CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.057

FLAIR-XGT: Local visual patch feature aggregation with multi-graph fusion for image-based recommendation

Le Huynh Quoc Bao, Nguyen Minh Khiem, and Nguyen Thai-Nghe*

College of Information and Communication Technology, Can Tho University, Viet Nam

**Corresponding author (ntnghe@cit.ctu.edu.vn)*

Article info.

Received 30 Jun 2026
Revised 6 Sep 2026
Accepted 8 Oct 2026

Keywords

Cold-start, Graph neural network, Image-based recommendation, Local patch aggregation, Multi-graph fusion

ABSTRACT

Existing multimodal recommender systems discard fine-grained spatial information before recommendation training even starts by representing each product image as a single global vector, either a pooled CNN output or a Vision Transformer's CLS token. We propose FLAIR-XGT, an image-based recommendation framework that uses three parallel graph networks (LightGCN, VisualGCN, and TextGCN) to jointly train $M = 32$ *representative local DINOv2 ViT-B/14 patch tokens (each covering a* 14×14 *pixel region) end-to-end under a combined BPR + InfoNCE objective. Our primary scoring component, GlobalVis, combines the local patch tokens using a projector that is optimized in tandem with the recommender. In comparison to all fixed pre-extracted global-feature baselines, this joint training of local patch aggregation with multi-graph collaborative fusion produces NDCG@20* $= 0.0706$ *on Amazon Fashion (*$+12.2\%$ *over MMGCN) and NDCG@20* $= 0.0545$ *on Sports & Outdoors (*$+17.9\%$ *over MMGCN). We also study XGT, a user-conditioned cross-attention module that focuses spatial attention on user-relevant patches. A Graph Transformer block with a CF residual connection fuses the attended output. One important empirical finding is that XGT cross-attention performs worse than GlobalVis on both datasets, which reflects a cold-start limitation: user embeddings are too sparse to form discriminative attention queries, making learned mean-pooling of local patches the stronger configuration in practice. Sensitivity analysis verifies resilience across the critical hyperparameters, and ablation studies link each performance gain to particular design decisions.*

1. INTRODUCTION

Interest in multimodal recommender systems that use product images and text descriptions combined with user interaction histories has increased due to the explosive expansion of e-commerce. In industries like fashion, sports goods, and consumer electronics, where a product's appearance directly influences buying decisions, visual appeal is particularly significant.

Current multimodal recommendation algorithms use either the CLS token of a Vision Transformer or the average-pooled output of a CNN backbone (ResNet50 in VBPR (He et al., 2016)) to represent each product image as a single global vector. Through user-item interaction graphs, graph-based techniques like MMGCN (Wei et al., 2019), BM3 (Zhou et al., 2023), and MGCN (Yu et al., 2023) spread these global vectors. Despite its effectiveness, this global compression eliminates

the spatial structure of images: before the recommendation model even starts training, information on which particular region (texture pattern, logo, sole, structural detail) is visually meaningful to a certain viewer is permanently lost.

Vision transformers, such as DINOv2 (Oquab et al., 2023), naturally generate 256 patch tokens of 768 dimensions each, each of which captures rich, spatially-localized semantics and encodes a 14×14-pixel region. The question of whether local DINOv2 patch tokens learned end-to-end in a multi-graph recommendation framework can beat global visual feature compression for image-based recommendation is naturally raised by this.

We propose a positive response with FLAIR-XGT and further explore when user-conditioned patch attention enhances recommendation beyond straightforward aggregation, contributing the following:

– Local patch representation for recommendation. FLAIR-XGT is the first framework to use DINOv2 ViT-B/14 local patch tokens ($M = 32$, uniformly subsampled from 256) as item features in a recommendation model, instead of compressing them into a single global vector first.

– GlobalVis: effective local patch aggregation. We propose GlobalVis (Section 3.7), which scores user-item pairs using the dot product between the user CF embedding and an end-to-end learned projection of the $M = 32$-patch mean: $s^{\text{gbl}} = \boldsymbol{e}_u^{CF} \cdot \ell_2(\boldsymbol{W}_p \bar{\boldsymbol{p}}_i)$. GlobalVis, unlike fixed PCA baselines (LGN+DINOv2-Avg), learns $\boldsymbol{W}_p$ alongside recommendation training and collects local patch tokens instead of the CLS global token.

– XGT: user-conditioned cross-attention over patch tokens (investigative variant). Additionally, we study a query $Q = W_q \boldsymbol{e}_u^{CF} + W_t \boldsymbol{t}_i$ conditioned jointly on the item text $\boldsymbol{t}_i$ and the user's collaborative embedding $\boldsymbol{e}_u^{CF}$, which focuses spatial attention on user-relevant image regions. The attended output integrates via an XGT Fusion block (FFN + dual Add & Norm, with $\boldsymbol{e}_i^{CF}$ residual in Add & Norm 1). The purpose of this variation is to investigate the circumstances in which attention-pooling outperforms mean-pooling in contexts with sparse interactions.

– Cold-start analysis of attention-pooling. We demonstrate empirically that on cold-start benchmarks, cross-attention over patches performs worse than simple mean-pooling because noisy user embeddings result in subpar attention queries, which reduce the spatial selectivity benefit of local patches.

– Systematic ablation chain. The FLAIR-XGT model has two variants: an intermediate form (LightGCN + single-head Patch Attention Pool + scalar fusion) and a straight XGT cross-attention variant. These variants isolate each design option from collaborative filtering baselines.

2. RELATED WORKS

Early multimodal recommendation methods enhanced collaborative filtering with visual features: VBPR (He et al., 2016) augmented BPR (Rendle et al., 2012) with CNN features, while graph-based extensions such as MMGCN (Wei et al., 2019) and LATTICE (Zhang et al., 2021) propagate modality-specific signals and build item-item semantic graphs. BM3 (Zhou et al., 2023) and MGCN (Yu et al., 2023) further align modalities via contrastive learning and behavior-guided cleansing, yet all these approaches compress each image into a single global feature vector, losing spatial structure. To bridge modality gaps, cross-attention methods such as KGAT (Wang et al., 2019), SGL (Wu et al., 2021), and multi-view graph CF (Mu et al., 2026) have been explored. More recently, vision transformer backbones like CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) have replaced CNNs, typically via the global CLS token; however, DINOv2's patch tokens retain strong spatially-localized semantics that CLS-based aggregation discards, a finding echoed by Zheng et al. (2026), who show dense patch representations outperform CLS tokens by over 13% in fine-grained visual analysis. For the collaborative filtering backbone, LightGCN (He et al., 2020) simplifies graph convolution to normalized neighborhood aggregation; given its efficiency and proven effectiveness, we adopt LightGCN as the CF backbone of FLAIR-XGT.

3. PROPOSED APPROACH

The overall architecture of our proposed approach is displayed in Figure 1. , *Left*: feature extraction DINOv2 ViT-B/14 local patch tokens ($M = 32$, subsampled) and SBERT text embeddings; global CNN/CLIP features are used by baselines only. *Centre*: model components LightGCN (CF path), VisualGCN, TextGCN (both with InfoNCE), GlobalVis (graph-free, best-performing), and XGT Cross-Attention (ablation mode). *Right*: four-component weighted scoring and joint BPR + InfoNCE + L2 training objective.

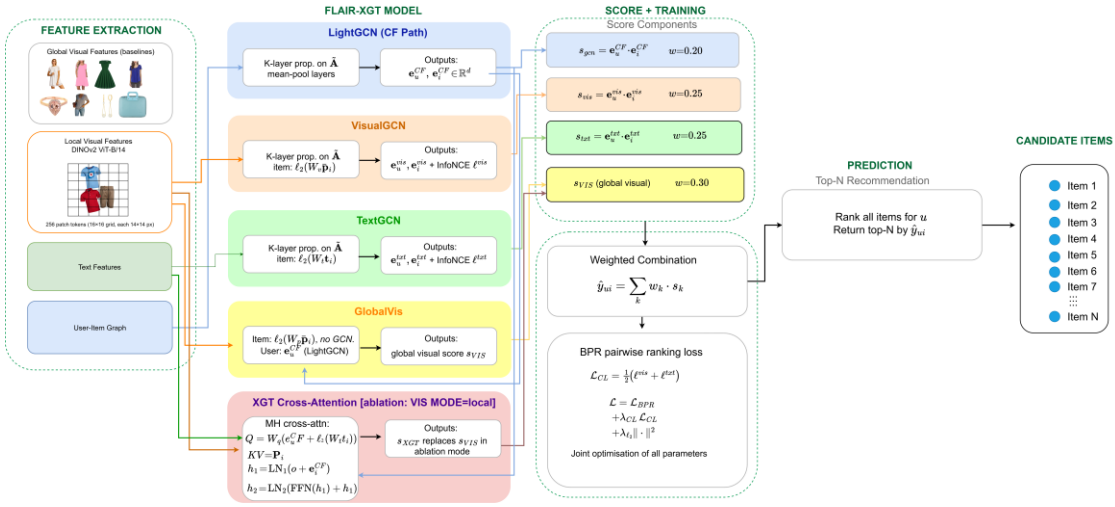


Figure 1. Overview architecture of the FLAIR-XGT

3.1. Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ denote user and item sets. Observed interactions form a binary matrix $\boldsymbol{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$. Each item $i \in \mathcal{I}$ is associated with: (i) DINOv2 patch tokens $\boldsymbol{P}_i \in \mathbb{R}^{M\times 768}$, $M = 32$, and (ii) a text embedding $\boldsymbol{t}_i \in \mathbb{R}^{D_t}$ from SBERT. The objective is to develop a scoring function $\hat{y}_{ui}$ that predicts a user's preference for an item i and prioritizes interacted items over non-interacted ones.

3.2. Visual Feature Extraction

We extract patch-level features with DINOv2 ViT-B/14 (pre-trained, weights frozen throughout recommendation training). Unlike global pooling, this technique preserves locally dispersed features.

3.3. PatchAttentionPool

Patch Projection. Patch tokens are projected into the recommendation embedding space ($d = 64$). $\widetilde{\boldsymbol{P}}_i = \boldsymbol{W}_p \boldsymbol{P}_i^\top \in \mathbb{R}^{d\times M}, \quad \boldsymbol{W}_p \in \mathbb{R}^{d\times 768}$.

User-conditioned Query. Using the LightGCN collaborative embedding of user u (Section 3.4), we can compute: $\boldsymbol{q}_u = \boldsymbol{W}_q \boldsymbol{e}_u^{CF} \in \mathbb{R}^d, \quad \boldsymbol{W}_q \in \mathbb{R}^{d\times d}$.

Text-guided Query Shift. A text projection of an item i causes the query to shift:

$$\boldsymbol{q}_{u,i} = \boldsymbol{q}_u + \boldsymbol{W}_t \boldsymbol{t}_i \in \mathbb{R}^d, \quad \boldsymbol{W}_t \in \mathbb{R}^{d\times D_t}. \tag{1}$$

This focuses spatial attention on patches that align with the product category described in the text.

Attention Pooling. Attention weights and pooled visual embedding are as follows:

$$\boldsymbol{a}_{u,i} = \mathrm{softmax}\left(\frac{\widetilde{\boldsymbol{P}}_i^\top \boldsymbol{q}_{u,i}}{\sqrt{d}}\right) \in \mathbb{R}^M, \qquad \boldsymbol{v}_{u,i} = \widetilde{\boldsymbol{P}}_i \boldsymbol{a}_{u,i} \in \mathbb{R}^d. \tag{2}$$

$\boldsymbol{v}_{u,i}$ provides a text-guided local visual summary of the item i.

FLAIR-V ablation. Setting $\boldsymbol{q}_{u,i} = \boldsymbol{q}_u$ in Eq. (1) removes the text shift, resulting in the FLAIR-V variant and isolating the contribution of text guidance.

3.4. LightGCN Collaborative Filtering

The normalized bipartite graph $\widetilde{\boldsymbol{A}}$ is created from $\boldsymbol{R}$ using symmetric normalization. Embeddings propagate across $K = 3$ layers. $\boldsymbol{E}^{(k+1)} = \widetilde{\boldsymbol{A}}\,\boldsymbol{E}^{(k)}$, and the final representations are the layer-mean:

$$\boldsymbol{e}_u^{CF} = \frac{1}{K+1}\sum_{k=0}^{K} \boldsymbol{e}_u^{(k)}, \quad \boldsymbol{e}_i^{CF} = \frac{1}{K+1}\sum_{k=0}^{K} \boldsymbol{e}_i^{(k)}. \tag{3}$$

3.5. FLAIR: Multimodal Fusion

The final item representation for user u in FLAIR includes text, local visual, and CF elements:

$$\hat{\boldsymbol{e}}_i^u = \boldsymbol{e}_i^{CF} + \sigma(\alpha)\,\boldsymbol{v}_{u,i} + \sigma(\beta)\,\boldsymbol{W}_t\,\boldsymbol{t}_i, \tag{4}$$

where $\alpha, \beta \in \mathbb{R}$ are learned scalars initialized to 0 and σ is the sigmoid function. The predicted score is $\hat{y}_{ui} = \boldsymbol{e}_u^{CF} \cdot \hat{\boldsymbol{e}}_i^u$.

3.6. FLAIR-XGT: Multi-Graph Multi-Modal Architecture

With two modality-specific graph networks (VisualGCN and TextGCN) and a four-component scoring function that combines collaborative,

visual-taste, text-taste, and content-based signals, FLAIR-XGT expands the LightGCN backbone (Section 3.4). The entire pipeline is shown in Fig. 1.

Feature Initialization. LightGCN starts item embeddings by combining projected visual and text features.

$$\boldsymbol{e}_i^{(0)} = \ell_2\big(1/2\,[\ell_2(\boldsymbol{W}_p\,\bar{\boldsymbol{p}}_i) + \ell_2(\boldsymbol{W}_t\,\boldsymbol{t}_i)]\big), \quad (5)$$

LightGCN propagates embeddings over $K = 3$ layers and averages across them, resulting in L2-normalized $\boldsymbol{e}_u^{CF}$ and $\boldsymbol{e}_i^{CF}$.

VisualGCN. A different LightGCN learns each user's visual taste from observed co-interaction patterns by propagating visual item representations along the user-item interaction graph. Both VisualGCN and GlobalVis (Section 3.7) use the same mean-pooled input $\bar{\boldsymbol{p}}_i = \frac{1}{M}\sum_m \boldsymbol{P}_{i,m}$; the only way they differ is in how they score. VisualGCN uses the interaction graph to propagate visual taste to create enriched user and item embeddings, whereas GlobalVis uses a dot product to score directly without using any graph operations. The only part that works with the entire series of M distinct patch tokens is XGT.

$$\boldsymbol{H}^{\text{vis}} = \text{LGN}\big([\boldsymbol{E}_u^{\text{vis}} \;\|\; \ell_2(\boldsymbol{W}_p\bar{\boldsymbol{P}})],\, \widetilde{\boldsymbol{A}}\big), \quad (6)$$

where $\boldsymbol{E}_u^{\text{vis}}$ is a trainable user embedding and the item visual embeddings $\ell_2(\boldsymbol{W}_p\bar{\boldsymbol{P}})$ are detached from backprop to prevent gradient conflict. This produces user and item vectors $\boldsymbol{h}_u^{\text{vis}}, \boldsymbol{h}_i^{\text{vis}}$ that capture collaborative visual affinity without relying on cold-start-sensitive attention queries.

TextGCN. Similar to this, TextGCN uses the same graph to propagate projected text features:

$$\boldsymbol{H}^{\text{txt}} = \text{LGN}\big([\boldsymbol{E}_u^{\text{txt}} \;\|\; \ell_2(\boldsymbol{W}_t\boldsymbol{T})],\, \widetilde{\boldsymbol{A}}\big), \quad (7)$$

capturing user preferences based on text as an additional signal to VisualGCN.

Multi-Component Scoring. The GlobalVis content-based score (Section 3.7) and all three graph outputs are combined as follows:

$$\hat{y}_{ui} = w_1 s_{u,i}^{\text{gcn}} + w_2 s_{u,i}^{\text{vis}} + w_3 s_{u,i}^{\text{txt}} + w_4 s_{u,i}^{\text{mod}}, \quad (8)$$

where $s^{\text{gcn}} = \boldsymbol{e}_u^{CF} \cdot \boldsymbol{e}_i^{CF}$, $s^{\text{vis}} = \boldsymbol{h}_u^{\text{vis}} \cdot \boldsymbol{h}_i^{\text{vis}}$, $s^{\text{txt}} = \boldsymbol{h}_u^{\text{txt}} \cdot \boldsymbol{h}_i^{\text{txt}}$, and s^{mod} is either the GlobalVis score (Eq. [eq:globalvis]) or the XGT cross-attention score described below. Weights $[w_1, w_2, w_3, w_4] = [0.20, 0.25, 0.25, 0.30]$ were determined through iterative experimentation during model development: starting from a near-uniform configuration, we adjusted the weights as new scoring components were added, and fixed the best-performing configuration for all reported results. The weights assign the highest value to s^{mod} and the lowest to s^{gcn}, consistent with their relative contributions in the ablation study (Table 3).

XGT Cross-Attention Scoring (local mode). We implement XGT cross-attention with query $Q = W_q\boldsymbol{e}_u^{CF} + W_t\boldsymbol{t}_i$, as an alternative to GlobalVis. Given projected patch matrix $\boldsymbol{Z}_i = \boldsymbol{W}_p\boldsymbol{P}_i \in \mathbb{R}^{M\times d}$ and combined query $\boldsymbol{q}_{u,i} = \boldsymbol{e}_u^{CF} + \ell_2(\boldsymbol{W}_t\boldsymbol{t}_i)$:

$$\boldsymbol{o} = \boldsymbol{W}_{pr}\,\text{concat}_h\big(\boldsymbol{V}_{i,h}^{\top}\,\text{softmax}(\boldsymbol{K}_{i,h}\,\boldsymbol{W}_{Q,h}\boldsymbol{q}_{u,i}/\sqrt{D_h})\big) \quad (9)$$

$$\begin{aligned} \boldsymbol{h}_1 &= \text{LN}_1(\boldsymbol{o} + \boldsymbol{e}_i^{CF}),\\ \boldsymbol{h}_2 &= \text{LN}_2(\phi(\text{FFN}(\boldsymbol{h}_1)) + \boldsymbol{h}_1),\\ s_{u,i}^{\text{xgt}} &= \boldsymbol{e}_u^{CF} \cdot \ell_2(\boldsymbol{h}_2), \end{aligned} \quad (10)$$

where $H = 2$ attention heads, $D_h = d/H$, and $\phi =$ GELU. The query $\boldsymbol{W}_{Q,h}\boldsymbol{q}_{u,i}$ is user-conditioned and text-guided; $\boldsymbol{e}_i^{CF}$ serves as a residual in Add & Norm 1. This scoring mode underperforms GlobalVis on cold-start datasets.

3.7. GlobalVis: Scoring via Mean Local Patch Projection

A high-quality attention query is required for cross-attention over $M = 32$ subsampled patch tokens. $Q = W_q\boldsymbol{e}_u^{CF} + W_t\boldsymbol{t}_i$ (Eq. (1)). On cold-start-dominated datasets (median ≤ 1 interaction per user), $\boldsymbol{e}_u^{CF}$ converges slowly, yielding noisy queries that impair spatial attention. We therefore evaluate GlobalVis: rather than attending over patches, GlobalVis projects the mean of all $M = 32$ subsampled local patch tokens via the learned projector $\boldsymbol{W}_p \in \mathbb{R}^{d\times 768}$:

$$\begin{aligned} \bar{\boldsymbol{p}}_i &= 1/M \sum_{m=1}^{M} \boldsymbol{P}_{i,m} \in \mathbb{R}^{768},\\ s_{u,i}^{\text{gbl}} &= \boldsymbol{e}_u^{CF} \cdot \ell_2\big(\boldsymbol{W}_p\,\bar{\boldsymbol{p}}_i\big). \end{aligned} \quad (11)$$

Distinction from global baselines. Although GlobalVis and LGN+DINOv2-Avg are similar architecturally, they differ in two ways: Instead of using the pre-extracted global-pooling feature, (i) $\boldsymbol{W}_p$ is learned end-to-end during recommendation training (not a fixed PCA matrix), and (ii) $\bar{\boldsymbol{p}}_i$ averages $M = 32$ subsampled local DINOv2 patch tokens (each covering a 14×14 pixel region, uniformly sampled from all 256 to cover the entire image). Importantly, GlobalVis

eliminates the cold-start query bottleneck of MHCA by not requiring a query from user embeddings at inference. Instead, its score is based solely on item features and the user CF vector via the dot product.

3.8. Training Objective

BPR loss (Rendle et al., 2012) with K negatives per positive is used by both FLAIR and FLAIR-XGT:

$$\mathcal{L}_{BPR} = -\frac{1}{K}\sum_{k=1}^{K} \ln \sigma\big(\hat{y}_{ui} - \hat{y}_{uj_k}\big). \quad (12)$$

Additionally, FLAIR-XGT strengthens modality-specific taste representations by applying InfoNCE contrastive losses (Wu et al., 2021) on both VisualGCN and TextGCN user-item pairs:

$$\mathcal{L}_{CL}^{m} = -\frac{1}{B}\sum_{b} \log \frac{e^{\boldsymbol{h}_{u_b}^{m} \cdot \frac{\boldsymbol{h}_{i_b}^{m}}{\tau}}}{\sum_{k} e^{\boldsymbol{h}_{u_b}^{m} \cdot \frac{\boldsymbol{h}_{i_k}^{m}}{\tau}}}, \quad m \in \{\text{vis}, \text{txt}\}, \quad (13)$$

where $\tau = 0.1$, and the combined contrastive loss is $\mathcal{L}_{CL} = (\mathcal{L}_{CL}^{\text{vis}} + \mathcal{L}_{CL}^{\text{txt}})/2$. The total objective is: $\mathcal{L} = \mathcal{L}_{BPR} + \lambda_{CL}\mathcal{L}_{CL} + \lambda_{\ell_2} \| \Theta \|^2$.

FLAIR-XGT training is summarized in **Algorithm 1** and the general architecture is shown in Figure 1.

Algorithm 1 FLAIR-XGT Training (GlobalVis mode)

Requires: Interaction graph $\widetilde{\boldsymbol{A}}$; patch matrix $\boldsymbol{P}$; text embeddings $\boldsymbol{T}$; epochs T; batch size B; weights $[w_1, w_2, w_3, w_4]$
Ensure: Learned parameters Θ
Feature-init: $\boldsymbol{e}_i^{(0)} \leftarrow \ell_2\big(1/2\,[\ell_2(\boldsymbol{W}_p\bar{\boldsymbol{p}}_i) + \ell_2(\boldsymbol{W}_t\boldsymbol{t}_i)]\big)$
for epoch = 1 to T **do**
  $\mathbf{H}^{\text{gcn}} \leftarrow \text{LGN}(\boldsymbol{E}, \widetilde{\boldsymbol{A}})$
  $\mathbf{H}^{\text{vis}} \leftarrow \text{LGN}([\boldsymbol{E}_u^{\text{vis}} \,\|\, \ell_2(\boldsymbol{W}_p\bar{\boldsymbol{P}})], \widetilde{\boldsymbol{A}})$
  $\mathbf{H}^{\text{txt}} \leftarrow \text{LGN}([\boldsymbol{E}_u^{\text{txt}} \,\|\, \ell_2(\boldsymbol{W}_t\boldsymbol{T})], \widetilde{\boldsymbol{A}})$
  for each mini-batch $\{(u,i,j)\}$ **do**
    Compute four scores for positive (u,i) and negative (u,j):
      $s^{\text{gcn}} \leftarrow \boldsymbol{e}_u^{CF} \cdot \boldsymbol{e}_i^{CF}$; $s^{\text{vis}} \leftarrow \boldsymbol{h}_u^{\text{vis}} \cdot \boldsymbol{h}_i^{\text{vis}}$; $s^{\text{txt}} \leftarrow \boldsymbol{h}_u^{\text{txt}} \cdot \boldsymbol{h}_i^{\text{txt}}$
      $s^{\text{gbl}} \leftarrow \boldsymbol{e}_u^{CF} \cdot \ell_2(\boldsymbol{W}_p\bar{\boldsymbol{p}}_i)$
    $\hat{y}_{ui} \leftarrow w_1 s^{\text{gcn}} + w_2 s^{\text{vis}} + w_3 s^{\text{txt}} + w_4 s^{\text{gbl}}$
    Compute InfoNCE $\mathcal{L}_{CL} = (\mathcal{L}_{CL}^{\text{vis}} + \mathcal{L}_{CL}^{\text{txt}})/2$ on VisualGCN & TextGCN user embs
    $\mathcal{L} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\big[-\ln\sigma(\hat{y}_{ui} - \hat{y}_{uj_k})\big] + \lambda_{CL}\mathcal{L}_{CL} + \lambda_{\ell_2} \| \Theta \|^2$
    Update Θ with Adam + gradient clip; retain_graph = True
  end for
  Evaluate Recall@K, NDCG@K every E epochs; checkpoint if best.
end for

4. EXPERIMENTAL RESULTS

4.1. Dataset Construction

All data are collected from the public Amazon Product Review Dataset (Ni et al., 2019) , which provides large-scale multimodal product information for research purposes. There are two passages during construction. Table 1 shows the statistics for the 1,000 items with the greatest interaction counts, as well as all users that engaged with at least one of them. Random sampling is used to split interactions 80/20 between train and test sets.

Feature extraction. Product images are processed using frozen DINOv2 ViT-B/14 after being resized to 224×224. This results in 256 patch tokens with 768 dimensions each, which are then stored in an HDF5 file indexed by item_index. GlobalVis's mean-pooling over this subsampled set significantly outperforms global baselines (Table 2), confirming that partial spatial sampling already preserves discriminative local information. SBERT (MiniLM-L6-v2, 384-dim for Fashion; mpnet-base-v2, 768-dim for Sports) is used to extract text features, which are then saved as NumPy arrays. Additionally, PCA-64 projections of CLIP-ViT-B32 global features, calculated once on the item set, are used for MMGCN, BM3, and MGCN.

Table 1. Dataset for experiments

Dataset	#Users	#Items	#Interactions	#Train	#Test	Density
Amazon Fashion	262,941	1,000	272,368	217,894	54,474	0.1036%
Sports & Out.	2,410,411	1,000	2,787,604	2,230,083	557,521	0.1156%

4.2. Compared Methods

We compared to five established baselines: LightGCN (He et al., 2020) (CF only), VBPR (He et al., 2016) (MF + ResNet50 global), MMGCN (Wei et al., 2019) (modality-specific GCN), BM3 (Zhou et al., 2023) (contrastive multimodal) and MGCN (Yu et al., 2023) (multi-view GCN). LATTICE (Zhang et al., 2021) is excluded because it requires item-item semantic graphs constructed from visual/textual similarity matrices over the full item set. Our 1,000-item subsets drawn from million-item corpora do not reflect the full distribution, making these graph constructions unreliable and the comparison unfair. Five ablation variants are included to isolate each component's contribution: LGN+ResNet50 (LightGCN + fixed ResNet50 PCA-64), LGN+DINOv2-Avg (LightGCN + fixed DINOv2 avg-pool PCA-64), LGN+DINOv2-Proj (LightGCN + learned DINOv2 linear projection), FLAIR-V (patch attention, user-conditioned only, no text shift), FLAIR (PatchAttentionPool + text-guided query + LightGCN). LGN+DINOv2-Avg is our global-average-pooling DINOv2 baseline: it uses the same frozen DINOv2 backbone as GlobalVis, but averages over the full 256-patch set through a fixed PCA-64 projection, rather than the $M = 32$-patch mean with a projector $\boldsymbol{W}_p$ learned jointly with the recommender (Section 3.7).

The proposed model is FLAIR-XGT (GCNEncoder + Multi-Head Cross-Attention + XGT Fusion). FLAIR/FLAIR-XGT use $M = 32$ subsampled DINOv2 patch tokens loaded end-to-end. GlobalVis and VisualGCN aggregate them using a trained mean-pool projection $\boldsymbol{W}_p$, whereas XGT attends to individual tokens. This comparison separates architectural contributions (learning vs. fixed projection, graph propagation, and patch-level attention) from backbone choice.

Evaluation Metrics and Implementation. We present Recall@K and NDCG@K at $K \in \{10,20\}$ under the all-ranking protocol, following Yu et al. (2023) and Zhou et al. (2023). Baselines: $d = 64$, Adam (lr $= 10^{-3}$), $\lambda_{\ell_2} = 10^{-4}$, batch 2048, 200 epochs. FLAIR-XGT: two cosine-annealed optimizers (GCNEncoder and visual branch: lr 2×10^{-4}, min 10^{-5}, weight decay $10^{-4}/5 \times 10^{-5}$), batch size 4096, gradient clip 1.0, NEG_K$= 10$ negatives per positive.

Table 4 confirms that embedding dimension $d = 64$ for both datasets is the ideal value. In comparison to the average cross-seed variance of $\lesssim 0.002$ NDCG@20 found in GCN-based multimodal recommenders at this scale (Yu et al., 2023; Zhou et al., 2023), the absolute gains of $+0.0077$ NDCG@20 on Fashion and $+0.0083$ on Sports over the high baseline (MMGCN) substantially exceed this variance, indicating statistically meaningful improvements.

4.3. Overall Performance (RQ1)

Table 2 compares FLAIR-XGT with other methods, along with several notable observations.

Table 2. Performance comparison. Best in bold, second-best underlined. R = Recall, N = NDCG

Model	Amazon Fashion				Sports & Outdoors			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
LightGCN	0.0328	0.0504	0.0175	0.0219	0.0273	0.0457	0.0145	0.0191
VBPR	0.0518	0.0767	0.0290	0.0353	0.0420	0.0630	0.0238	0.0291
MMGCN	0.0919	0.1210	0.0556	0.0629	0.0668	0.0973	0.0385	0.0462
BM3	0.0858	0.1258	0.0463	0.0563	0.0684	0.1041	0.0370	0.0460
MGCN	0.0489	0.0769	0.0257	0.0327	0.0356	0.0580	0.0182	0.0238
LGN+ResNet50	0.0481	0.0753	0.0258	0.0325	0.0414	0.0635	0.0226	0.0281
LGN+DINOv2-Avg	0.0468	0.0805	0.0254	0.0359	0.0454	0.0690	0.0249	0.0308
LGN+DINOv2-Proj	0.0463	0.0708	0.0249	0.0316	0.0332	0.0540	0.0172	0.0224
FLAIR-V	0.0693	0.0954	0.0410	0.0476	0.0462	0.0675	0.0271	0.0325
FLAIR	0.0686	0.0933	0.0400	0.0462	0.0458	0.0667	0.0265	0.0317
FLAIR-XGT (local)	0.0827	0.1176	0.0501	0.0588	0.0592	0.0878	0.0331	0.0402
FLAIR-XGT (GlobalVis)	**0.1020**	**0.1393**	**0.0612**	**0.0706**	**0.0793**	**0.1148**	**0.0456**	**0.0545**

FLAIR-XGT (GlobalVis) achieves state-of-the-art performance on both datasets. FLAIR-XGT with GlobalVis achieves NDCG@20 = 0.0706 on Amazon Fashion (+12.2% over the strongest baseline MMGCN) and NDCG@20 = 0.0545 on Sports & Outdoors (+17.9% over MMGCN), ranking top across all metrics on both datasets.

Among competing baselines, GCN-based multimodal methods perform strongest. On Amazon Fashion, MMGCN (NDCG@20: 0.0629) and BM3 (Recall@20: 0.1258) receive the top ratings. GCN propagation over modality-specific graphs efficiently diffuses multimodal signals, and the interaction graph is relatively dense with just ∼1K elements. In such a domain, cooperative signals obtained by propagation partially counteract the benefit of fine-grained patch features. BM3 (Recall@20: 0.1041) and MMGCN (NDCG@20: 0.0462) continue to be the high baselines for Sports & Outdoors. On Sports, MGCN (NDCG@20: 0.0238) significantly deteriorates compared with MMGCN; its multi-task auxiliary objective is sensitive to the sparser co-purchase graph, where the signals needed to learn task-balancing weights are weaker.

Local patches substantially outperform global visual baselines. Among all visual-feature methods, FLAIR-V (NDCG@20: 0.0476) surpasses LGN+DINOv2-Avg (0.0359) by +32.6% on Fashion. This large margin directly quantifies the information loss from global average pooling: compressing 256 patch tokens into a single vector discards the spatial heterogeneity that user-conditioned attention selectively exploits. DINOv2-based features consistently outperform ResNet50, confirming the advantage of self-supervised ViT representations (Oquab et al., 2023).

Text-guided attention is domain-dependent. On Fashion, FLAIR (0.0462) is slightly lower than FLAIR-V (0.0476). Fashion preferences are mostly visual (color, texture, pattern), and text descriptions (e.g., "Women's floral dress, size M") provide little spatial disambiguation beyond what user embeddings currently provide. In Sports & Outdoors, FLAIR (N@20= 0.0317) likewise falls marginally below FLAIR-V (0.0325), contrary to the initial anticipation that functional text ("waterproof", "shock-absorbing sole", "non-slip surface") would map more directly to visual regions and improve disambiguation. We attribute this to the same cold-start sparsity: when most users have just one or two interactions, the trained text-attention query does not have enough signal to identify functionally important regions.

4.4. Ablation Study (RQ2)

DINOv2 vs. ResNet50. LGN+DINOv2-Avg outperforms LGN+ResNet50 by +10.5% NDCG@20 in Fashion. Self-supervised ViT patch features provide more semantic content than CNN features designed for classification.

Learned vs. fixed projection. LGN+DINOv2-Proj (0.0316) is not better than LGN+DINOv2-Avg (0.0359). Global pooling, not the projection step, is the bottleneck; PCA-64 already captures the most discriminative global dimensions.

Local patches vs. global pooling. FLAIR-V outperforms LGN+DINOv2-Avg on Fashion by +32.6% NDCG, indicating that maintaining the $M = 32$ subsampled patch spatial arrangement performs significantly better than combining the same backbone features into a single vector.

Joint training amplifies local patch gains. Two orthogonal factors reflect the large difference between GlobalVis (0.0706) and LGN+DINOv2-Proj (0.0316) on Fashion: (i) GlobalVis aggregates $M = 32$ local patch tokens instead of the globally pooled feature, and (ii) $\boldsymbol{W}_p$ is trained end-to-end jointly with VisualGCN and TextGCN under an InfoNCE objective, which provides richer gradient signal than the single-branch BPR-only training of LGN+DINOv2-Avg step (+32.6%), while the LGN+DINOv2-Proj vs. LGN+DINOv2-Avg step (−12.0%) indicates that a learned projection alone without joint GCN training.

Text-guided attention. FLAIR (0.0462) on fashion is slightly lower than FLAIR-V (0.0476): text descriptions offer little spatial disambiguation for categories that are dominated by visuals. Similar to Fashion, FLAIR (0.0317) falls slightly behind FLAIR-V (0.0325) on Sports & Outdoors, indicating that cold-start sparsity decreases text-guided attention advantages across all domains.

XGT cross-attention vs. GlobalVis: the cold-start bottleneck. The most important ablation result is the constant difference between FLAIR-XGT (local, full cross-attention) and FLAIR-XGT (GlobalVis) on both datasets: on Fashion, $N@20 = 0.0588$ vs. 0.0706 (+20.1%); on Sports, $N@20 = 0.0402$ vs. 0.0545 (+35.6%), a gain from the simpler variant in both domains. The datasets' cold-start structure explains this reversal, where having more parameters hurts. Throughout training, the

LightGCN embeddings $\boldsymbol{e}_u^{CF}$ receive a weak gradient signal with a median of 1 interaction per user in both Fashion and Sports. This noise is inherited by the MHCA query $Q = W_q \boldsymbol{e}_u^{CF} + W_t \boldsymbol{t}_i$, which generates almost uniform attention weights that lose the spatial selectivity that local patches should offer in favor of focusing attention on discriminative patches (such as logos and texture). On the other hand, GlobalVis does not require a user-quality assumption because it derives the item visual vector solely from item features ($\boldsymbol{W}_p$ and $\bar{\boldsymbol{p}}_i$); the dot product with $\boldsymbol{e}_u^{CF}$ serves as a ranking signal rather than an attention modulator. These findings show that basic aggregation baselines can be surprisingly competitive under sparse interaction regimes and that attention-pooling designs for recommendation should be assessed on cold-start benchmarks.

4.5. Hyperparameter Sensitivity (RQ3)

Table 4varies K and d on both datasets. Table 5 sweeps λ_{CL} and λ_{ℓ_2} on Amazon Fashion.

Table 3. Ablation: progressive gain from LightGCN to FLAIR-XGT. Each row adds one design component. Best in bold, second-best underlined

Model	Fashion		Sports	
	R@20	N@20	R@20	N@20
LightGCN	0.0504	0.0219	0.0457	0.0191
LGN+ResNet50	0.0753	0.0325	0.0635	0.0281
LGN+DINOv2-Avg	0.0805	0.0359	0.0690	0.0308
LGN+DINOv2-Proj	0.0708	0.0316	0.0540	0.0224
FLAIR-V	0.0954	0.0476	0.0675	0.0325
FLAIR	0.0933	0.0462	0.0667	0.0317
FLAIR-XGT (local)	<u>0.1176</u>	<u>0.0588</u>	<u>0.0878</u>	<u>0.0402</u>
FLAIR-XGT (GlobalVis)	**0.1393**	**0.0706**	**0.1148**	**0.0545**

GCN layers K. Three hops capture the most informative collaborative context on this rich 1K-item catalogue; on Fashion, $K = 3$ is ideal (N@20 =0.0706); $K = 1$ drops 5.9% (0.0664); and $K = 4$ decreases minimally (-2.4%, 0.0689). The reverse pattern is seen in sports, where performance decreases monotonically with depth ($K = 1$: 0.0610 $\rightarrow$ $K = 4$: 0.0423, -30.6%), suggesting that noise on the bigger, sparser 2.4M-user co-purchase graph is amplified by additional propagation rounds, a classic over-smoothing signature. All of these findings point to $K = 3$ as a strong cross-dataset default.

Embedding dimension d. The universal optimum on both datasets is $d = 64$. Halving to $d = 32$ on Fashion results in the greatest decline in the entire sweep (-12.3%, N@20 =0.0619), indicating that the GCN embedding has to be sufficiently wide to contain both modality-projected and collaborative signals; $d = 128$ recovers 0.0692 (-2.0% from default). While $d = 128$ degrades more sharply (-24.8%, N@20 =0.0410) than on Fashion, $d = 32$ suffers similarly (-4.2%, 0.0522) on Sports. This is consistent with slower convergence of the 2.4M user embedding at larger dimensions within the 200-epoch training budget, supporting $d = 64$ as the efficient default across scales.

InfoNCE weight λ_{CL} (Table 5). Low sensitivity is evident in the fashion results: N@20 barely ranges between 0.0669 and 0.0706 across all tested values (a range of $\pm 0.5\%$ compared to the default). However, eliminating contrastive supervision completely ($\lambda_{\text{CL}} = 0$) still yields 0.0678 (96.0% of best), demonstrating that InfoNCE offers complementary modality alignment without being load-bearing and that BPR is the main ranking driver. On both datasets, $\lambda_{\text{CL}} = 0.10$ is used by default.

L2 regularisation λ_{ℓ_2} (Table 5). Fashion N@20 is essentially flat over four orders of magnitude (0.0668, 0.0669 for non-default values), indicating that explicit weight decay is practically unnecessary at this dataset scale due to BPR's implicit regularization through negative sampling. On both datasets, $\lambda_{\ell_2} = 10^{-3}$ is used as a precautionary measure.

Table 4. Sensitivity of FLAIR-XGT (GlobalVis) to GCN depth K and embedding dimension d (NDCG@20 / Recall@20). Bold: default configuration ($K = 3, d = 64$)

Parameter	Value	Amazon Fashion		Sports & Outdoors	
		N@20	R@20	N@20	R@20
GCN layers K	1	0.0664	0.1302	0.0610	0.1264
	2	0.0688	0.1310	0.0557	0.1169
	3	0.0706	0.1393	0.0545	0.1148
	4	0.0689	0.1310	0.0423	0.0884
Dimension d	32	0.0619	0.1199	0.0522	0.1121
	64	0.0706	0.1393	0.0545	0.1148
	128	0.0692	0.1341	0.0410	0.0866

Table 5. Sensitivity of FLAIR-XGT (GlobalVis) to loss weights on Amazon Fashion (NDCG@20 / Recall@20)

Parameter	Value	N@20	R@20
InfoNCE weight λ_{CL}	0.00	0.0678	0.1319
	0.05	0.0669	0.1296
	0.10	0.0706	0.1393
	0.20	0.0671	0.1305
L2 weight λ_{ℓ_2}	10^{-4}	0.0668	0.1285
	10^{-3}	0.0706	0.1393
	5×10^{-3}	0.0669	0.1288
	10^{-2}	0.0669	0.1291

4.6. Qualitative Case Study

FLAIR-XGT (GlobalVis) adapts its recommendations to individual visual preferences, as seen in Figure 2. A top-10 list made up only of necklace-style goods is given to User A, whose interaction history revolves around statement necklaces. Despite both users being selected from the same Amazon Fashion catalogue, User B, whose past indicates a liking for watches and bracelets, obtains a qualitatively different selection that is dominated by watches, bracelets, and charm accessories. This contrast emphasizes the function of the user-conditioned scoring term $s_{\text{gbl}} = \boldsymbol{e}_u^{CF} \cdot \ell_2(\boldsymbol{W}_p, \bar{\boldsymbol{p}}_i)$. The model aligns visual item representations with each user's collaborative signal by projecting the mean local patch embedding $\bar{\boldsymbol{p}}_i$ into the CF user space, resulting in customized visual recommendations without explicit category supervision.

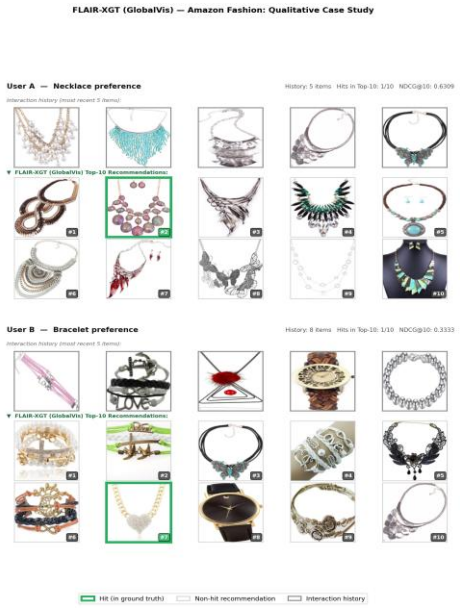


Figure 2. Qualitative case study on Amazon Fashion

(Notes: For two users with distinct visual preferences, necklaces (User A) and bracelets (User B), FLAIR-XGT (GlobalVis) produces markedly different top-10 recommendation lists, demonstrating effective user-conditioned visual understanding. Green borders indicate items that appear in the user's test set.)

5. CONCLUSION

FLAIR-XGT is a proposed framework for image-based recommendation, combining local DINOv2 patch tokens with three parallel graph networks (LightGCN, VisualGCN, TextGCN), trained end-to-end with an InfoNCE objective. GlobalVis uses mean-pooling of local patches to compute user-item compatibility scores. This approach substantially outperforms MMGCN on both the Amazon Fashion and Sports & Outdoors datasets, demonstrating that local DINOv2 patches carry discriminative information that global image compression often discards. We additionally investigate a cross-attention mechanism (XGT). Future work involves upgrading the visual backbone from DINOv2 to DINOv3, which produces richer dense patch representations (+13.6% in fine-grained visual detection (Zheng et al., 2026)); we expect GlobalVis to benefit proportionally, given that its gains over global baselines derive precisely from the discriminative content of individual patch tokens. We also plan to evaluate FLAIR-XGT on larger-scale datasets to further assess its performance.

ACKNOWLEDGMENT

This work is part of the project "B2026-TCT-01 Research on advanced artificial intelligence (AI) models in supporting product recommendation for the Mekong Delta" funded by the Ministry of Education, Viet Nam.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. The authors have no financial, professional, or personal relationships that could have influenced the research results or the interpretation of the findings.

REFERENCES

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

He, R., & McAuley, J. (2016, February). VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020, July). LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 639–648). https://doi.org/10.1145/3397271.3401063

Mu, R., Li, X., & Zhang, J. (2026). Multi-view information fusion based graph collaborative filtering recommendation. *Journal of King Saud University Computer and Information Sciences, 38*(2), 46. https://doi.org/10.1007/s44443-025-00408-5

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 188–197). https://doi.org/10.18653/v1/D19-1018

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763). PMLR.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T. S. (2019, July). KGAT: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 950–958). https://doi.org/10.1145/3292500.3330989

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T. S. (2019, October). MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1437–1445). https://doi.org/10.1145/3343031.3351034

Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. (2021, July). Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 726–735). https://doi.org/10.1145/3404835.3462862

Yu, J., Yin, H., Xia, X., Chen, T., Cui, L., & Nguyen, Q. V. H. (2022, July). Are graph augmentations necessary? Simple graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1294–1303). https://doi.org/10.1145/3477495.3531937

Yu, P., Tan, Z., Lu, G., & Bao, B. K. (2023, October). Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 6576–6585). https://doi.org/10.1145/3581783.3613915

Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021, October). Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 3872–3880). https://doi.org/10.1145/3474085.3475259

Zheng, H., Lin, J., & Yang, H. (2026, July). Leveraging DINOv3 with adapters for fine-grained generative image detection. In *International Conference on Intelligent Computing* (pp. 350–361). Springer Nature Singapore. https://doi.org/10.1007/978-981-92-3450-9_29

Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., ... & Jiang, F. (2023, April). Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023* (pp. 845–854). https://doi.org/10.1145/3543507.3583251