VIP5: Towards Multimodal Foundation Models for Recommendation

Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, Yongfeng Zhang
Department of Computer Science, Rutgers University, NJ 08854, US
{sg1309,juntao.tan,shuchang.syt.liu,zuohui.fu,yongfeng.zhang}@rutgers.edu

ABSTRACT

Computer Vision (CV), Natural Language Processing (NLP), and Recommender Systems (RecSys) are three prominent AI applications that have traditionally developed independently, resulting in disparate modeling and engineering methodologies. This has impeded the ability for these fields to directly benefit from each other's advancements. With the increasing availability of multimodal data on the web, there is a growing need to consider various modalities when making recommendations for users. While text data, such as tags and descriptions, have been widely utilized, visual information has received limited attention in recommender systems despite its importance in decision-making in daily life, particularly in domains such as fashion and retail. Previous approaches to incorporating multimodal information in recommender systems have been limited to matching-based recommendation tasks and are not well-suited to other tasks or modalities without substantial modification of the original architecture. With the recent emergence of foundation models, large language models have emerged as a potential general-purpose interface for unifying different modalities and problem formulations. In light of this, we propose the development of a multimodal foundation model by considering both visual and textual modalities under the P5 recommendation paradigm (VIP5) to unify various modalities and recommendation tasks. This will enable the processing of vision, language, and personalization information in a shared architecture for improved recommendations. To achieve this, we introduce multimodal personalized prompts to accommodate multiple modalities under a shared format. Additionally, we propose a parameter-efficient training method for foundation models, which involves freezing the backbone and fine-tuning lightweight adapters, resulting in improved recommendation performance and increased efficiency in terms of training time and memory usage.

KEYWORDS

Recommender Systems; Multimodal Foundation Model; Parameter-efficient Tuning; Personalized Prompt; Unified Framework

1 INTRODUCTION

With rapid growth, recommender systems have gradually become an indispensable element in people's daily lives. With more time spent on the Web, people reveal their interests through richer modalities than before. Beyond traditional personalization information such as ratings, tags and text reviews, multimodal information such as images, GIFs, music, and videos enjoy growing popularity especially for young generations [32]. In response to the trend, current recommendation systems [7, 24, 45, 74, 77] consider more diverse contents when making recommendation decisions to users.

Historically, the technical developments for processing different types of information are mostly spread across different research communities. For example, the Recommender System (RecSys) community mostly focused on personalization information, especially various types of user interactions such as user clicks, likes and purchases; the Computer Vision (CV) community mostly focused on various visual information such as images and videos; while the Natural Language Processing (NLP) community mostly focused on various textual information such as sentences and documents. Such relatively independent development of techniques results in very different modeling and engineering methodology for different tasks, which makes it difficult for different fields to directly benefit from each other's technical advancements.

Fortunately, recent advances in large foundation models unfold a promising route for building general-purpose models and unifying diverse modalities, so that one single architecture can handle visual, textual and personalized information at the same time. As a pioneering work, GPT-3 [5] can perform in-context learning, enabling it to solve brand-new problems given few-shot demonstration examples as prompts. Similarly, CLIP [47] maintains superior zero-shot generalization ability when shifting to an out-of-distribution visual domain if provided with appropriate prompt. With more and more emergent abilities [68] revealed in foundation models, they become not only a popular backbone to finetune downstream tasks [2, 3, 52, 67] but also an effective training scheme for unifying multiple modalities in a shared interface [6, 8, 9, 29, 64]. Following the trend in language and vision domains, P5 [18] and M6-Rec [10] put forward the concept of personalized foundation models for recommendation and propose to pretrain on instructional prompts to accommodate various recommendation tasks under a shared model and training objective.

While there are large models for language [5, 48, 49], vision [47, 73], and recommendation [10, 18] domains separately, in this work, we take one step further and aim to unify the above foundation models to jointly process multi-modality information sources for personalization and recommendation. To this end, we propose the Multimodal Foundation Model (MFM), which provides the following advantages for recommender systems: 1) MFM provides multimodal personalized prompts for supporting all modalities' connections to the recommendation foundation model. Specifically, to construct multimodal personalized prompts, MFM employs a mapping network to transfer features from other modalities into the corresponding tokens. By this step, multimodal features are projected to the same manifold space of the backbone foundation model. 2) The MFM framework provides the ability of *Parameter-efficient tuning* rather than *Pre-training* in existing recommendation foundation models such as P5 [18]. Different from the pre-training step of P5 that updates all the parameters in the backbone foundation model – which is impractical when the size of foundation model grows explosively – MFM only finetunes a small proportion

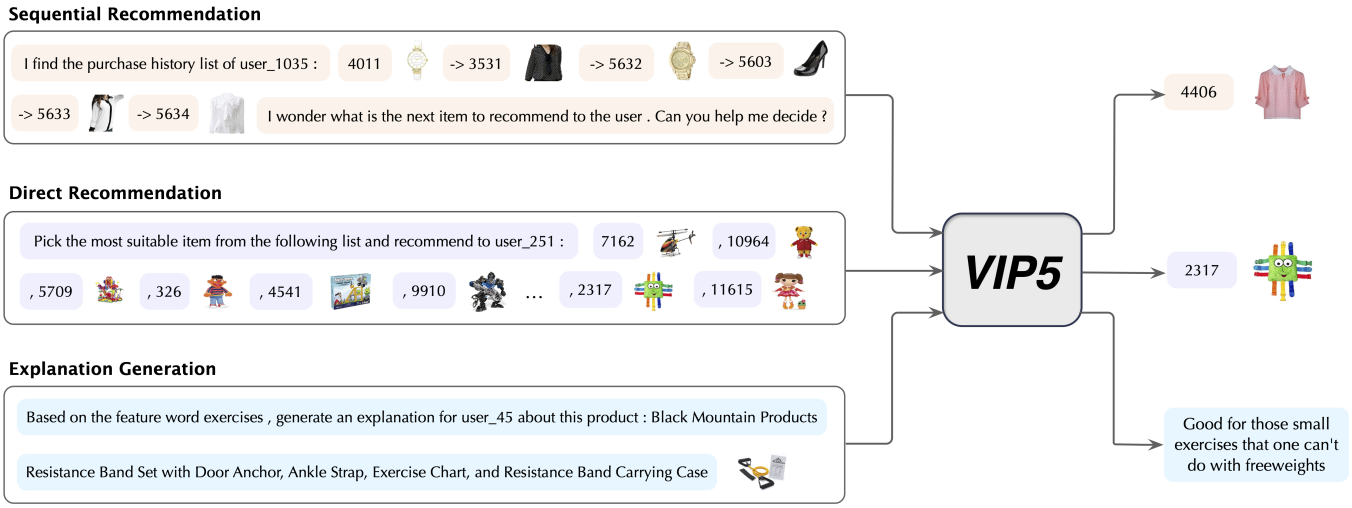

Figure 1: An example task scope of VIP5 covering three popular recommendation tasks. Based on multimodal personalized prompts (*left*) that interleave language and visual tokens, VIP5 is able to transfer all task and all modalities into a unified sequence format, and generates target outputs (*right*) according to certain task descriptions. VIP5 treats large language models as a fixed general-purpose interface and finetunes extra visual and language processing layers to achieve the ability for handling various recommendation tasks.

of extra lightweight adapter modules during training while maintaining the large language model backbone fixed. 3) With the ability of multi-modality learning and parameter-efficient tuning, MFM further improves the performance of recommendation foundation models with both less training time and less memory usage, making it easier to train and deploy foundation models for recommendation. Overall, our key contributions can be outlined as follows:

- We propose the MFM framework to unify CV, NLP, and RecSys foundation models and facilitate recommendation with multimodal information.
- We introduce multimodal personalized prompts to adapt multimodality information into a shared tokenized space with textual, visual and personalization inputs.
- We develop adapter-based parameter-efficient tuning for MFM to achieve a better recommendation performance and training efficiency.
- Based on the experimental results, MFM outperforms strong baselines on three task groups while saving substantial training time and memory usage.

2 RELATED WORK

Parameter-efficient Tuning. The rapid development of large pretrained models has benefited both downstream NLP and vision tasks via finetuning. However, as a conventional learning paradigm, finetuning all parameters in a colossal model becomes impractical when the model size scales up substantially. In contrast, parameter-efficient tuning approaches [13] that only finetune a small proportion of parameters have emerged as a more prevalent practice. On language tasks, Adapter [26], LoRA [27], Compacter [43], Parallel Adapter [20], and $(IA)^3$ [39] are proposed to adapt pretrained language models to downstream tasks in a more efficient way. Following this trend, parameter-efficient tuning methods targeting at multimodal tasks such as WiSE-FT [71], CLIP-Adapter [15], Tip-Adapter [76], VL-Adapter [57], and Ladder Side-Tuning [56] have also brought more memory saving and robustness when fine-tuning vision-language models.

Prompt Learning. Prompt learning [40] gradually emerges as a popular paradigm to control the behavior of large language models since it can effectively adapt a pretrained model to downstream tasks in either zero-shot or few-shot style. The success of GPT series [5, 48] attracts the first wave of interests on the topic. The in-context learning capability of GPT-3 [5] inspires many efforts on automatic prompt search or generation [16, 30, 54, 79] to achieve higher-quality discrete prompts. However, it is naturally hard to optimize these approaches in a discrete space. To solve this issue, soft prompt based approaches such as Prefix-Tuning [37], Prompt-Tuning [33], CoOp [82], and Visual-Prompt Tuning [28] are proposed to leverage additional trainable continuous embeddings as prefix to conduct finetuning on downstream tasks. While achieving better scalability and generalization ability, the learned soft prompts are more difficult to interpret than discrete prompts. To accommodate all above merits, instruction prompts that directly describe different tasks via natural language instructions are adopted by a lot of methods [3, 46, 52, 67, 70], highlighting significant improvements on unseen tasks. Recently, new techniques such as Chains-of-Thought Prompting [69] and Self-Consistency decoding strategy [66] further boost the performance of large language models [58] as well as provide explainable reasoning paths.

Large Recommendation Models. Motivated by the success of large language models, the RecSys community started to pay more attention to recommendation model's generalization ability and transferability. For instance, Transformers4Rec [11] and BERT4Rec

[55] introduce a pipeline to utilize popular NLP Transformer architectures for sequential and session-based recommendation tasks. In addition, UniSRec [25] learns generalizable item and sequence representations based on item description texts rather than explicit item IDs. Similarly, TransRec [63] facilitates transferable content-based recommendation by utilizing user feedback with mixture-of-modality items and representing user preferences and items via modality encoders such as BERT and ResNet. Inspired by the prompt learning paradigm, PEPLER [35] proposes to learn personalized continuous prompts to represent user and item IDs and generates natural language explanations to justify recommendations. In contrast, M6-Rec [10] converts all user behavior information to plain text sequences and feeds them into a Transformer encoder. M6-Rec then designs a task-specific training loss for each downstream task and conducts finetuning. Apart from previous efforts, P5 [18] leverages not only instruction-based finetuning to represent personalized fields for users and items but also describes various tasks via natural language instructions. Hence, P5 can unify various recommendation tasks into a shared encoder-decoder architecture and a joint training objective.

Multimodal Recommendation. It is imperative to develop multimodal recommender systems because most of the information on the Web is multimodal. Current approaches to multimodal recommendation can be divided into three categories. The most common usage is to leverage multimodal content as side information to assist recommendation decisions. For example, VBPR [22] proposes using visual features to supplement user feedback and improve matching-based recommendations. PiNet [45] proposes to cover more personalized visual preferences about users. It simultaneously learns heterogeneous visual features with semantic and collaborative information and then fuses different visual information through a dual-gating module. Another stream of approaches focus on providing recommendations along with correlated visual explanations. These methods usually work in domains where visual information is important to user behavior patterns, such as fashion [7, 24, 62], travel [17], and food [45]. Furthermore, several recent approaches have been proposed to discover the rich intra-item and inter-item semantic structures from multimodal contents to facilitate better item representations and thus enhance recommendation performances [12, 74, 75].

Language as General-Purpose Interfaces. As a powerful medium, language can be used to describe most things. With the development of modern language models, language has gradually evolved as a general-purpose interface. As a result, many methods are springing up to connect a broad range of tasks and various modalities into a shared foundation language model. To name a few, Vokenization [59] and VaLM [65] introduce visual information as auxiliary tokens and conduct self-supervised learning to assist NLP tasks that require language understanding abilities. VL-T5 [9] and Pix2Seq [6] further formulates vision-language tasks and object detection as a sequence-to-sequence based conditional language generation problem, respectively. UniTAB [72] then unifies the previous two frameworks to facilitate object-level fine-grained grounding and reasoning in vision-language tasks without a pre-trained object detector. To solve the text-to-image generation task, DALL-E [50] employs a pre-trained VQ-VAE [60] encoder to transfer images into latent tokens and learns a Transformer decoder to model text and image tokens autoregressively. OFA [64] and PaLI [8] subsume the advantages of aforementioned methods and create omnipotent vision-language foundation models featured by text-to-image generation and multilingual capabilities, respectively. Moreover, MetaLM [19] and Flamingo [2] take a pre-trained vision backbone to perceive visual inputs and seamlessly dock interleaved multimodal inputs on a pre-trained language backbone. With merely few-shot in-context learning, they can achieve much better performances than fully fine-tuned models. Beyond vision and language modalities, recent approaches such as VIMA [29], InstructRL [38], and SayCan [1] further extend large language models to the robotics domain. These models enable an embodied agent to follow instructions to accomplish a broad spectrum of real-world robotic tasks, which exhibit better scalability and generalization ability than conventional strategies.

3 MFM PARADIGM WITH MULTIMODAL PERSONALIZED PROMPTS

We introduce the proposed MFM paradigm in this section. In Section 3.1, we incorporate multimodal signals into personalized prompts. In Section 3.2, we elaborate how to conduct parameter-efficient tuning with adapters based on multimodal personalized prompts.

3.1 Multimodal Personalized Prompts

A personalized prompt is a prompt that contains personalized fields for both users and items [18, 35]. The format of such fields can be either ID numbers or detailed descriptions. In our work, we aim to develop foundation models as a general-purpose interface to connect all available modalities that could be helpful for eliciting user preferences. To facilitate this end, we propose "multimodal personalized prompts". Technically, we consider textual, visual, and personalization information as three example modalities in our multimodal personalized prompts (depicted in Figure 2).

Given an item image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W are the image height and width, we first adopt a visual encoder such as CLIP image branch [47] to extract its feature $x \in \mathbb{R}^{d_v}$, where d_v represents the visual feature dimension. To connect the image feature to other text-based tokens in a personalized prompt, as illustrated in Figure 2(c), we design a mapping network f with two linear layers to transfer the original image feature to k image tokens: $p_1, \ldots, p_k = f(x)$. Then we append the image tokens to their corresponding item tokens to construct a multimodal personalized field $\mathcal{M}$:

$$\mathcal{M} : \underbrace{w_1 \cdots w_m}_{\text{item tokens}}, \underbrace{p_1, \ldots, p_k}_{\text{image tokens}} . \tag{1}$$

We create a collection of 29 multimodal personalized prompts covering three important task families – *sequential recommendation*, *direct recommendation*, and *explanation*. The full list of prompts is provided in Figure 8, 9, and 10. Based on the collection of multimodal personalized prompts, we use the multimodal personalized field $\mathcal{M}$ as in Eq.(1) to substitute the item field in the prompt. It is worth noting that the prompts for sequential and direct recommendation usually contain more than one multimodal personalized fields.

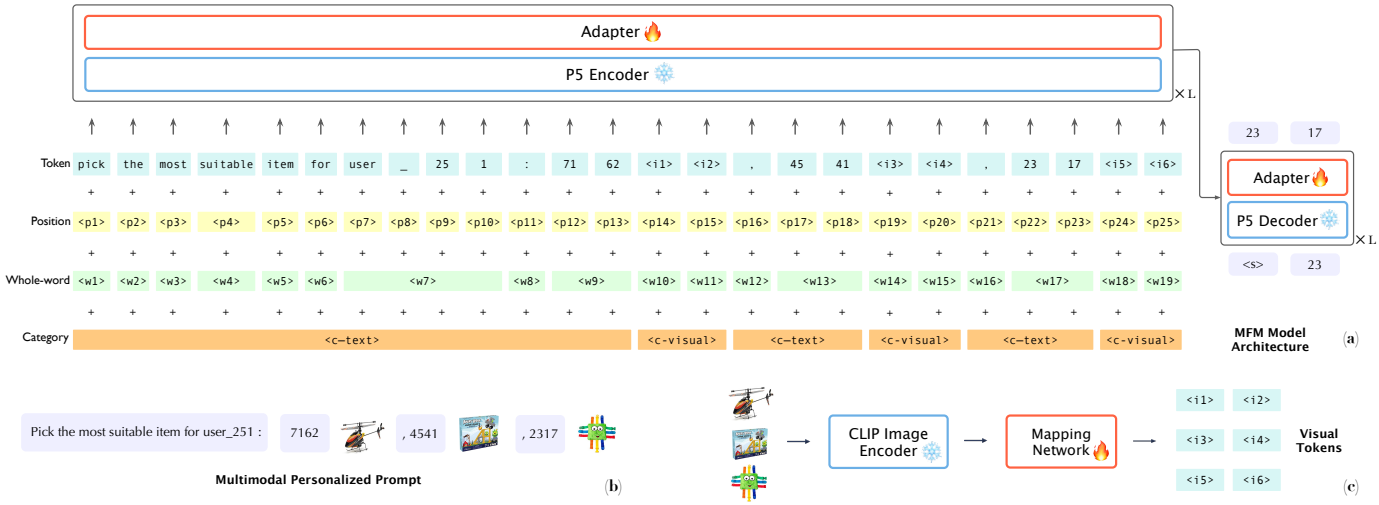

Figure 2: An illustration of the VIP5 framework. VIP5 is built on an encoder–decoder Transformer model that takes in textual inputs as well as image inputs to produce responses or make recommendation decisions. In the figure, a fire symbol represents training with parameter update while a snow flake symbol stands for the frozen parameters.

3.2 Parameter-efficient Tuning with Adapters

Our MFM framework is a Transformer-based [61] encoder–decoder architecture, as shown in Figure 2(a). For a tokenized multimodal token sequence $\mathbf{S}$, we first apply position encoding $\mathcal{P}$ and whole-word embedding $\mathcal{W}$ on $\mathbf{S}$ to help the model better recognize the absolute positions of input tokens and important user/item fields (e.g., "user_251" is split into 4 separate tokens ["user", "_", "25", "1"], but they share the same whole-word embedding "⟨w7⟩"). Besides, we adopt an additional category embedding $\mathcal{C}$ to identify whether a token is textual or visual. Afterwards, we feed the resulting sequence into the L-layered text encoder $\mathcal{E}$ and decoder $\mathcal{D}$ modules.

Except for multimodal personalized prompts, we propose parameter-efficient tuning with adapters for computation- and memory- efficient model training. More specifically, we insert additional adapters to the foundation model backbone. During training, we keep the parameters of the backbone frozen and purely update the parameters of these lightweight adapter modules. Such parameter-efficient tuning strategy can largely reduce the ratio of trainable parameters, thus cost less training time and memory usage during training. In addition, only tuning a small part of additional parameters can overcome the efficiency concern caused by the longer sequence when incorporating visual tokens into text-based personalized prompts. More importantly, fine-tuning the whole foundation model backbone would lead to the over-fitting for some easier tasks especially when the backbone model size is huge, while parameter-efficient tuning can benefit from both the training efficiency and the power of large foundation models.

Formally, if we denote the input sequence for the i-th layer of text encoder as $\mathbf{S}_i = [s_1, \cdots, s_n]$, in traditional Transformer, $\mathbf{S}_i$ will go through one self-attention block and a feed-forward network. While in MFM, we insert adapters [26, 57] in both the self-attention block and the feed-forward network, the exact position is after each module and before the LayerNorm [4]. The whole process can be written as:

$$\mathbf{S}_{i+1} = A_2\left(\text{FFN}\left(A_1\left(\text{Attention}\left(\mathbf{S}_i\mathbf{W}_\text{Q}, \mathbf{S}_i\mathbf{W}_\text{K}, \mathbf{S}_i\mathbf{W}_\text{V}\right)\right)\right)\right), \quad (2)$$

where $\mathbf{W}_\text{Q}, \mathbf{W}_\text{K}, \mathbf{W}_\text{V} \in \mathbb{R}^{d \times d_h}$ are weight matrices for projecting query, key, and value, respectively, $d_h = d/h$ is the dimensionality for each head. The Attention function is defined as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}. \quad (3)$$

Besides, FFN is a feed-forward module consisting of two fully-connected layers. A_1 and A_2 are the feature adapters after the self-attention and feed-forward network. They are both bottleneck fully-connected layers with an activation function in between. We can represent these adapters as:

$$A = f_\text{up}\left(\sigma\left(f_\text{down}(\mathbf{S}_i)\right)\right) + \mathbf{S}_i, \quad (4)$$

where f_down and f_up are the down-sampling and up-sampling layers of an adapter, and σ is the GELU activation function [23]. Similar to text encoder, we also adopt adapters between the cross-attention block and its LayerNorm layer inside text decoder.

MFM utilizes the conditional token generation loss for all three recommendation tasks. After encoding the input multimodal personalized prompts into a contextualized latent sequence with $\mathcal{E}$, the text decoder $\mathcal{D}$ autoregressively predict next tokens conditioned on the already generated tokens $\mathbf{y}_{<j}$ and the input text $\mathbf{t}$. In summary, MFM adopts the following training objective to perform parameter-efficient tuning with adapters:

$$\mathcal{L}_\theta = -\sum_{j=1}^{|\mathbf{y}|} \log P_\theta\left(\mathbf{y}_j \mid \mathbf{y}_{<j}, \mathbf{t}\right). \quad (5)$$

After training, we perform inference with MFM based on given multimodal personalized prompts. For sequential and direct recommendation task groups, we create a list of candidate items for recommendation via beam search. For explanation task group, we simply apply greedy decoding for text generation.

Table 1: Detailed statistics of the datasets used in our paper.

Dataset	Clothing	Sports	Beauty	Toys
#Users	39,387	35,598	22,363	19,412
#Items	23,033	18,357	12,101	11,924
#Reviews	278,677	296,337	198,502	167,597
#Photos	22,299	17,943	12,023	11,895
#Sparsity (%)	0.0307	0.0453	0.0734	0.0724

4 EXPERIMENTS

In this section, we provide the performance comparison between the MFM framework and representative approaches for different task groups. We conduct extensive experiments and ablation studies to explore the following research questions:

- **RQ1:** Does the proposed parameter-efficient MFM framework perform well when compared with baseline methods across the three task groups?
- **RQ2:** When conducting parameter-efficient tuning, which parts should we insert adapters and perform finetuning? In addition, will different adapter reduction rates affect the performance and efficiency of MFM?
- **RQ3:** Does visual information play an important role in multimodal personalized prompts? What if we change the number of image tokens and the type of visual encoder?

4.1 Experimental Setups

Datasets. We use four real-world datasets collected from *Amazon* platform [21, 44] to conduct the experiments and ablation studies in our paper, namely *Clothing, Shoes & Jewelry*, *Sports & Outdoors*, *Beauty*, and *Toys & Games*. These Amazon datasets[1] all provide available user purchase records, user reviews, item descriptions, and item images. In Table 1, we provide detailed statistics about the four datasets.

Tasks and Metrics. In this paper, we cover three popular recommendation tasks groups, i.e., A) sequential recommendation, B) direct recommendation, and C) explanation generation to perform all the experiments. We adopt the same pre-processing steps and train/validation/test splits as in [18]. For sequential recommendation task group, the last and the second last items of each user's interaction history sequence are adopted as the ground-truth of test and validation splits, respectively. The remaining items in the interaction history sequence used as training data. For direct recommendation task group, we use the same train/validation/test splits of sequential recommendation for generating the 100 candidate lists to choose from [80]. For the explanation generation task group, we follow the 8:1:1 random split style and extract explanations of ratings with the Sentires library [78].

We use Hit Ratio (HR@k) and Normalized Discounted Cumulative Gain (NDCG@k) to evaluate the performance on sequential recommendation and direct recommendation task groups, while text generation metrics such as BLEU and ROUGE are adopted for explanation generation task. Note that across all tables in this paper, **bold** numbers highlight the best approach on each metric.

[1] http://jmcauley.ucsd.edu/data/amazon/links.html

Implementation Details. We employ the pre-trained P5-small checkpoint as the backbone of our MFM framework since the P5-small achieves better performance than P5-base in most cases [18]. The encoder and decoder of MFM both have 6 Transformer blocks, with a 512-dimension embedding size and attentions of 8 heads. Since we need to process visual information in MFM, we adopt the image branch of CLIP [47] as the visual encoder of the MFM framework. To speed up the training process, we extract the image features in advance with CLIP visual encoders. Similar to P5, SentencePiece [53] tokenizer with a vocabulary size of 32,100 is adopted to generate sub-word units as input tokens. By default, the mapping network serves as the image tokenizer in our framework and the number of image tokens is set to 2, while the adapters have a reduction factor of 8 for the bottleneck dimension.

For each task group, all multimodal personalized prompts but the last one is used for training MFM, while Prompts A-3/A-9, Prompts B-5/B-8, and Prompts C-3/C-12 are used for evaluation purpose, where A-3, B-5, C-3 are used to test the model performance under seen prompts in training, while A-9, B-8, C-12 are used to test the model performance under zero-shot unseen prompts. MFM is trained for 10 epochs with a batch size of 36 on four NVIDIA A100 GPUs. The learning rate is set to 1×10^{-3} and AdamW [41] is selected as the optimizer. With more image tokens contained in the multimodal personalized prompts, we enlarge the maximum length of input tokens to be 1024. During inference, we set the beam size B to 20 for sequential recommendation and direct recommendation tasks that require generating a list of candidate items.

Comparison Baselines. To make performance comparisons, we consider a collection of baselines for each task group. For all the three task groups, we include the P5 model [18] as a baseline so as to compare with existing foundation models for recommendation. P5 pre-trains all the three tasks with predefined text-based personalized prompts through autoregressive language modeling loss, and performs inference with either greedy decoding or beam search strategy to generate expected outputs. Besides P5, we also compare with several task-specific approaches in the following.

For sequential recommendation, we compare with **HGN** [42], **SASRec** [31], and **S^3-Rec** [81] as baseline methods. Specifically, HGN considers both the long and short terms of user behaviors and models such properties with a hierarchical gating network. SASRec is a classical baseline of sequential recommendation, which combines self-attention with recurrent neural networks (RNNs). In SASRec, RNNs are used to capture long-term user interests while self-attention mechanism can help the model focus on relatively important factors to avoid noisy signals. In order to improve the representation ability of sequential recommendation models, S^3-Rec designs a series of self-supervised auxiliary tasks to maximize the mutual information among different signals such as users' purchase sequences, items, and their attributes.

For direct recommendation, we compare with **BPR-MF** [51] which is one of the most representative collaborative filtering methods. **BPR-MLP** is also used as a baseline after adding an additional multi-layer perceptron to the original MF model. The above methods all directly take users' implicit feedback to model their preferences without other external knowledge about the users and items. Since we focus on incorporating multimodal information into

Table 2: Performance comparison on sequential recommendation.

Methods	Sports				Beauty			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
HGN	0.0189	0.0120	0.0313	0.0159	0.0325	0.0206	0.0512	0.0266
SASRec	0.0233	0.0154	0.0350	0.0192	0.0387	0.0249	0.0605	0.0318
S^3-Rec	0.0251	0.0161	0.0385	0.0204	0.0387	0.0244	0.0647	0.0327
P5 *(A-3)*	0.0272	0.0169	0.0361	0.0198	0.0503	0.0370	0.0659	0.0421
MFM *(A-3)*	**0.0412**	**0.0345**	**0.0475**	**0.0365**	**0.0556**	**0.0427**	**0.0677**	**0.0467**
P5 *(A-9)*	0.0258	0.0159	0.0346	0.0188	0.0490	0.0358	0.0646	0.0409
MFM *(A-9)*	0.0392	0.0327	0.0456	0.0347	0.0529	0.0413	0.0655	0.0454
Methods	Clothing				Toys			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
HGN	0.0107	0.0071	0.0175	0.0092	0.0321	0.0221	0.0497	0.0277
SASRec	0.0107	0.0066	0.0194	0.0095	0.0463	0.0306	0.0675	0.0374
S^3-Rec	0.0076	0.0045	0.0135	0.0063	0.0443	0.0294	0.0700	0.0376
P5 *(A-3)*	0.0478	0.0376	0.0554	0.0401	0.0655	0.0570	0.0726	0.0593
MFM *(A-3)*	**0.0603**	**0.0564**	**0.0632**	**0.0573**	**0.0662**	**0.0577**	**0.0749**	**0.0604**
P5 *(A-9)*	0.0455	0.0359	0.0534	0.0385	0.0631	0.0547	0.0701	0.0569
MFM *(A-9)*	0.0569	0.0531	0.0597	0.0540	0.0641	0.0556	0.0716	0.0580

Table 3: Performance comparison on direct recommendation.

Methods	Sports					Beauty				
	HR@1	HR@5	NDCG@5	HR@10	NDCG@10	HR@1	HR@5	NDCG@5	HR@10	NDCG@10
BPR-MF	0.0314	0.1404	0.0848	0.2563	0.1220	0.0311	0.1426	0.0857	**0.2573**	0.1224
BPR-MLP	0.0351	0.1520	0.0927	0.2671	0.1296	0.0317	0.1392	0.0848	0.2542	0.1215
VBPR	0.0262	0.1138	0.0691	0.2060	0.0986	0.0380	0.1472	0.0925	0.2468	0.1245
P5 *(B-5)*	0.0574	0.1503	0.1050	0.2207	0.1276	0.0601	0.1611	0.1117	0.2370	0.1360
MFM *(B-5)*	0.0606	0.1743	0.1185	0.2539	0.1441	0.0580	0.1598	0.1099	0.2306	0.1327
P5 *(B-8)*	0.0567	0.1514	0.1049	0.2196	0.1269	0.0571	0.1566	0.1078	0.2317	0.1318
MFM *(B-8)*	**0.0699**	**0.1882**	**0.1304**	**0.2717**	**0.1572**	**0.0615**	**0.1655**	**0.1147**	0.2407	**0.1388**
Methods	Clothing					Toys				
	HR@1	HR@5	NDCG@5	HR@10	NDCG@10	HR@1	HR@5	NDCG@5	HR@10	NDCG@10
BPR-MF	0.0296	0.1280	0.0779	0.2319	0.1112	0.0233	0.1066	0.0641	0.2003	0.0940
BPR-MLP	0.0342	0.1384	0.0858	0.2327	0.1161	0.0252	0.1142	0.0688	0.2077	0.0988
VBPR	0.0352	0.1410	0.0877	**0.2420**	0.1201	0.0337	0.1294	0.0808	**0.2199**	0.1098
P5 *(B-5)*	0.0320	0.0986	0.0652	0.1659	0.0867	0.0418	0.1219	0.0824	0.1942	0.1056
MFM *(B-5)*	0.0481	0.1287	0.0890	0.1992	0.1116	0.0428	0.1225	0.0833	0.1906	0.1051
P5 *(B-8)*	0.0355	0.1019	0.0688	0.1722	0.0912	0.0422	0.1286	0.0858	0.2041	0.1099
MFM *(B-8)*	**0.0552**	**0.1544**	**0.1058**	0.2291	**0.1297**	**0.0433**	**0.1301**	**0.0875**	0.2037	**0.1110**

recommendation, we naturally take **VBPR** [22] to compare with since it integrates visual signals into latent factors for preference prediction.

For explanation generation, we inherent the baselines of P5 – **Attn2Seq** [14], **NRT** [36], and **PETER** [34] to make comparisons. Specifically, Attn2Seq proposes an attention-based attribute-to-sequence model to generate textual explanations based on a given user–item pair and a user rating score. In the model, an attribute encoder is used to transfer inputs into vectors and then generate explanation sentences via an LSTM architecture. Different from Attn2Seq, NRT learns to simultaneously predict rating scores and create user reviews according to the input user and item IDs. Following the setting of NRT, PETER recently employs a modified attention mask to the vanilla Transformer block [61] to generate both the ratings and the rating explanations for given user–item pairs. When providing a hint feature word as assistive input, PETER becomes its variant approach **PETER+**. We also take PETER+ as our explanation generation baseline.

4.2 Performance on Different Task Groups (RQ1)

In this section, we conduct parameter-efficient tuning for MFM on multimodal personalized prompts from all the three task groups.

Table 4: Performance comparison on explanation generation (numbers are in percentage %).

Methods	Sports				Beauty			
	BLUE4	ROUGE1	ROUGE2	ROUGEL	BLUE4	ROUGE1	ROUGE2	ROUGEL
Attn2Seq	0.5305	12.2800	1.2107	9.1312	0.7889	12.6590	1.6820	9.7481
NRT	0.4793	11.0723	1.1304	7.6674	0.8295	12.7815	1.8543	9.9477
PETER	0.7112	12.8944	1.3283	9.8635	1.1541	14.8497	2.1413	11.4143
P5 *(C-3)*	0.6212	11.8539	2.0707	9.0189	1.0230	14.3242	2.0761	10.9085
MFM *(C-3)*	**1.0639**	**14.8628**	**2.1012**	**11.1059**	**1.2850**	**17.7492**	**2.3482**	**12.9170**
PETER+	**2.4627**	24.1181	5.1937	18.4105	**3.2606**	25.5541	5.9668	19.7168
P5 *(C-12)*	1.3144	22.9182	4.9976	17.1976	1.6313	24.6267	4.9623	18.6423
MFM *(C-12)*	2.3003	**24.4887**	**5.5057**	**18.6610**	2.8390	**26.0513**	**6.0159**	**20.4387**
Methods	Clothing				Toys			
	BLUE4	ROUGE1	ROUGE2	ROUGEL	BLUE4	ROUGE1	ROUGE2	ROUGEL
Attn2Seq	0.6296	11.4588	1.2558	9.0429	1.6238	13.2245	2.9942	10.7398
NRT	0.4599	10.1480	0.9720	8.2434	1.9084	13.5231	3.6708	11.1867
PETER	0.7204	12.1836	1.3912	9.7084	1.9861	14.2716	3.6718	11.7010
P5 *(C-3)*	0.7569	12.2833	1.8116	9.6023	1.4522	12.6100	**3.8144**	10.1450
MFM *(C-3)*	**1.1904**	**14.1685**	**2.0308**	**10.8488**	**2.3241**	**15.3006**	3.6590	**12.0421**
PETER+	**3.6204**	28.4342	7.7994	22.4059	**4.7919**	28.3083	9.4520	22.7017
P5 *(C-12)*	1.8811	27.7922	7.3203	21.5462	2.6216	27.8984	9.0076	21.6136
MFM *(C-12)*	3.2581	**28.9059**	**8.5168**	**22.8807**	3.9293	**28.9225**	**9.5441**	**23.3148**

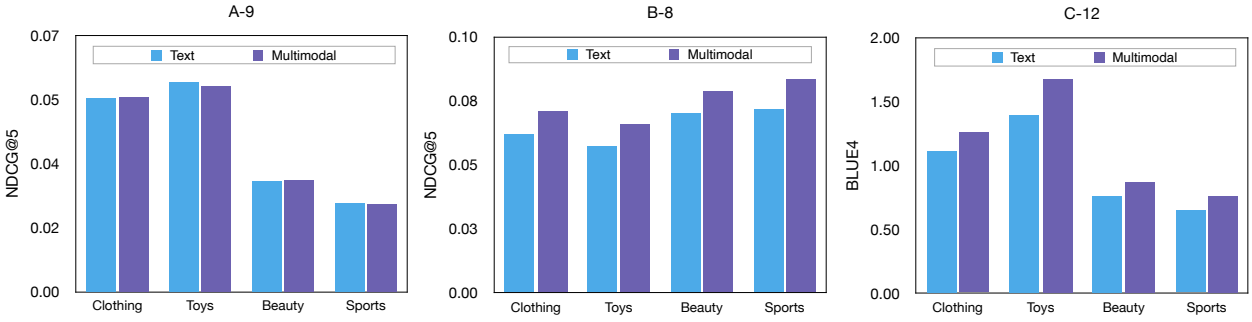

Figure 3: Performance comparison between text-based prompt and multimodal prompt.

For each task group, one seen prompt during training and one unseen prompt is selected to perform evaluation. The performance comparison with the aforementioned baselines are provided in Table 2, Table 3, and Table 4.

Sequential Recommendation. As shown in Table 2, we adopt Prompt A-3 and Prompt A-9 to evaluate the performances of different approaches. From the table, we can see that MFM is able to achieve better performances than all sequential recommendation baselines on all the four experiment datasets, among which a relatively large gap can be observed on *Sports* and *Clothing* datasets. The results show that our parameter-efficient tuning strategy works effectively on the sequential recommendation task group.

Direct Recommendation. For the direction recommendation task group, we use Prompt B-5 and Prompt B-8 as input multimodal personalized prompts to evaluate different methods. We present performance comparison in Table 3, where we can find that MFM is capable of beating all baselines on *Sports*. On *Toys*, *Beauty*, and *Clothing* datasets, MFM achieves slightly lower HR@10 performances but still outperforms all baselines on other metrics.

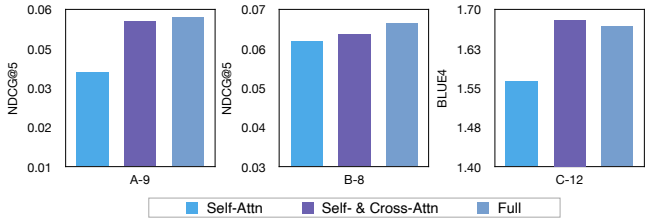

Figure 4: Performance comparison among only activating adapters in self-attention blocks, both self-attention and cross-attention blocks, and full finetuning.

Explanation Generation. Table 4 illustrates the performance comparison for explanation generation task group. In the table, Prompt C-12 are applied to evaluate all methods under hint feature

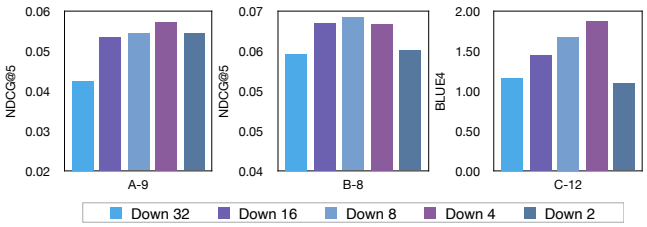

Figure 5: Ablation study on the *downsample reduction rate*.

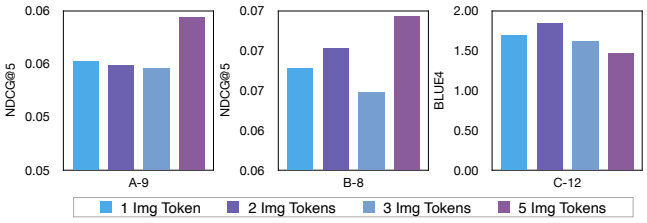

Figure 6: Ablation study on the *image token number*.

word setup, while Prompt C-3 targets at direct explanation generation with only the given user–item pair. The experimental results indicate that MFM outperforms other baselines when equipped with the multimodal personalized Prompt C-3. For Prompt C-12, MFM achieves superior performances than P5 across all datasets in terms of all metrics and has the highest ROUGE1, ROUGE2, ROUGEL scores of the four experimental datasets.

4.3 Ablation on Parameter-efficient Tuning (RQ2)

In this section, we discuss how to conduct parameter-efficient tuning with adapters to show the impact of different parameter-efficient tuning choices.

How to conduct parameter-efficient tuning. We first try three approaches of fine-tuning: 1) inserting adapters in self-attention blocks of Transformer architecture and only fine-tuning them, 2) fine-tuning the adapters in both self-attention and cross-attention blocks, 3) fully fine-tuning all parameters. For this ablation, we conduct all experiments on *Toys* with ResNet-101 visual features, a reduction rate of 8, and a single image token in multimodal prompt. As illustrated in Figure 4, we can see that fine-tuning adapters in all attention blocks is necessary to achieve better (Prompt C-12) or comparable (Prompt A-9 & B-8) results with full fine-tuning. Moreover, according to Table 5, the former saves 21.2% time and 18.1% memory usage during training compared to the latter, showing both the effectiveness and efficiency of our MFM framework.

On adapter reduction rate. The reduction rate is an important hyper-parameter for adapters. When decreasing the reduction rate, the hidden dimension of bottleneck layers will increase correspondingly, resulting in a higher percentage of trainable parameters. We select five different values of reduction rates and perform all experiments with ResNet-101 visual features and a single image token in multimodal prompt. In Figure 5, we can see that 4 and 8 are suitable reduction rates for all the three task groups.

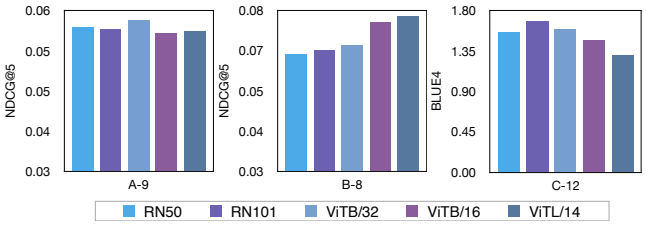

Figure 7: Ablation study on the *visual encoder type*.

4.4 Ablation on Visual Components (RQ3)

In this section, we aim to explore whether visual information matters for different task groups. We also estimate the influence of the number of image tokens and the visual encoder type.

Text-based vs. multimodal personalized prompts. To compare text-based and multimodal personalized prompts, we set the number of image tokens to 0 and 1, respectively, and conduct experiments on all four datasets with a reduction rate of 8 and ResNet-101 visual features. From Figure 3, we can see that by introducing visual signals into personalized prompts, improvements are observed on all datasets for direct recommendation task group (Prompt B-8). This is in line with our expectation that the visual appearance of an item is an important factor when people making choices from a list of candidates. For sequential recommendation, visual information did not bring obvious performance improvements, indicating that the purchase sequence itself is more significant for predicting the next item. In terms of explanation generation, visual information also exerts positive impacts on all of the experiment datasets especially for *Toys* dataset.

On the number of image tokens. To verify the influence of the number of image tokens, we choose four different numbers (1, 2, 3, 5) and conduct additional experiments on *Toys* with a reduction rate of 8 and ResNet-101 visual features. According to Figure 6, enabling 5 image tokens in multimodal personalized prompt achieves the best performance on Prompt A-9 and Prompt B-8, while 2 image tokens perform the best for Prompt C-12. However, longer visual prompt results in more training time (e.g., 5 image tokens take 60.8% more time than 2 image tokens). Therefore, we choose 2 image tokens as default setting considering the trade-off.

On visual encoder type. The type of visual encoder is another factor that could decide the representation ability of a multimodal personalized prompt. In view of this, we vary the type of CLIP visual branch from the following architectures – ResNet50, ResNet101, ViT-B/32, ViT-B/16, ViT-L/14 (in an ascending order of visual encoder ability according to CLIP [47]). All experiments are performed on *Toys* with a reduction rate of 8 and a single image token in multimodal prompt. The results are reported in Figure 7. Similar to our previous conclusions, visual information matters most for direct recommendation, so we can observe a continuous performance gain when we gradually substitute a better visual encoder. However, for sequential recommendation and explanation generation, changing a better visual encoder does not always lead to a better performance. This is most likely because the purchase sequence is more important than visual information for predicting the next item in sequential recommendation, resulting in similar performances under different visual encoders. As for explanation generation, hint

Table 5: Comparison of different training strategies in terms of trainable parameter (%), training time (min), and memory usage (GB) on the *Toys* dataset.

Methods/Metrics	Time/Epoch	Trainable Param.	Memory Usage
Self-Attn	10.55	2.97	27.4
Self- & Cross-Attn	11.10	3.58	29.0
Full (P5)	14.08	100	35.6

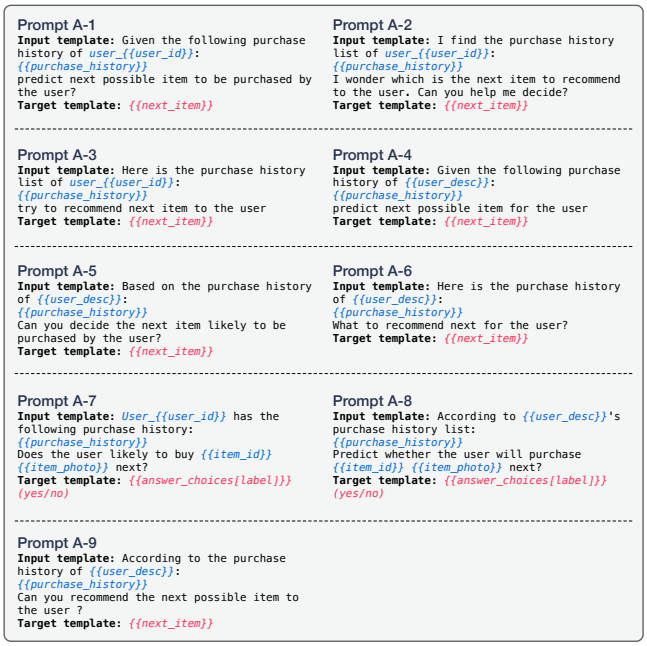

Figure 8: Multimodal personalized prompts for Task Group A: Sequential Recommendation.

words play a significant role on the generated sentences and the compatibility between the hint word and the visual embedding varies for different visual encoders. However, MFM is still better than the best baseline under most visual encoders.

5 CONCLUSIONS AND FUTURE WORK

This paper presents a multimodal foundation model under the VIP5 framework to unify vision, language, and personalization information into a parameter-efficient multimodal foundation recommendation model. We integrate visual signals with text and personalization information by designing multimodal personalized prompts to improve recommendation with diverse modalities. With the parameter-efficient tuning strategy, we only need to update a small proportion of adapters while obtaining better trade-off between recommendation performance, training efficiency and memory usage. Through extensive experiments and ablation studies, we show the effectiveness of our framework and show that multimodality information is a helpful signal to assist various recommendation tasks. In the future, we plan to scale up the backbone model size, include more diverse modality types, and explore better prompt strategies such as chain-of-thought and self-consistency.

APPENDIX

From Figure 8 to Figure 10, we provide a detailed list of 29 multimodal personalized prompts used in our paper that covers three recommendation tasks.

Figure 9: Multimodal personalized prompts for Task Group B: Direct Recommendation.

Figure 10: Multimodal personalized prompts for Task Group C: Explanation Generation.

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex

Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

[3] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. In *International Conference on Learning Representations*.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[6] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. 2022. Pix2seq: A Language Modeling Framework for Object Detection. In *International Conference on Learning Representations*. https://openreview.net/forum?id=e42KbIw6Wb

[7] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* (2022).

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 1931–1942. https://proceedings.mlr.press/v139/cho21a.html

[10] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. *arXiv preprint arXiv:2205.08084* (2022).

[11] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4Rec: Bridging the Gap between NLP and Sequential/Session-Based Recommendation. In *Fifteenth ACM Conference on Recommender Systems*. 143–153.

[12] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2022. Leveraging Content-Style Item Representation for Visual Recommendation. In *European Conference on Information Retrieval*. Springer, 84–92.

[13] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904* (2022).

[14] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *EACL*.

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).

[16] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *ACL-IJCNLP*.

[17] Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022. Improving Personalized Explanation Generation through Visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

[18] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the Sixteenth ACM Conference on Recommender Systems*.

[19] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language Models are General-Purpose Interfaces. *arXiv preprint arXiv:2206.06336* (2022).

[20] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=0RDcd5Axok

[21] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[22] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[23] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[24] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. *arXiv preprint arXiv:1905.12862* (2019).

[25] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[26] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.

[27] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[28] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. *arXiv preprint arXiv:2203.12119* (2022).

[29] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. VIMA: General Robot Manipulation with Multimodal Prompts. *arXiv preprint arXiv:2210.03094* (2022).

[30] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

[31] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[32] Roberta Katz, Sarah Ogilvie, Jane Shaw, and Linda Woodhead. 2022. *Gen Z, explained: The art of living in a digital age*. University of Chicago Press.

[33] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *EMNLP*.

[34] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4947–4957.

[35] Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems* (2022).

[36] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 345–354.

[37] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*.

[38] Hao Liu, Lisa Lee, Kimin Lee, and Pieter Abbeel. 2022. Instruction-Following Agents with Jointly Pre-Trained Vision-Language Models. *arXiv preprint arXiv:2210.13431* (2022).

[39] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638* (2022).

[40] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021).

[41] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[42] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.

[43] Rabeeh Karimi mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=bqGK5PyI6-N

[44] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[45] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3460–3468.

[46] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing Instructional Prompts to GPTk's Language. *Findings of the Association for Computational Linguistics: ACL 2022* (2022).

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019).

[49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[51] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[52] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1715–1725.

[54] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*.

[55] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[56] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *arXiv preprint arXiv:2206.06522* (2022).

[57] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5227–5237.

[58] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261* (2022).

[59] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2066–2080.

[60] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[62] Dhruv Verma, Kshitij Gulati, Vasu Goel, and Rajiv Ratn Shah. 2020. Fashionist: Personalising outfit recommendation for cold-start scenarios. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4527–4529.

[63] Jie Wang, Fajie Yuan, Mingyue Cheng, Joemon M Jose, Chenyun Yu, Beibei Kong, Zhijin Wang, Bo Hu, and Zang Li. 2022. TransRec: Learning Transferable Recommendation from Mixture-of-Modality Feedback. *arXiv preprint arXiv:2206.06190* (2022).

[64] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 23318–23340. https://proceedings.mlr.press/v162/wang22al.html

[65] Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022. Visually-Augmented Language Modeling. *arXiv preprint arXiv:2205.10178* (2022).

[66] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[67] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

[68] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD

[69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of 36th Conference on Neural Information Processing Systems (NeurIPS)*.

[70] Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E Peters. 2020. Learning from Task Descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1361–1375.

[71] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7959–7971.

[72] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling. In *ECCV*.

[73] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[74] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[75] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2021. Latent Structures Mining with Contrastive Modality Fusion for Multimedia Recommendation. *arXiv preprint arXiv:2111.00678* (2021).

[76] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification. In *ECCV*.

[77] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1449–1458.

[78] Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014. Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification. In *SIGIR*.

[79] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493* (2022).

[80] Wayne Xin Zhao, Zihan Lin, Zhichao Feng, Pengfei Wang, and Ji-Rong Wen. 2022. A revisiting study of appropriate offline evaluation for top-N recommendation algorithms. *ACM Transactions on Information Systems* 41, 2 (2022), 1–41.

[81] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.

[82] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* (2022), 1–12.