quality**, which evaluates the realism and diversity of generated items; and (ii) **compatibility and retrieval accuracy**, which measures how effectively these generations support downstream retrieval. For generation quality, we employ the Fréchet Inception Distance (FID) and the Learned Perceptual Image Patch Similarity (LPIPS), in line with previous works [12].

FID captures the alignment between the distribution of generated and real garments, while LPIPS reflects perceptual similarity at the feature level, which is particularly relevant for assessing fine-grained visual attributes such as texture, color, and style that influence compatibility. Lower FID and LPIPS values indicate higher realism and perceptual similarity, respectively. For retrieval, we follow [5] and report the Area Under the ROC Curve (AUC) for pairwise compatibility prediction and the Mean Reciprocal Rank (MRR) for complementary item retrieval. These metrics directly reflect the objectives of compatibility modeling: AUC evaluates the model's ability to rank compatible pairs above incompatible ones, while MRR measures how effectively the system retrieves the true compatible item at top positions in a realistic recommendation setting. All experiments treat non-matching bottoms in the test set as negatives to ensure a consistent and reproducible protocol.

4.3. Implementation details

All models were implemented in PyTorch and trained on a single NVIDIA A10 GPU (24 GB VRAM) with an AMD EPYC-Rome CPU and 64 GB RAM. We re-implemented baseline methods from their original papers because their source code was unavailable. Hyperparameters were aligned with the configurations reported in their respective publications for FashionVC and ExpReduced. For FashionTaobaoTB, hyperparameters were explored within comparable ranges to ensure fairness.

All images were resized to 128×128 pixels. The generator in GeCo was trained for 200 epochs with a learning rate of 2×10^{-4}, batch size 64, and regularization coefficient $\lambda = 100$. The noise and bottleneck dimensions of the UNet were set to 1024 and 512, respectively. The compatibility stage was trained for 50 epochs using a learning rate of 1×10^{-4} (decayed by 0.1 every eight epochs) and a batch size of 64. Feature extraction relied on a pre-trained ResNet-18 backbone fine-tuned jointly with the projection layers. The retrieval stage uses lightweight projection heads that map features into a 128-dimensional compatibility space. Concretely, the query branch receives the concatenation of the top and generated-template features, while the candidate-bottom branch is projected separately into the same space. Both heads are intentionally shallow, as additional architectural complexity was not necessary in our implementation to obtain stable optimization and competitive ranking performance.

We explored multiple weighting configurations for α, β, and τ and report the best-performing settings for each dataset. GeCo achieved the best compatibility-modeling and retrieval performance with $\alpha = 0.75$, $\beta = 0.75$, and $\tau = 0.5$ on FashionVC, $\alpha = 0.25$, $\beta = 0.5$, and $\tau = 0.1$ on ExpReduced, and $\alpha = 0.25$, $\beta = 0.5$, and $\tau = 0.5$ on FashionTaobaoTB. On all datasets, GeCo obtained the best performance with $\gamma = 0.01$. Overall, more than 2,000 experiments were conducted to ensure robustness across configurations.

4.4. Baselines

We evaluate GeCo against a diverse set of baselines. All models operate under the same *visual-only* setting, using identical training, validation, and test splits, uniform preprocessing, and common evaluation metrics to ensure comparability. **Random** assigns uniform compatibility scores to all top–bottom pairs, providing a lower bound on performance. **Popularity** ranks each bottom by its frequency of occurrence with tops in the training data [3], serving as a frequency-based heuristic baseline. **BPR-DAE** [3] models compatibility as a pairwise preference ranking task using a BPR loss on visual embeddings. Originally multimodal, it is adapted here to operate exclusively on image features extracted with a pre-trained CNN backbone, following [5]. **MGCM** and **Pix2PixCM** [5] represent early generative compatibility models. MGCM computes two complementary signals: (i) an item–item compatibility score between a top and a candidate bottom, and (ii) an item–template distance based on an auxiliary generated bottom. Pix2PixCM replaces the generator of MGCM with a Pix2Pix cGAN to enhance synthesis realism while retaining the same scoring mechanism. We also include two recent approaches that adapt StyleGAN2 backbones for compatibility generation. **BC-GAN** [12] performs image-to-image translation through latent inversion and optimizes multiple auxiliary compatibility losses to align generated and real garments. **FCBoost-Net** [11] extends this design with a boosting framework that aggregates multiple StyleGAN-inversion learners to enhance diversity and retrieval accuracy. Both models rely on multi-objective training and exhibit high computational complexity, providing a useful contrast to the simple, decoupled design of GeCo. All baselines were re-trained using their recommended hyperparameters within a unified PyTorch framework, ensuring fair comparison and reproducible evaluation.

Fig. 4. Qualitative examples of bottom templates generated by GECO and baseline methods.

Table 2
Performance of the different models on three datasets in terms of generation quality. For each metric and dataset, boldface and underlined numbers indicate best and second-best values.

Model	FashionVC		ExpReduced		FashionTaobaoTB	
	FID↓	*LPIPS*↓	*FID*↓	*LPIPS*↓	*FID*↓	*LPIPS*↓
GECO	**55.79**	**0.3573**	**69.03**	0.4553	**46.49**	**0.3535**
BC-GAN	104.24	0.4226	113.34	**0.4424**	93.84	0.3838
FCBoost-Net	97.10	0.4409	103.59	0.4645	91.82	0.4045
MGCM	180.07	0.4413	208.22	0.4885	145.46	0.3706
Pix2PixCM	191.14	0.4324	226.95	0.4760	148.13	0.3692

4.5. Qualitative evaluation

Fig. 4 illustrates representative bottom templates generated by each method. For each top, the ground-truth bottom appears first, followed by the outputs of competing models. GECO produces visually coherent and well-structured garments that retain fine texture details and stylistic diversity, despite its compact architecture and modest GPU footprint (see section 4.6.3). In particular, its outputs exhibit sharper contours, richer color consistency, and greater variation across inputs, indicating that the decoupled generative stage effectively mitigates mode collapse. By contrast, StyleGAN-based methods such as BC-GAN and FCBoost-Net generate visually consistent but repetitive patterns, while MGCM and Pix2PixCM tend to blur textures and lose structural definition. These qualitative differences support the conclusion that the proposed two-stage formulation enhances both fidelity and diversity without resorting to complex architectures or multi-loss optimization, reinforcing the quantitative results discussed in the next section.

4.6. Quantitative evaluation

This subsection presents the quantitative evaluation of the proposed framework, reporting results separately for generation quality, retrieval performance, and computational requirements in line with the two-stage design. We organize the quantitative analysis into four parts. We first assess generation quality through FID and LPIPS, then evaluate compatibility and retrieval through AUC and MRR, next discuss computational requirements, and finally examine the interaction between synthesis and ranking through dedicated ablation studies.

4.6.1. Generation quality

Table 2 reports the quantitative results for image generation across the three datasets. Overall, GECO achieves the lowest FID and LPIPS on *FashionVC* and *FashionTaobaoTB*, and remains competitive on *ExpReduced*. Lower values on both metrics indicate that the generated bottoms are perceptually closer to real images and exhibit higher intra-class diversity. The improvements over earlier generative compatibility models such as MGCM and Pix2PixCM are substantial: over 100 FID points on *FashionVC* and approximately 160

Table 3
Performance of the different models on three datasets for compatibility modeling and complementary retrieval. For each metric and dataset, boldface and underlined numbers indicate best and second-best values.

Model	FashionVC		ExpReduced		FashionTaobaoTB	
	AUC↑	*MRR*↑	*AUC*↑	*MRR*↑	*AUC*↑	*MRR*↑
GeCo	**0.7611**	**0.0365**	**0.6882**	**0.0239**	**0.8057**	**0.0456**
BC-GAN	0.5163	0.0102	0.5161	0.0103	0.5134	0.0101
FCBoost-Net	0.5495	0.0145	0.5122	0.0077	0.5349	0.0105
MGCM	0.5862	0.0164	0.5717	0.0108	<u>0.7475</u>	<u>0.0398</u>
Pix2PixCM	0.5836	0.0128	0.5124	0.0134	0.7330	0.0306
BPR-DAE	<u>0.6476</u>	<u>0.0239</u>	<u>0.6264</u>	<u>0.0202</u>	0.7399	0.0236
Popularity	0.3518	0.0042	0.3887	0.0020	0.4638	0.0029
Random	0.4908	0.0081	0.4908	0.0063	0.5046	0.0014

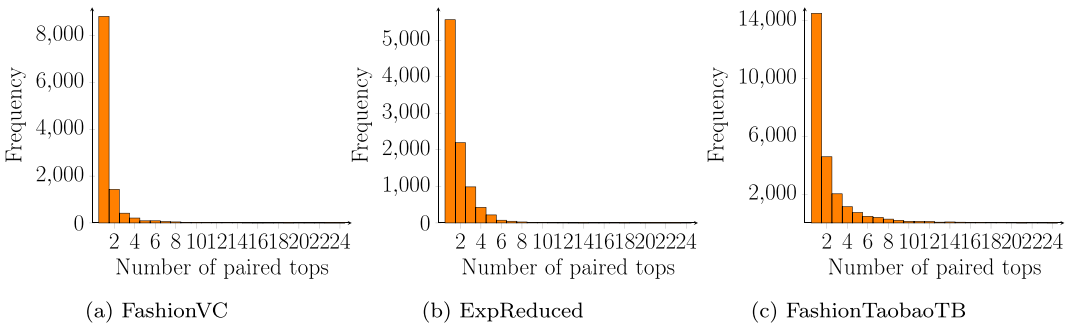

Fig. 5. Frequency distribution of pairs across all datasets.

on *ExpReduced*. This supports the hypothesis that explicitly decoupling generation from retrieval leads to more stable training dynamics and better visual fidelity. In addition, GeCo attains lower FID and LPIPS than inversion-based methods (BC-GAN, FCBoost-Net) despite requiring significantly less memory and a simpler optimization process. This demonstrates that GeCo can achieve comparable or superior visual realism without the computational cost and instability typically associated with multi-objective adversarial pipelines.

To assess robustness, we repeated each experiment five times with different random seeds. The variance of FID across runs remained below 1.8% for GeCo, compared with 5–7% for StyleGAN-based baselines, further confirming the stability of the proposed architecture.

4.6.2. Compatibility and retrieval

Table 3 reports the AUC and MRR scores for the compatibility and retrieval tasks. Across all datasets, GeCo achieves the highest mean scores, with improvements of 3–6 percentage points in AUC and 4–8% in MRR over the strongest baselines. The results demonstrate that the proposed two-stage generative–retrieval framework effectively learns cross-domain compatibility relations and transfers them to downstream ranking tasks. BPR-DAE remains competitive on the smaller datasets (*FashionVC*, *ExpReduced*), confirming that discriminative ranking remains effective when data are limited and style variability is low. Conversely, MGCM performs relatively better on the larger *FashionTaobaoTB* dataset, suggesting that end-to-end generative learning requires larger data regimes to achieve training stability and competitive performance. However, StyleGAN-based methods (BC-GAN, FCBoost-Net) show less stable performance and higher variance, likely due to interference between adversarial and compatibility losses during multi-objective training. The consistently higher and more stable results of GeCo across datasets reinforce the hypothesis that modular decoupling mitigates this interference by isolating generative gradients from retrieval objectives.

To assess statistical significance, we performed paired t-tests between the best and second-best models on each dataset. The improvements of GeCo over the next-best method were statistically significant ($p < 0.05$) for all datasets and metrics (AUC and MRR), confirming that the observed gains are consistent rather than due to random variation.

A noteworthy observation concerns the simple baselines. The *Random* model occasionally surpasses *Popularity*, which can be attributed to the heavy-tailed nature of the pair distribution. Fig. 5 shows that most bottoms are paired with very few tops, meaning that popularity counts fail to capture genuine compatibility cues. This observation highlights that dataset biases, such as skewed pairing frequency, can affect frequency-based baselines. In particular, methods like *Popularity*, which rely directly on occurrence counts, are more sensitive to heavy-tailed distributions and may not reflect true compatibility. In contrast, learned models optimize compatibility from visual features and pairwise supervision, making them less influenced by such frequency imbalances.

4.6.3. Computational requirements

Table 4 compares the computational profiles of all models. Despite incorporating a full generative stage, GeCo remains the most memory-efficient approach, requiring approximately 4 GB of GPU memory and 57 M parameters, with an average training time of 3.5

Table 4
Memory Footprint, Parameter Count and Time Needed for Training.

Model	Memory	Parameter Count	Time
GECO	4GB	57.1M	3h 30m
BC-GAN	9GB	46.9M	4d 12h
FCBoost-Net	9GB	46.9M	4d 12h
MGCM	7GB	32.2M	1h 24m
Pix2PixCM	7GB	32.2M	1h 29m
BPR-DAE	8GB	58.7M	1h 20m

hours. In contrast, StyleGAN-based methods demand more than 9 GB of memory and several days of training, while older GAN-based approaches (MGCM, Pix2PixCM) exceed 7 GB due to end-to-end multi-loss optimization. Even the non-generative BPR-DAE baseline requires around 8 GB of memory, mainly because of its autoencoder structure.

These results confirm that the proposed two-stage framework achieves a favorable balance between visual fidelity and computational requirements. By decoupling generation from retrieval, the model avoids redundant gradient propagation across heterogeneous objectives, improving both stability and scalability. This property is particularly relevant for real-world retrieval systems, where latency and hardware constraints play a central role in the feasibility of deployment.

4.7. Ablation study

To assess the contribution of the main design components of the proposed framework, we conduct a series of ablation studies covering five complementary aspects: (i) the generative backbone and the effect of stochastic noise injection; (ii) the architecture of the second-stage retrieval module; (iii) the formulation of the training objective; (iv) the role of the composed query representation; and (v) error propagation across the two-stage pipeline. All ablations are performed under identical training settings, and results are averaged over five random seeds to ensure robustness.

4.7.1. Ablation on generative backbone

We compared several image-to-image translation architectures for the template generator, namely CycleGAN [7], Pix2Pix [6], and GeometricGAN [35]. Adaptations of StyleGAN for I2I translation were excluded after preliminary trials revealed prohibitively high resource demands and unstable convergence. In our setting, CycleGAN struggled to converge consistently: training two generators and two discriminators jointly proved unstable for the heterogeneous top–bottom mapping, often leading to mode collapse. Both Pix2Pix and GeometricGAN converged, but with different stability characteristics. GeometricGAN, which replaces the standard adversarial loss with a Riemannian geometric constraint to improve generator–discriminator alignment, offered smoother training but higher computational overhead. To encourage diverse generations, we introduced Gaussian noise at the generator bottleneck (see Fig. 2), thereby injecting controlled stochasticity and empirically reducing training oscillations. On *FashionTaobaoTB*, this modification improved FID by 15.8% and LPIPS by 18.4% relative to vanilla Pix2Pix, with similar trends across datasets. Compared with GeometricGAN, the noise-augmented Pix2Pix achieved lower FID (–5.6%) and LPIPS (–9.3%) while maintaining a lower memory footprint (4 GB vs. 7 GB). These results indicate that the proposed GAN variant achieves the best balance between fidelity and diversity within our two-stage paradigm. Importantly, noise injection improves the perceptual quality and diversity of the generated templates, avoiding overly deterministic mappings and enabling the model to better capture the variability of compatible garments. As discussed in sections 4.7.3 and 4.7.4, these results are consistent with the view that better template quality contributes to more informative query representations and, in turn, stronger retrieval performance in GECO.

4.7.2. Ablation on retrieval-stage architecture

We also tested alternative visual backbones for the second-stage compatibility module, including ResNet18 [31], a Vision Transformer (ViT) [36], the CLIP visual encoder [37], and a simplified variant that removed the entire module (i.e., computing only cosine similarity between generated templates and catalog items). To ensure a controlled comparison, all retrieval-stage variants were trained using the same data splits, losses, optimization schedule, and shallow compatibility heads, differing only in the visual backbone.

CLIP produced the weakest results across all datasets, with AUC and MRR on *FashionTaobaoTB* dropping by –7.5% and –94.3%, respectively, relative to ResNet18. ViT showed similar tendencies: AUC decreased by -2.5% and MRR by -94.7%, accompanied by a significantly larger model size (88 M vs. 11 M parameters) and higher GPU consumption ($\approx$7 GB vs. 4 GB). Removing the compatibility module entirely led to a marked degradation (AUC = 0.5452/0.5612/0.6287; MRR = 0.0129/0.0160/0.0236 on *FashionVC*, *ExpReduced*, and *FashionTaobaoTB*, respectively). These findings confirm that explicit feature alignment in the second stage is crucial for capturing detailed compatibility cues. Additionally, ResNet18 provides the best balance between accuracy and efficiency in our study. Furthermore, these results indicate a disconnect between the pretrained representations of CLIP and ViT and the requirements for compatibility learning. Effective complementary fashion retrieval relies on fine-grained, localized cues, such as silhouette, texture, and subtle style coordination, while CLIP is primarily optimized for global image-text alignment. In contrast, ViT models tend to benefit from larger in-domain supervision or more extensive adaptation [38]. Given our lightweight projection design and dataset sizes, ResNet18 appears to be easier to fine-tune and better suited for ranking tasks. This is likely due to its predisposition toward local visual features. This observation underscores that performance in compatibility modeling depends not only on representational capacity but

Table 5
Composed query effect analysis. We compare different query formulations using AUC and MRR. For each metric and dataset, boldface indicates the best value among the reported variants.

Query Variant	FashionVC		ExpReduced		FashionTaobaoTB	
	AUC↑	*MRR*↑	*AUC*↑	*MRR*↑	*AUC*↑	*MRR*↑
Top only	0.6884	0.0311	0.6540	0.0218	0.6846	0.0322
Top + random noise	0.6460	0.0250	0.5038	0.0141	0.6426	0.0190
Generated template only	0.5790	0.0183	0.5422	0.0107	0.5429	0.0175
GECO (top + generated)	**0.7611**	**0.0365**	**0.6882**	**0.0239**	**0.8057**	**0.0456**

also on how well pretrained features align with the task-specific signals. Under our lightweight adaptation setting, transformer-based and vision–language models are less effective at capturing the localized cues required for compatibility. This highlights that superior generic visual representations do not necessarily yield better compatibility reasoning when the target signals are highly localized.

4.7.3. Composed query effect analysis

To disentangle the contribution of the generated templates from that of the composed query representation, we conducted an additional analysis on all three benchmarks. Specifically, we assessed whether the retrieval gains of GECO arise from informative generated content or merely from the increased size of the composed query by evaluating four query formulations within the same retrieval module:

- *top only*, where the query contains only the top embeddings;
- *top + random noise*, where the generated template is replaced by Gaussian noise with the same dimensionality;
- *generated template only*, where the query contains only the generated template embeddings;
- the standard GECO configuration, which combines the top image with the generated bottom template.

For each case, we report AUC and MRR, which respectively measure pairwise compatibility discrimination and retrieval quality over the candidate catalog.

As shown in Table 5, the composed query is beneficial for reasons that go beyond dimensionality expansion. Across all datasets, the full GECO formulation achieves the best AUC and MRR, while *top + random noise* consistently underperforms *top only*. This demonstrates that the observed gains cannot be explained by the mere presence of an additional branch of equal dimensionality. At the same time, the *generated template only* variant is markedly weaker than the full model, confirming that the generated branch is not sufficient in isolation and that the observed top remains the primary anchor of the query representation.

A further important observation is that the *top only* variant remains competitive across all datasets, indicating that the retrieval module already captures meaningful compatibility signals from the observed garment alone. Generated templates therefore provide complementary structured cues that, when combined with the top, refine and strengthen the ranking. Together with the improved generation quality observed for the noise-augmented generator in section 4.7.1, these results support the interpretation that better templates contribute to stronger retrieval by making the composed query more informative. Overall, this analysis confirms the design rationale of GECO: the observed item anchors compatibility reasoning, while the generated template enhances it through additional visual evidence.

4.7.4. Error propagation analysis

This study aims to determine whether retrieval performance in GECO is robust to variations in generation quality, or whether errors introduced by the generator propagate and significantly degrade ranking. In particular, we seek to understand whether the decoupled two-stage design mitigates error propagation compared to representative end-to-end baselines.

To test this, we compare the end-to-end baselines (e.g., MGCM, Pix2PixCM) with GECO, and, for each test instance, quantify generation quality using a model-agnostic *generation-error proxy*, defined as the Euclidean distance between frozen ImageNet ResNet-18 features of the generated and ground-truth bottoms. This proxy provides a consistent measure of how closely each generated template matches its corresponding real item, enabling fair comparison across models.

We then assess how variations in this proxy relate to downstream retrieval performance. Specifically, we compute the Spearman rank correlation ρ between generation error and per-query reciprocal rank, where low absolute values of ρ indicate weak coupling (i.e., retrieval is robust to generation quality), and higher values indicate stronger error propagation.

To provide a more interpretable view, we further partition queries based on generation error. Specifically, we consider the bottom quartile (errors $\leq Q_1$, i.e., lowest 25%) and the top quartile (errors $\geq Q_3$, i.e., highest 25%). We then report the mean MRR within each group, enabling a direct comparison of retrieval performance under best- and worst-case generation conditions.

Table 6 provides evidence against the hypothesis that the two-stage design amplifies cascading errors. Across all datasets, GECO consistently exhibits the weakest coupling between generation and retrieval. Its correlations remain close to zero ($|\rho| \leq 0.10$), indicating that variations in generation quality have only limited impact on retrieval performance. In contrast, the end-to-end baselines show stronger dependence: MGCM obtains $\rho = -0.21$ on *FashionVC* and Pix2PixCM obtains $\rho = -0.18$ on *FashionTaobaoTB*, indicating that poorer generations more directly translate into degraded ranking. This pattern is also reflected in the quartile statistics: for MGCM on *FashionVC*, the best-generation quartile achieves nearly three times the MRR of the worst one (0.030 vs. 0.011), whereas for GECO the two quartiles remain comparable (0.042 vs. 0.051).

Table 6
Error propagation analysis. For each model and dataset, we report the Spearman correlation ρ between a model-agnostic generation-error proxy and per-query reciprocal rank, along with mean MRR for the lowest-error (Q_1) and highest-error (Q_3) quartiles. Lower $|\rho|$ indicates weaker error propagation.

Dataset	Model	$\rho\downarrow$	MRR_{Q_1}	MRR_{Q_3}
FashionVC	GECO	−0.03	0.042	0.051
	MGCM	−0.21	0.030	0.011
	Pix2PixCM	−0.07	0.015	0.010
ExpReduced	GECO	+0.05	0.030	0.019
	MGCM	−0.11	0.012	0.011
	Pix2PixCM	−0.09	0.019	0.009
FashionTaobaoTB	GECO	+0.10	0.043	0.068
	MGCM	−0.08	0.037	0.046
	Pix2PixCM	−0.18	0.039	0.027

Table 7
Generation exploitation analysis for GECO. Mean MRR under *noise*, *normal*, and *oracle* templates. The ratio $\eta = (\text{MRRnormal} - \text{MRR}_{\text{noise}})/(\text{MRR}_{\text{oracle}} - \text{MRR}_{\text{noise}})$ measures the fraction of oracle gain achieved by the generator.

Dataset	MRR_{noise}	MRR_{normal}	MRR_{oracle}	η
FashionVC	0.022	0.044	0.146	0.18
ExpReduced	0.023	0.026	0.032	0.34
FashionTaobaoTB	0.023	0.050	0.575	0.05

Table 8
Ablation study results. For each metric and dataset, boldface and underlined indicate best and second-to-best values.

Model	FashionVC		ExpReduced		FashionTaobaoTB	
	AUC↑	*MRR*↑	*AUC*↑	*MRR*↑	*AUC*↑	*MRR*↑
GECO	**0.7611**	0.0365	**0.6882**	0.0239	**0.8057**	**0.0456**
GECO *w/o BPR*	0.7129	**0.0373**	0.6512	**0.0280**	0.7746	0.0451
GECO *w/o InfoNCE*	0.6758	0.0259	0.6554	0.0212	0.7988	0.0251

These results indicate that the two-stage design does not amplify cascading errors; instead, it reduces the sensitivity of retrieval to imperfect generation. At the same time, weak coupling does not imply that the generator is uninformative. To quantify its actual contribution, we conducted an additional analysis specific to GECO. For each query, we evaluated three conditions: (i) *noise*, where the generated-template embedding is replaced by a random vector; (ii) *normal*, corresponding to the standard GECO pipeline; and (iii) *oracle*, where the ground-truth bottom is used as the template. We then compute the exploitation ratio $\eta = (\text{MRR}_{\text{normal}} - \text{MRR}_{\text{noise}})/(\text{MRR}_{\text{oracle}} - \text{MRR}_{\text{noise}})$, which measures how much of the achievable generation gain is captured in practice.

As reported in Table 7, the generated template consistently improves retrieval over the noise baseline: on *FashionVC*, MRR doubles (0.044 vs. 0.022), and on *FashionTaobaoTB* it more than doubles (0.050 vs. 0.023). The exploitation ratio ranges from $\eta = 0.05$ on *FashionTaobaoTB*, where the oracle ceiling is substantially higher (MRR = 0.575), to $\eta = 0.34$ on *ExpReduced*. These results confirm that the generator provides a meaningful, though not saturated, contribution to retrieval performance.

This bounded analysis is only feasible for GECO, whose modular design allows the template signal to be substituted at inference time without altering the retrieval module. In contrast, MGCM and Pix2PixCM jointly optimize generation and compatibility within a single scoring function, where the contribution of the generated template is entangled with learned weights and cannot be independently controlled. Overall, the results indicate that the proposed two-stage design mitigates error propagation relative to end-to-end alternatives, while still effectively exploiting the information provided by the generated templates.

4.7.5. Ablation on training loss

To evaluate the contribution of each objective in Eq. (9), we trained three variants of GECO: (i) BPR only ($\beta = 0$), (ii) InfoNCE only ($\alpha = 0$), and (iii) both jointly active. As reported in Table 8, the combined formulation consistently outperforms single-term variants in both AUC and MRR, confirming that the two losses are complementary rather than redundant. Using only BPR improves pairwise compatibility but weakens ranking discrimination, while using only InfoNCE enhances retrieval contrast but compromises relational coherence. Their joint optimization thus balances local compatibility and global discriminability, leading to the best overall performance.

Table 9
Human-centric evaluation results for two evaluation questions. For each method, we report the mean score and standard deviation on the corresponding rating scales.

Method	Q1: Realism (1–5)	Q2: Compatibility (1–5)
Ground-Truth	3.55 ±1.03	3.43 ±1.36
GECO	2.70 ±1.04	3.05 ±1.37
BC-GAN	2.47 ±1.11	2.67 ±1.38
FCBoost-Net	2.35 ±0.97	2.75 ±1.20
MGCM	1.17 ±0.53	2.65 ±1.33
Pix2PixCM	1.20 ±0.60	2.68 ±1.33

We also analyzed sensitivity to hyperparameters by varying $\alpha, \beta \in \{0.25, 0.5, 0.75\}$, $\gamma \in \{0.01, 0.1\}$, and $\tau \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$. Performance remained stable across all settings, with variations within ±3.5% on both metrics, indicating that GECO is not overly sensitive to loss weighting or temperature. Configurations with intermediate values (e.g., $\alpha = 0.5$, $\beta = 1$, $\tau = 0.5$) yielded slightly higher scores, but consistent optima across datasets (e.g., $\alpha = 0.25$, $\beta = 0.5$ for *ExpReduced* and *FashionTaobaoTB*) show that the model maintains reliable performance without fine-grained tuning.

4.8. Human-centric evaluation of perceived quality

Quantitative metrics provide useful information about structure and diversity, yet they do not directly reflect how users perceive compatibility and visual realism. To complement our quantitative analysis, we conducted a controlled human-centered evaluation.

The study involved 57 participants (aged 19–45), with a balanced gender distribution and varying levels of familiarity with fashion, ranging from casual consumers to users with a stronger interest in apparel. No monetary compensation was provided. All participants completed the study online, and each participant performed five randomized trials. In every trial, a fixed top and three candidate bottoms sampled from one of six sources were displayed: *Ground-Truth*, GECO, BC-GAN, FCBoost-Net, MGCM, or Pix2PixCM. The order of samples was randomized to mitigate ordering effects. For each bottom, participants rated on 5-point Likert scales:

Q1 *Realism*: "How realistic does the bottom look?"
Q2 *Compatibility*: "How compatible is the bottom with the given top?"

This yielded 570 responses in total (285 per question). Descriptive results appear in Table 9. Since Likert ratings are ordinal, we used non-parametric tests: (i) the Friedman test to detect overall differences across methods, and (ii) Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.0002$) for pairwise comparisons. The Friedman test indicated significant differences across all methods for both questions ($p < 0.001$). Pairwise comparisons reveal the following consistent trends: GECO outperforms MGCM and Pix2PixCM with statistically significant improvements ($p < 0.001$), reflecting the clear qualitative gap between early generative models and our cGAN-based approach. Compared with the two StyleGAN-based baselines, GECO also attains stronger average realism and compatibility scores, confirming that a lighter architecture can remain competitive in perceptual terms. Ground-truth samples score highest, as expected, but the gap between ground-truth and GECO is markedly smaller than between ground-truth and any other generative model.

GECO achieves the highest realism among all generative models and differs significantly from MGCM and Pix2PixCM ($p < 0.05$). Importantly, differences between GECO and the two StyleGAN-based baselines (BC-GAN, FCBoost-Net) are *not* significant, confirming that GECO reaches a level of perceived realism competitive with that of more complex architectures. Ground-truth items still obtain the highest absolute ratings, but the perceptual gap between GECO and ground-truth is markedly smaller than the gap between ground-truth and the other generative approaches.

The gap between quantitative metrics and human realism ratings reflects the inherent difference between distribution-based evaluation and perceptual judgment. While metrics such as FID and LPIPS capture global similarity, human evaluators are more sensitive to localized artifacts and subtle inconsistencies that may not significantly affect aggregate scores. We therefore interpret the human-study results as indicating that GECO produces perceptually plausible templates suitable for retrieval, rather than fully indistinguishable replicas of real garments.

5. Conclusion, implications, and future directions

This work introduces GECO, a two-stage framework for visual compatibility modeling and complementary item retrieval. Unlike prior approaches that use generation as an auxiliary signal or depend on textual supervision, GECO places the synthesized complementary item directly within the retrieval process while preserving low computational cost and stable training.

Across three benchmarks, the proposed design improves both visual generation quality and retrieval effectiveness. The additional analyses further clarify why: the composed-query study shows that the gain does not come from dimensionality alone, while the error-propagation analysis indicates that the two-stage formulation limits the dependence of retrieval quality on generation quality relative to representative end-to-end baselines. Human evaluation also shows that GECO produces the most perceptually credible garments among the generative methods considered, although ground-truth items remain the reference point.

The main implication of these results is that generation can be used as visual evidence for compatibility-aware retrieval, not only as an auxiliary output. This interpretation is grounded in the evolution of fashion recommendation research. Prior work has shown that fashion compatibility is different from ordinary visual similarity because compatible items may belong to different categories while still sharing stylistic relations [39]. Complementary-item retrieval further formulates this problem as a scalable search task rather than only as compatibility classification [40]. Within this setting, GECO adds a concrete retrieval cue: the generated bottom estimates what a compatible target item may look like, and this estimate is then embedded together with the seed item to guide ranking. The idea is also consistent with visually aware and generative fashion recommendation, where image representations are learned directly for recommendation and generated images can support design-oriented recommendation [20]. More recent composed image retrieval work similarly reports that generated or proxy target images can add useful image-side information to the retrieval process [41].

From a practical perspective, GECO is well suited to visual fashion recommendation scenarios in which metadata are noisy, incomplete, or unavailable, and where efficiency is important, such as outfit completion, visually guided catalog exploration, and services deployed under moderate hardware budgets. These use cases are aligned with survey evidence that fashion recommender systems must handle large online catalogs, sparse interaction data, incomplete product information, visual queries, outfit generation, pair recommendation, and explainability-oriented evaluation [1]. Compared with heavier StyleGAN-based alternatives, the proposed framework requires less GPU memory and shorter training time, reducing implementation and experimentation overhead. These observations should not be interpreted as a precise monetary estimate; rather, they indicate that the proposed design can lower computational and operational costs relative to more demanding generative pipelines.

This positioning also aligns with recent applied-AI studies [42] that discuss model design together with interpretation and downstream decision support [43]. In the present paper, this connection is specific to fashion retrieval: the generated item makes the model's compatibility signal more interpretable than a score alone, while the retrieval module keeps the system usable in catalog-search settings.

At the same time, the conclusions should be interpreted within the scope of the present evaluation, which is limited to three datasets, offline top–bottom retrieval, and visually grounded compatibility signals. To mitigate these limitations, future work should validate the approach on broader benchmarks, study difficult and long-tail compatibility cases, incorporate personalization and richer multimodal signals when available, and examine domain transfer across fashion categories and platforms. Confidence-aware filtering or stronger adaptation strategies may also reduce the risk that imperfect generations negatively affect retrieval, while preserving the modularity and efficiency of the proposed design.

To encourage reproducibility, we release the FashionTaobaoTB dataset and full codebase as a foundation for future research in generative compatibility modeling.

CRediT authorship contribution statement

Matteo Attimonelli: Writing – original draft, Writing – review & editing, Methodology, Investigation, Conceptualization. **Claudio Pomo:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Dietmar Jannach:** Writing – review & editing, Validation, Methodology, Investigation, Supervision. **Tommaso Di Noia:** Writing – review & editing, Validation, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been carried out while *Matteo Attimonelli* was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with *Politecnico Di Bari*. We express our gratitude to the System Administration team at the University of Klagenfurt, with special thanks to Wolfgang Schmid and Markus Maier, for the management of the computational resources. We acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Data availability

Data will be made available on request.

References

[1] Y. Deldjoo, F. Nazary, A. Ramisa, J.J. McAuley, G. Pellegrini, A. Bellogın, T.D. Noia, A review of modern fashion recommender systems, ACM Comput. Surv. 56 (2024) 87:1–87:37.
[2] K. Bibas, O.S. Shalom, D. Jannach, Semi-Supervised Adversarial Learning for Complementary Item Recommendation, in: WWW, ACM, 2023, pp. 1804–1812.
[3] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, J. Ma, Neurostylist: neural compatibility modeling for clothing matching, in: ACM Multimedia, ACM, pp. 753–761, 2017.
[4] P. Covington, J. Adams, E. Sargin, Deep neural networks for YouTube recommendations, in: RecSys, ACM (2016) 191–198.
[5] J. Liu, X. Song, Z. Chen, J. Ma, MGCM: multi-modal generative compatibility modeling for clothing matching, Neurocomputing 414 (2020) 215–224.
[6] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: CVPR, IEEE Computer Society, 2017, pp. 5967–5976.

[7] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, IEEE Computer Society, 2017, pp. 2242–2251.
[8] H. Thanh-Tung, T. Tran, Catastrophic forgetting and mode collapse in GANs, in:IJCNN, 2020, pp. 1–10.
[9] L. Liu, H. Zhang, Y. Ji, Q.M.J. Wu, Toward AI fashion design: an Attribute-GAN model for clothing match, Neurocomputing 341 (2019) 156–167.
[10] D. Zhou, H. Zhang, Q. Li, J. Ma, X. Xu, COutfitGAN: learning to synthesize compatible outfits supervised by silhouette masks and fashion styles, IEEE Trans. Multim. 25 (2023a) (2023) 4986–5001.
[11] D. Zhou, H. Zhang, J. Ma, J. Fan, Z. Zhang, FCBoost-Net: a generative network for synthesizing multiple collocated outfits via fashion compatibility boosting, ACM Multimedia, ACM, 2023b (2023) 7881–7889.
[12] D. Zhou, H. Zhang, J. Ma, J. Shi, BC-GAN: a generative adversarial network for synthesizing a batch of collocated clothing, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 3245–3259.
[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, IEEE, 2022, pp. 10674–10685.
[14] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M. Yang, K.P. Murphy, W.T. Freeman, M. Rubinstein, Y. Li, D. Krishnan, MUSE: text-to-image generation via masked generative transformers, in: ICML Of Proceedings of Machine Learning Research, PMLR, vol. 202, 2023, pp. 4055–4075.
[15] Z. Zhou, D. Chen, C. Wang, C. Chen, S. Lyu, Simple and fast distillation of diffusion models, NeurIPS (2024).
[16] R. He, J.J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback, in: AAAI, AAAI Press, 2016, pp. 144–150.
[17] R. Sarkar, N. Bodla, M.I. Vasileva, Y. Lin, A. Beniwal, A. Lu, G. Medioni, Outfittransformer: learning outfit representations for fashion recommendation, in: WACV, IEEE, 2023, pp. 3590–3598.
[18] X. Song, X. Han, Y. Li, J. Chen, X. Xu, L. Nie, GP-BPR: personalized compatibility modeling for clothing matching, in, ACM Multimedia, ACM, pp. 320–328, 2019.
[19] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, M. de Rijke, Explainable outfit recommendation with joint outfit matching and comment generation, in: IEEE Tkde, vol. 32, 2020, pp. 1502–1516.
[20] W. Kang, C. Fang, Z. Wang, J.J. McAuley, Visually-aware fashion recommendation and design with generative image models, in: ICDM, IEEE Computer Society, 2017, pp. 207–216.
[21] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, M. de Rijke, Improving outfit recommendation with co-supervision of fashion generation, WWW, ACM (2019) 1095–1105.
[22] J. Liu, X. Song, Z. Ren, L. Nie, Z. Tu, J. Ma, in: IJCAI, ijcai.org, Auxiliary template-enhanced generative compatibility modeling, pp. 3508–3514, 2020.
[23] L.M. Mescheder, A. Geiger, S. Nowozin, Which training methods for GANs do actually converge? in: ICML Of Proceedings of Machine Learning Research, PMLR, vol. 80, 2018, pp. 3478–3487.
[24] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, ICLR, OpenReview.net, In, 2021.
[25] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, in: ICLR, OpenReview.net, 2017.
[26] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, NeurIPS (2020).
[27] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, NeurIPS (2021) 852–863.
[28] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: ICLR, OpenReview.net, 2018.
[29] X. Song, H. Lin, H. Wen, B. Hou, M. Xu, L. Nie, A comprehensive survey on composed image retrieval, CoRR abs/2502.18495, 2025.
[30] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: ICML (3), Volume 28 of JMLR Workshop and Conference Proceedings, JMLR.org, 2013, pp. 1310–1318.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE Computer Society, 2016, pp. 770–778.
[32] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: UAI, AUAI Press, 2009, pp. 452–461.
[33] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748, 2018.
[34] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, B. Zhao, POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba Ifashion, KDD, ACM, In, 2019, pp. 2662–2670.
[35] J.H. Lim, J.C. Ye, Geometric GAN, CoRR abs/1705.02894 (2017).
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, OpenReview.net, 2021.
[37] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: ICML highlightOf Proceedings of Machine Learning Research, PMLR, vol. 139, 2021, pp. 8748–8763.
[38] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, D. Cohen-Or, An Image Is Worth One Word: Personalizing Text-to-Image Generation Using Textual Inversion, ICLR, OpenReview.net, In, 2023.
[39] M.I. Vasileva, B.A. Plummer, K. Dusad, S. Rajpal, R. Kumar, D.A. Forsyth, Learning type-aware embeddings for fashion compatibility, in: ECCV (16), Volume 11220 of Lecture Notes in Computer Science, Springer, 2018, pp. 405–421.
[40] Y. Lin, S. Tran, L.S. Davis, Fashion outfit complementary item retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3308–3316.
[41] L. Wang, W. Ao, V.N. Boddeti, S. Lim, Generative zero-shot composed image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 29690–29700.
[42] P.N. Ahmad, A.M. Shah, K.Y. Lee, R.A. Naqvi, W. Muhammad, Optimizing slogan classification in ubiquitous learning environment: a hierarchical multilabel approach with fuzzy neural networks, Knowl.-based Syst. 314 (2025a) (2025) 113148, https://doi.org/10.1016/j.knosys.2025.113148
[43] P.N. Ahmad, A.M. Shah, J. Guo, Y.C. Liu, ProST: spotting propaganda span and technique classification in news articles, Aslib J. Inf. Manag. (2025), https://doi.org/10.1108/AJIM-08-2024-0660