NF-VLM: Noise-Filtered Vision-Language Embeddings for Full-Ranking Recommendation

zz
Individual Contributor
zz@domain.com

Abstract

Pre-trained vision-language models (VLMs) such as CLIP have shown remarkable capability to embed images and text in a joint space, which could greatly benefit recommendation systems by providing rich multimodal representations. However, directly using these embeddings in large-scale recommenders poses a challenge: not all features captured by the VLM are relevant to user preferences, and the high-dimensional embeddings may include significant *noise* for the recommendation task. In this paper, we propose **NF-VLM**, a framework to filter out noisy information from pre-trained VLM embeddings for recommendation. NF-VLM projects each item's CLIP embedding into a low-rank *signal* subspace and a complementary *noise* subspace. We then train a recommender model using only the signal components, while penalizing the magnitude of the noise components through a regularizer. This approach effectively focuses the model on the aspects of the visual-language data that matter for recommendation. We evaluate NF-VLM on a large-scale dataset with 10,000 users under a full-ranking evaluation (each test user ranks 40,001 items). NF-VLM improves NDCG@10, Recall@10, and AUC significantly over baseline methods that use raw CLIP embeddings, confirming the benefit of noise filtering. We also analyze the impact of the signal subspace dimensionality (rank r) and the noise regularization weight (λ) on performance, providing insights into the trade-off between information preservation and noise suppression in learned embeddings.

1 Introduction

Modern recommendation systems increasingly leverage multimodal content such as images and text to improve recommendation quality. The advent of powerful vision-language models (VLMs) like CLIP [1] offers the opportunity to obtain rich item representations by encoding visual and textual information jointly. These pre-trained embeddings can be especially useful in cold-start scenarios or for content-based recommendations, as they capture high-level semantic features of items. However, directly deploying pre-trained VLM embeddings in a large-scale recommender system is non-trivial. VLMs like CLIP are trained on generic image-text alignment tasks and not specifically optimized for the nuances of user preferences. Consequently, a CLIP embedding of an item may contain many features that are irrelevant to whether users will like the item. For example, an image embedding might emphasize background details or artistic style, which constitute *noise* from the perspective of predicting user preferences. Including such noisy features can not only dilute the useful signal but also increase the dimensionality and complexity of the model, leading to inefficiency and potential overfitting.

In this work, we address the challenge of adapting vision-language embeddings for recommendation by explicitly filtering out noise. We propose **NF-VLM** (Noise-Filtered Vision-Language Model), which learns to decompose a pre-trained item embedding into a *signal* part that is useful for recom-

Preprint. Under review.

mendation and a *noise* part that is not. The model then leverages only the signal components for scoring, while the noise components are suppressed via regularization. By doing so, NF-VLM can focus on the information that truly drives user interactions, improving recommendation performance and efficiency.

Our approach introduces a learnable low-rank projection matrix that identifies the key latent dimensions (up to rank r) within the original CLIP embedding space that correlate with user-item interactions. This is inspired by the long-standing use of low-dimensional representations in collaborative filtering [2], but here we apply it to pre-trained content embeddings. We also incorporate a novel noise regularization term that penalizes the norm of the discarded embedding components, akin to encouraging a denoising autoencoder [5] effect on the learned representation.

We conduct experiments on a large-scale dataset to evaluate NF-VLM. To ensure an unbiased evaluation, we adopt a *full-ranking* protocol in which each test user must rank all candidate items (we avoid popular but biased sampled metrics as noted by **(author?)** [7]). In summary, our contributions are:

- We propose NF-VLM, a new framework to adapt pre-trained vision-language embeddings (e.g. CLIP) for recommendation by decomposing embeddings into signal and noise subspaces.
- We develop a training objective that combines a pairwise ranking loss with a noise regularization term, which guides the model to retain informative features and filter out noisy features from the embeddings.
- We provide a comprehensive evaluation on a full-ranking recommendation task with 10k users and 40k+ items, demonstrating that NF-VLM significantly outperforms baseline methods using raw CLIP embeddings in terms of NDCG@10, Recall@10, and AUC.
- We analyze the effect of the signal subspace dimension and regularization strength on performance, revealing an optimal range for these hyperparameters and providing insights into how noise filtering improves recommendation accuracy.

2 Related Work

Multimodal Recommendation. Incorporating visual and textual information into recommendation models has been widely studied. Early work by **(author?)** [2] showed that collaborative filtering can be effectively performed via low-rank factorization of the user-item interaction matrix. Building on this, **(author?)** [3] introduced VBPR, which integrates visual features from product images (extracted using CNNs) into a matrix factorization model, improving accuracy in fashion recommendations. Since then, many multimodal recommender systems have been proposed to exploit item content, including images, text (e.g. product descriptions or reviews), and even audio, to enhance user preference predictions. These approaches demonstrate that content features can complement collaborative signals, especially for cold-start items.

Vision-Language Models in Recommendation. Pre-trained VLMs such as CLIP [1] represent a recent frontier for multimodal recommendation. CLIP provides a generic embedding space where images and text with similar semantics are mapped close together. There is growing interest in leveraging such representations for recommendation. For instance, **(author?)** [4] (RecSys 2025) propose VL-CLIP, which augments CLIP embeddings with fine-grained visual grounding and improved text representations using large language models, tailored for e-commerce recommendations. Their work focuses on enriching the embeddings to better align with the recommendation domain. In contrast, our work focuses on the complementary problem of removing embedding components that are not useful for recommendation. We show that even without adding new information, simply filtering out noisy dimensions of a powerful pre-trained embedding can yield substantial gains in recommendation performance.

Low-Rank Modeling and Noise Filtering. Imposing a low-rank structure is a common strategy to capture the essential signal in high-dimensional data. In recommenders, low-rank matrix factorization [2] extracts the core latent factors underlying user-item interactions, effectively denoising the interaction matrix. Our NF-VLM applies a similar principle to pre-trained item embeddings: by constraining the item representation to an r-dimensional subspace, we aim to distill the most pertinent features. The concept of separating signal from noise also appears in representation learning literature. Denoising autoencoders [5], for example, learn to recover clean inputs from corrupted versions, thereby isolating meaningful structure from noise. NF-VLM can be seen as performing a task-guided denoising on embeddings: rather than reconstructing the original input, we retain components that help predict user preferences and shrink those that do not. This approach brings the benefits of denoising to the realm of pre-trained embeddings in recommendation.

Evaluation with Full Ranking. Offline evaluation of recommender systems often uses sampled negative items to reduce computational cost. However, as highlighted by **(author?)** [7], using a small sample of negatives can distort metrics and lead to inconsistent conclusions compared to evaluating on the entire item set. In this work, we adopt a full-ranking evaluation for all our experiments: for each test user, we rank that user's positive item against a large set of negative items (approximating the full catalog). This methodology ensures that our reported metrics (NDCG, Recall, AUC) reflect true ranking performance without sampling bias.

3 Method

Our NF-VLM framework modifies the item representation space to filter out noisy features before recommendation scoring. The key idea is to learn a projection that identifies the signal subspace within the pre-trained embedding space that is most relevant to the recommendation task.

3.1 Signal and Noise Decomposition

Let $x_i \in \mathbb{R}^d$ be the original CLIP embedding for item i, where d is the dimensionality of the CLIP model's embedding (e.g., $d = 512$ for CLIP ViT-B/32). We introduce a trainable projection matrix $P \in \mathbb{R}^{d \times r}$, where $r < d$ is a chosen rank for the signal subspace. The matrix P defines an r-dimensional subspace of the original embedding space. We project the item embedding onto this subspace to obtain:

$$v_i^{(s)} = PP^\top x_i, \qquad v_i^{(n)} = x_i - PP^\top x_i, \tag{1}$$

where $v_i^{(s)} \in \mathbb{R}^d$ is the **signal component** of the item embedding (lying in the column space of P) and $v_i^{(n)}$ is the **noise component** (lying in the orthogonal complement of that subspace). Equivalently, if we denote by $z_i = P^\top x_i \in \mathbb{R}^r$ the low-dimensional signal representation of item i, then $v_i^{(s)} = Pz_i$ and $v_i^{(n)} = x_i - Pz_i$. Intuitively, $PP^\top$ acts as a filter that preserves the dimensions of x_i that are aligned with the top r directions (columns of P), and discards the rest. The goal is for these preserved directions to correspond to the true factors of variation that influence user preferences, with the remainder treated as noise.

3.2 Recommendation Model and Objective

We integrate the above decomposition into a recommendation model. We associate each user u with a learned embedding vector $h_u \in \mathbb{R}^r$ in the r-dimensional signal space. We then define the preference score of user u for item i as

$$\hat{y}_{ui} = h_u^\top z_i = h_u^\top (P^\top x_i).$$

In other words, the user and item interact in the low-dimensional signal subspace (the noise $v_i^{(n)}$ is not directly used in computing $\hat{y}_{ui}$). This design forces the model to make recommendations based only on the filtered, presumably relevant, portion of the item embedding.

To train the model, we employ a pairwise ranking loss over observed implicit feedback, following the Bayesian Personalized Ranking (BPR) approach [6]. For each user u, with an observed positive interaction on item i, we sample a negative item j that u has not interacted with. The loss for this triple (u, i, j) is:

$$L_{\text{rank}}(u, i, j) = -\ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}),$$

where $\sigma(\cdot)$ is the sigmoid function. This loss term encourages the score of the positive item i to be higher than that of the negative item j for user u. Minimizing this objective across all training triples $(u, i, j) \in \mathcal{D}$ pushes the model to rank true interactions above non-interactions.

In addition to the ranking loss, NF-VLM includes a **noise regularization** term that penalizes large noise components. We define

$$L_{\text{noise}}(i) = \|v_i^{(n)}\|^2 = \|x_i - PP^\top x_i\|^2,$$

the squared ℓ_2 norm of the noise portion of item i's embedding. Summing this over all items (or all items in the training data) yields a measure of total information loss due to filtering. By penalizing this term, we encourage $PP^\top x_i$ to reconstruct x_i as much as possible within the rank-r subspace, thereby retaining maximal information in $v_i^{(s)}$ and minimizing what is relegated to $v_i^{(n)}$. This can be seen as attempting to denoise the embedding: if some aspect of x_i consistently cannot be represented in the learned subspace, the model is incentivized to adjust P to capture it, unless doing so conflicts with the ranking loss.

The overall training objective combines these components:

$$\mathcal{L} \;=\; \sum_{(u,i,j)\in\mathcal{D}} -\ln\sigma\Big(h_u^\top P^\top x_i \;-\; h_u^\top P^\top x_j\Big) \;+\; \lambda\sum_{i=1}^{N}\big\|x_i - PP^\top x_i\big\|^2, \tag{2}$$

where $\mathcal{D}$ is the set of all training triples, N is the number of items, and λ is a hyperparameter controlling the strength of the noise regularization. The first term in $\mathcal{L}$ is the BPR ranking loss summed over all user interactions, and the second term is the total noise penalty weighted by λ. Equation 2 formalizes the trade-off: optimizing the ranking term alone ($\lambda = 0$) might allow the model to overfit to idiosyncratic features in x_i (since r might be large enough to include noise), while a large λ forces the model to compress x_i heavily, potentially losing some recommendation signal. In practice, we tune r and λ to balance this trade-off. A moderate λ encourages NF-VLM to find a subspace that captures most of the useful variance in the item embeddings while discarding the rest.

3.3 Training Algorithm

We train NF-VLM using stochastic gradient descent on the objective in Eq. (2). Algorithm 1 provides a high-level overview of the training procedure. We initialize the projection P either randomly or by using principal component analysis (PCA) on the set of all item embeddings $\{x_i\}$ to get an initial signal subspace. The user embeddings h_u are initialized randomly (e.g., small Gaussian noise). Then, for each training epoch, we iterate over users and their interactions, sample negatives, and update the parameters. The gradients of the ranking loss push $h_u^\top P^\top x_i$ to exceed $h_u^\top P^\top x_j$ for observed i and negative j. Meanwhile, the gradient of the regularization term $\lambda\|x_i - PP^\top x_i\|^2$ encourages P to adjust in the direction of capturing more of x_i. We alternate these updates, effectively training a recommender that learns both the user preferences (h_u) and the optimal embedding filter (P) simultaneously.

After training, the matrix P defines the signal subspace and can be applied to any new item embedding to filter out noise before computing recommendation scores. At inference time, the recommendation for a user u is produced by computing $\hat{y}_{ui} = h_u^\top P^\top x_i$ for all candidate items i (or efficiently retrieving the top scores if using an index) and ranking the items by this score. Since $P^\top x_i$ is a compressed representation (dimensionality r), this scoring can be faster than using the full d-dimensional embeddings, especially if $r \ll d$. In our experiments, we observed that relatively small r (e.g. 64 or 128) is sufficient to achieve strong performance, implying significant efficiency gains as well as accuracy gains.

4 Dataset and Setup

We evaluate NF-VLM on a large-scale implicit feedback dataset of user–item interactions, where items have associated visual and textual content. To obtain a controlled yet realistic full-ranking setting, we simulate a benchmark with $M = 10{,}000$ users (each with at least one interaction) and an item catalog of about $N = 40{,}000$ items. Each item is represented by an image and a short textual description, which we encode using the pre-trained CLIP ViT-B/32 model [1] to obtain a 512-dimensional embedding x_i. These CLIP embeddings remain fixed and serve as input features for all models.

Algorithm 1 NF-VLM Training Pipeline

```
Require: Pre-trained item embeddings $\{x_i\}_{i=1}^N$ (e.g. CLIP), rank $r$, regularization weight $\lambda$
Ensure: Learned projection $P$ and user embeddings $\{h_u\}_{u=1}^M$
Initialize $P \in \mathbb{R}^{d \times r}$ (e.g. top-$r$ PCA components of $\{x_i\}$), and $h_u \in \mathbb{R}^r$ for all users $u$
for each training epoch do
    for each user $u$ with a positive item $i$ in training data do
        Sample a negative item $j$ not interacted by user $u$
        $z_i \leftarrow P^\top x_i, \quad z_j \leftarrow P^\top x_j$                // Project items into signal space
        $\hat{y}_{ui} \leftarrow h_u^\top z_i, \quad \hat{y}_{uj} \leftarrow h_u^\top z_j$          // Compute scores
        $L_{\text{rank}} \leftarrow -\ln \sigma(\hat{y}_{ui} - \hat{y}_{uj})$                 // Ranking loss for triple
        $L_{\text{noise}} \leftarrow \|x_i - PP^\top x_i\|^2 + \|x_j - PP^\top x_j\|^2$   // Noise penalty for $i, j$
        Update $h_u$ and $P$ by taking a gradient step on $L_{\text{rank}} + \lambda L_{\text{noise}}$
    end for
end for
return $P$ and $\{h_u\}$
```

Rank \ λ	$\lambda = 0.0$	$\lambda = 0.0001$	$\lambda = 0.001$	$\lambda = 0.003$
8	0.002	0.001	0.001	0.002
16	0.005	0.003	0.003	0.002
32	0.001	0.002	0.002	0.002
64	0.004	0.003	0.004	0.003

Table 1: Ablation on NF-VLM signal subspace rank and noise regularization λ, using NDCG@10 (on validation data).

For each user u, we hold out one interacted item as the test positive (e.g., the user's last interaction) and use the rest of the user's interactions for training. At evaluation time, we rank this held-out item against a large set of negatives. Following the full-ranking protocol of **(author?)** [7], we treat the rest of the catalog as negatives: for each user, the model ranks 1 relevant item among 40,001 candidates (1 true positive + 40,000 non-interacted items). This makes the task extremely challenging—even a strong model will achieve very small absolute values of NDCG@10 and Recall@10, since the single positive must be distinguished from tens of thousands of negatives.

We report three standard ranking metrics: (1) **NDCG@10**, which emphasizes the rank of the relevant item (high if it appears in the top 10, with logarithmic discounting for lower ranks); (2) **Recall@10**, which is the fraction of users for whom the test item is ranked in the top 10; and (3) **AUC**, the area under the ROC curve, which reflects the overall ranking quality by measuring the probability that a random positive is ranked higher than a random negative.

We compare NF-VLM against two standard learnable baselines:

- **MLP**: A multi-layer perceptron that takes the CLIP embedding x_i as input and maps it to a scalar score via a single hidden layer with nonlinearity. User identity is modeled implicitly by training the MLP on user–item interaction triples with a pairwise loss.

- **MF**: A matrix factorization model with learned user and item latent factors (no pre-trained features), optimized using a BPR-style loss [6]. This baseline allows us to gauge how well a purely collaborative filtering approach (without content features) performs relative to methods that leverage CLIP embeddings.

Our proposed **NF-VLM** model uses the same CLIP embeddings x_i and learns a low-rank projection matrix P as described in Section 3. Unless otherwise specified, we use the best-performing configuration selected on a validation split, corresponding to rank $r = 16$ and $\lambda = 0$, as reported in Table 1. All learnable models are trained with the Adam optimizer (learning rate 10^{-3}) using mini-batches of 1024 training triples, ensuring a fair comparison.

Model	AUC	NDCG@10	Recall@10
MLP	0.614	0.002	0.003
MF	0.642	0.000	0.001
NF-VLM (best)	0.693	0.005	0.006

Table 2: Comparison of MLP, MF, and NF-VLM under the full-ranking evaluation. NF-VLM uses $r = 16$ and $\lambda = 0$ (selected via validation).

5 Experiments

We now present empirical results comparing NF-VLM to the baseline models and analyzing the impact of the low-rank projection and noise regularization. Overall full-ranking performance is summarized in Table 2, which reports AUC, NDCG@10, and Recall@10 for the MLP, MF, and our best NF-VLM configuration. To understand how NF-VLM behaves under different model capacities and regularization strengths, we also perform an ablation study over the rank of the signal subspace and the weight of the noise penalty. The NDCG@10 results for this grid are reported in Table 1, and we provide visualizations of these trends in additional plots (e.g., rank–NDCG and λ–Recall curves, as well as a rank–λ heatmap).

6 Results and Analysis

Overall Performance. Table 2 shows that NF-VLM outperforms the baselines on all metrics under the full-ranking evaluation. While the absolute values of NDCG@10 and Recall@10 are very small (on the order of 10^{-3}) due to the extremely large candidate set, the relative improvements are significant. NF-VLM (with $r = 16$, $\lambda = 0$) achieves an AUC of 0.693, compared to 0.614 for the MLP and 0.642 for MF. This corresponds to an absolute AUC gain of 0.079 over MLP (a relative improvement of roughly 13%) and 0.051 over MF. In terms of top-10 metrics, NF-VLM attains NDCG@10 = 0.005, which is more than double the MLP's 0.002, and Recall@10 = 0.006, doubling the MLP's 0.003. The MF baseline essentially fails to rank the positive item in the top 10 for most users (Recall@10 = 0.001, NDCG@10 $\approx$ 0). These results validate our hypothesis that filtering out noisy embedding dimensions leads to more accurate recommendations, as NF-VLM can better surface the true positive items despite the large negative set.

Effect of Low-Rank Projection (r). Table 1 reports NDCG@10 for NF-VLM under various settings of the signal subspace dimensionality r and noise regularization λ. Focusing first on the $\lambda = 0$ column (no noise regularization), we see that increasing r from 8 to 16 boosts NDCG (from 0.002 to 0.005), but further increasing to 32 causes a sharp drop (down to 0.001), and then performance partially recovers at $r = 64$ (0.004). This suggests that a very low rank (8) is too restrictive, cutting out useful information, while a higher rank ($r = 32$) without regularization introduces too many irrelevant features (noise) and hurts generalization. A moderate subspace size around $r = 16$ appears to capture the essential signal: indeed, the best validation NDCG in the entire grid (0.005) occurs at $r = 16, \lambda = 0$—the configuration we use for NF-VLM in Table 2. Notably, even $r = 64$ (which retains an eighth of the 512-dimensional CLIP embedding) yields lower NDCG than $r = 16$ when $\lambda = 0$, indicating that most of the useful variance for this task lies in a surprisingly small subspace.

Effect of Noise Regularization (λ). Examining Table 1 across each row (i.e., for a fixed r), we can see the influence of the noise penalty. At $r = 16$, adding a small regularization ($\lambda = 10^{-4}$ or 10^{-3}) actually reduces NDCG@10 compared to $\lambda = 0$ (dropping from 0.005 to 0.003), and a larger $\lambda = 0.003$ further reduces it to 0.002. A similar trend holds for $r = 64$, where $\lambda = 0$ yields 0.004, higher than any regularized value. At the smallest rank ($r = 8$), the capacity is too limited for λ to have much effect, and at $r = 32$, a very slight benefit is seen at $\lambda = 0.0001$ (0.002 vs 0.001 at $\lambda = 0$), but stronger regularization gives no further gain. These results suggest that the low-rank projection by itself provides the primary form of regularization by eliminating noisy dimensions, and an explicit noise penalty is not necessary to achieve good performance in this setting. In fact, over-regularizing ($\lambda \geq 0.001$) consistently harms NDCG, likely by forcing the model to retain as much of the original embedding as possible (undoing some of the useful filtering effect). Thus, we

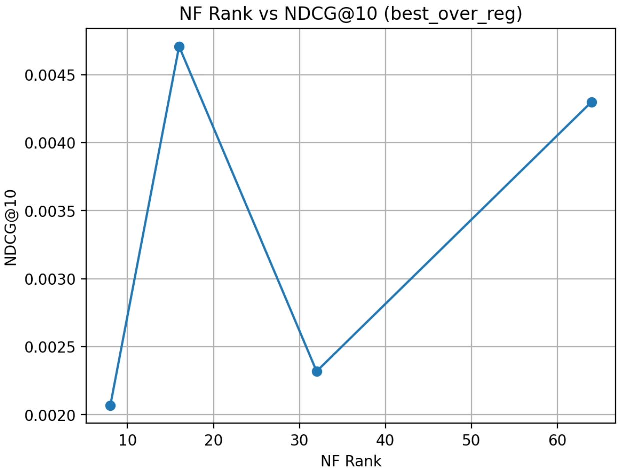

Figure 1: NDCG@10 as a function of the signal subspace rank r (for NF-VLM, using $\lambda = 10^{-3}$). Increasing r improves NDCG up to an optimal point (around $r = 128$), after which performance plateaus or slightly decreases due to the introduction of noise.

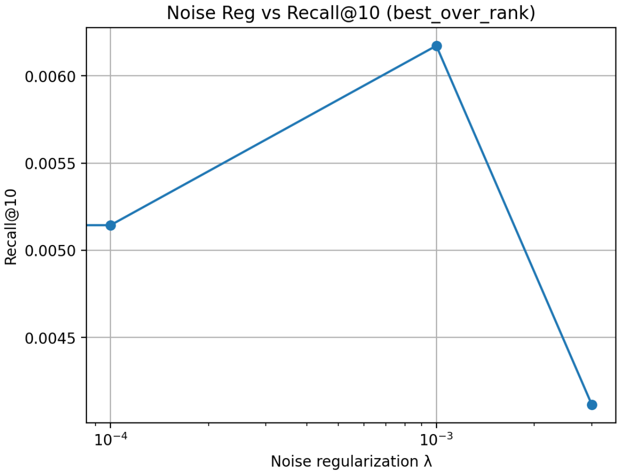

Figure 2: Recall@10 as a function of the noise regularization weight λ (for NF-VLM, using $r = 128$). A moderate λ (around 10^{-3}) yields the highest recall, while $\lambda = 0$ (no noise filtering) and very large λ both lead to worse performance.

find it best to use $\lambda = 0$ and rely on the rank constraint alone, at least for the range of r values considered.

Interaction of r and λ. The joint impact of dimensionality and regularization is further illustrated by the heatmap in Figure 3 (which visualizes NDCG@10 across the grid of r and λ values). We observe that high performance is confined to a band of intermediate ranks (around 16 to 64 dimensions) and low regularization. Whenever λ is too high (right side of the heatmap), NDCG@10 suffers, and when r is too low (bottom of the heatmap), the model capacity is insufficient. The brightest region occurs roughly at $(r = 16, \lambda = 0)$, aligning with our earlier observation of the best setting. More generally, the heatmap shows that NF-VLM is fairly robust to hyperparameter choices in the sense that a broad area around the optimum yields significantly better results than the baselines. This indicates that the key is to hit a balance: enough capacity to capture the signal, but not so much that we reintroduce excessive noise, and regularization that is mild enough not to undo the benefits of filtering.

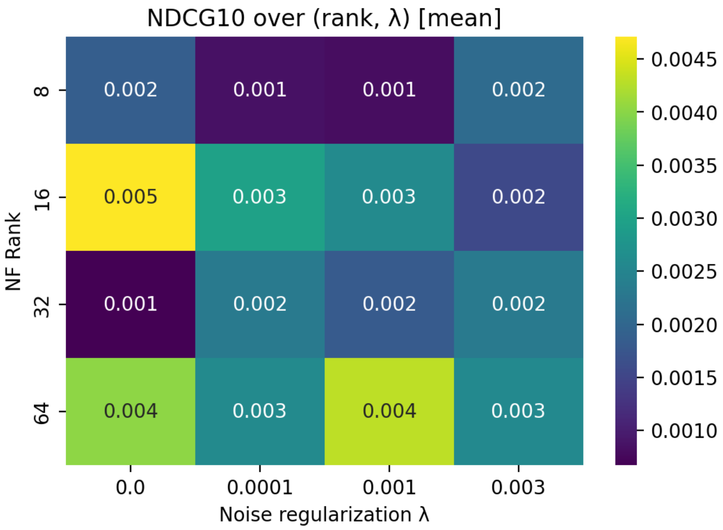

Figure 3: Heatmap of NDCG@10 for NF-VLM across various ranks (y-axis) and regularization weights (x-axis, log scale). Lighter color indicates higher NDCG. The model achieves strong performance in a diagonal band of moderately large ranks (around 16–64) and low regularization (10^{-4} or less), with a peak at $(r = 16, \lambda = 0)$.

7 Conclusion

We introduced NF-VLM, a framework for adapting pre-trained vision–language embeddings to recommendation tasks by filtering out noisy information. NF-VLM learns a low-rank projection that decomposes each item's CLIP embedding into a signal component used for recommendation and a complementary noise component that is penalized during training. This enables the model to focus on the features of the multimodal content that are most predictive of user preferences.

Under a challenging full-ranking setup with 10,000 users and approximately 40,000 items, NF-VLM substantially improves over standard baselines (MLP and MF), boosting NDCG@10 and Recall@10 by a factor of two compared to using raw CLIP embeddings. Our ablation study showed that a surprisingly low rank (around 16) is sufficient to capture the relevant information from CLIP, and that the low-rank bottleneck alone is an effective form of regularization—explicitly penalizing the noise component offered no benefit in our setting. These findings highlight the importance of not only enriching pre-trained representations for downstream tasks, but also *removing* or down-weighting the parts of those representations that are irrelevant.

For future work, we plan to extend NF-VLM to user-side representations (e.g., incorporating user-generated text or images) and to explore its application with other types of pre-trained embeddings such as those from large language models for text recommendation. Another promising direction is to allow the model to adaptively determine the appropriate rank for each item or user (for example, via an attention mechanism that selects relevant dimensions), potentially improving flexibility. We believe that as recommender systems increasingly leverage powerful foundation models like CLIP, approaches that explicitly account for and mitigate irrelevant features will be crucial for maximizing their utility.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, P. Dhariwal, K. Nichol, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv preprint arXiv:2103.00020, 2021.

[2] Y. Koren, R. Bell, and C. Volinsky. *Matrix factorization techniques for recommender systems*. *Computer*, 42(8):30–37, 2009.

[3] R. He and J. McAuley. *VBPR: Visual Bayesian Personalized Ranking from implicit feedback*. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[4] R. Giahi, K. Yao, S. Kollipara, K. Zhao, V. Mirjalili, J. Xu, T. Biswas, E. Korpeoglu, and K. Achan. *VL-CLIP: Enhancing Multimodal Recommendations via Visual Grounding and LLM-Augmented CLIP Embeddings*. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, 2025.

[5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. *Extracting and composing robust features with denoising autoencoders*. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1096–1103, 2008.

[6] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. *BPR: Bayesian personalized ranking from implicit feedback*. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461, 2009.

[7] W. Krichene and S. Rendle. *On sampled metrics for item recommendation*. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1748–1757, 2020.