

CROSSAN: Towards Efficient and Effective Adaptation of Multiple Multimodal Foundation Models for Sequential Recommendation

Junchen Fu, Yongxin Ni, Joemon M. Jose, Ioannis Arapakis, Kaiwen Zheng, Youhua Li, and Xuri Ge

Abstract—In this paper, we explore a less-studied yet practically important problem: how to efficiently and effectively adapt multiple (>2) multimodal foundation models (MFMs) for the sequential recommendation task. To this end, we propose a plug-and-play Cross-modal Side Adapter Network (CROSSAN), which leverages a fully decoupled side adapter-based paradigm to achieve efficient and scalable adaptation. Compared to the state-of-the-art efficient approaches, CROSSAN reduces training time by over 30%, GPU memory consumption by 20%, and trainable parameters by over 57%, while enabling effective cross-modal learning across diverse modalities. To further enhance multimodal fusion, we introduce the Mixture of Modality Expert Fusion (MOMEF) mechanism. Extensive experiments on public benchmarks demonstrate that CROSSAN consistently outperforms existing methods, achieving 6.7%–8.1% performance improvements when adapting four foundation models with raw modalities. Moreover, the overall performance continues to improve as more MFMs are incorporated. We will release our code and datasets to facilitate future research.

Index Terms—Recommender Systems, Multimodal Foundation Models, Efficient Adaptation, Sequential Recommendation, CROSSAN, MOMEF

I. INTRODUCTION

Multimodal Foundation Models (MFMs) have advanced rapidly, with models like ViT [1], BERT [2], GPT [3], VideoMAE [4], and AST [5] demonstrating exceptional performance in representing a wide range of raw modalities. At the same time, the increasing availability of recommendation datasets [6] containing raw multimodal data (e.g., images, text, video, audio, etc.) provides a natural avenue for exploring how these powerful models can be effectively adapted for sequential recommendation tasks.

One intuitive approach that preserves enough information, is to adapt these models to the raw modalities of recommendation datasets [7], [8], [9]. Adaptation methods for MFMs that leverage raw modality information, such as fine-tuning and parameter-efficient fine-tuning (PEFT), are generally recognized for their ability to achieve better performance compared to traditional feature-based approaches [7], [8], [10], [6].

J. Fu, K. Zheng, and J. Jose are with the School of Computing Science, University of Glasgow, Glasgow, UK (e-mail: j.fu.3@research.gla.ac.uk, k.zheng.1@research.gla.ac.uk, and joemon.jose@glasgow.ac.uk). Y. Ni is with Boston University, Boston, United States (email: niyongxin2016@gmail.com). I. Arapakis is with Telefonica Scientific Research, Barcelona, Spain (email: arapakis.ioannis@gmail.com). Y. Li is with City University of Hong Kong, Hong Kong, China. (e-mail: youhuali2-c@my.cityu.edu.hk) X. Ge is with the School of Artificial Intelligence at Shandong University, Jinan, Shandong, China. (e-mail: xuri.ge@sdu.edu.cn) (*Corresponding author: Yongxin Ni and Xuri Ge.*)

However, these approaches have been largely sidelined due to the central challenge of the significant computational costs associated with existing adaptation methods for multiple MFMs. Both full fine-tuning and PEFT techniques, such as Adapters [11], [12] and LoRA [13], become increasingly expensive as the number of modality encoders grow¹. This computational bottleneck makes it more challenging to investigate the adaptation of additional MFMs, leaving the potential of these models for sequential recommendation largely unexplored.

Recently, the side adapter paradigm [15], [16], [17] has garnered significant attention for its superior efficiency in adaptation compared to traditional adapter-based or LoRA-based approaches. This efficiency stems from its fully decoupled design, which eliminates the need for gigantic computational graphs. However, existing studies on side adapters have primarily focused on single or dual-modal scenarios, leaving their potential for scalability to additional modalities and the integration of multiple multimodal foundation models (MFMs) unexplored. This gap highlights the need for further research to extend their applicability to more complex and diverse multimodal setups.

Motivated by the aforementioned challenges, this study aims to explore the under-explored area of adapting multiple (>2) multimodal foundation models (MFMs) for sequential recommendation and to propose a solution that enables efficient and effective adaptation. To this end, we introduce the Cross-modal Side Adapter Network (CROSSAN), a novel plug-and-play approach that addresses the computational inefficiencies of existing adaptation methods for multiple MFMs in item representation. To enhance multimodal interactions and improve overall effectiveness, we propose a cross-modal side adapter network, building on key insights from our preliminary study (section II). In our analysis, we demonstrate that cross-modal interaction enhances mutual information compared with independent side adapters. Moreover, in contrast to existing methods that fuse item representations through a fully connected layer [15], we adopt the Mixture of Modality Expert Fusion (MOMEF). This approach incorporates a fine-grained gating mechanism that enables the adaptive integration of input modalities, providing a straightforward yet highly effective solution for capturing intricate multimodal interactions and enhancing representation fidelity. CROSSAN offers a scalable, efficient, and effective solution, achieving superior perfor-

¹Recommender systems are typically retrained on a daily or weekly basis [14], making costly training paradigms impractical in practice.

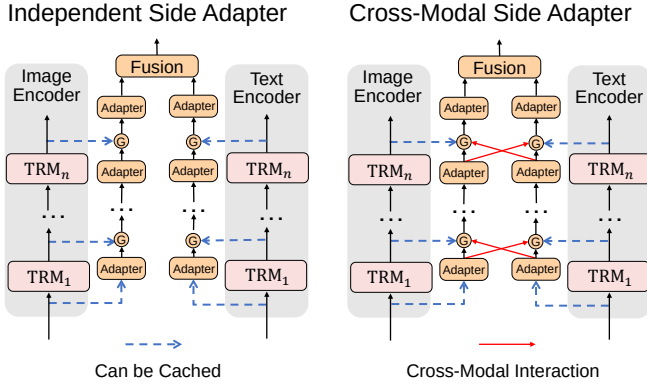


Fig. 1: Independent vs. Cross-modal Side Adapter. During training, only the adapters and fusion layer are updated, while the rest of the foundation models remain frozen.

mance on public datasets. It outperforms existing efficient adaptation approaches, with its performance improvement becoming even more significant as more MFMs are integrated. Our contributions are listed below:

- To achieve enhanced multimodal representation learning while maintaining high efficiency, we introduce CROSSAN, a simple yet effective, cross-modality side adapter method.
- Building upon traditional concatenation-based fusion for multimodal item representation, we explore different fusion strategies and show that the Mixture of Modality Expert Fusion (MOMEF) mechanism effectively integrates item representations across modalities, leading to improved recommendation performance.
- Through extensive experiments, we demonstrate that the CROSSAN delivers superior performance and efficiency on the public dataset. These results highlight the potential of leveraging multiple MFMs with raw modalities for the multimodal sequential recommendation, paving the way for future research in this direction.

II. PRELIMINARY STUDY: CROSS OR INDEPENDENT?

To maximize efficiency, we adopt a recently advanced, fully decoupled, side adapter paradigm with a caching strategy [15], [16], [17], where the adapters are positioned outside the transformer models. This paradigm, while extensively studied for single-modality adaptation, remains underexplored in the context of multiple multimodal foundation models (MFMs). Therefore, we investigate two potential approaches: implementing the adapters either in a cross-modal configuration or independently, as illustrated in Figure 1. Preliminary experiments conducted on two commonly used combined modalities (text and image), using the Microlens-100K dataset, indicate that the cross-modal approach outperforms the independent method across four evaluation metrics of Hit ratio and NDCG (Table I). To understand this performance improvement, we further provide a theoretical analysis in section III.

TABLE I: Performance comparison between INDSAN (Independent Side Adapter) and CROSSAN (Cross-Modal Side Adapter). “*” denotes that the improvements are significant at the level of 0.05 with a paired T-test.

Metric	INDSAN	CROSSAN
HR@10	0.0957	0.0970*
HR@20	0.1373	0.1393*
NDCG@10	0.0521	0.0537*
NDCG@20	0.0626	0.0644*
Mutual Information	0.0129	0.2001*

III. THEORETICAL ANALYSIS

In this section, we provide a theoretical analysis demonstrating why the proposed Cross-Modal Side Adapter (CROSSAN) outperforms the Independent Side Adapter (INDSAN). For clarity, we focus on a representative scenario involving text and image modalities.

A. Mutual Information as an Evaluation Metric

Let $Z^{(v)}$ and $Z^{(t)}$ denote the learned visual and textual representations at a given network layer. The mutual information (MI) [18] between these representations is defined based on entropy as:

$$I(Z^{(v)}; Z^{(t)}) = H(Z^{(v)}) - H(Z^{(v)} | Z^{(t)}), \quad (1)$$

where $H(\cdot)$ denotes the entropy, which measures the uncertainty or information content of a random variable. Specifically, $H(Z^{(v)})$ quantifies the total uncertainty of the visual representations, while $H(Z^{(v)} | Z^{(t)})$ captures the remaining uncertainty in $Z^{(v)}$ after observing the textual representations $Z^{(t)}$.

A higher mutual information $I(Z^{(v)}; Z^{(t)})$ implies that the textual representations $Z^{(t)}$ provide more information about the visual representations $Z^{(v)}$, thus reducing their uncertainty. This reflects stronger dependency and more effective cross-modal integration, indicating that the learned representations capture complementary and aligned information across modalities, which is desirable for downstream tasks. Many studies show that maximizing mutual information enhances representation quality [19], [20], [21]. Thus, we use it as the evaluation metric for multimodal learning.

B. Independent vs. Cross-Modal Adapters for Item Representation

Based on the definition of mutual information in terms of entropy [18], where $H(\cdot)$ denotes entropy, the **INDSAN** model produces approximately independent modality-specific representations, denoted as $Z^{(v)}$ and $Z^{(t)}$. Specifically, we assume that the joint distribution of these representations satisfies $p(Z^{(v)}, Z^{(t)}) \approx p(Z^{(v)})p(Z^{(t)})$, indicating near-independence between modalities. This near-independence implies that one modality provides minimal information about the other, leading to $H(Z^{(v)} | Z^{(t)}) \approx H(Z^{(v)})$ and, consequently, negligible mutual information:

$$I_{\text{INDSAN}}(Z^{(v)}, Z^{(t)}) = H(Z^{(v)}) - H(Z^{(v)} | Z^{(t)}) \approx 0.$$

In contrast, **CROSSAN** establishes dependency through cross-modal interactions at each layer, where

$$Z_l^{(v)} = f_l^{(v)}(X, Z_{l-1}^{(t)}), \quad Z_l^{(t)} = g_l^{(t)}(Y, Z_{l-1}^{(v)}).$$

This ensures that $p(Z_l^{(v)}, Z_l^{(t)}) \neq p(Z_l^{(v)})p(Z_l^{(t)})$, meaning the uncertainty of the final representation $Z_L^{(v)}$ is reduced when $Z_L^{(t)}$ is known, i.e., $H(Z_L^{(v)} | Z_L^{(t)}) < H(Z_L^{(v)})$. Consequently, the mutual information in **CROSSAN** is inherently positive:

$$I_{\text{CROSSAN}}(Z_L^{(v)}, Z_L^{(t)}) = H(Z_L^{(v)}) - H(Z_L^{(v)} | Z_L^{(t)}) \geq 0.$$

This yields an upper bound mutual information than **INDSAN**. This continuous cross-layer interaction aligns with prior findings, indicating that deeper, multi-level interactions enhance shared information across modalities [22], [23], [24].

To empirically validate our theoretical analysis, we evaluate the final outputs of the textual and visual adapters using the optimal checkpoints for both **INDSAN** and **CROSSAN**, processing all items from the MicroLens-100K dataset. The results corroborate our conclusion, demonstrating that **CROSSAN** achieves significantly higher mutual information than **INDSAN**. Specifically, as illustrated in Table I, **CROSSAN** yields mutual information values more than 15 times greater than those observed with **INDSAN**. This notable difference supports the previous analysis and provides insights into the underlying reasons for **CROSSAN**'s performance improvement.²

IV. METHODOLOGY

In this work, we introduce **CROSSAN**, a plug-and-play approach designed to adapt multiple MFMs. **CROSSAN** is designed to provide effective multimodal representations in a scalable and efficient manner, offering a general solution for the adaptation of multiple MFMs. The overview of **CROSSAN** is shown in Figure 2. Our key innovation lies in designing a novel fully cross-modal gating between each modality's side adapters, dedicated to learning rich mutual information between modalities. Furthermore, to enable more effective multimodal fusion, we propose a Mixture of Modality Expert Fusion (MOMEF) network, which dynamically combines the multimodal outputs of all towers to achieve fine-grained fusion for each item.

Problem Formulation. Given a recommendation dataset $\mathcal{D} = \{\mathcal{U}, \mathcal{V}\}$, where $\mathcal{U}$ and $\mathcal{V}$ represent the set of users and items respectively, our objective in a multimodal sequential recommendation task is to predict the next item that a user u will interact with, based on their past n interactions. For multimodal recommendation, each item v can have representations from M different modalities, such as text (v^{text}), image (v^{image}), video (v^{video}), and audio (v^{audio}). Although additional modalities can be incorporated, depending on the application, in this paper, we mainly focus on these four standard modalities.

²In the following section, we primarily adopt the state-of-the-art **IISAN** as the main comparison counterpart. In terms of mutual information, **IISAN** is theoretically equivalent to **INDSAN** due to the absence of dependencies between side adapters. The measured mutual information of **IISAN** is 0.0127, which is approximately the same as that of **INDSAN**. Compared to **INDSAN**, **IISAN** introduces an additional set of independent inter-modal side adapters that capture static multimodal information. However, dependencies among the side adapters are still absent.

Following [15], [7], we process each of the modalities using their corresponding pre-trained foundational models, such as BERT [2] for text, ViT [1] for images, VideoMAE [4] for videos, and Audio Spectrogram Transformer (AST) [5] for audio. By leveraging these pre-trained MFMs backbones, we obtain the hidden states for each modality (e.g., h_i^{text} , h_i^{image} , h_i^{video} , and h_i^{audio}) from their transformer layers (TRM_i).

Cross-modal Gating. We employ a simple yet effective cross-modal gating mechanism, which differs from the traditional cross-attention mechanism proposed in [25]. Unlike cross-attention, which relies on computationally intensive attention mechanisms, our approach is based on straightforward tunable weights, offering a lightweight and efficient alternative. Specifically, for each modality at the i -th layer, the input to the adapter combines the output of the adapter from the $(i-1)$ -th layer with the hidden states from the corresponding MFM. For instance, taking the side adapter's text modality at the i -th layer as an example, we define its input as follows:

$$h_i^t = \text{Adapter}_i^t \left(\sum_{m \in M} \alpha_i^m h_{i-1}^m + \alpha_i^t h_{i-1}^{\text{BERT}} \right) \quad (2)$$

where $\sum_{m \in M} \alpha_i^m + \alpha_i^t = 1$.

Here, M represents the set of modalities considered in this work, specifically $M = \{\text{text}, \text{image}, \text{video}, \text{audio}\}$, and t denotes the text modality. The parameter α controls the learnable weight for each layer. We utilize the adapter block design proposed by [11], as it has proven to be highly effective in sequential recommendation tasks [12], [15].

Mixture-of-Modality Expert Fusion (MOMEF). Furthermore, to enhance multimodal representation fusion while maintaining efficiency, we adopt a MOMEF method, drawing inspiration from the mixture-of-experts paradigm [26], [27], which dynamically combines the multimodal outputs of all towers to achieve a fine-grained fusion for each item. Specifically, each modality $m \in M$ has its own expert network producing an output vector $f_m(i)$ for item i . We compute importance scores $w_m(i)$ using a gating network based on fully connected layers and select the top- k modalities $\text{TopK}(i)$ with the highest scores for each item. The final multimodal item representation e^{final} is then obtained by dynamically weighting and combining the outputs of these top- k modality experts:

$$e^{\text{final}} = \sum_{m \in \text{TopK}(i)} w_m(i) \cdot f_m(i) \quad (3)$$

where $\sum_{m \in \text{TopK}(i)} w_m(i) = 1$. This approach allows MOMEF to focus on the most pertinent modalities output for each item, leading to a more effective and fine-grained multimodal fusion.

Subsequently, the vector e^{final} is fed into the sequential encoders to compute the final predicted score $\hat{y}_{ui}$ for user u and item i , which is calculated as the product of the sequential encoder output and the corresponding item embedding. Note that our entire framework maintains high efficiency, as all trainable modules are primarily composed of linear layers and gating mechanisms, without relying on attention mechanisms.

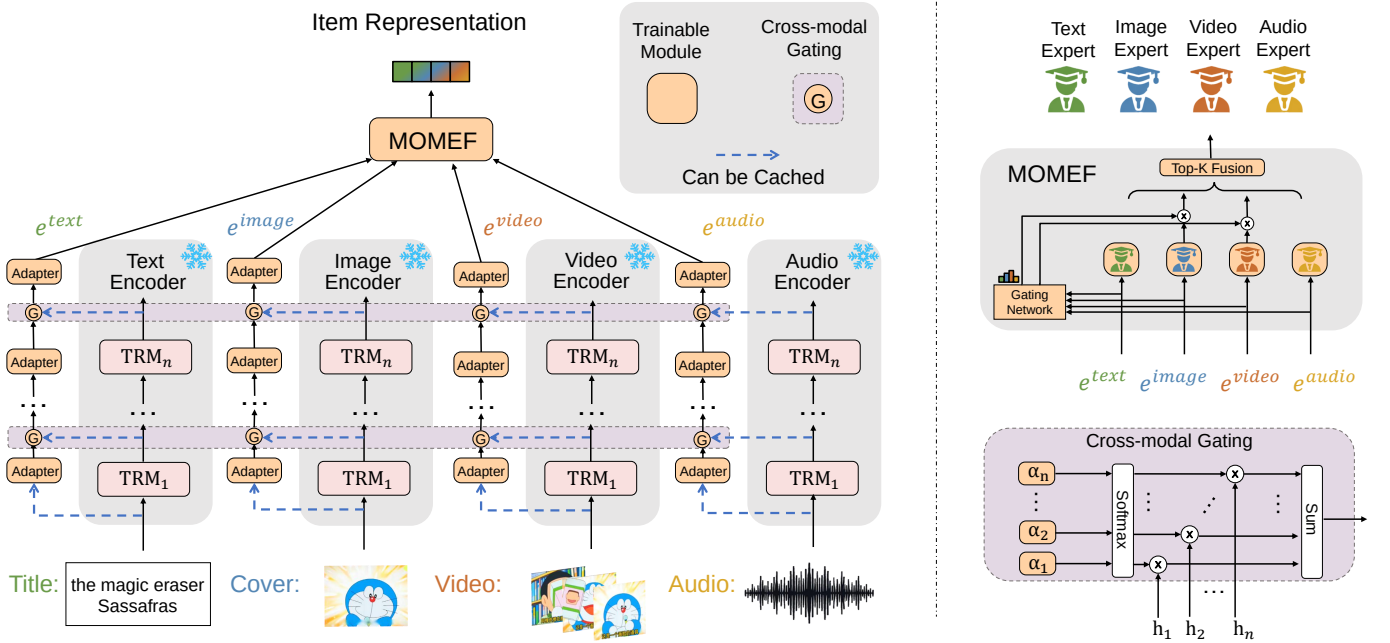


Fig. 2: CROSSAN Overview. The example utilizes four MFMs, each paired with its respective side adapters for adaptation. A cross-modal gating mechanism is applied, combining the outputs of each modality’s side adapter at the same layer-level with gated fusion, ensuring effective interaction across modalities. MOMEF (Mixture of Modality Expert Fusion) treats each input modality as an expert. The gating network assigns probabilities to each expert, selecting the top-k experts based on these probabilities for further processing.

Loss Function. Regarding training, we adopt the commonly used in-batch debiased Cross-Entropy loss function $\mathcal{L}_{CE}$ [28], [7], [29], [15], [6], which is defined as:

$$D_{ui} = \exp(\hat{y}_{ui} - \log(p_i)) + \sum_{j \in [B], j \neq I_u} \exp(\hat{y}_{uj} - \log(p_j)) \quad (4)$$

$$\mathcal{L}_{CE} = - \sum_{u \in \mathcal{U}} \sum_{i \in [2, \dots, n+1]} \log \frac{\exp(\hat{y}_{ui} - \log(p_i))}{D_{ui}} \quad (5)$$

where p_i represents the popularity of item i , I_u denotes the set of items interacted by user u , and B is the batch size. The item $n + 1$ refers to the predicted item for user u .

V. EXPERIMENT SETUP

Dataset. To assess the effectiveness of adapting multiple MFMs for recommendation tasks, we consider datasets that contain more than two raw modalities. Specifically, we use the publicly available Microlens-100K and the Microlens-50K datasets provided in [6].³ The statistical details of the dataset are presented in Table II. Further evaluations on datasets with only two modalities are provided in the section VII.

Evaluation. Based on previous studies [6], [7], [12], [15], our approach implements a leave-one-out evaluation strategy. Specifically, the final item in the interaction sequence is set aside for testing, the second-to-last item is used for validation, and the rest of the sequence is employed for training. To

evaluate our model’s performance, we consider the HR (Hit Ratio) and NDCG (Normalized Discounted Cumulative Gain) metrics, which are aligned with previous studies [6], [7], [12]. Unless otherwise suggested, all reported results correspond to the test set. We also note that the predicted item is evaluated against the entire set of items [30].

Implementation Details. We employ “bert-base-uncased”, “vit-base-patch16-224”, “MIT/ast-finetuned-audioset-10-10-0.4593”, and “MCG-NJU/videomae-base” from Huggingface⁴ as the text, image, audio, and video encoders, respectively.⁵ Our choice is informed by previous research in the field [6], [15], [7], [31], [32]. For video processing, we use the first three seconds of footage and extract 16 frames for VideoMAE, following its original setup [4], with corresponding audio processed by AST. For side adapters, we employ LayerDrop, dropping half the layers of each foundation model for efficiency [15], [16]. We utilize a transformer-based sequential encoder to model user sequences, following the approach outlined in [33], [7]. The hidden dimension of the sequential encoder is set to 64 after a search in {32, 64, 128}, with two Transformer blocks and attention heads following [12], [15]. The learning rate is optimized between 1e-5 and 1e-3, keeping dropout at 0.1 [6]. We search batch sizes from 32 to 1024, selecting the largest based on GPU memory limits. Adapter hidden dimensions and LoRA ranks are tuned

⁴<https://huggingface.co/>

³To the best of our knowledge, Microlens is the only publicly available dataset that includes three or more raw modalities (e.g., text, images, video, and audio). We leave the exploration of additional datasets for future work as and when more modality-rich datasets become available.

⁵The exploration of using LLMs as encoders is beyond the scope of this paper, primarily due to the challenge of managing asymmetry across multiple MFMs. However, we provide a preliminary investigation of advanced LLM and LVM models in the section VII.

TABLE II: Dataset Description.

Dataset	Users	Items	Interaction	Raw Content
MicroLens-100K	100,000	19,738	719,405	Text&Image&Video&Audio
MicroLens-50K	50,000	19,099	339,511	Text&Image&Video&Audio

between 32 and 8192. Hyperparameters are determined by tuning on validation data, and all results are reported on the test set. All experiments are completed on an A6000 GPU.

VI. EXPERIMENT

Our evaluation addresses the following research questions:

- **RQ1:** How effective is CROSSAN compared with existing adaptation approaches, and does adapting more MFMs improve its performance compared to state-of-the-art efficient adaptation approaches?
- **RQ2:** How does CROSSAN’s efficiency compare to state-of-the-art adaptation approaches?
- **RQ3:** How does each component affect the overall performance?
- **RQ4:** How does the hyperparameter affect CROSSAN?

A. Effectiveness Evaluation (RQ1)

We compare our approach against the popular efficient adaptation method Full finetuning, Adapter [11], LoRA [13], and the state-of-the-art IISAN [15], as well as its intuitive extension to support additional modalities. IISAN and IISAN-E serve as our primary multimodal adaptation baselines for two reasons: (1) with two modalities, IISAN achieves competitive performance compared to other methods, and (2) our limited GPU memory (48GB) prevents us from using full fine-tuning or adapter-based approaches on more than two modalities, given their already high requirements for just two (see Table V). Given that the original IISAN implementation is limited to two modalities, we extend its structure by incorporating additional intra-SAN layers to handle more multimodal foundation models (MFMs). In this expanded IISAN-E, we use a gated sum to combine the hidden states of all MFMs into the inter-SAN.

As shown in Table III, CROSSAN achieves the best performance on the MicroLens-100k, outperforming all other adaptation approaches. Furthermore, we observe a progressive improvement in performance as more modalities from MFMs are incorporated. This trend becomes especially evident when compared to IISAN and its extension, IISAN-E, where the relative improvement increases with the introduction of additional modalities, as adding more MFMs to IISAN-E does not consistently result in better performance. This highlights the scalability and efficacy of CROSSAN over existing state-of-the-art efficient adaptation methods. To further validate these findings, we evaluate CROSSAN on the MicroLens-50K dataset (Table III). The relative improvements remain consistent, with one exception: the T+I+V configuration shows comparable improvement to T+I. However, the overall trend of increasing performance with more MFMs and raw modalities persists. These results reaffirm the superiority of CROSSAN.

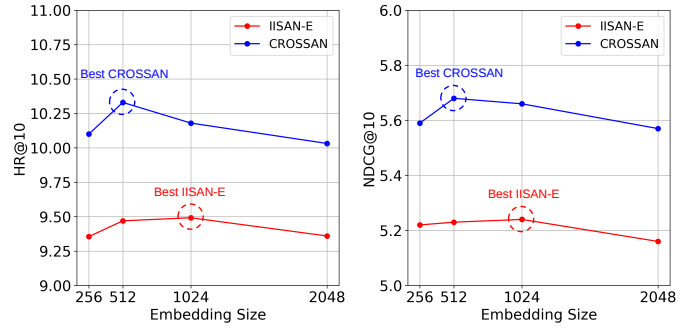


Fig. 3: Optimal Embedding dimension for IISAN-E and CROSSAN

To further compare CROSSAN with additional state-of-the-art multimodal sequential recommendation models (MM-SRs) beyond IISAN, we also include a recent state-of-the-art HM4SR model in our evaluation, as summarized in Table IV. Across all evaluation metrics, CROSSAN consistently outperforms these strong baselines, demonstrating its superior effectiveness. These results further highlight the advantages of adapting raw modalities through multimodal foundation models (MFMs).

(Answer to RQ1): After extensive evaluation, we conclude that CROSSAN demonstrates superior scalability compared to the state-of-the-art efficient adaptation approaches. This is evidenced by its more substantial performance improvement when additional raw modalities are incorporated.

B. Efficiency Evaluation (RQ2)

In this section, we explore the efficiency of CROSSAN in terms of three dimensions: Training time, GPU memory, and Trainable Parameters, following the work by [15]. We primarily focus on the **training-time** and **GPU Memory**, since they are the most important aspects of efficiency in practical settings. Due to computational limitations, we were only able to evaluate the efficiency of traditional adaptation approaches on image and text modalities⁶. Therefore, we primarily compare CROSSAN with IISAN-E (The intuitive extension of IISAN [15]), as other methods are too computationally intensive and not suitable for direct comparison with CROSSAN. In the following, we default to reporting efficiency based on the best performance.

As shown in Table V, traditional adaptation methods with only two modalities reach the maximum of our available GPU memory. While IISAN-E reduces GPU memory, our proposed method, CROSSAN, is even more efficient and achieves better performance (see Table III). We attribute this observation to two key factors: (1) CROSSAN reduces the computational burden by using only one adapter tower per modality in each Multimodal Foundation Model (MFM), whereas IISAN-E requires an additional inter-modal adapter tower; (2) Hyperparameter tuning revealed that IISAN-E achieves its best per-

⁶Adopting full fine-tuning or PEFT for the four encoders is nearly impossible given our GPU resources (we have access to only one A6000 GPU). Therefore, we leave this investigation for future work or for institutions with more extensive computational resources.

TABLE III: Performance comparison of CROSSAN on Microlens-100K and Microlens-50K with four types of raw modalities (Text, Image, Video and Audio, denoted as T, I, V, and A, respectively). “*” indicates that the improvements of the best models compared with previous state-of-the-art methods are significant at the level of 0.05 with paired T-test. ‘Relative Improvement’ is computed in comparison to the corresponding IISAN and its extension.

Model	Microlens-100K				Microlens-50K			
	HR@10	NDCG@10	HR@20	NDCG@20	HR@10	NDCG@10	HR@20	NDCG@20
Full Finetuning (T+I)	0.0934	0.0499	0.1363	0.0607	0.0772	0.0408	0.1129	0.0498
Adapter (T+I)[11]	0.0962	0.0514	0.1376	0.0618	0.0765	0.0407	0.1132	0.0500
LoRA (T+I)[13]	0.0866	0.0462	0.1298	0.0571	0.0644	0.0331	0.0975	0.0415
IISAN (T+I)[15]	0.0960	0.0526	0.1366	0.0628	0.0771	0.0421	0.1121	0.0509
IISAN-E (T+I+A)	0.0953	0.0523	0.1353	0.0623	0.0777	0.0422	0.1137	0.0513
IISAN-E (T+I+V)	0.0939	0.0517	0.1341	0.0619	0.0775	0.0428	0.1150	0.0522
IISAN-E (T+I+V+A)	0.0949	0.0524	0.1350	0.0625	0.0790	0.0430	0.1135	0.0517
CROSSAN (T+I) (ours)	0.0999*	0.0553*	0.1428*	0.0661*	0.0806*	0.0431*	0.1177*	0.0524*
CROSSAN (T+I+A) (ours)	0.1012*	0.0557*	0.1445*	0.0666*	0.0811*	0.0444*	0.1188*	0.0539*
CROSSAN (T+I+V) (ours)	0.1006*	0.0557*	0.1428*	0.0663*	0.0808*	0.0437*	0.1183*	0.0531*
CROSSAN (T+I+V+A) (ours)	0.1033*	0.0568*	0.1452*	0.0673*	0.0847*	0.0462*	0.1222*	0.0557*
Relative Improvement								
Text+Image	+3.86%	+4.99%	+4.43%	+5.03%	+4.25%	+2.24%	+4.81%	+2.99%
Text+Image+Audio	+5.87%	+6.17%	+6.39%	+6.43%	+4.17%	+4.95%	+4.34%	+4.88%
Text+Image+Video	+6.70%	+7.18%	+6.04%	+6.74%	+4.06%	+2.04%	+2.74%	+1.75%
Text+Image+Video+Audio	+8.10%	+7.70%	+7.00%	+7.15%	+6.73%	+6.97%	+7.13%	+7.21%

TABLE IV: Performance comparison for state-of-the-art multimodal sequential recommendation approaches MicroLens datasets with respect to Hit Ratio (HR@K), and NDCG (N@K). The best results are in bold.

Method	HR@10	NDCG@10	HR@20	NDCG@20
MicroLens-100K				
HM4SR[34] (WWW’25)	0.0807	0.1207	0.0437	0.0538
IISAN[15] (SIGIR’24)	0.0960	0.1366	0.0526	0.0628
CROSSAN(T+I)	0.0999	0.1428	0.0553	0.0661
CROSSAN (T+I+V+A)	0.1033	0.1452	0.0568	0.0673
MicroLens-50K				
HM4SR[34] (WWW’25)	0.0661	0.0976	0.0367	0.0446
IISAN[15] (SIGIR’24)	0.0771	0.1121	0.0421	0.0509
CROSSAN (T+I)	0.0806	0.1177	0.0431	0.0524
CROSSAN (T+I+V+A)	0.0847	0.1222	0.0462	0.0557

TABLE V: Efficiency Comparison. TT, GM, and TP stand for Training Time (seconds/epoch), GPU Memory (MB), and Trainable Parameters, respectively. A lower value for each metric indicates an improvement in efficiency. We demonstrate the improvement over IISAN-E.

Method	TT (↓)	GM (↓)	TP (↓)
Full Finetuning (T+I)	3,278	45,886	194,897,216
Adapter(T+I)	2,856	35,652	38,017,088
LoRA(T+I)	3,110	36,902	37,992,512
IISAN-E _{best} (T+I+V+A)	213	5,556	58,432,824
IISAN-E _{same} (T+I+V+A)	207	4,476	30,889,784
CROSSAN (T+I+V+A) (ours)	144	4,272	24,741,456
Improvement _{best}	+32.39%	+23.11%	+57.66%
Improvement _{same}	+30.43%	+4.56%	+19.90%

formance with an embedding size of 1024, while CROSSAN reaches optimal performance with a hidden dimension of only

TABLE VI: Ablation Study on Fusion Method. D-Gated and S-Gated Fusion represents Dynamic Gated and Static Gated Fusion. Concat Fusion refers to the direct concatenation of all modalities.

Dataset	Method	HR@10	NDCG@10	HR@20	NDCG@20
Microlens-100K	MOMEF	0.1033	0.0568	0.1452	0.0673
	D-Gated Fusion	0.0994	0.0551	0.1416	0.0657
	S-Gated Fusion	0.0972	0.0532	0.1394	0.0638
	Concat Fusion	0.0965	0.0531	0.1380	0.0636
Microlens-50K	MOMEF	0.0847	0.0462	0.1222	0.0557
	D-Gated Fusion	0.0796	0.0430	0.1196	0.0530
	S-Gated Fusion	0.0807	0.0437	0.1171	0.0529
	Concat Fusion	0.0816	0.0442	0.1181	0.0534

TABLE VII: Ablation Study on Cross- Vs. Independent-Modal. H and N represent the Hit Ratio and NDCG.

Dataset	Modality	HR@10	NDCG@10	HR@20	NDCG@20
Microlens-100K	Cross-modal	0.1033	0.0568	0.1452	0.0673
	Independent	0.0993	0.0546	0.1418	0.0654
Microlens-50K	Cross-modal	0.0847	0.0462	0.1222	0.0557
	Independent	0.0815	0.0439	0.1169	0.0528

512, as shown in Figure 3. This results in IISAN-E having more trainable parameters, which contributes to its reduced efficiency. Even when IISAN is configured with the same embedding dimension as CROSSAN, it remains less efficient due to the additional inter-modal adapter tower. (**Answer to RQ2**): Although CROSSAN achieves superior performance, it also exceeds the efficiency of the state-of-the-art IISAN model.

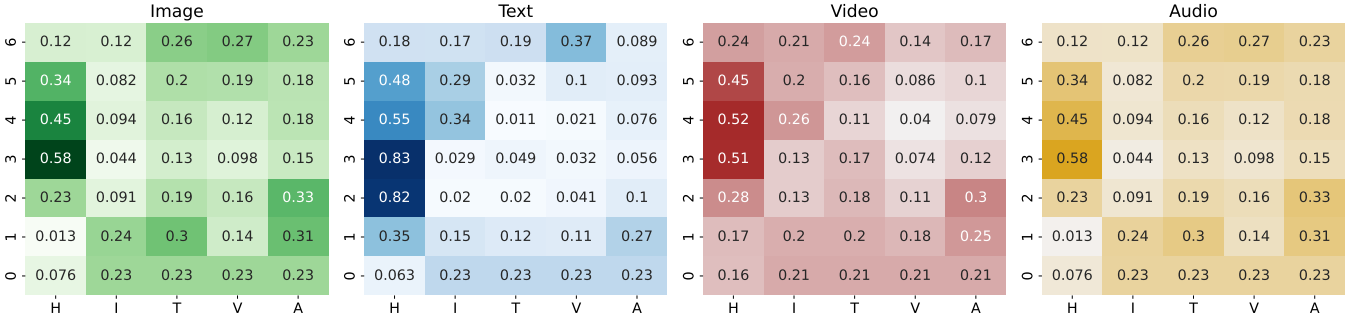


Fig. 4: The heatmap visualization of CROSSAN’s gates, where deeper colors indicate higher values. The y-axis corresponds to the layers of the side adapter, ranging from bottom (0) to top (6). “H” denotes the hidden state of the current layer’s modality, while “I”, “T”, “V”, and “A” represent the gate values for the previous side adapter layer’s outputs corresponding to their respective modality.

C. Ablation Study (RQ3)

In this section, we present an ablation study focusing on two key components of CROSSAN: the fusion mechanism, MOMEF, and cross-modal interaction.

For the fusion mechanism, we compare MOMEF against three commonly used fusion methods: (1) Concat Fusion, which is widely adopted in existing literature [35], [15], [8]; (2) Static Gated Fusion (S-Gated), where a learnable gate is assigned to each modality and remains fixed after training, applying the same weights to all items; and (3) Dynamic Gated Fusion (D-Gated), which leverages a fully connected layer to generate different weights for each item, allowing more flexibility based on task-specific inputs. The proposed method, MOMEF, utilizes dynamic gating of input modalities to selectively activate the top-k modalities for each item. This approach offers a more fine-grained mechanism, allowing not only the weighting of modalities but also the precise selection of the most useful modalities.

As shown in Table VI, MOMEF outperforms all other methods, demonstrating its superior ability to fuse multiple modalities efficiently. While the D-gated method achieves the second-best results on the Microlens-100K dataset but it underperforms on the Microlens-50K dataset across three metrics when compared to direct concat. This suggests that other approaches may be less robust across different datasets.

Regarding cross-modal interaction, the results in Table VII further confirm our preliminary findings (section II): incorporating cross-modal interactions significantly enhances multimodal learning, validating its importance in achieving better performance. Additionally, we observed that the gating mechanism is crucial; without it, the model struggles to train properly, leading to a collapse in performance. We present the heatmap in Figure 4, where we observe that the gate values for the hidden states in the corresponding modality tower are significantly larger in the middle layers, while the values in the lower layers are relatively smaller. This emphasizes the importance of the middle layers in MFMs. **(Answer to RQ3):** Both MOMEF and cross-modal interaction contribute to the overall performance of CROSSAN, with each approach demonstrating its usefulness.

TABLE VIII: Top-K experts for CROSSAN.

Dataset	Top-K	HR@10	NDCG@10	HR@20	NDCG@20
Microlens-100K	4	0.0992	0.0545	0.1422	0.0653
	3	0.1002	0.0558	0.1459	0.0673
	2	0.1033	0.0568	0.1452	0.0673
	1	0.1014	0.0560	0.1444	0.0668
Microlens-50K	4	0.0821	0.0444	0.1189	0.0536
	3	0.0847	0.0462	0.1222	0.0557
	2	0.0815	0.0438	0.1197	0.0534
	1	0.0822	0.0451	0.1178	0.0541

D. Hyperparameter Analysis (RQ4)

In this section, we explore three key hyperparameters: (1) learning rate, (2) hidden dimension, and (3) number of experts. The former two are fundamental parameters commonly explored in adapter-based recommendation models [12], while the number of experts is introduced by MOMEF. To optimize the model’s performance, we conduct an extensive hyperparameter search for these settings.

Embedding Dimension. As shown in Figure 5, the performance of CROSSAN demonstrates a clear dependency on the embedding dimension size. A relatively large embedding dimension is essential for effective adaptation. In contrast, smaller embedding leads to noticeable drops in performance. However, due to the efficiency of CROSSAN’s adapter-based architecture, increasing the embedding dimension does not result in significant computational overhead, such as extended training time or excessive GPU memory usage. Notably, when scaling the embedding size up to 8192, performance remains stable, suggesting that CROSSAN maintains its effectiveness as long as the embedding size exceeds a certain threshold.

Learning Rate. Figure 5 illustrates the effect of learning rate on model performance. CROSSAN’s performance appears to remain stable once the learning rate exceeds $5e-5$. However, using a smaller learning rate (e.g., $1e-5$) results in suboptimal performance. This highlights the necessity of fine-tuning this hyperparameter within a suitable range to achieve optimal performance.

Top-K experts. The Top-K is a new hyperparameter introduced through MOMEF. As shown in Table VIII, the optimal number of experts differs across datasets. For example, the

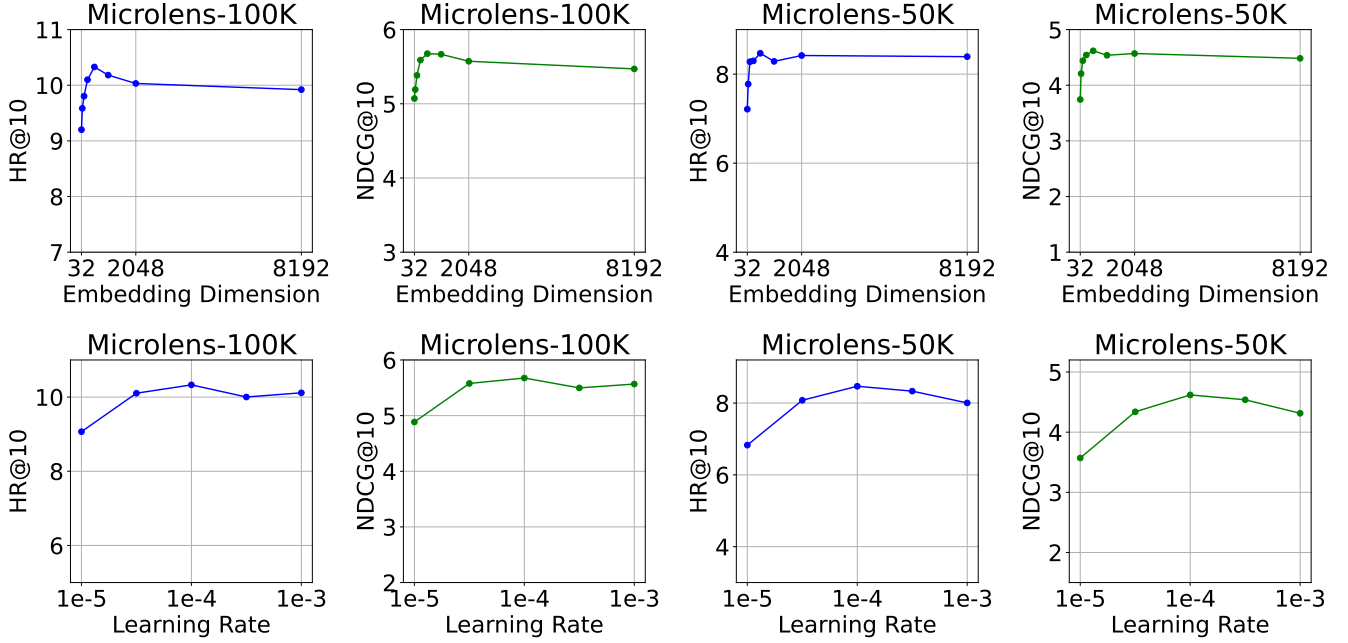


Fig. 5: The first row of figures depicts CROSSAN’s embedding dimension adaptation, where the embedding dimensions of the side adapters range from 32 to 8192. The second row of figures illustrates CROSSAN’s learning rate adaptation, with the learning rate varying between 1e-5 to 1e-3.

Microlens-100K dataset performs best with two experts, while three experts yield the best results for the Microlens-50K dataset. These findings indicate that both very large and small numbers of experts are ineffective.

(Answer to RQ4): Based on the experiments, we conclude two observations from the hyperparameter analysis: (1) The embedding dimension and learning rate for CROSSAN should be set within a large range to ensure stable performance. (2) Top-K experts are dataset-specific and typically lie within a moderate range.

VII. MORE EVALUATIONS

Exploration of More Datasets. To further investigate the effectiveness of CROSSAN, we conducted experiments on three Amazon datasets used in [15]. We compared the performance of CROSSAN against full fine-tuning and various PEFT methods, including Adapter, LoRA, BitFit, and IISAN. Our results show that CROSSAN achieves the best adaptation performance across all three datasets using only two modalities as described in Table IX. Although the improvement over IISAN on these datasets is not on a large margin, CROSSAN offers greater efficiency by requiring only two side adapter towers, compared to the three needed by IISAN. Consequently, even with just two modalities, CROSSAN proves to be an optimal solution for these datasets.

Exploration of LLM and LVM. To assess the impact of incorporating a recent large language model (LLM)⁷ and large vision-language model (LVM)⁸, we follow the implementation

of [36] by replacing the original text tower. This substitution yields a slight performance improvement for CROSSAN when using Qwen-VL-7B, while Qwen-7B performs marginally worse in terms of NDCG@10 and marginally improves the other metrics (Table X). Nonetheless, while the implementation of LLMs can provide slight performance gains, the simple BERT-based text tower used in our main paper already delivers strong and competitive results.

More Investigation on Independent Variant. In some cases, the independent side-adapter variant still has merit, as it avoids layer-wise cross-modal operations and may be more suitable for asymmetrical architectures [36]. Understanding why it underperforms compared to the cross-modal approach is therefore meaningful. A natural hypothesis is that the independent variant underperforms because it lacks sufficient interaction during layer-wise fusion, and the sparse fusion in MOMEF may not fully realize its potential. To investigate this, we first examine the effect of varying the number of experts to mitigate the sparsity factor, as summarized in Table XII. The results indicate that the independent approach requires a relatively larger number of experts in MOMEF. In contrast, the cross-modal method reduces the optimal number of experts by promoting a more unified feature representation, thanks to its higher mutual information.

To further enhance the final fusion interaction, we adopt a commonly used cross-attention fusion module to the independent variant. Unlike MOMEF, this module performs dense computation (see Table XI). The results support our hypothesis: cross-attention yields small improvements over MOMEF fusion in the independent variant (e.g., HR@10 improves from 0.0993 to 0.1008), aligning with the intuition

⁷<https://huggingface.co/Qwen/Qwen2-7B>

⁸<https://huggingface.co/Qwen/Qwen2-VL-7B>

TABLE IX: Performance (%) comparison across methods on three datasets. Best results in bold, second-best underlined.

Method	Scientific		Instrument		Office	
	HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10
FFT	6.62	4.06	8.76	6.76	6.30	4.58
Adapter	6.61	3.91	8.82	6.83	6.65	4.85
LoRA	6.62	4.09	8.43	6.64	6.55	4.78
BitFit	6.37	3.76	8.65	6.71	6.78	4.87
IISAN	6.83	4.14	9.06	7.01	6.80	4.92
CROSSAN	6.90	4.22	9.08	7.05	6.83	4.95

TABLE X: Performance comparison on MicroLens-100K datasets. The best results are in bold. The second best is underlined.

Method	H@10	N@10	H@20	N@20
CROSSAN	0.1033	0.1452	<u>0.0568</u>	0.0673
CROSSAN (QWen-7B)	0.1033	0.1464	<u>0.0564</u>	<u>0.0674</u>
CROSSAN (QWen-VL-7B)	0.1033	<u>0.1460</u>	0.0570	0.0677

TABLE XI: Performance Comparison of Independent vs. Cross-Modal Architectures with Cross-Attention Fusion on the MicroLens-100K Dataset.

Modality	H@10	N@10	H@20	N@20
Independent (w/ MOMEF Fusion)	0.0993	0.0546	0.1418	0.0654
Independent (w/ Cross-attn Fusion)	0.1008	0.0558	0.1423	0.0663
Cross-modal (w/ MOMEF Fusion)	0.1033	0.0568	0.1452	0.0673
Cross-modal (w/ Cross-attn Fusion)	0.1014	0.0561	0.1435	0.0668

that cross-attention can help in this setting. However, when integrated into CROSSAN’s interaction layers, performance slightly declined (e.g., Hit@10: 0.1033 $\rightarrow$ 0.1014), suggesting that cross-attention is not always advantageous when strong cross-modal interactions are already present. These findings may provide useful insights for future work that adopts the independent variant, highlighting that cross-attention can be beneficial compared with MOMEF.

VIII. RELATED WORK

Multimodal Foundation Models (MFMs). Recent advances in multimodal learning leverage pre-trained models to enhance downstream task performance while reducing pre-training costs [37], [38], [39], [40], [41]. BERT [2] pioneered the pretraining and fine-tuning paradigm in NLP, while Vision Transformer (ViT) [1] adapted this approach for image classification. CLIP [42] bridged vision and language through contrastive learning, enabling robust zero-shot capabilities. Large-scale models like GPT-4 [43], T5 [44], and multimodal variants like DALL-E [45] and Flamingo [46] expanded the ability to process diverse modalities. In video representation learning, models such as SlowFast-R50 [47], MViT-b [48], and VideoMAE [4] effectively capture both temporal and spatial features. Meanwhile, audio models like Wave2Vec [49] and the Audio Spectrogram Transformer (AST) [5] enhance audio classification by operating on spectrograms. Together, these models highlight the increasing strength of multimodal learning across a variety of tasks.

Efficient Adaptation of MFMs in RS. The RS field has progressively investigated the incorporation of diverse modalities

TABLE XII: Performance comparison of Independent and Cross-modal methods.

Method	Top-K	H@10	N@10	H@20	N@20
Independent	1	0.0973	0.0535	0.1404	0.0643
	2	0.0993	0.0546	0.1418	0.0654
	3	0.0993	0.0546	0.1432	0.0657
	4	0.0989	0.0540	0.1411	0.0646
Cross-modal	1	0.1014	0.0560	0.1444	0.0668
	2	0.1033	0.0568	0.1452	0.0673
	3	0.1002	0.0558	0.1459	0.0673
	4	0.0992	0.0545	0.1422	0.0653

to improve the effectiveness of recommendations [50], [51], [52], [53], [54], [55], [56], [57], [7], [58], [59], [60], [61], [62], [63], [58], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]. Recent studies [77], [7], [6], [8] have demonstrated the superiority of the MoRec framework using end-to-end learning, showing it significantly outperforms traditional approaches that rely on offline feature extraction. For instance, [8] and [15] highlighted the advantages of end-to-end training, which jointly leverages both image and text modalities, compared to methods that employ a single modality. Despite the strong performance of raw content learning, a major drawback of these approaches is the continued dependence on full fine-tuning of large multimodal encoders, leading to performance inefficiencies.

Parameter-efficient fine-tuning (PEFT) methods have made strides in addressing this issue, as shown in works like M6-Rec [78], Tallrec [68], and AdapterRec [12], which demonstrate that PEFT techniques can achieve competitive performance with significantly reduced overhead. However, many PEFT methods still rely on established approaches, often overlooking practical efficiency concerns.

Much of the existing research predominantly focuses on traditional Adapter or LoRA-based solutions and is limited to single-modality adaptation due to the inefficiency of these adaptation approaches. A recent study, IISAN [15], introduced a structure that utilizes independent adapters within each tower, complemented by a single inter-adapter for image and text adaptation. However, the inter-tower has a fixed input and lacks sufficient cross-modal interaction. In contrast, CROSSAN emphasizes cross-modal side adapters with joint learning and updates. Additionally, IISAN considers only image and text scenarios, leaving many other existing modalities unexplored. Upon attempting to expand this method, we concluded that it lacked the desirable scalability for additional modalities. To the best of our knowledge, cross-modality side adapters are largely underrepresented in the existing literature on recommender systems. Furthermore, CROSSAN

investigates the novel area of scalable and efficient adaptation for more (>2) MFMs with raw modalities, which will facilitate future research.

IX. CONCLUSION

CROSSAN provides a scalable, efficient, and effective approach for adapting multiple multimodal foundation models (MFMs) in sequential recommendation tasks. By incorporating cross-modal side adapters along with the Mixture of Modality Expert Fusion (MOMEF), CROSSAN achieves superior performance. Extensive experimental results validate the approach's ability to improve recommendation effectiveness as additional modalities are integrated, demonstrating its superiority over existing methods.

Future research can extend CROSSAN to various multimodal tasks, including multimodal classification [79], retrieval [80], and generative modeling [81], where the integration of diverse data modalities is critical. These directions offer significant potential to improve both the efficiency and performance of multimodal learning across a broad range of applications, positioning CROSSAN as a general adaptation approach for future research.

REFERENCES

- [1] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [3] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [4] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078–10093, 2022.
- [5] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.
- [6] Y. Ni, Y. Cheng, X. Liu, J. Fu, Y. Li, X. He, Y. Zhang, and F. Yuan, "A content-driven micro-video recommendation dataset at scale," *arXiv preprint arXiv:2309.15379*, 2023.
- [7] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni, "Where to go next for recommender systems? id-vs. modality-based recommender models revisited," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, p. 2639–2649.
- [8] Y. Li, H. Du, Y. Ni, P. Zhao, Q. Guo, F. Yuan, and X. Zhou, "Multimodality is all you need for transferable recommender systems," *arXiv preprint arXiv:2312.09602*, 2023.
- [9] C. Wu, F. Wu, T. Qi, and Y. Huang, "Empowering news recommendation with pre-trained language models," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1652–1656.
- [10] Q. Liu, J. Hu, Y. Xiao, J. Gao, and X. Zhao, "Multimodal recommender systems: A survey," *arXiv preprint arXiv:2302.03883*, 2023.
- [11] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.
- [12] J. Fu, F. Yuan, Y. Song, Z. Yuan, M. Cheng, S. Cheng, J. Zhang, J. Wang, and Y. Pan, "Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 208–217.
- [13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [14] Y. Zhang, F. Feng, C. Wang, X. He, M. Wang, Y. Li, and Y. Zhang, "How to retrain recommender system? a sequential meta-learning method," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1479–1488.
- [15] J. Fu, X. Ge, X. Xin, A. Karatzoglou, I. Arapakis, J. Wang, and J. M. Jose, "Lisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 687–697.
- [16] Y.-L. Sung, J. Cho, and M. Bansal, "Lst: Ladder side-tuning for parameter and memory efficient transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12991–13005, 2022.
- [17] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.
- [18] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [19] W. Guo, H. Huang, X. Kong, and R. He, "Learning disentangled representation for cross-modal retrieval with deep mutual information estimation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1712–1720.
- [20] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1893–1902.
- [21] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, and W. M. Wells, "Multimodal representation learning via maximization of local mutual information," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 273–283.
- [22] H. Wu, X. Zhang, J. Tian, S. Wu, C. Dou, Y. Sun, and Z. Feng, "A mutual information-based disentanglement framework for cross-modal retrieval," in *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*. Springer, 2021, pp. 585–596.
- [23] C. Gu, J. Bu, X. Zhou, C. Yao, D. Ma, Z. Yu, and X. Yan, "Cross-modal image retrieval with deep mutual information maximization," *Neurocomputing*, vol. 496, pp. 166–177, 2022.
- [24] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 671–15 680.
- [25] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.
- [26] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.
- [27] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.
- [28] X. Yi, J. Yang, L. Hong, D. Z. Cheng, L. Heldt, A. Kumthekar, Z. Zhao, L. Wei, and E. Chi, "Sampling-bias-corrected neural modeling for large corpus item recommendations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 269–277.
- [29] W. Ji, X. Liu, A. Zhang, Y. Wei, Y. Ni, and X. Wang, "Online distillation-enhanced multi-modal transformer for sequential recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 955–965.
- [30] W. Krichene and S. Rendle, "On sampled metrics for item recommendation," in *KDD*, 2020.
- [31] S. M. Jain, "Hugging face," in *Introduction to transformers for NLP: With the hugging face library and models to solve problems*. Springer, 2022, pp. 51–67.
- [32] Z. Dong, X. Liu, B. Chen, P. Polak, and P. Zhang, "Musechat: A conversational music recommendation system for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 775–12 785.

- [33] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," *arXiv preprint arXiv:2210.12316*, 2022.
- [34] S. Zhang, L. Chen, D. Shen, C. Wang, and H. Xiong, "Hierarchical time-aware mixture of experts for multi-modal sequential recommendation," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 3672–3682.
- [35] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021.
- [36] J. Fu, X. Ge, X. Xin, A. Karatzoglou, I. Arapakis, K. Zheng, Y. Ni, and J. M. J. Joemon, "Efficient and effective adaptation of multimodal foundation models in sequential recommendation," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2025.
- [37] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang, "Generative pretraining in multimodality," *arXiv preprint arXiv:2307.05222*, 2023.
- [38] Z. Chen, Y. Zhang, Y. Fang, Y. Geng, L. Guo, X. Chen, Q. Li, W. Zhang, J. Chen, Y. Zhu *et al.*, "Knowledge graphs meet multi-modal learning: A comprehensive survey," *arXiv preprint arXiv:2402.05391*, 2024.
- [39] Z. Zhang, S. Cao, and Y.-X. Wang, "Tamm: Triadapter multi-modal learning for 3d shape understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 413–21 423.
- [40] Y. He, J. Fu, K. Zheng, S. Xu, F. Chen, J. Li, J. M. Jose, and X. Ge, "Double-filter: Efficient fine-tuning of pre-trained vision-language models via patch&layer filtering," in *Forty-second International Conference on Machine Learning*, 2025.
- [41] X. Ge, J. Fu, F. Chen, S. An, N. Sebe, and J. M. Jose, "Towards end-to-end explainable facial action unit recognition via vision-language joint learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8189–8198.
- [42] A. Radford, J. W. Kim *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [43] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [45] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.
- [46] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.
- [47] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.
- [48] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6824–6835.
- [49] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [50] L. Zhang, X. Zhou, Z. Zeng, and Z. Shen, "Multimodal pre-training for sequential recommendation via contrastive learning," *ACM Transactions on Recommender Systems*, vol. 3, no. 1, pp. 1–23, 2024.
- [51] Y. Ye, Z. Zheng, Y. Shen, T. Wang, H. Zhang, P. Zhu, R. Yu, K. Zhang, and H. Xiong, "Harnessing multimodal large language models for multimodal sequential recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 13 069–13 077.
- [52] X. Zhou, "Mmrec: Simplifying multimodal recommendation," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, 2023, pp. 1–2.
- [53] H. Hu, W. Guo, Y. Liu, and M.-Y. Kan, "Adaptive multi-modalities fusion in sequential recommendation systems," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 843–853.
- [54] Q. Liu, J. Zhu, Y. Yang, Q. Dai, Z. Du, X.-M. Wu, Z. Zhao, R. Zhang, and Z. Dong, "Multimodal pretraining, adaptation, and generation for recommendation: A survey," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6566–6576.
- [55] Z. Tang, Z. Huan, Z. Li, X. Zhang, J. Hu, C. Fu, J. Zhou, and C. Li, "One model for all: Large language models are domain-agnostic recommendation systems," *arXiv preprint arXiv:2310.14304*, 2023.
- [56] C. Wu, F. Wu, T. Qi, and Y. Huang, "Mm-rec: multimodal news recommendation," *arXiv preprint arXiv:2104.07407*, 2021.
- [57] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1405–1414.
- [58] Y. Wei, Y. Xu, L. Zhu, J. Ma, and J. Huang, "Fumner: A fine-grained self-supervised momentum distillation framework for multimodal recommendation," *Information Processing & Management*, vol. 61, no. 5, p. 103776, 2024.
- [59] R. Li, W. Deng, Y. Cheng, Z. Yuan, J. Zhang, and F. Yuan, "Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights," *arXiv preprint arXiv:2305.11700*, 2023.
- [60] J. Wu, X. Yu, K. He, Z. Gao, and T. Gong, "Promise: A pre-trained knowledge-infused multimodal representation learning framework for medication recommendation," *Information Processing & Management*, vol. 61, no. 4, p. 103758, 2024.
- [61] W. Wei, C. Huang, L. Xia, and C. Zhang, "Multi-modal self-supervised learning for recommendation," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 790–800.
- [62] J. Wang, Z. Zeng, Y. Wang, Y. Wang, X. Lu, T. Li, J. Yuan, R. Zhang, H.-T. Zheng, and S.-T. Xia, "Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6548–6557.
- [63] Y. Cheng, Y. Pan, J. Zhang, Y. Ni, A. Sun, and F. Yuan, "An image dataset for benchmarking recommender systems with raw pixels," *arXiv preprint arXiv:2309.06789*, 2023.
- [64] Y. Liu, E. Yang, Y. Dang, G. Guo, Q. Liu, Y. Liang, L. Jiang, and X. Wang, "Id embedding as subtle features of content and structure for multimodal recommendation," *arXiv preprint arXiv:2311.05956*, 2023.
- [65] Z. Qu, R. Xie, C. Xiao, Y. Yao, Z. Liu, F. Lian, Z. Kang, and J. Zhou, "Thoroughly modeling multi-domain pre-trained recommendation as language," *arXiv preprint arXiv:2310.13540*, 2023.
- [66] H. Hu, Q. Liu, C. Li, and M.-Y. Kan, "Lightweight modality adaptation to sequential recommendation via correlation supervision," *arXiv preprint arXiv:2401.07257*, 2024.
- [67] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu, "Once: Boosting content-based recommendation with both open-and closed-source large language models," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 452–461.
- [68] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "Tallrec: An effective and efficient tuning framework to align large language model with recommendation," *arXiv preprint arXiv:2305.00447*, 2023.
- [69] X. Zhang, B. Xu, F. Ma, C. Li, L. Yang, and H. Lin, "Beyond co-occurrence: Multi-modal session-based recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [70] Z. Huan, K. Ding, A. Li, X. Zhang, X. Min, Y. He, L. Zhang, J. Zhou, L. Mo, J. Gu *et al.*, "Exploring multi-scenario multi-modal ctr prediction with a large scale dataset," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 1232–1241.
- [71] X. Wang, H. Chen, Z. Pan, Y. Zhou, C. Guan, L. Sun, and W. Zhu, "Automated disentangled sequential recommendation with large language models," *ACM Transactions on Information Systems*.
- [72] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, H. Zhang, Y. Liu, C. Wu, X. Li, C. Zhu *et al.*, "How can recommender systems benefit from large language models: A survey," *arXiv preprint arXiv:2306.05817*, 2023.
- [73] N. Khan and D. S. Sisodia, "Exploiting diffusion-based structured learning for item interactions representations in multimodal recommender systems," *Information Processing & Management*, vol. 62, no. 3, p. 104075, 2025.
- [74] P. Zhou, C. Liu, J. Ren, X. Zhou, Y. Xie, M. Cao, Z. Rao, Y.-L. Huang, D. Chong, J. Liu *et al.*, "When large vision language models meet multimodal sequential recommendation: An empirical study," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 275–292.
- [75] Y. Qu, L. Qu, T. Chen, Q. V. H. Nguyen, and H. Yin, "Efficient multimodal streaming recommendation via expandable side mixture-of-experts," *arXiv preprint arXiv:2508.05993*, 2025.
- [76] J. Fu, X. Ge, X. Xin, H. Yu, Y. Feng, A. Karatzoglou, I. Arapakis, and J. Jose, "The 1st erel@ mir workshop on efficient representation learning

for multimodal information retrieval,” in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2149–2152.

- [77] S. Elsayed, L. Brinkmeyer, and L. Schmidt-Thieme, “End-to-end image-based fashion recommendation,” *arXiv preprint arXiv:2205.02923*, 2022.
- [78] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, “M6-rec: Generative pretrained language models are open-ended recommender systems,” *arXiv preprint arXiv:2205.08084*, 2022.
- [79] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, “Multimodal classification of remote sensing images: A review and future directions,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.
- [80] D. Rafailidis, S. Manolopoulou, and P. Daras, “A unified framework for multimodal retrieval,” *Pattern Recognition*, vol. 46, no. 12, pp. 3358–3370, 2013.
- [81] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, and P. Fung, “Enabling multimodal generation on clip via vision-language knowledge distillation,” *arXiv preprint arXiv:2203.06386*, 2022.



Junchen Fu is a third-year PhD candidate under the supervision of Prof. Joemon Jose at the School of Computing Science, University of Glasgow. He has contributed to several leading AI conferences and journals, including ICML, SIGIR, WWW, WSDM, CIKM, MM, CVPR, and TPAMI. He also serves as the program committee member and invited reviewer for premier conferences and journals, such as ICLR, SIGIR, KDD, WSDM, CIKM, AAAI, MM, TKDE, TOIS, TMM, and TORS. His research primarily focuses on enhancing the efficiency of adapting

multimodal foundation models for recommendation.



Yongxin Ni is currently a PhD student at Boston University. He received the B.E. degree in Computer Science and Technology from Shenzhen University, the M.Tech. degree in Intelligent Systems from National University of Singapore. He has contributed to several leading AI conferences and journals, including CVPR, SIGIR, WWW, WSDM, CIKM, AAAI, and TPAMI. He also serves as reviewers for top-tier conferences and journals, such as ICLR, SIGIR, WSDM, CIKM, AAAI, MM, TKDE, TOIS, TMM, TORS, IP&M, and TIP. His research interests include deep learning, multi-modality representation and recommender systems.

include deep learning, multi-modality representation and recommender systems.



Joemon Jose is a Professor with the School of Computing Science, University of Glasgow, Glasgow, and a member of the Information Retrieval Group. He has published over 300 papers with more than 10,000 Google Scholar citations, and an H-index of 51. He leads the Multimedia Information Retrieval group which investigates research issues related to the above themes. His research focuses around the following three themes: social media analytics, multimodal interaction for information retrieval, and multimedia mining and search.



Ioannis Arapakis is a Principal Research Scientist at Telefónica Research, specializing in behavior interpretation algorithms for user modeling in both offline and online contexts, with a focus on web search. He holds a Ph.D. in Information Retrieval from the University of Glasgow, and an M.Sc. in Information Technology from the Royal Institute of Technology. Previously, he was a Research Scientist at Yahoo Labs, where he worked on data mining, IR, HCI, and multimedia mining projects.



Kaiwen Zheng is a third-year Ph.D. candidate in the School of Computing Science at the University of Glasgow, under the supervision of Prof. Joemon Jose. His research primarily focuses on multimodal expression recognition. He has published in leading AI conferences and journals, including ICML, WWW, and ICME, and serves as an invited reviewer for premier venues such as AAAI, IEEE Transactions on Affective Computing (TAFFC), ICME, and CIKM.



(LLMs) and agents.

Youhua Li is currently pursuing his Ph.D. in Decision Analytics and Operations at the City University of Hong Kong. He received his M.S. degree in Computer Science and Technology from ShanghaiTech University, and his B.S. degree in Automation Engineering from the University of Electronic Science and Technology of China. He has published in top-tier conferences and journals such as CVPR, TPAMI, ICDE, WWW, WSDM, and SIGIR. His research interests span data-driven decision making and operations optimization, large language models



conferences and journals, such as NeurIPS, MM, ICLR, TKDE, TOIS, IJCV, etc.

Xuri Ge is now a tenure-track assistant professor in the school of artificial intelligence, Shandong University. He earned his PhD at the University of Glasgow (UK) and received M.S. degree from Xiamen University (China). His current research interests include computer vision, multimodal representation learning, and information retrieval. He has contributed to several leading conferences and journals, including NeurIPS, ICML, SIGIR, ACM MM, CIKM, WSDM, ACM TIST, and IP&M, etc.

He serves as the PC member and reviewer for