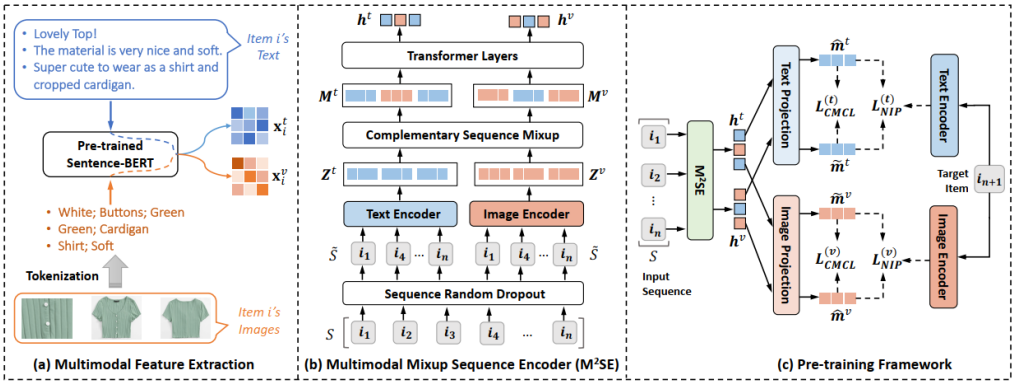
MSM4SR

1. 动机与解决方案

1. 模态差异：不同模态的内容对齐挑战→将 image 转化为 text
2. 领域差异：用户兴趣不仅受到物品模态内容影响还受到时间影响→使用对比学习损失

2. 模型框架



主要包括三个部分：

1. 模态特征提取器：获取初始模态特征；
2. 多模态混合序列编码器：将物品的多模态内容与用户行为序列结合；
3. 对比预训练：基于对比学习微调

2.1. 模态特征提取

对文本

$$\mathbf{x}_i^t = stack\left[\text{BERT}(t_1^i), \text{BERT}(t_2^i), \cdots, \text{BERT}(t_{|\mathcal{T}_i|}^i)\right], \quad (1)$$

每个 t 是一个句子

对图片，首先用预训练的模型如 CLIP 将图片用 N 个关键词描述，然后再通过和文本一样的处理。

$$\begin{aligned} f(w) &= \text{sim}(\text{CLIP}(v_\ell^i), \text{CLIP}(w)) \quad \forall w \in \mathcal{D}, \\ \mathbf{v}_\ell^i &= \text{BERT}(concat(\text{TopN}(\{f(w_1), \cdots, f(w_{|\mathcal{D}|})\}, N))), \\ \mathbf{x}_i^v &= stack[\mathbf{v}_1^i, \mathbf{v}_2^i, \cdots, \mathbf{v}_{|\mathcal{V}_i|}^i], \end{aligned} \quad (2)$$

2.2. 多模态混合序列编码器

包括四个部分：

1. **序列随机丢弃**：$S \xrightarrow{drop} \tilde{S}$，以一定概率随机 drop 序列中的物品。
2. **文本视觉编码器**：编码器由一个注意力层（两层的线性转换）和 MoE 构成。以文本为例
   - 注意层：

$$\alpha^t = \text{softmax}\big((\mathbf{x}_i^t \mathbf{W}_1^t + \mathbf{b}_1^t)\mathbf{W}_2^t + b_2^t\big),$$
$$\mathbf{e}_i^t = \sum_{j=1}^{|\mathcal{T}_i|} \alpha_j^t \mathbf{x}_i^t[j,:], \qquad (3)$$

   这里是不同句子的注意力加权和。

   - MoE：

$$E_k(\mathbf{e}_i^t) = \text{LayerNorm}(\text{Dropout}(\mathbf{e}_i^t \mathbf{W}_k^t)),$$
$$\mathbf{g}^t = \text{softmax}(\mathbf{e}_i^t \mathbf{W}_3^t), \qquad (4)$$

混合专家网络（Mixture of Experts， MoE）是一种用于处理复杂非线性问题的神经网络结构，由多个子网络（称为“专家”）组成，每个专家负责处理输入数据的一个子集，然后将专家的输出进行加权混合，得到最终的输出结果：

$$\mathbf{z}_i^t = \sum_{k=1}^{O} g_k^t E_k(\mathbf{e}_i^t), \qquad (5)$$

其中，g_k是门控路由从路由向量$g \in R^k$得到的第 k 个专家的相应组合权重。Text Encoder 最后输出序列中所有 z 的拼接 $\mathbf{Z}^t = stack[\mathbf{z}_1^t, \tilde{\mathbf{z}}_2^t, \cdots, \mathbf{z}_{|\tilde{S}|}^t]$.

3. **互补序列混合**：$Z^t, Z^v \xrightarrow{Mix} M^t, M^v$。以概率$\rho \leq 0.5$随机替换$Z^t$和$Z^v$的物品嵌入
4. **TRM 层**：M 加上位置编码输入到 L 层的 TRM 中，将输出的最后一行作为输入的两种混合模态表征：$\mathbf{h}^t$ and $\mathbf{h}^v$

2.3. 预训练

预训练包括两个目标：（1）特定模态下下一个物品预测；（2）跨模态对比学习前者是为了捕获混合序列与下一个物品在某个模态下的联系，后者是为了校准不

同模态空间下的表征差异。

在获得这两个任务的损失函数之前，将经过编码器得到的两个混合模态输出通过线性变换分别映射到文本空间和视觉空间：

$$\begin{aligned} \widehat{\mathbf{m}}_j^t &= \mathbf{h}_j^t \mathbf{W}_t + \mathbf{b}_t, & \widetilde{\mathbf{m}}_j^t &= \mathbf{h}_j^v \mathbf{W}_t + \mathbf{b}_t, \\ \widehat{\mathbf{m}}_j^v &= \mathbf{h}_j^v \mathbf{W}_v + \mathbf{b}_v, & \widetilde{\mathbf{m}}_j^v &= \mathbf{h}_j^t \mathbf{W}_v + \mathbf{b}_v, \end{aligned} \tag{6}$$

2.3.1. 特定模态下下一个物品预测

正样本：物品的模态表征 $\mathbf{z}_j^t$；负样本：同批次内其他物品的模态表征

$$\mathcal{L}_{\mathrm{NIP}}^{(t)} = -\sum_{j=1}^{|\mathcal{B}|} \log \frac{f(\widehat{\mathbf{m}}_j^t, \mathbf{z}_j^t) + f(\widetilde{\mathbf{m}}_j^t, \mathbf{z}_j^t)}{\sum_{j'=1}^{|\mathcal{B}|} [f(\widehat{\mathbf{m}}_j^t, \mathbf{z}_{j'}^t) + f(\widetilde{\mathbf{m}}_j^t, \mathbf{z}_{j'}^t)]}, \tag{7}$$

同理视觉。

2.3.2. 跨模态对比学习

正样本：$\widehat{\mathbf{m}}_j^t, \widetilde{\mathbf{m}}_j^t$ 互为正

样本；负样本：其他物品的映射

$$\ell(\widehat{\mathbf{m}}_j^t, \widetilde{\mathbf{m}}_j^t) = \log \frac{f(\widehat{\mathbf{m}}_j^t, \widetilde{\mathbf{m}}_j^t)}{\sum_{j'=1}^{|\mathcal{B}|} f(\widehat{\mathbf{m}}_j^t, \widetilde{\mathbf{m}}_{j'}^t) + \sum_{j' \neq j}^{|\mathcal{B}|} f(\widehat{\mathbf{m}}_j^t, \widehat{\mathbf{m}}_{j'}^t)},$$

$$\mathcal{L}_{\mathrm{CMCL}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{|\mathcal{B}|} (\ell(\widehat{\mathbf{m}}_j^t, \widetilde{\mathbf{m}}_j^t) + \ell(\widetilde{\mathbf{m}}_j^t, \widehat{\mathbf{m}}_j^t)). \tag{9}$$

同理视觉。

最终的损失函数：

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\mathrm{NIP}}^{(t)} + \mathcal{L}_{\mathrm{NIP}}^{(v)} + \lambda(\mathcal{L}_{\mathrm{CMCL}}^{(t)} + \mathcal{L}_{\mathrm{CMCL}}^{(v)}), \tag{11}$$

2.4. 序列推荐微调

去掉随即 drop 和混合模态操作，旨在训练 ID embedding 和多模态序列编码器的参数。

$$\widehat{\mathbf{y}}^{(\mathcal{S})} = \mathrm{softmax}\big(\mathbf{h}^t(\mathbf{F}^t + \mathbf{E})^\top + \mathbf{h}^v(\mathbf{F}^v + \mathbf{E})^\top\big), \tag{12}$$

$$\mathcal{L}_{\text{finetune}} = -\sum_{(\mathcal{S},i) \in \widetilde{\mathcal{B}}} \log\big(\widehat{\mathbf{y}}^{(\mathcal{S})}(i)\big), \tag{13}$$

F 是所有物品的 text 或 image 表征。h 是经过 TRM 输出的序列表征。

3. 实验

3.1. 对比实验

Dataset	Metric	LightGCN	GRCN	DualGNN	MV-RNN	SASRec	S^3-Rec	SINE	DIF-SR	SASRecM	MSM4SR	%Imp
Pantry	R@5	0.0270	0.0365	0.0321	0.0157	0.0277	0.0315	0.0297	0.0300	0.0369	**0.0405**	9.76%
	R@10	0.0460	0.0552	0.0485	0.0276	0.0457	0.0535	0.0534	0.0473	0.0600	**0.0673**	12.17%
	R@20	0.0774	0.0856	0.0739	0.0467	0.0722	0.0845	0.0873	0.0736	0.0934	**0.1040**	11.35%
	N@5	0.0176	0.0229	0.0202	0.0101	0.0147	0.0187	0.0167	0.0163	0.0223	**0.0235**	2.62%
	N@10	0.0236	0.0289	0.0254	0.0134	0.0204	0.0257	0.0243	0.0219	0.0298	**0.0321**	7.72%
	N@20	0.0315	0.0366	0.0318	0.0184	0.0271	0.0335	0.0329	0.0284	0.0382	**0.0414**	8.38%
Arts	R@5	0.0543	0.0546	0.0596	0.0299	0.0704	0.0715	0.0667	0.0712	0.0816	**0.0854**	4.66%
	R@10	0.0726	0.0741	0.0788	0.0446	0.0910	0.0961	0.0935	0.0899	0.1099	**0.1184**	7.73%
	R@20	0.0967	0.0999	0.1033	0.0661	0.1125	0.1250	0.1237	0.1126	0.1430	**0.1570**	9.79%
	N@5	0.0381	0.0386	0.0433	0.0184	0.0442	0.0467	0.0404	0.0449	0.0525	**0.0531**	1.14%
	N@10	0.0440	0.0448	0.0495	0.0232	0.0509	0.0546	0.0491	0.0510	0.0616	**0.0637**	3.41%
	N@20	0.0501	0.0513	0.0557	0.0283	0.0563	0.0619	0.0567	0.0567	0.0699	**0.0735**	5.15%
Office	R@5	0.0325	0.0556	0.0518	0.0259	0.0841	0.0823	0.0837	0.0857	0.0850	**0.0968**	12.95%
	R@10	0.0518	0.0714	0.0661	0.0416	0.1025	0.1027	0.1059	0.1039	0.1060	**0.1206**	13.77%
	R@20	0.0752	0.0911	0.0843	0.0641	0.1222	0.1254	0.1305	0.1241	0.1316	**0.1480**	12.46%
	N@5	0.0219	0.0408	0.0385	0.0159	0.0558	0.0575	0.0546	0.0561	0.0584	**0.0721**	23.46%
	N@10	0.0281	0.0460	0.0431	0.0210	0.0617	0.0641	0.0618	0.0620	0.0652	**0.0797**	22.23%
	N@20	0.0339	0.0509	0.0477	0.0266	0.0667	0.0698	0.0680	0.0671	0.0716	**0.0866**	20.95%

3.2. 消融实验

Methods	Pantry						Office					
	R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20
MSM4SR	**0.0405**	**0.0673**	**0.1040**	**0.0235**	**0.0321**	**0.0414**	**0.0968**	**0.1206**	**0.1480**	**0.0721**	**0.0797**	**0.0866**
$MSM4SR_{ResNet}$	0.0387	0.0647	0.1007	0.0234	0.0317	0.0408	0.0924	0.1159	0.1435	0.0673	0.0749	0.0818
$MSM4SR_{w/o\ NIP}$	0.0292	0.0501	0.0816	0.0179	0.0245	0.0324	0.0863	0.1062	0.1302	0.0573	0.0637	0.0697
$MSM4SR_{w/o\ CMCL}$	0.0390	0.0662	0.1030	0.0223	0.0310	0.0403	0.0875	0.1095	0.1349	0.0578	0.0649	0.0713
$MSM4SR_{w/o\ Mixup}$	0.0380	0.0649	0.1014	0.0220	0.0307	0.0399	0.0957	0.1192	0.1480	0.0692	0.0768	0.0841
$MSM4SR_{w/o\ Pre\text{-}train}$	0.0359	0.0595	0.0920	0.0203	0.0286	0.0369	0.0883	0.1094	0.1335	0.0597	0.0665	0.0726
$MSM4SR_{E2E}$	0.0355	0.0605	0.0937	0.0209	0.0290	0.0373	0.0821	0.1013	0.1243	0.0527	0.0589	0.0647

w/o Pre-train：删除预训练任务，直接基于多模态微调设置从头训练所提的模型；E2E：端到端的优化，即将预训练损失和微调损失相加，进行训练。两个损失学习不同的方面，预训练学习不同模态之间的相互作用而微调损失强调使用交叉熵的推荐任务，一起优化可能使模型无法收敛到最优解。

Dataset	Model	R@10	R@20	N@10	N@20
Pantry	GRCN	0.0552	0.0856	0.0289	0.0366
	MSM4SR-V	0.0596	0.0928	0.0290	0.0373
	MSM4SR-T	0.0649	0.1001	0.0318	0.0407
	MSM4SR	**0.0673**	**0.1040**	**0.0321**	**0.0414**
Arts	S^3-Rec	0.0961	0.1250	0.0546	0.0619
	MSM4SR-V	0.1018	0.1345	0.0581	0.0663
	MSM4SR-T	0.1144	0.1523	**0.0652**	**0.0748**
	MSM4SR	**0.1184**	**0.1570**	0.0637	0.0735
Office	S^3-Rec	0.1027	0.1254	0.0641	0.0698
	MSM4SR-V	0.1153	0.1415	0.0764	0.0830
	MSM4SR-T	0.1186	0.1464	0.0772	0.0841
	MSM4SR	**0.1206**	**0.1480**	**0.0797**	**0.0866**

-V -T 只是在微调的公式（12）去掉相应模态部分，但是还是在多模态上进行过预训练。