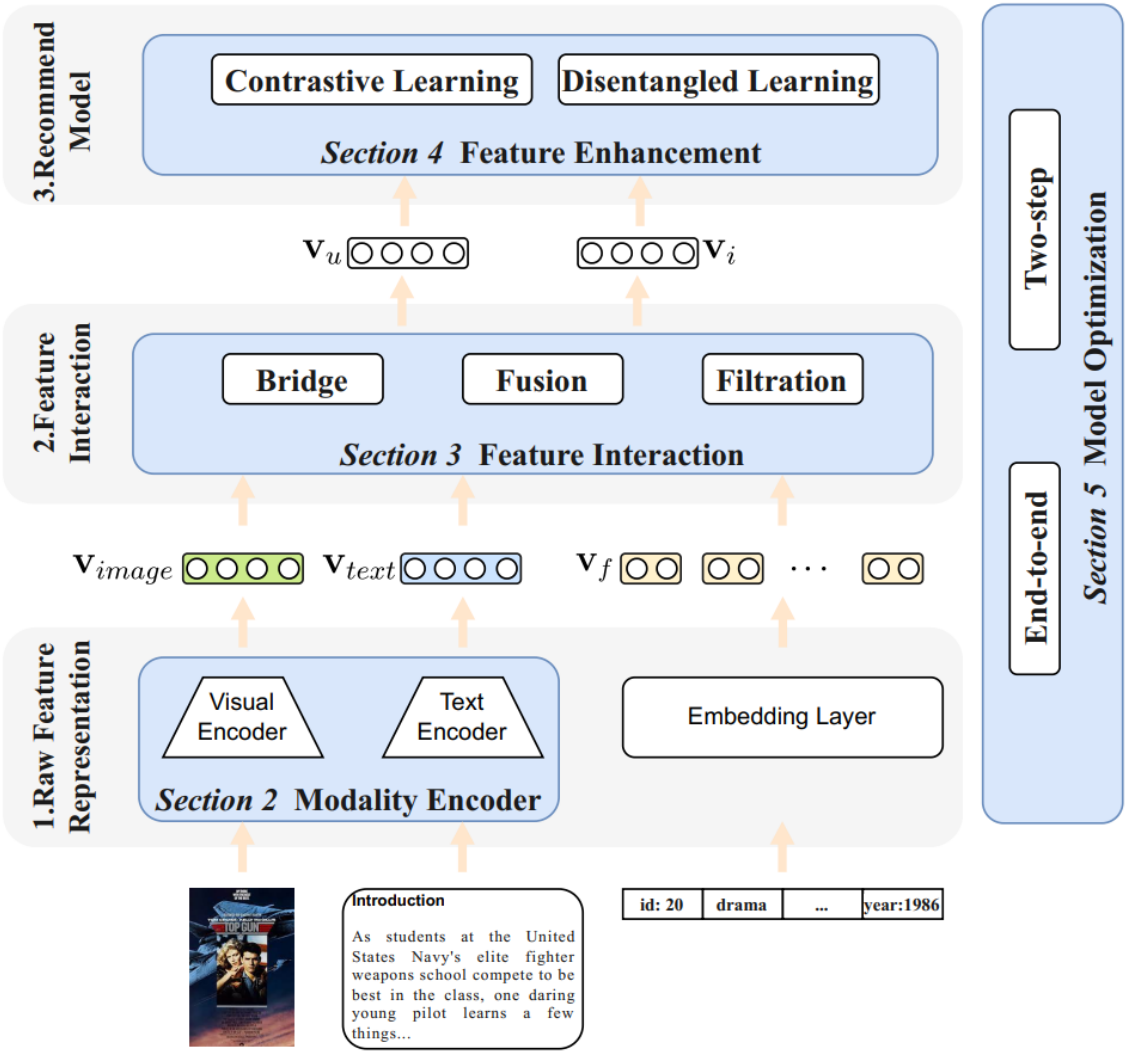
多模态推荐系统综述介绍

刘启东
西安交通大学 & 香港城市大学
liuqidong@stu.xjtu.edu.cn
2024年6月24日

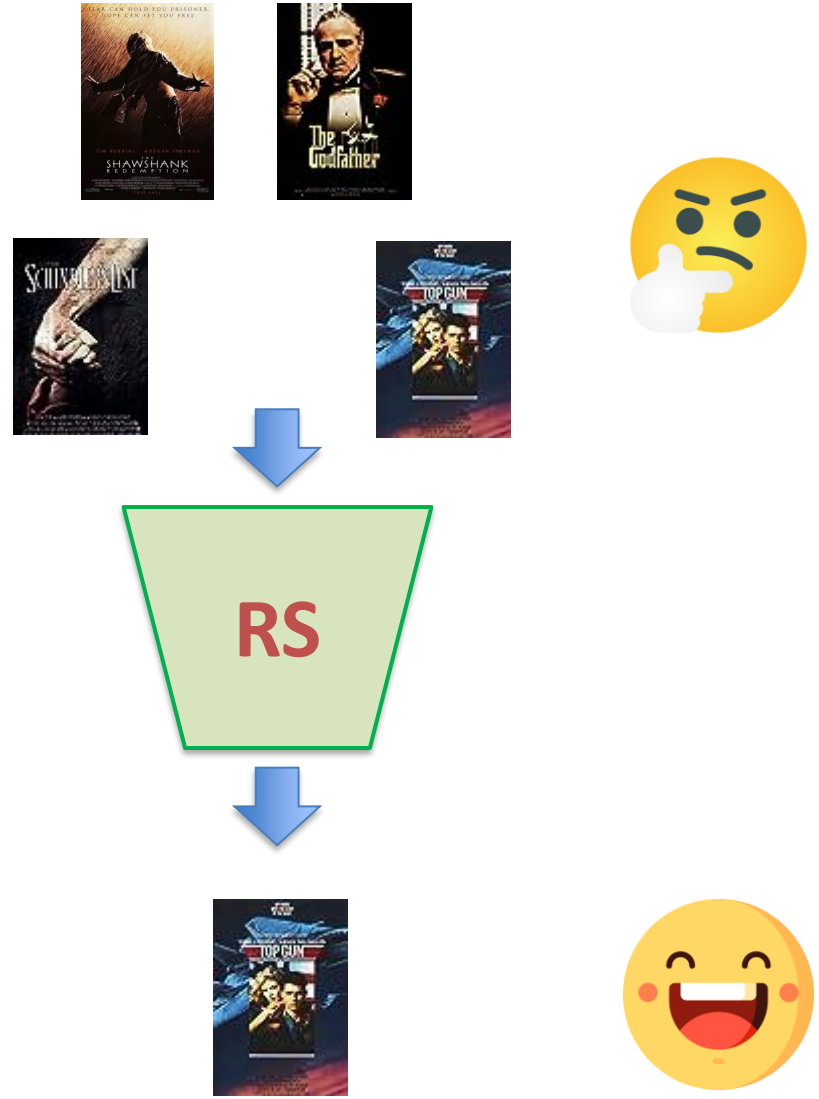
大纲

- 背景和流程
- 模态编码器
- 特征交互
- 特征增强
- 模型优化
- 未来的方向与讨论



背景介绍

- **推荐系统 (Recommender Systems)**
  - 根据用户的兴趣为其推荐合适的物品
  - 可以用于缓解 信息过载问题



背景介绍

- **推荐系统 (Recommender Systems)**
  - 根据用户的兴趣为其推荐合适的物品
  - 可以用于缓解 信息过载问题

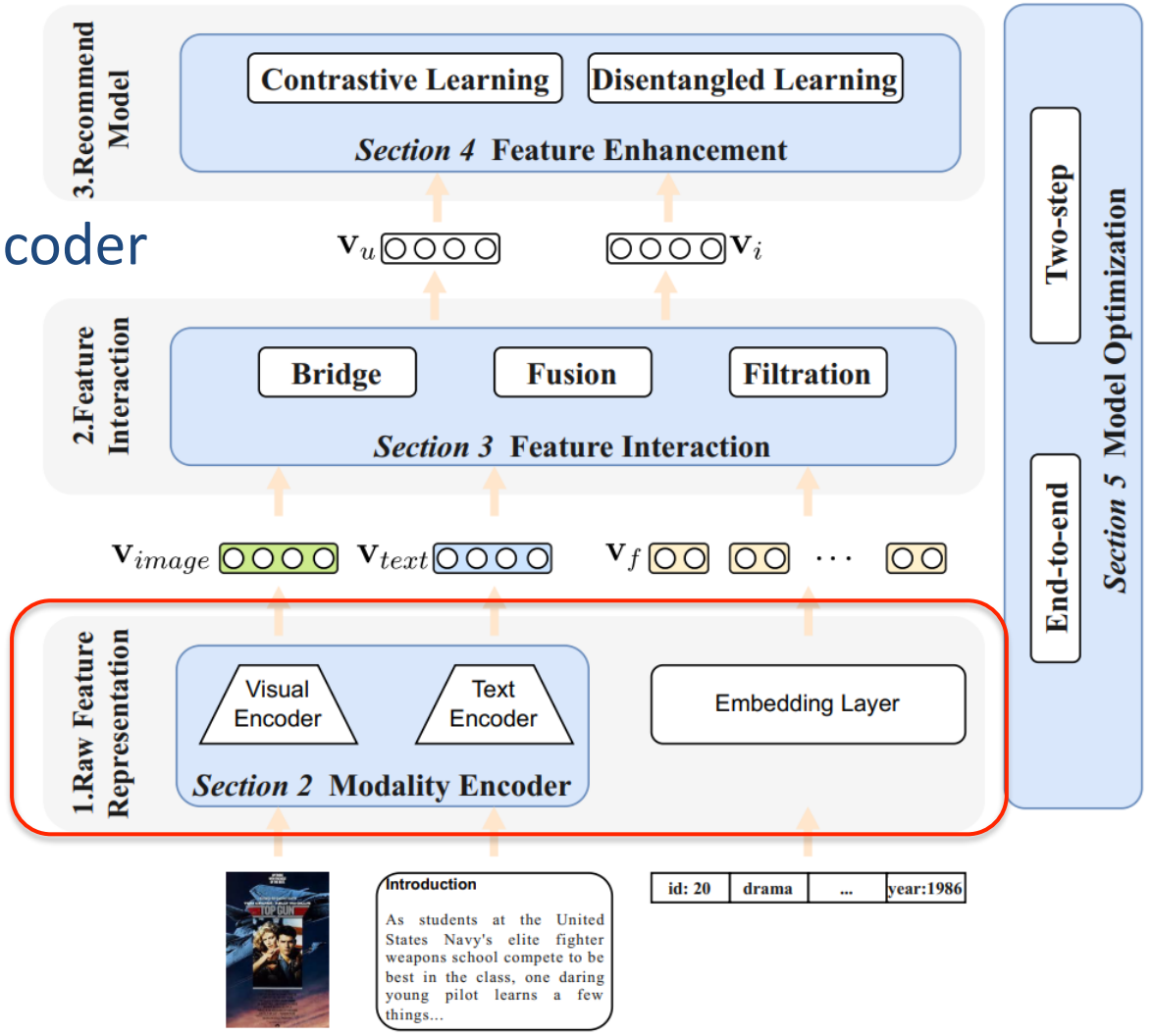
- **多模态推荐系统 (Multimodal Recommender Systems)**
  - 在推荐过程中利用多模态特征
  - 有效缓解 数据稀疏问题
  - 增强推荐系统的 语义理解能力

As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.

Genres: drama, action

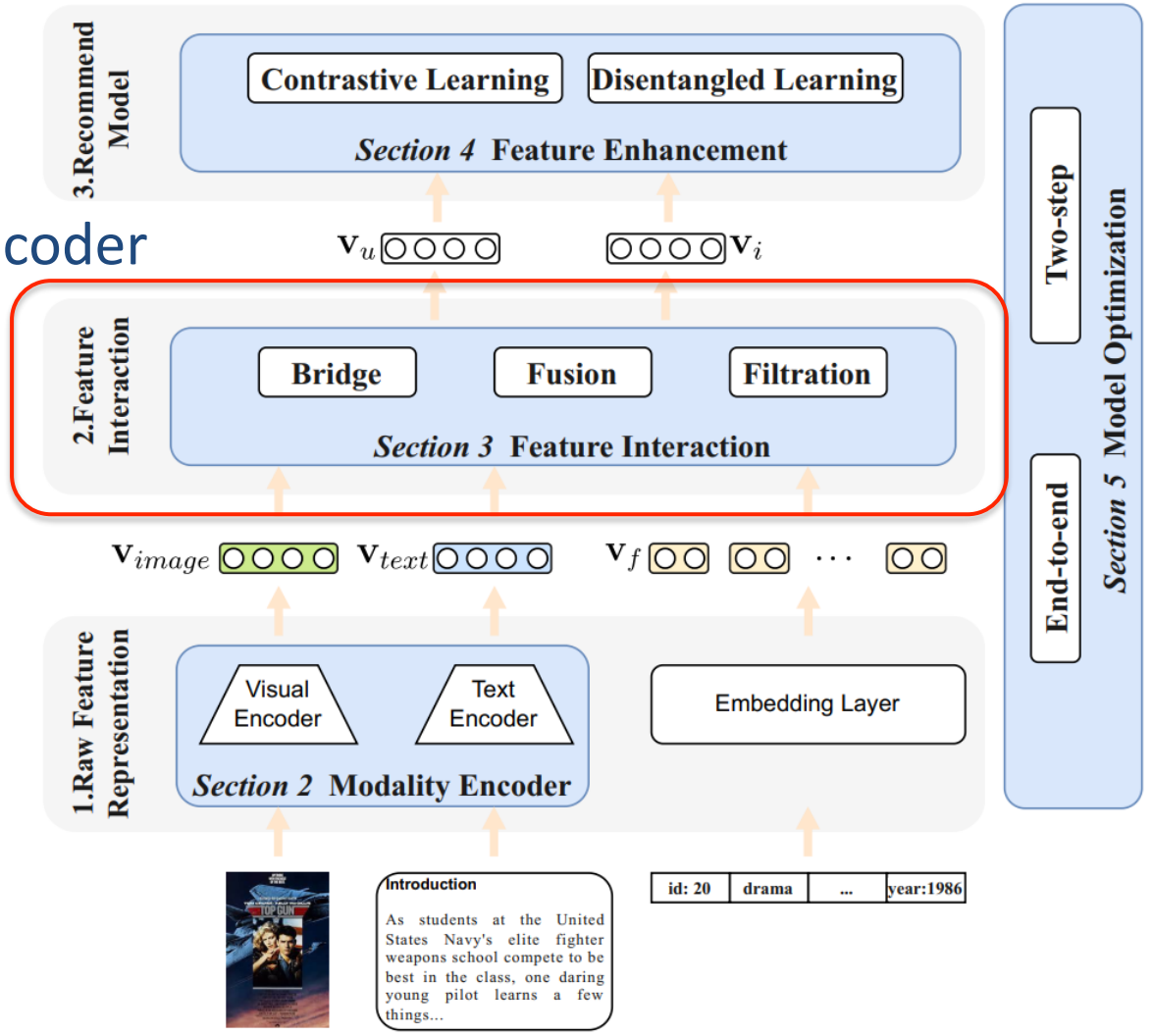
多模态推荐系统流程

- **原始特征表征**
  - 表格特征 – Embedding Layer
  - 多模态特征 (poster, intro) – Modality Encoder



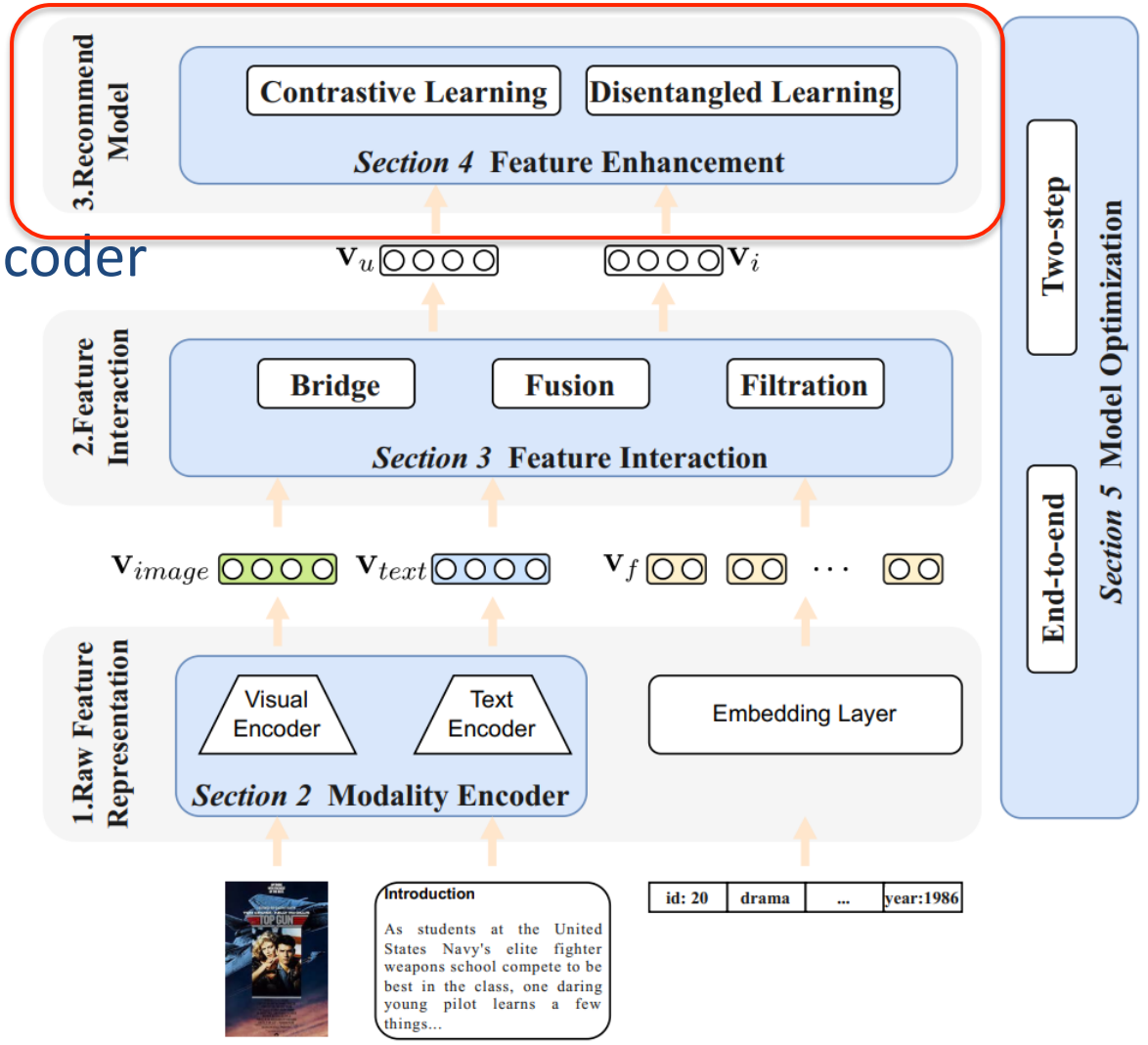
多模态推荐系统流程

- **原始特征表征**
  - 表格特征 – Embedding Layer
  - 多模态特征 (poster, intro) – Modality Encoder
- **特征交互**
  - 将不同的模态的特征映射到同一空间
  - 产生 用户表征 和 物品表征



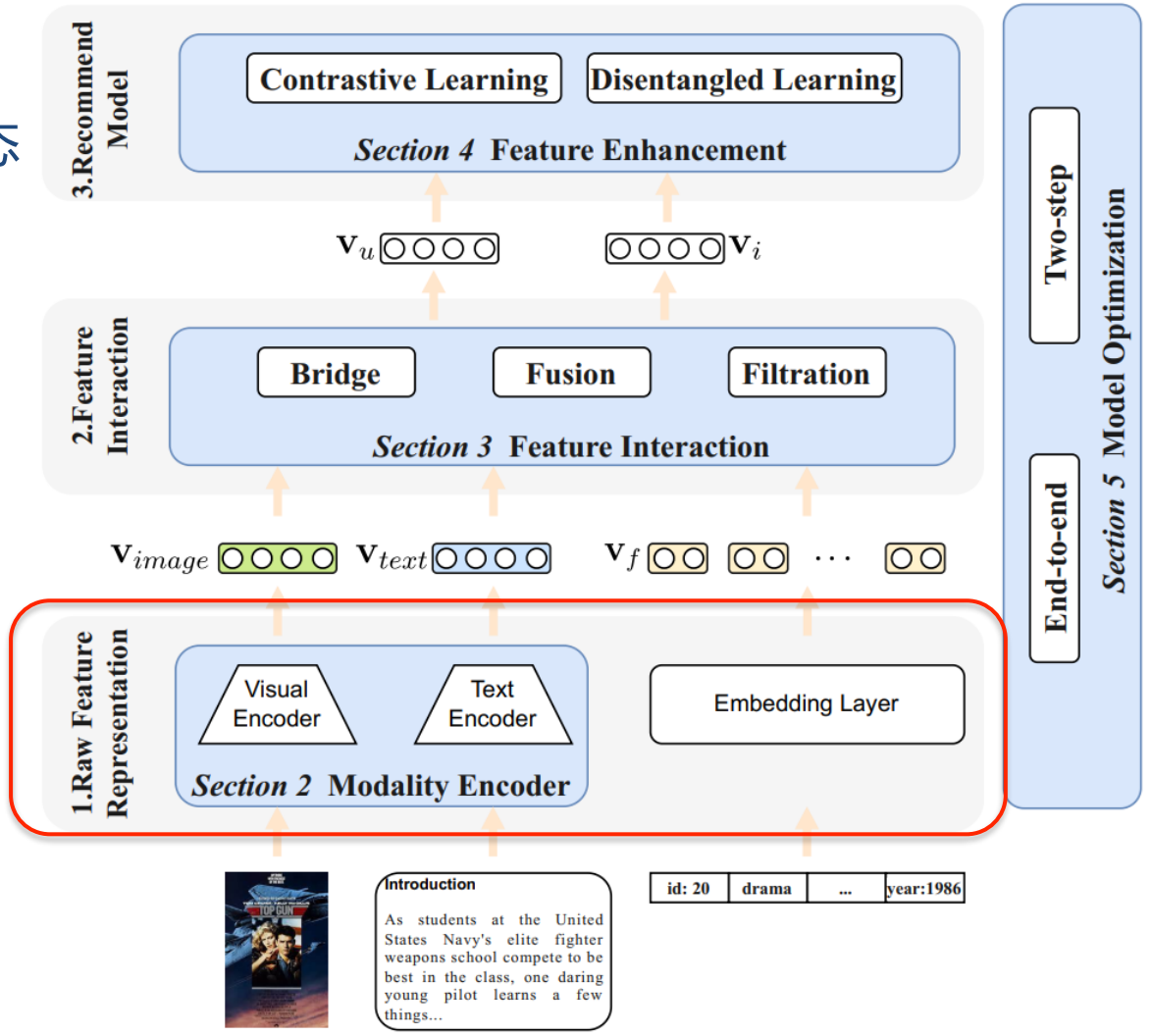
多模态推荐系统流程

- **原始特征表征**
  - 表格特征 – Embedding Layer
  - 多模态特征 (poster, intro) – Modality Encoder
- **特征交互**
  - 将不同的模态的特征映射到同一空间
  - 产生 用户表征 和 物品表征
- **推荐**
  - 增强用户与物品表征
  - 给出推荐结果



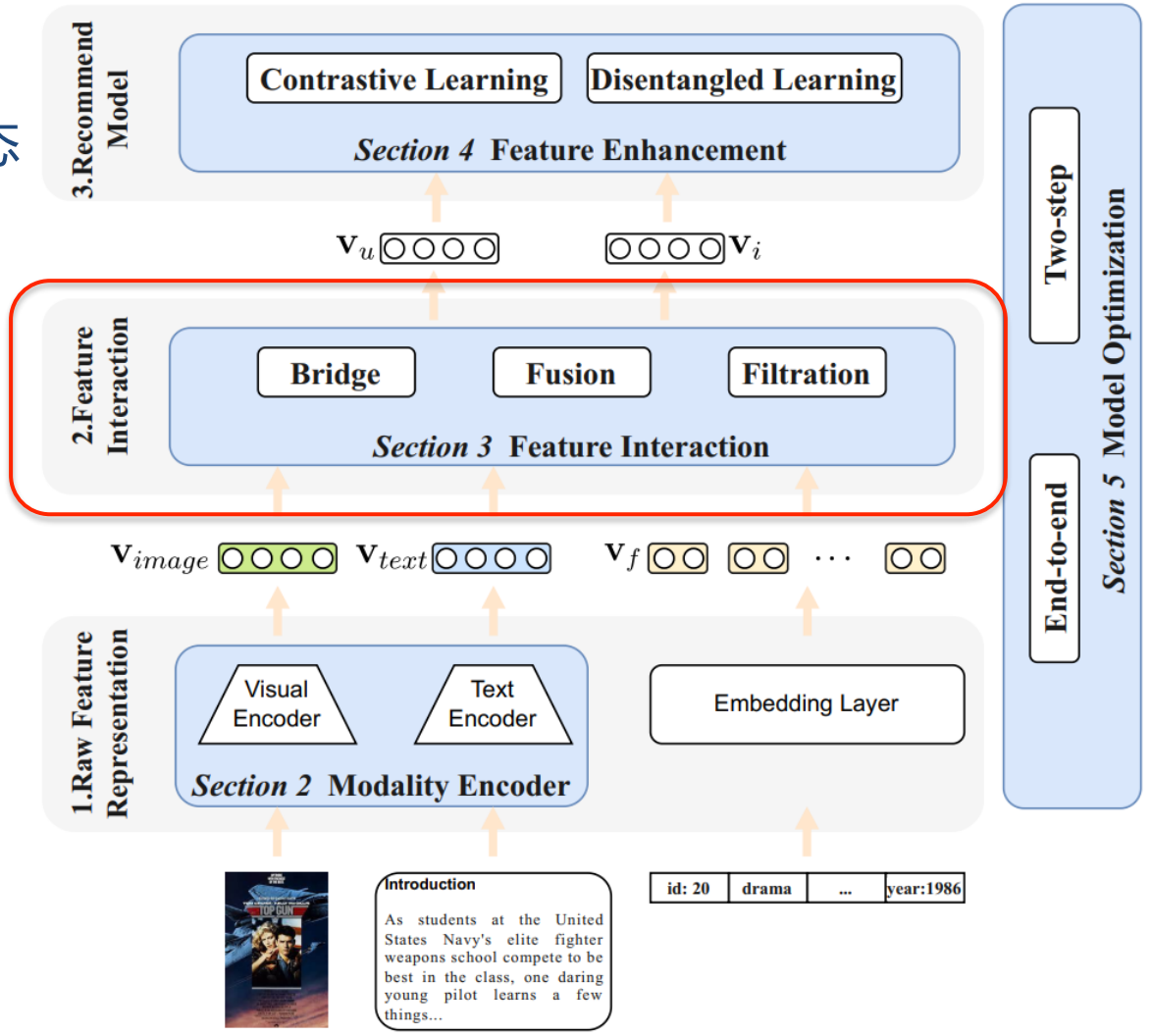
分类法

- **模态编码器**
  - 挑战: 对于原始多模态输入，如何从复杂的模态特征中抽取表征
  - 特殊性：一般RS直接使用embedding layer，MRS需要text和image encoders



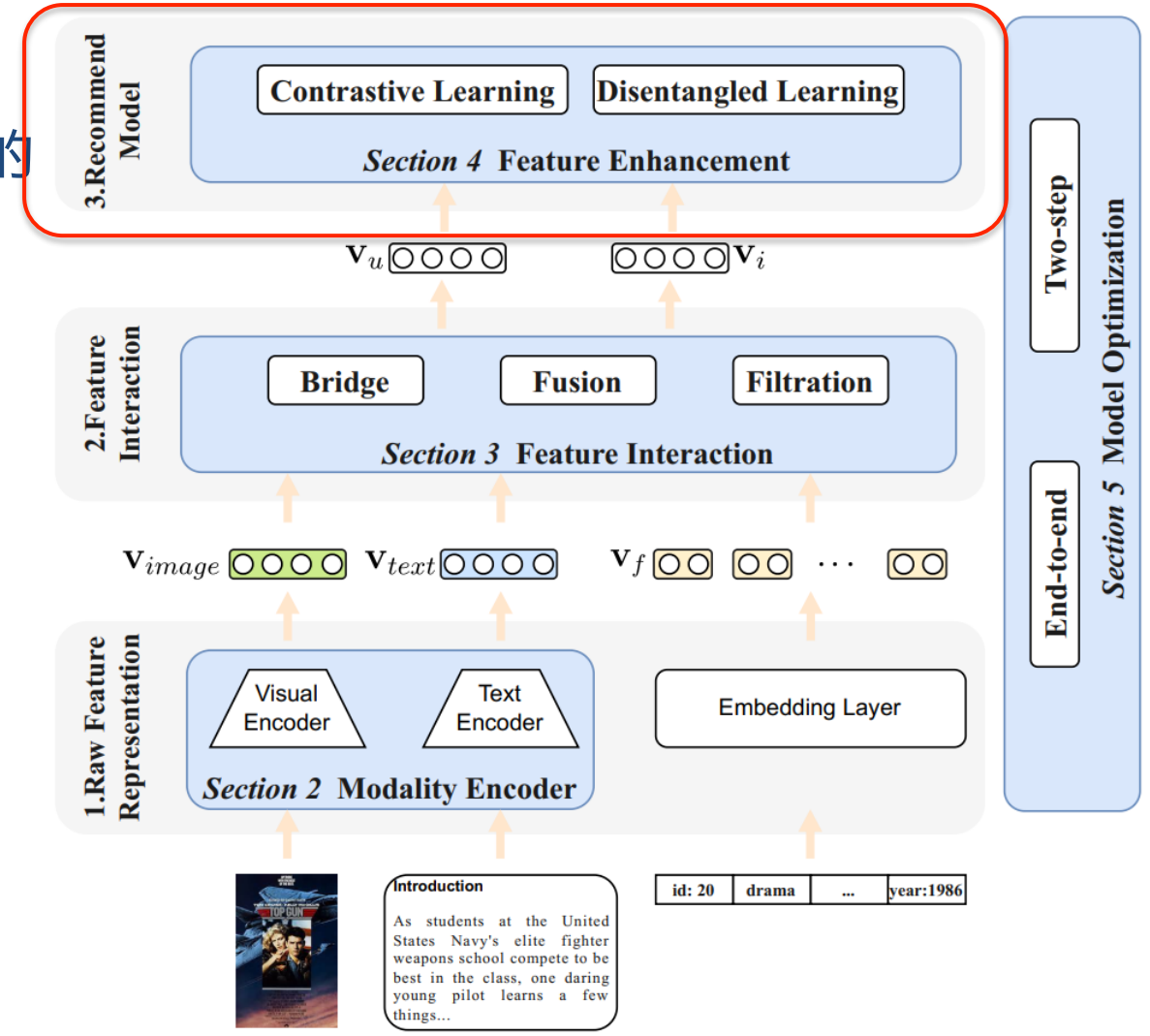
分类法

- **模态编码器**
  - 挑战: 对于原始多模态输入，如何从复杂的模态特征中抽取表征
  - 特殊性：一般RS直接使用embedding layer，MRS需要text和image encoders
- **特征交互技术**
  - 挑战：对于特征交互步骤，关键在于如何将不同语义空间的多模态特征融合，以及如何捕获用户对不同模态的偏好
  - 特殊性：一般RS关注高阶特征组合，MRS关注不同模态特征对齐与融合



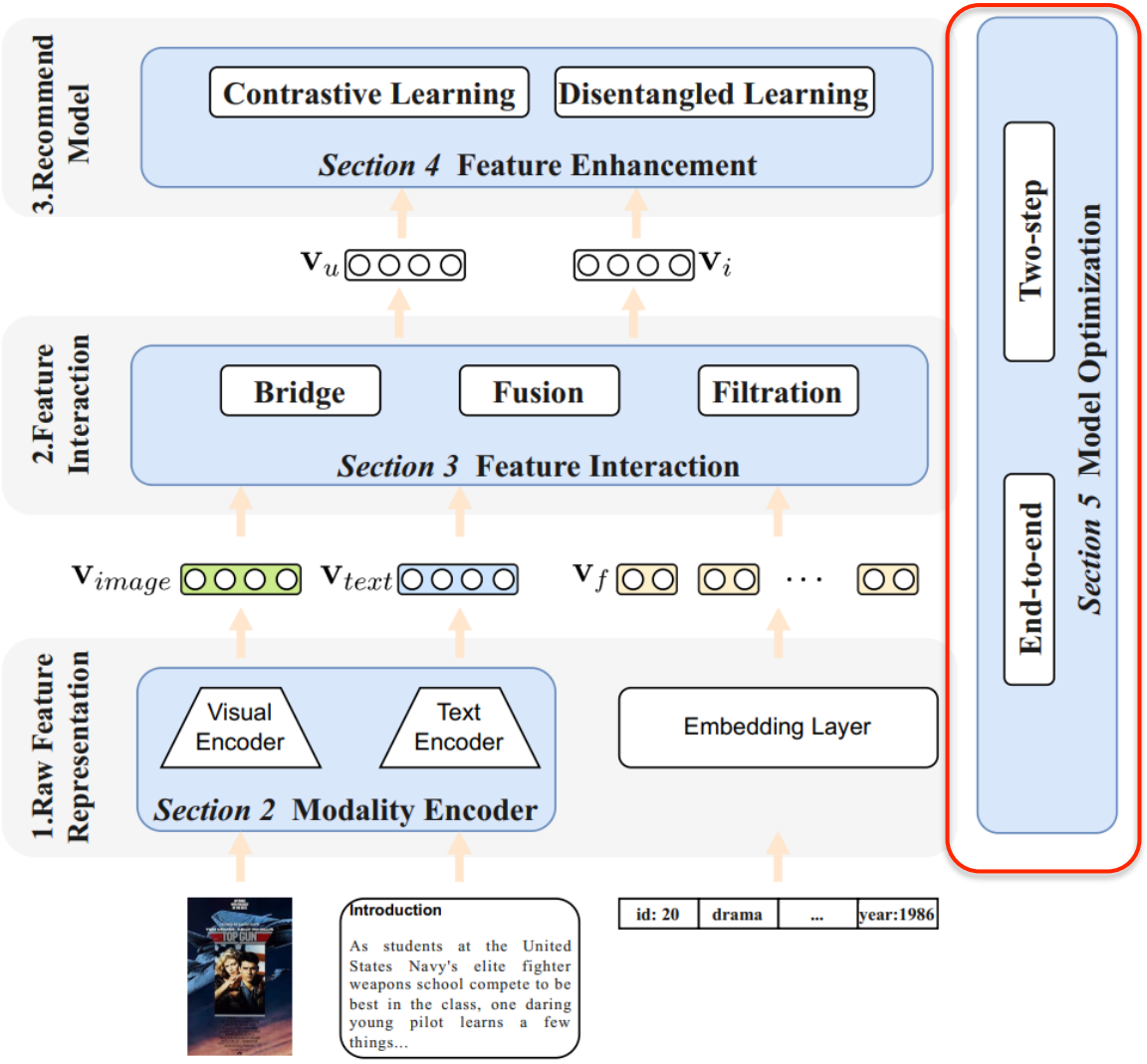
分类法

- **特征增强**
  - 挑战: 对于最后的推荐步骤，如何在数据稀疏的条件下学习更加语义丰富的特征
  - 特殊性：一般RS只对单一模态特征增强，MRS需要对多种模态特征增强



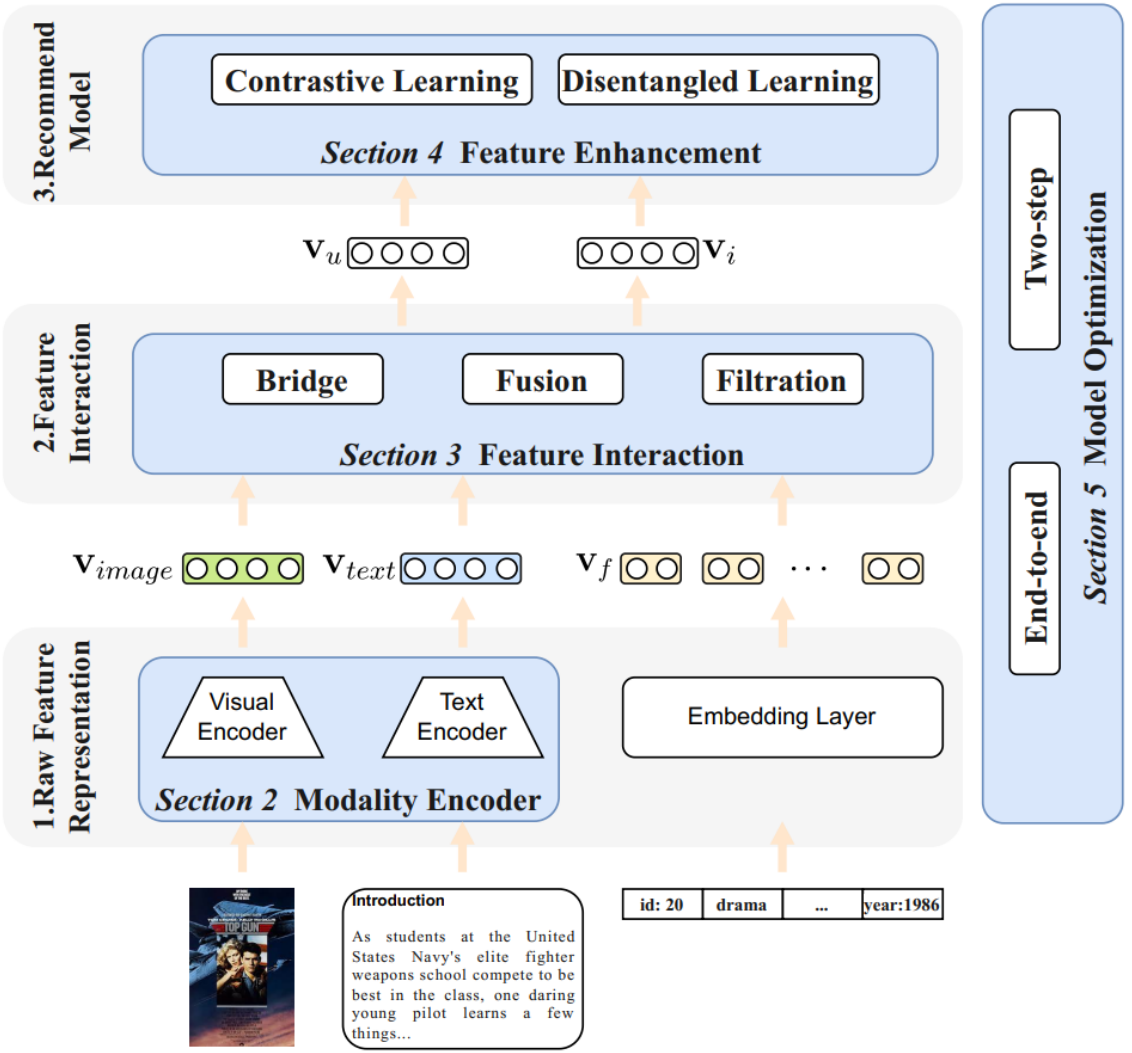
分类法

- **特征增强**
  - 挑战: 对于最后的推荐步骤，如何在数据稀疏的条件下学习更加语义丰富的特征
  - 特殊性：一般RS只对单一模态特征增强，MRS需要对多种模态特征增强
- **模型优化**
  - 挑战：对于整个MRS的训练，如何同时优化轻量的RS模型和参数庞大的特征编码器
  - 特殊性：一般RS的embedding layer占据了大部分参数，MRS则是特征编码器



大纲

- 背景和流程
- 模态编码器
- 特征交互
- 特征增强
- 模型优化
- 未来的方向与讨论



模态编码器

- **模态编码器**
  - 从 多模态原始特征 中提取压缩的、有效的特征
  - 三类编码器:
    - 视觉编码器：处理图像特征，如电影推荐中的海报等
    - 文本编码器： 处理文本特征，如新闻推荐中的文章内容
    - 其他编码器：处理视频、音频等特征，很多随数据集一起开源，如Kwai[1]

[1]. Su, Runze, et al. "Themes informed audio-visual correspondence learning." arXiv preprint arXiv:2009.06573 (2020).

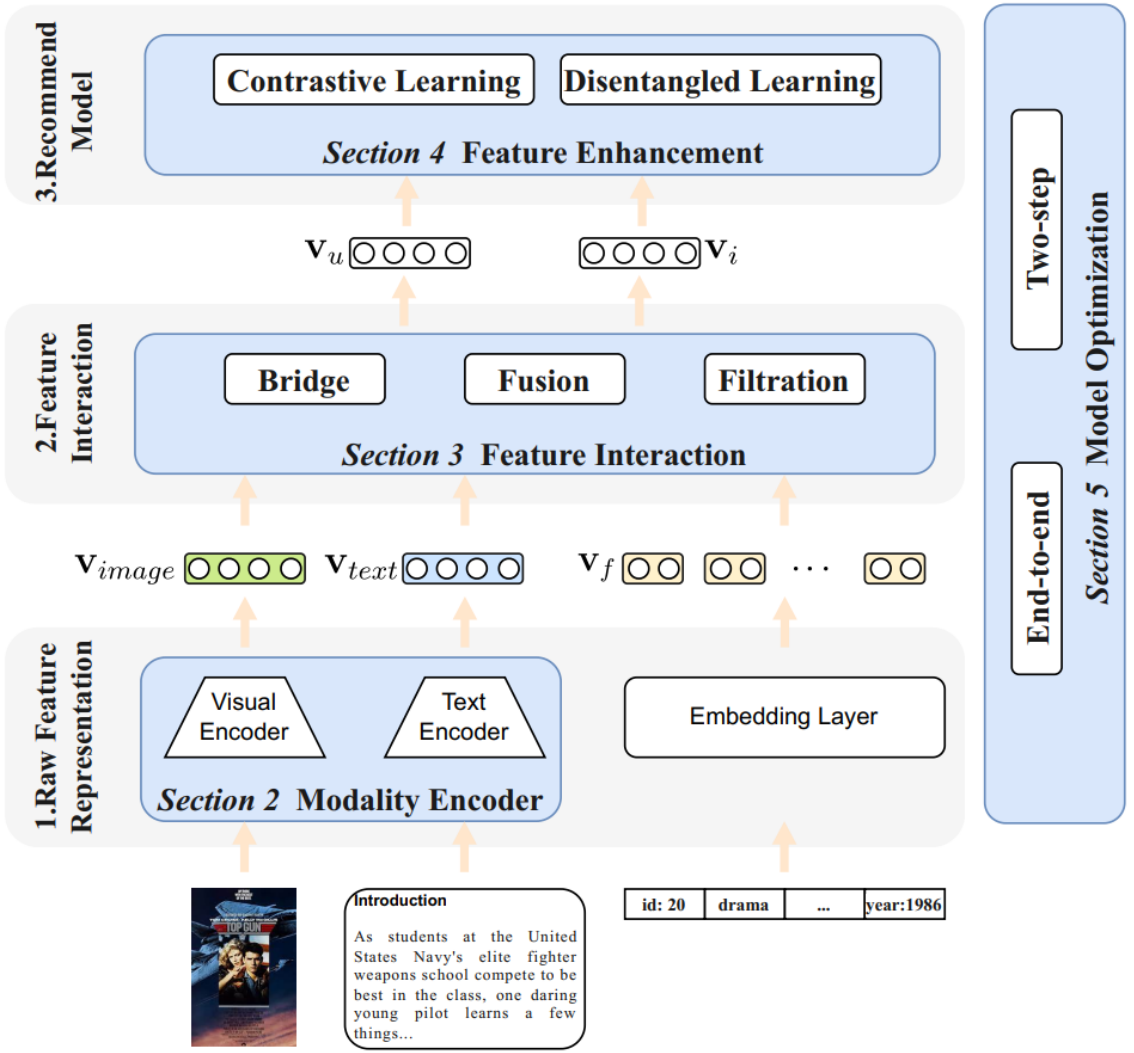
- **模态编码器**
  - 从 多模态原始特征 中提取压缩的、有效的特征
  - 三类编码器: 视觉编码器、文本编码器、其他编码器

Table 1: Category for Modality Encoder

Modality	Category	Related Works
Visual Encoder	CNN	[5], [33], [20], [69] [42], [57], [6], [4], [31], [29], [28], [14], [32], [34],
	ResNet	[35], [44], [43], [36], [54], [49], [60], [59], [55], [17]
	Transformer	[7], [13]
Textual Encoder	Word2vec	[33], [1], [49], [57], [59], [55], [14]
	RNN	[62], [28]
	CNN	[58], [4]
	Sentence-transformer	[20], [67], [66], [42], [73]
	Bert	[35], [44], [30], [43], [36], [54], [7], [60], [38], [31], [13]
Other Modality Encoder	Published Feature	[50], [67], [66], [65], [59], [55], [64]

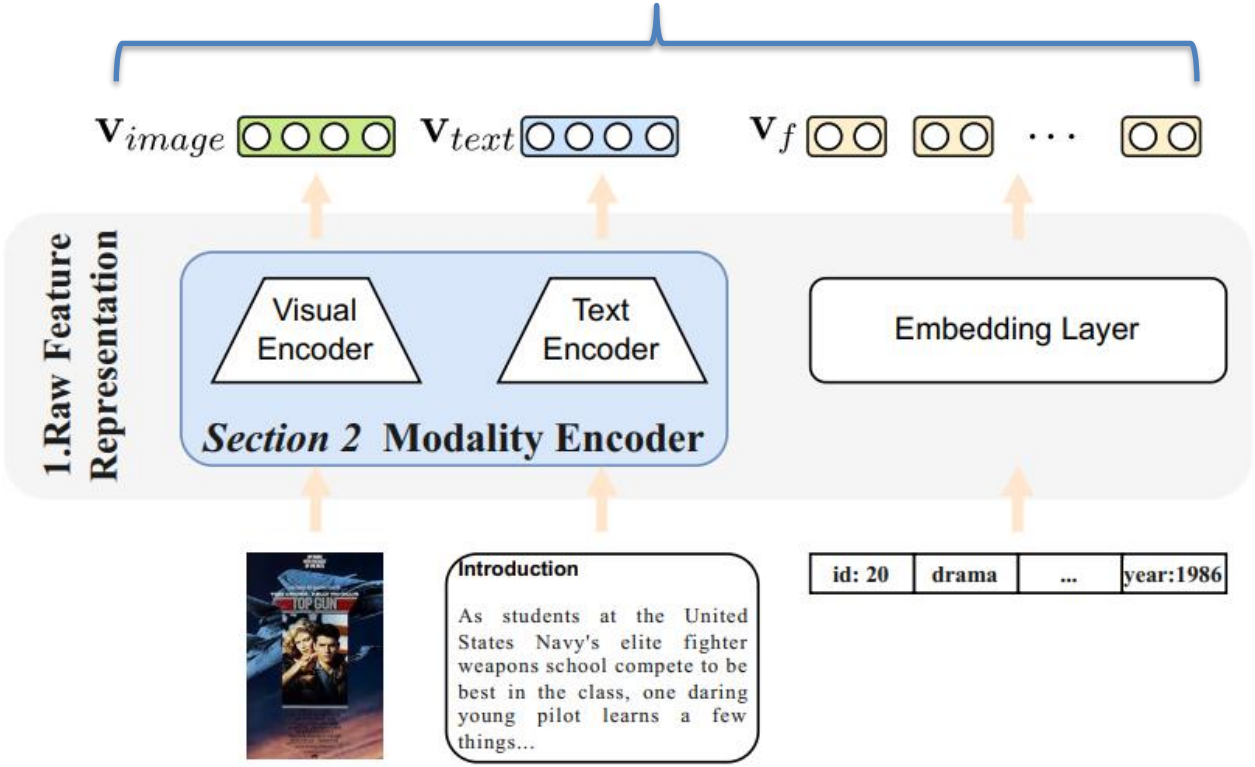
大纲

- 背景和流程
- 模态编码器
- 特征交互
  - 连接
  - 融合
  - 过滤
- 特征增强
- 模型优化
- 未来的方向与讨论



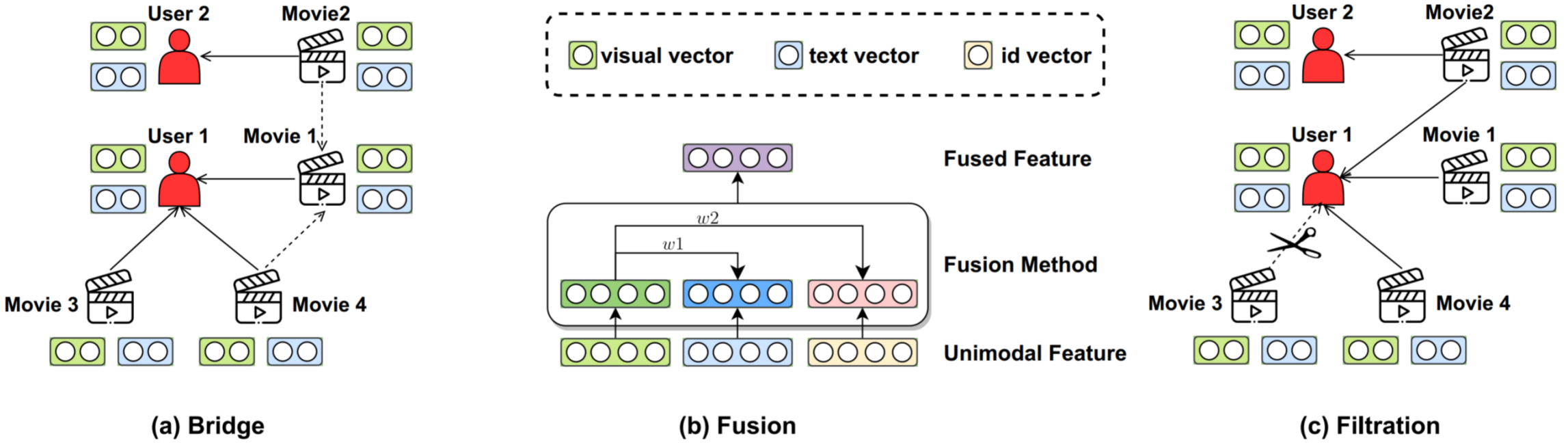
特征交互

- MRS所面临的挑战
  - 稀疏的推荐交互数据
  - 由编码器得到的多模态特征处于不同的语义空间



特征交互

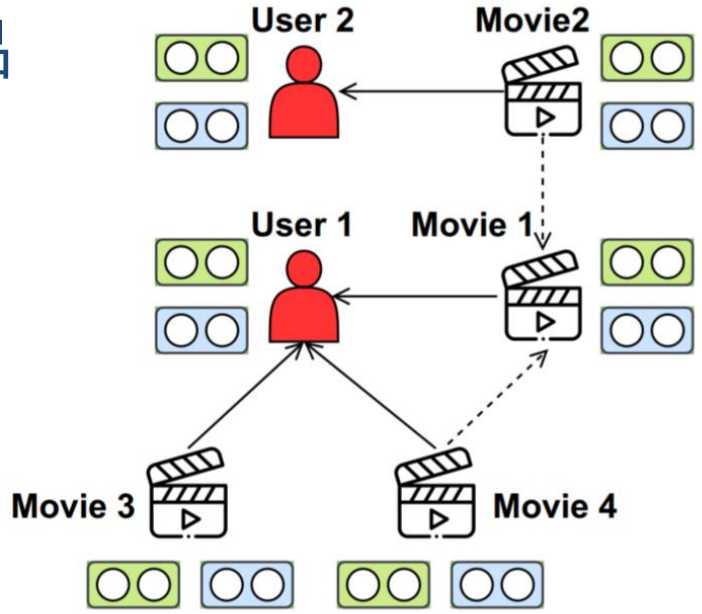
- **三种交互方式**
  - 连接 (Bridge): 捕获用户和物品间的模态关系
  - 融合 (Fusion): 结合对不同模态的偏好
  - 过滤 (Filtration): 过滤噪声数据



(a) Bridge (b) Fusion (c) Filtration

特征交互

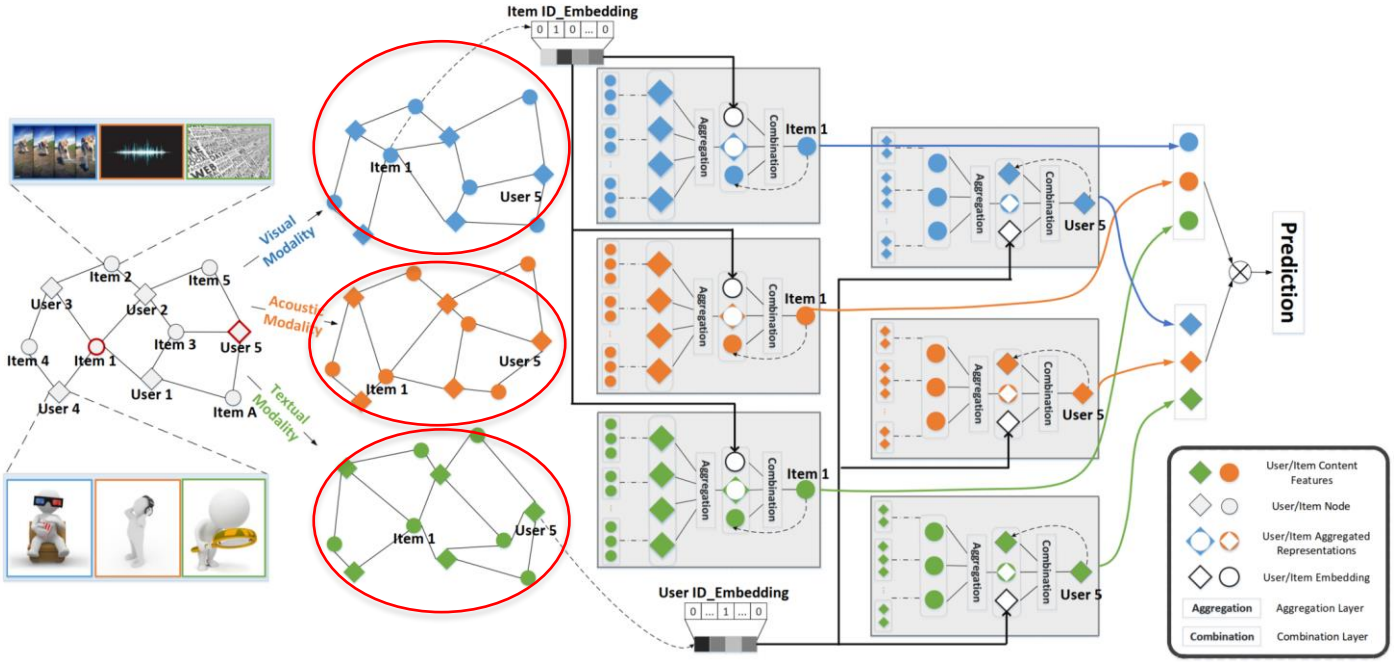
- **连接 (Bridge)**
  - 定义：构建多模态信息交互的通道，捕获用户和物品间的关系。通常使用图建模来实现
  - 分类：
    - 用户-物品图 (User-item Graph)
    - 物品-物品图 (Item-item Graph)
    - 知识图谱 (Knowledge Graph)



(a) Bridge

特征交互

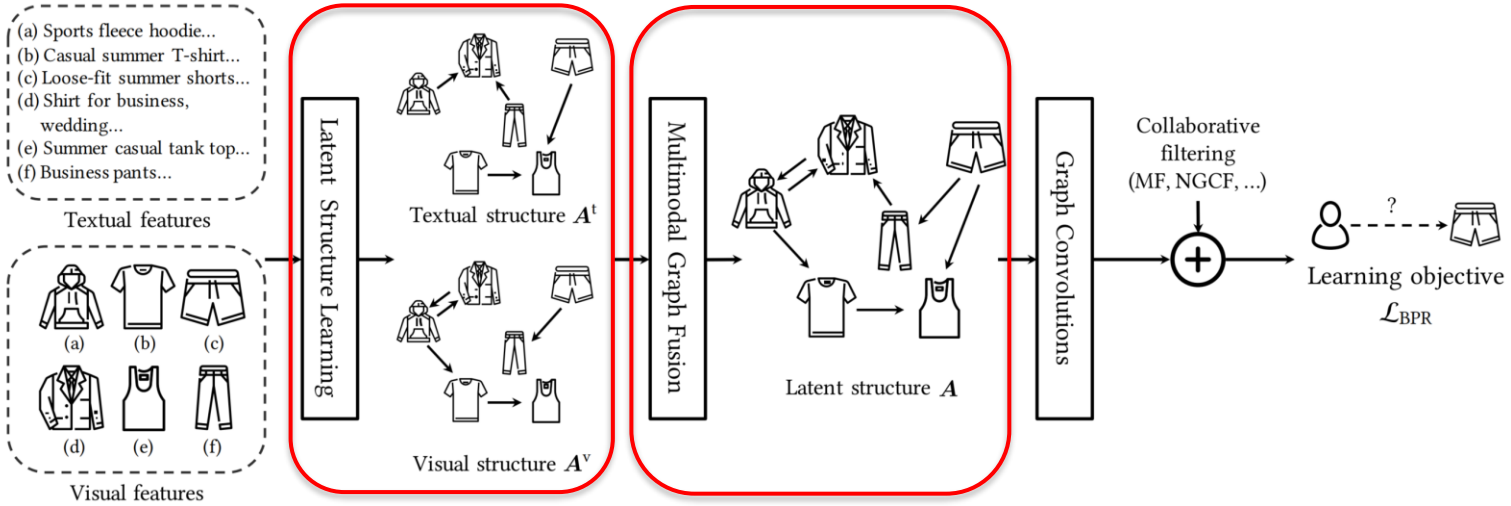
- **用户-物品图 (User-Item Graph)**
  - 利用用户和物品之间的信息交换，捕获用户对于不同模态的偏好
- MMGCN (MM'19)
  - 根据用交互，对每一个模态构建user-item bipartite graph
  - 在每一个模态图上进行GCN
  - 拼接不同模态表征并推荐



Wei, Yinwei, et al. "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video." ACM MM. 2019.

特征交互

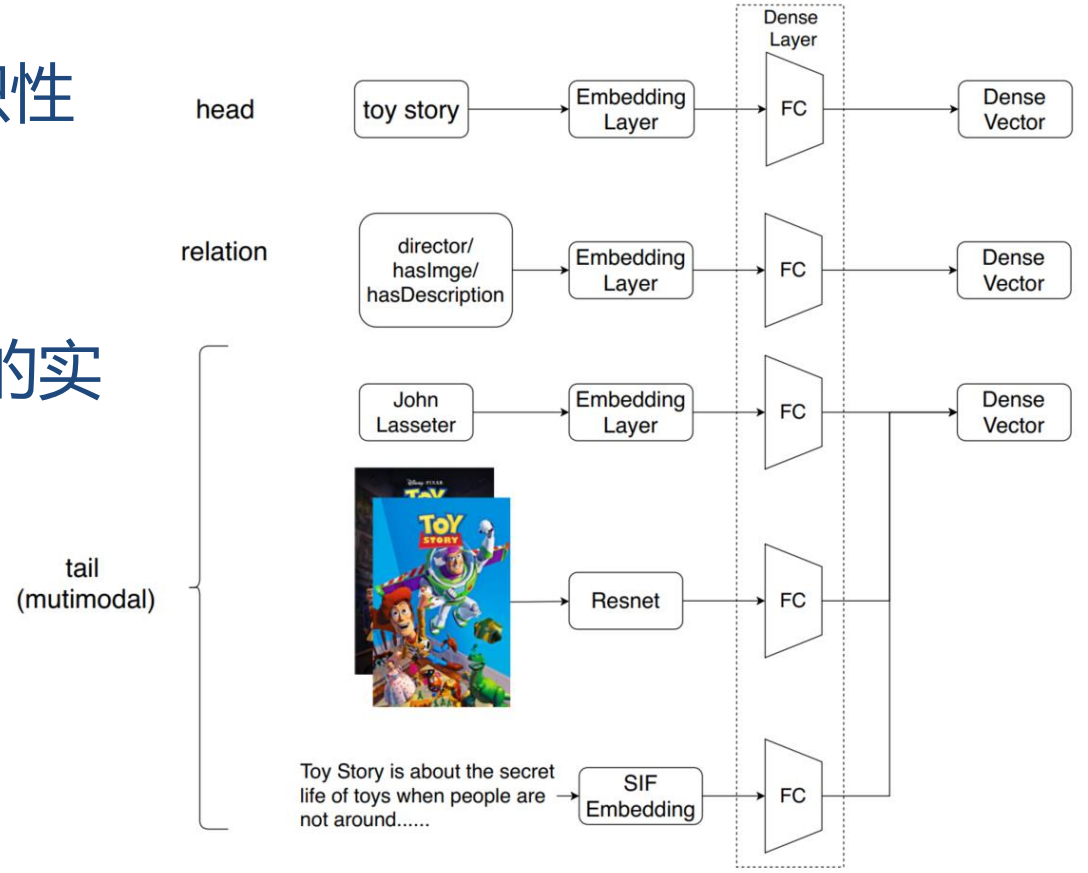
- **物品-物品图 (Item-Item Graph)**
  - 利用物品和物品之间的多模态关系，学习更好的物品表征
- LATTICE (MM'21)
  - 利用物品的模态表征，通过聚类方法，对每一个模态构建item-item graph
  - 通过可学习权重，捕获不同模态的重要性，并将模态图融合
  - 由融合的item-item graph学习物品表征，并应用于下游推荐任务中



Zhang, Jinghao, et al. "Mining latent structures for multimedia recommendation." ACM MM. 2021.

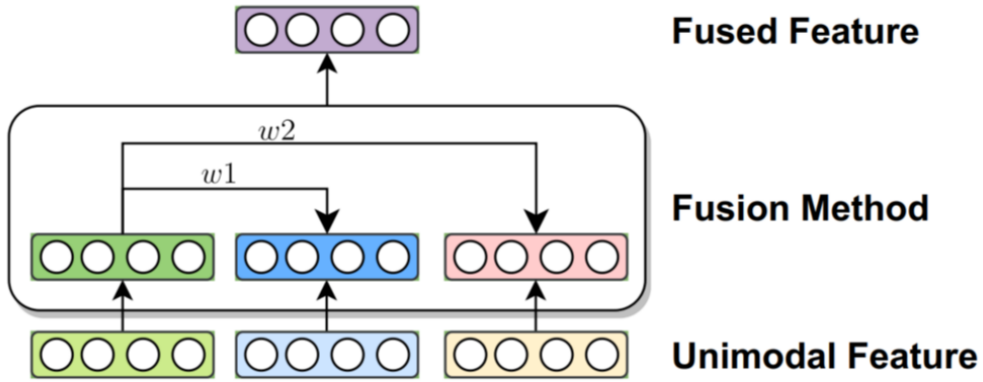
特征交互

- **知识图谱 (Knowledge Graph)**
  - 利用多模态知识图谱(MKG)中的物品间知识性关系，学习更好的物品表征
- MKGAT (CIKM'20)
  - MKG Encoder: 用于学习多模态知识图谱中的实体表征。其中尾实体结合了多模态特征
  - Stage 1: 使用KGAT训练实体表征
  - Stage 2: 将实体表征结合到RS中用于推荐



Sun, Rui, et al. "Multi-modal knowledge graphs for recommender systems." CIKM. 2020.

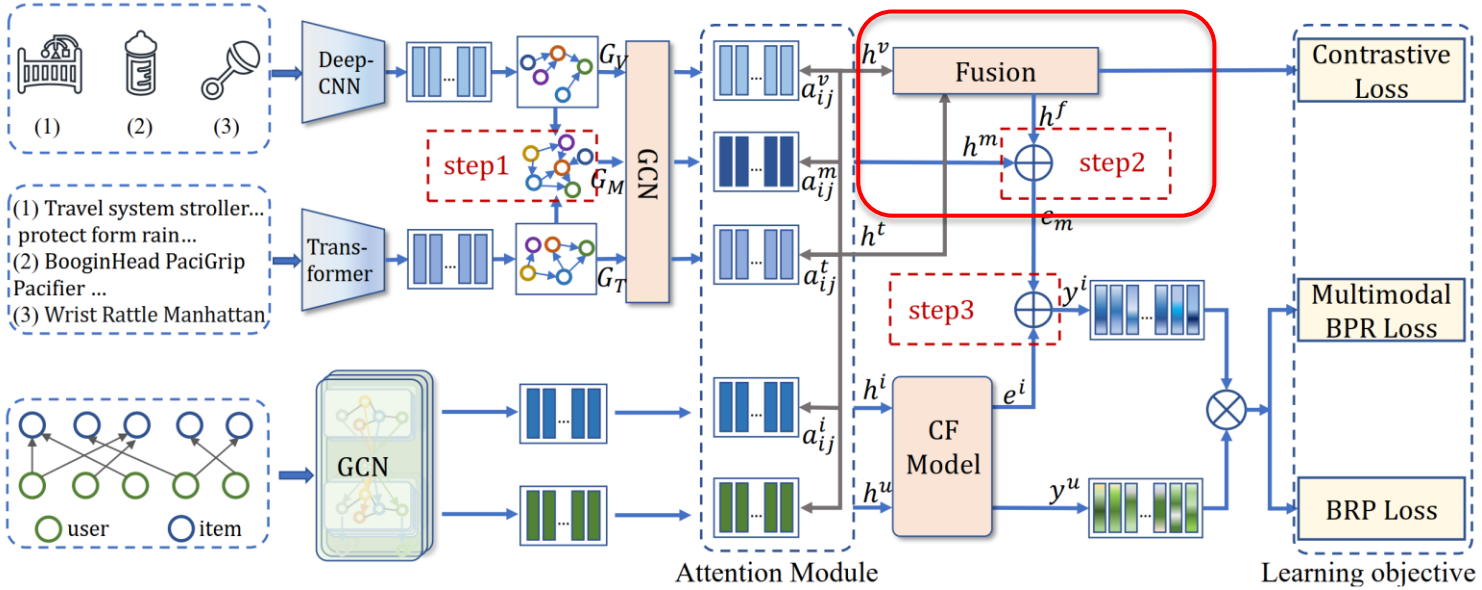
- **融合 (Fusion)**
  - 定义：不同模态的信息量和种类有很大不同，因此fusion关注的是物品内的模态关系。通常使用注意力机制来实现
  - 分类：
    - 粗粒度注意力 (Coarse-grained Attention)
    - 细粒度注意力 (Fine-grained Attention)
    - 结合注意力 (Combined Attention)
    - 其他方法 (Other Methods)



(b) Fusion

特征交互

- **粗粒度融合 (Coarse-grained Attention)**
  - 交互层面的融合，注意力通常关注的是交互之间的关系
- TMFUN (SIGIR'23)
  - 利用Attention机制聚合图表征学习得到的节点表征
  - 先进行文本和视觉模态表征融合，再进行模态和ID表征融合

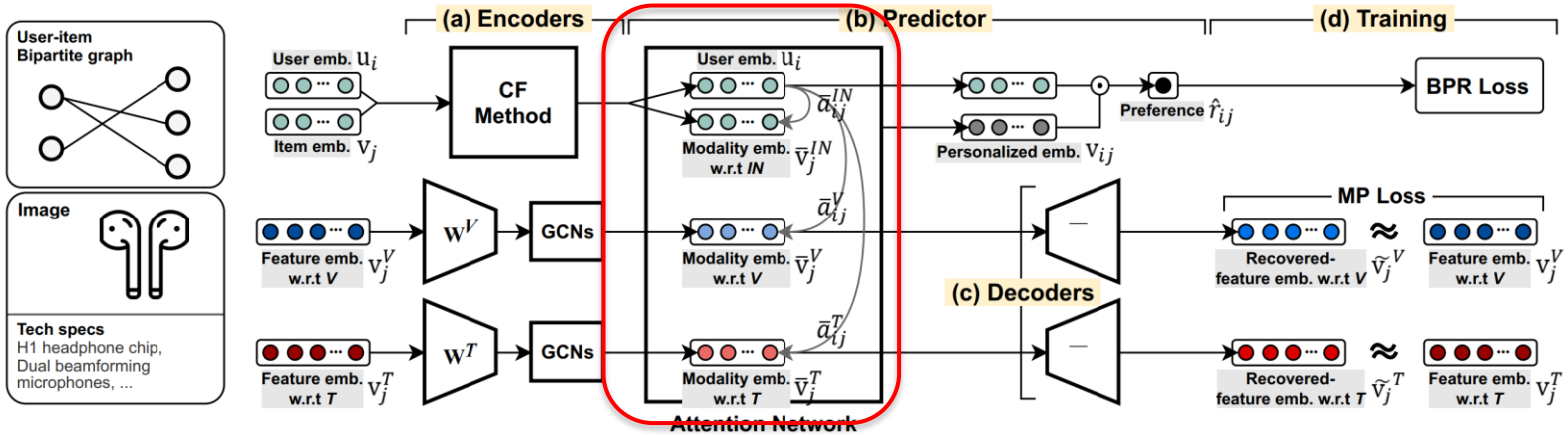


$$h^f = \mu h^v + (1-\mu)h^t$$

Zhou, Yan, et al. "Attention-guided multi-step fusion: a hierarchical fusion network for multimodal recommendation." SIGIR. 2023.

特征交互

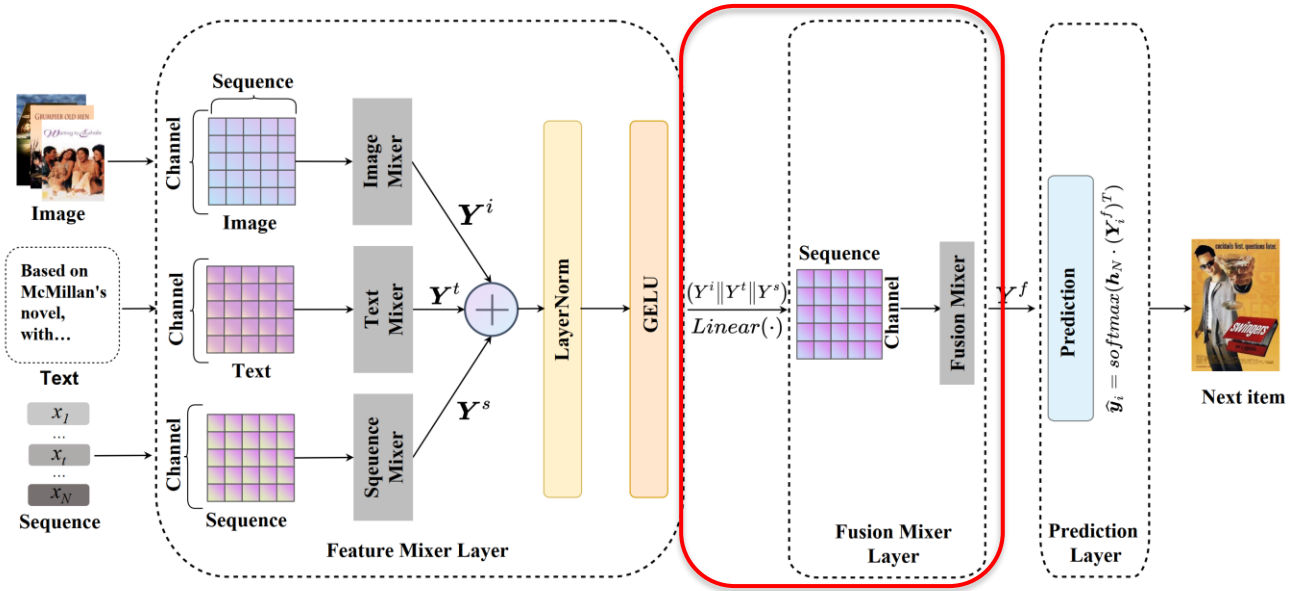
- **细粒度融合 (Fine-grained Attention)**
  - 模态层面的融合，注意力通常关注的是模态之间的关系
- MARIO (CIKM'22)
  - 使用CGN得到模态表征，CF得到ID表征
  - 基于模态表征和ID表征的注意力机制，为每个interaction学习各自的模态权重并融合
  - 融合后的物品表征和用户表征进行推荐



$$\bar{a}_{ij}^{m} = \frac{\exp(a_{ij}^{m})}{\sum_{m \in \mathcal{M}} \exp(a_{ij}^{m})}, \quad \text{where} \quad a_{ij}^{m} = \frac{\mathbf{u}_i \odot \bar{\mathbf{v}}_j^{m}}{\sqrt{d}},$$

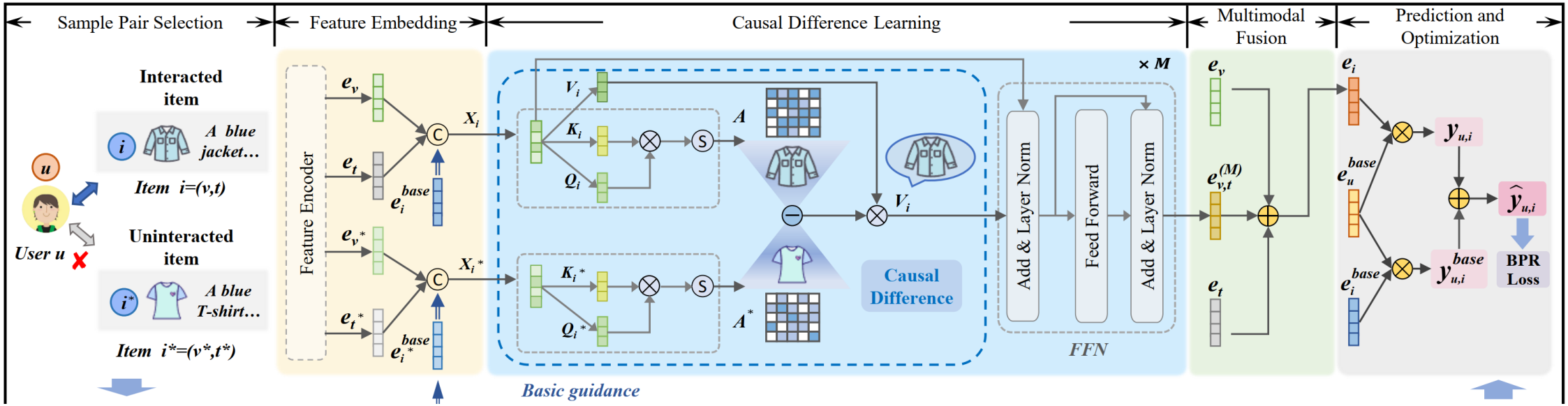
Kim, Taeri, et al. "MARIO: modality-aware attention and modality-preserving decoders for multimedia recommendation." CIKM. 2022.

- **其他方法 (Other Methods)**
  - 使用拼接、门控机制、MOE、MLP Mixer进行模态融合
- MMMLP (WWW'23)
  - 对每个模态的表征序列进行MLP Mixer，学习序列信息
  - 拼接所有模态的表征，进行MLP Mixer，自适应地融合不同模态信息



Liang, Jiahao, et al. "MMMLP: multi-modal multilayer perceptron for sequential recommendations." WWW. 2023.

- **过滤 (Filtration)**
  - 定义：过滤模态特征内含有的噪声或者交互数据中含有的噪声
- MCLN (SIGIR'23)
  - 视觉特征与文本特征之间的虚假关联，造成了对于偏好学习的噪声
  - 使用因果学习来解耦出噪声特征



Li, Shuaiyang, et al. "Multimodal Counterfactual Learning Network for Multimedia-based Recommendation." SIGIR. 2023.

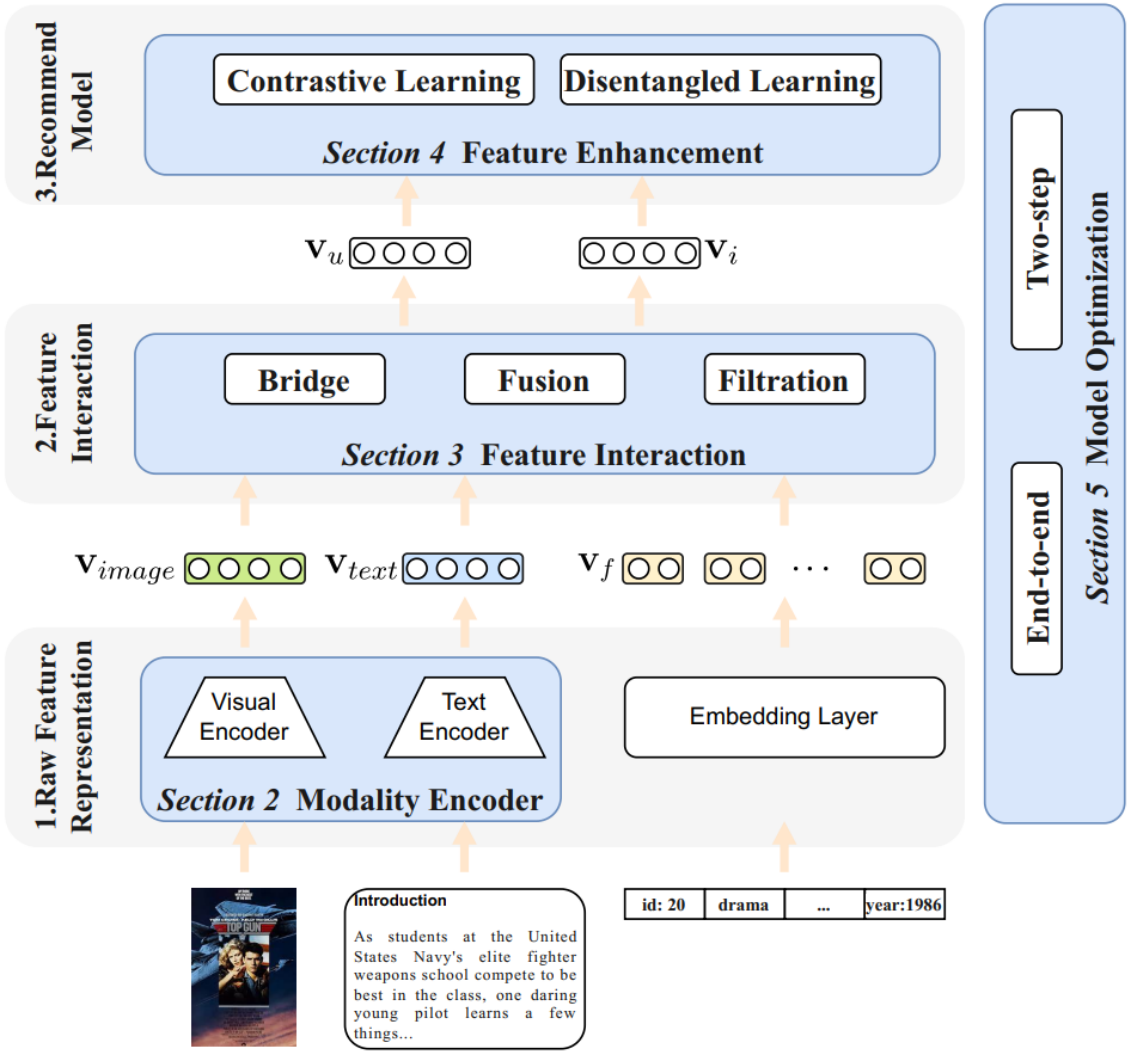
特征交互

- 多种交互技术结合
  - 这三类特征交互方法关注的是多模态推荐中的不同方面，因此也有很多工作结合了多种交互方法

Interaction	Goal	Category	Related Works
Bridge	Capture inter-relationship between users and items.	User-item Graph	[50], [43], [55], [59], [64]
		Item-item Graph	[67], [66], [42], [36], [2], [40], [67]
		Knowledge Graph	[58], [52], [1], [54], [7], [49], [32]
Fusion	Combine various preference to modalities.	Coarse-grained Attention	[37], [44], [35], [6]
		Fine-grained Attention	[5], [20], [61], [27], [50], [43], [36], [7], [14], [31], [20], [29], [25], [32], [4], [17],[28], [75], [60], [26]
		Combined Attention	[33], [30], [13]
		Other Fusion Methods	[57], [3], [38], [38], [69], [62]
Filtration	Filter out noisy data	Filtration	[49], [73], [65], [34], [63]

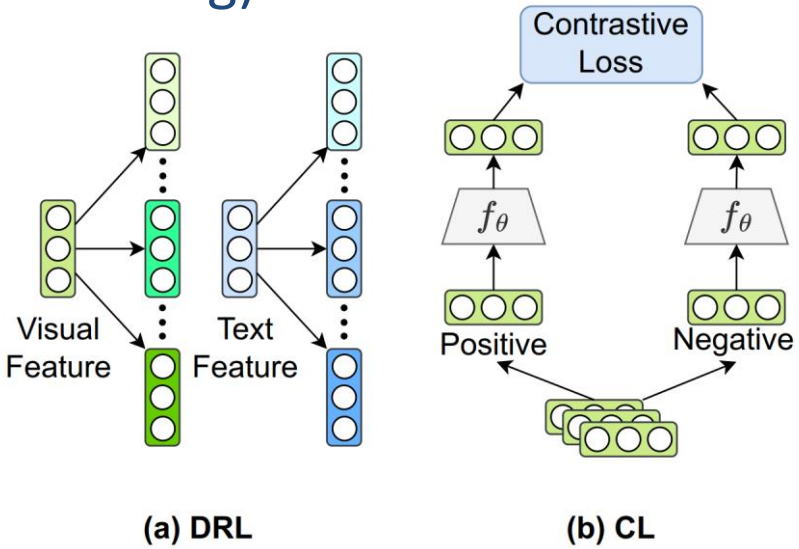
大纲

- 背景和流程
- 模态编码器
- 特征交互
- 特征增强
  - 解耦表征学习
  - 对比学习
- 模型优化
- 未来的方向与讨论



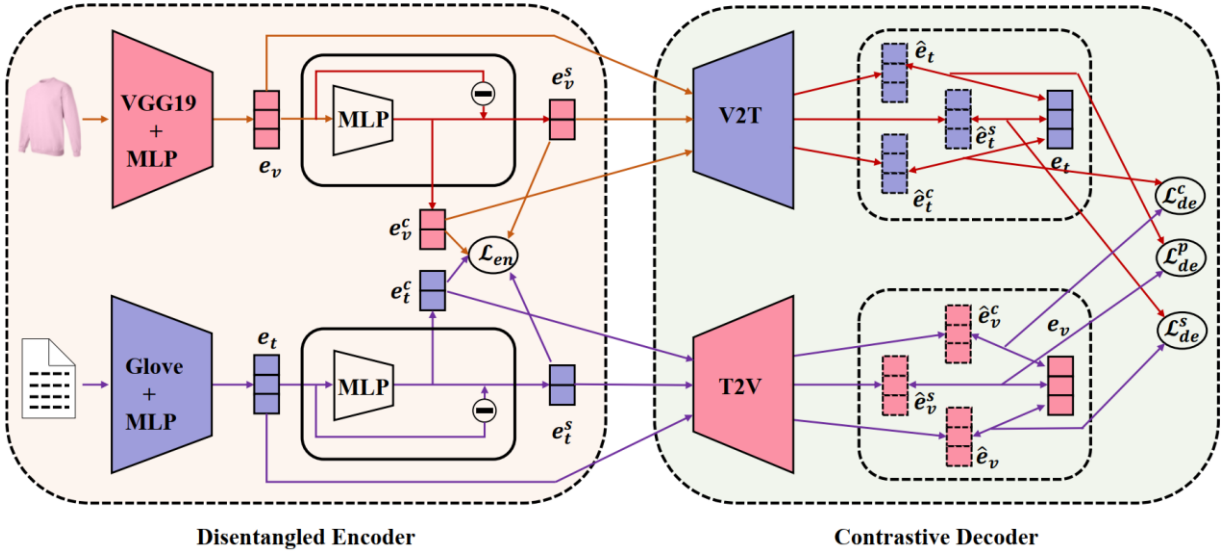
特征增强

- **目标**
  - 抽取不同模态表征中相同和共有语义信息
  - 通过表征增强，缓解推荐系统存在的数据稀疏问题
- **分类**
  - 解耦表征学习 (Disentangled Representation Learning)
  - 对比学习 (Contrastive Learning)



特征增强

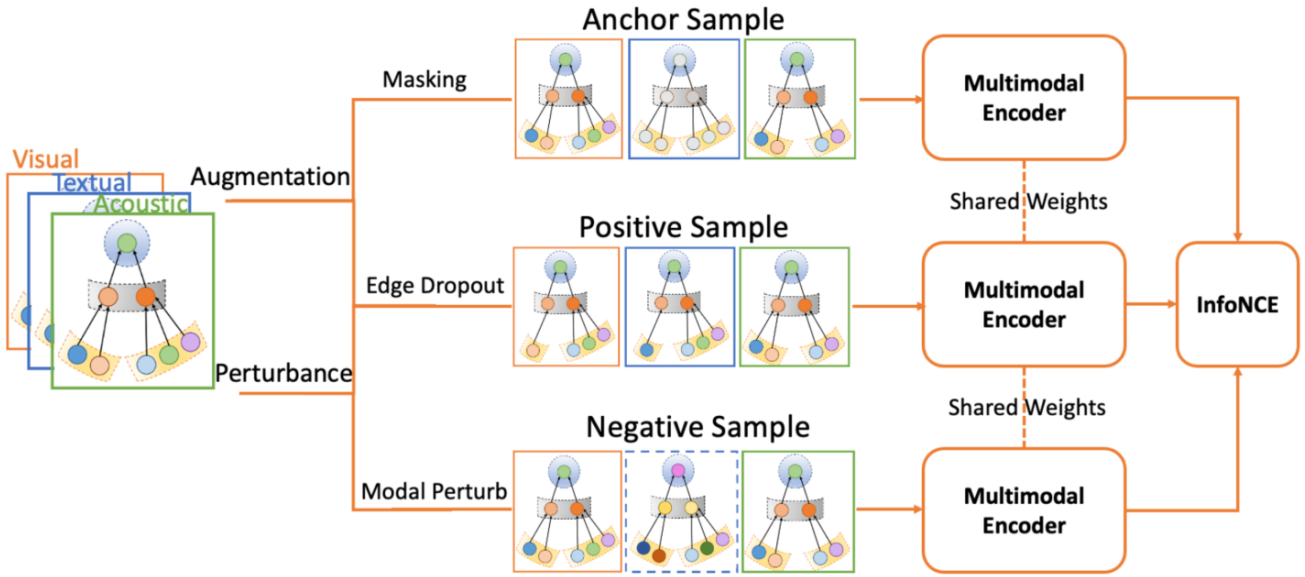
- **解耦表征学习 (Disentangled Representation Learning)**
  - 捕获用户对不同模态的不同偏好，挖掘细粒度的物品表征
- PAMD (WWW'22)
  - 解耦模态公有特征和特有特征，添加一个解耦学习损失
  - 通过对比学习对齐不同模态的公有特征，推远不同模态的特有特征



Han, Tengyue, et al. "Modality matches modality: Pretraining modality-disentangled item representations for recommendation." WWW. 2022.

特征增强

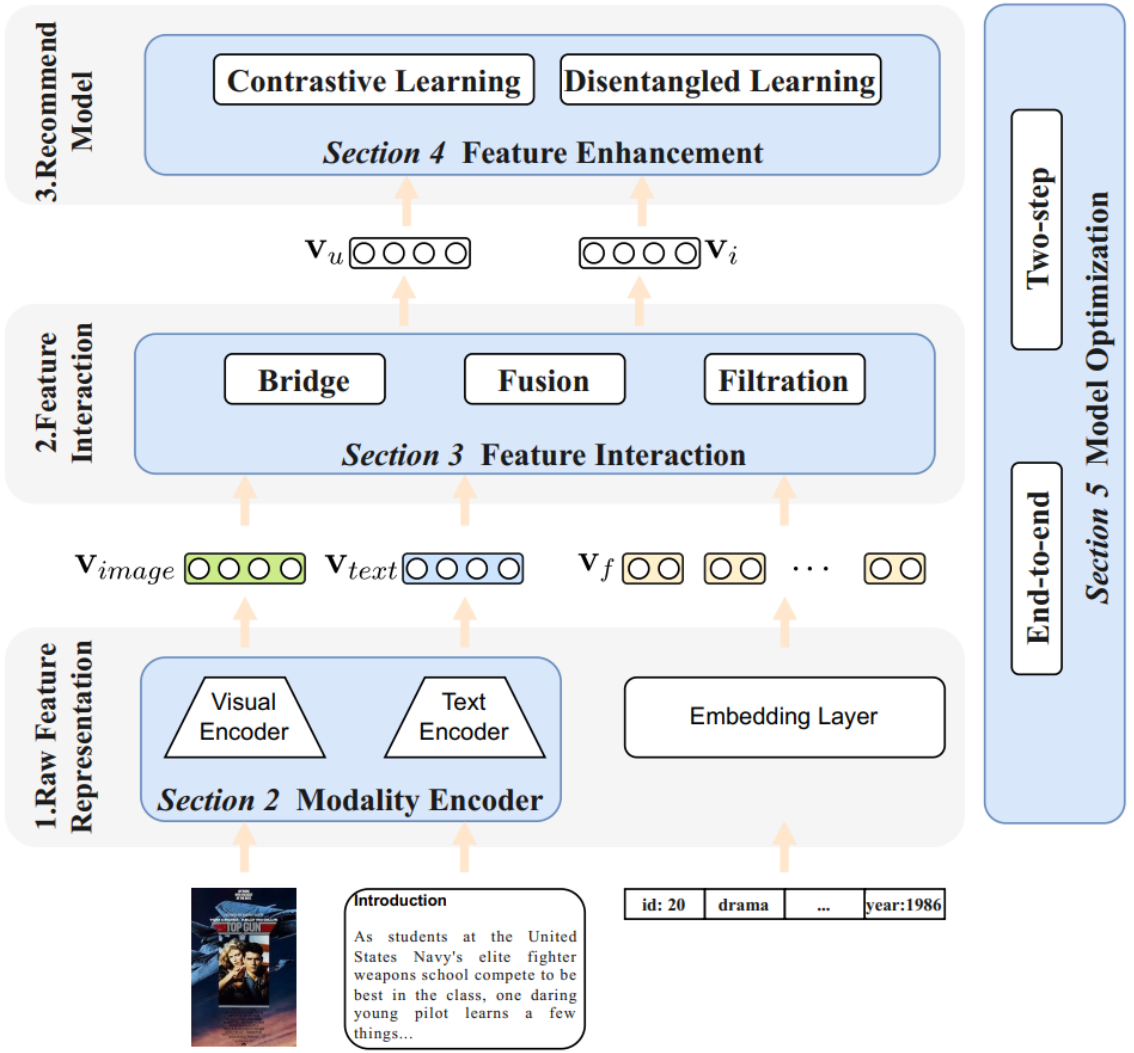
- **对比学习 (Contrastive Learning)**
  - 通过数据增强的方法增强模态表征，对齐不同模态的表征
- MMGCL (SIGIR'22)
  - Modal Perturb: 随机mask掉一个模态表征，得到融合后的表征
  - Edge Dropout: 随机dropout掉user-item graph上的边，得到融合后的表征



Yi, Zixuan, et al. "Multi-modal graph contrastive learning for micro-video recommendation." SIGIR. 2022.

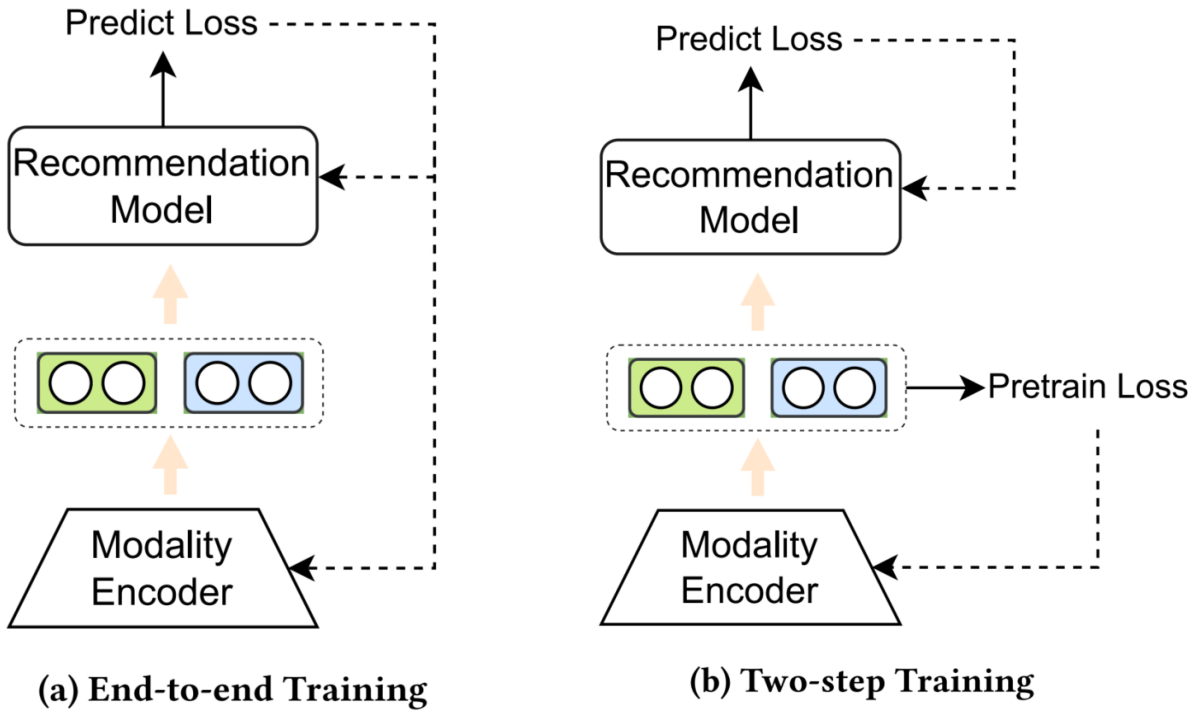
大纲

- 背景和流程
- 模态编码器
- 特征交互
- 特征增强
- 模型优化
  - 端到端
  - 两阶段
- 未来的方向与讨论



模型优化

- **挑战**
  - 轻量化的推荐模型与结构复杂的模态编码器存在不平衡
  - 这一不平衡导致训练资源消耗大以及推荐效果次优
- 解决方案
  - 端到端续联
  - 两阶段训练

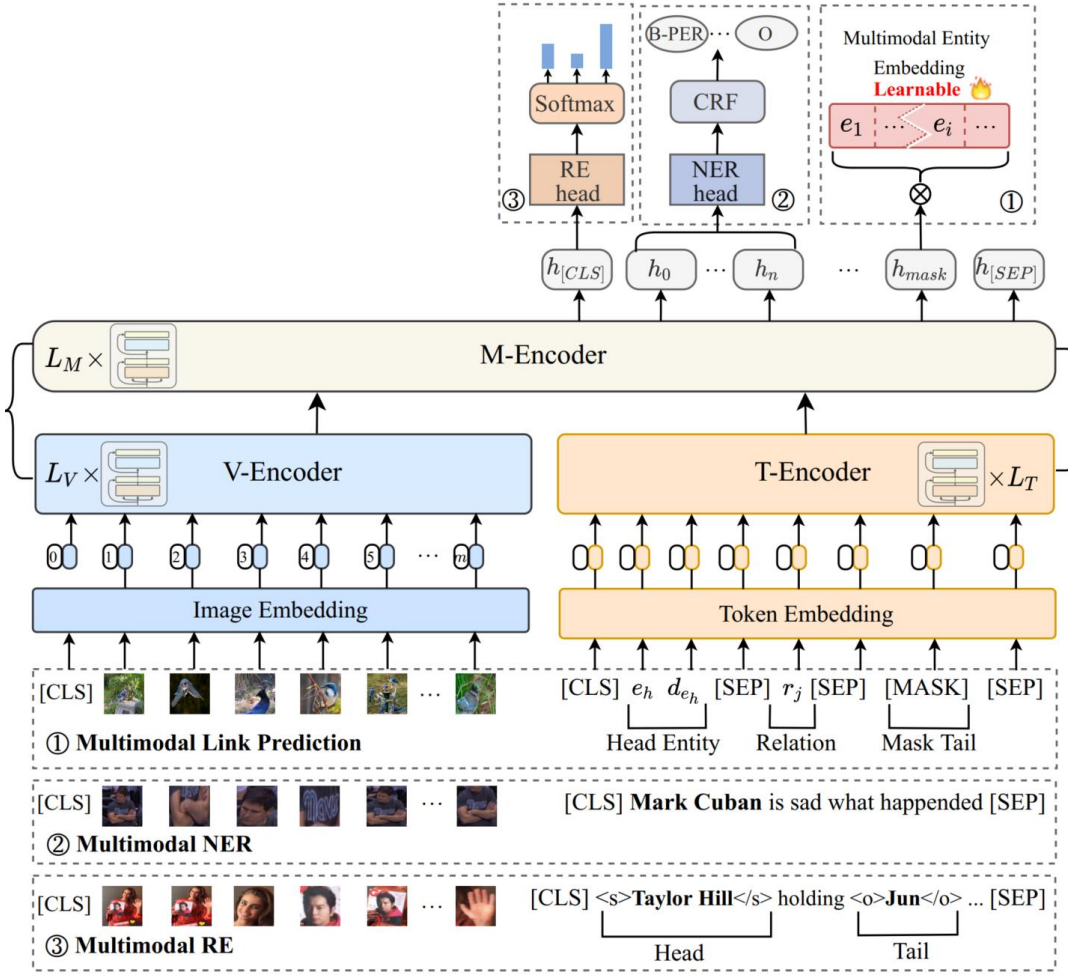


(a) End-to-end Training

(b) Two-step Training

模型优化

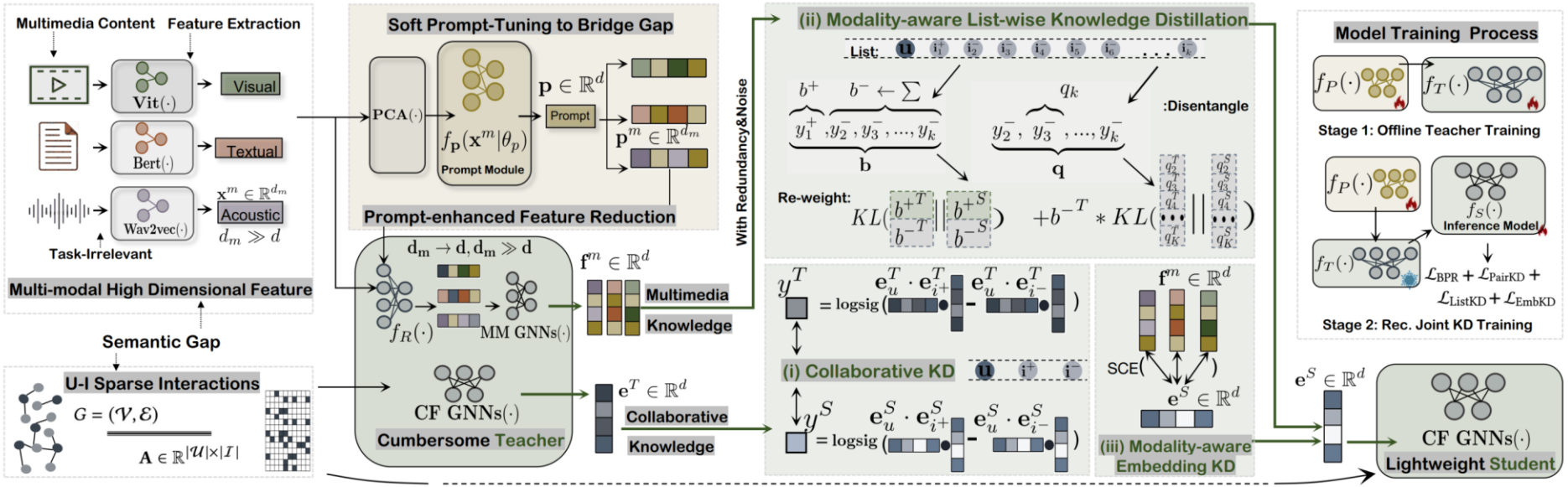
- **端到端 (End-to-end)**
  - 推荐模型与模态编码器一同训练。优点是可以更好适配RS，缺点是训练损耗大
- MKGformer (SIGIR'22)
  - 为了缓解不同模态间存在的不相关性问题，提出对模态编码器的深度融合
  - M-Encoder: 融合ViT和BERT的最后几层transformer层
  - 端到端训练



Chen, Xiang, et al. "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion." SIGIR. 2022.

模型优化

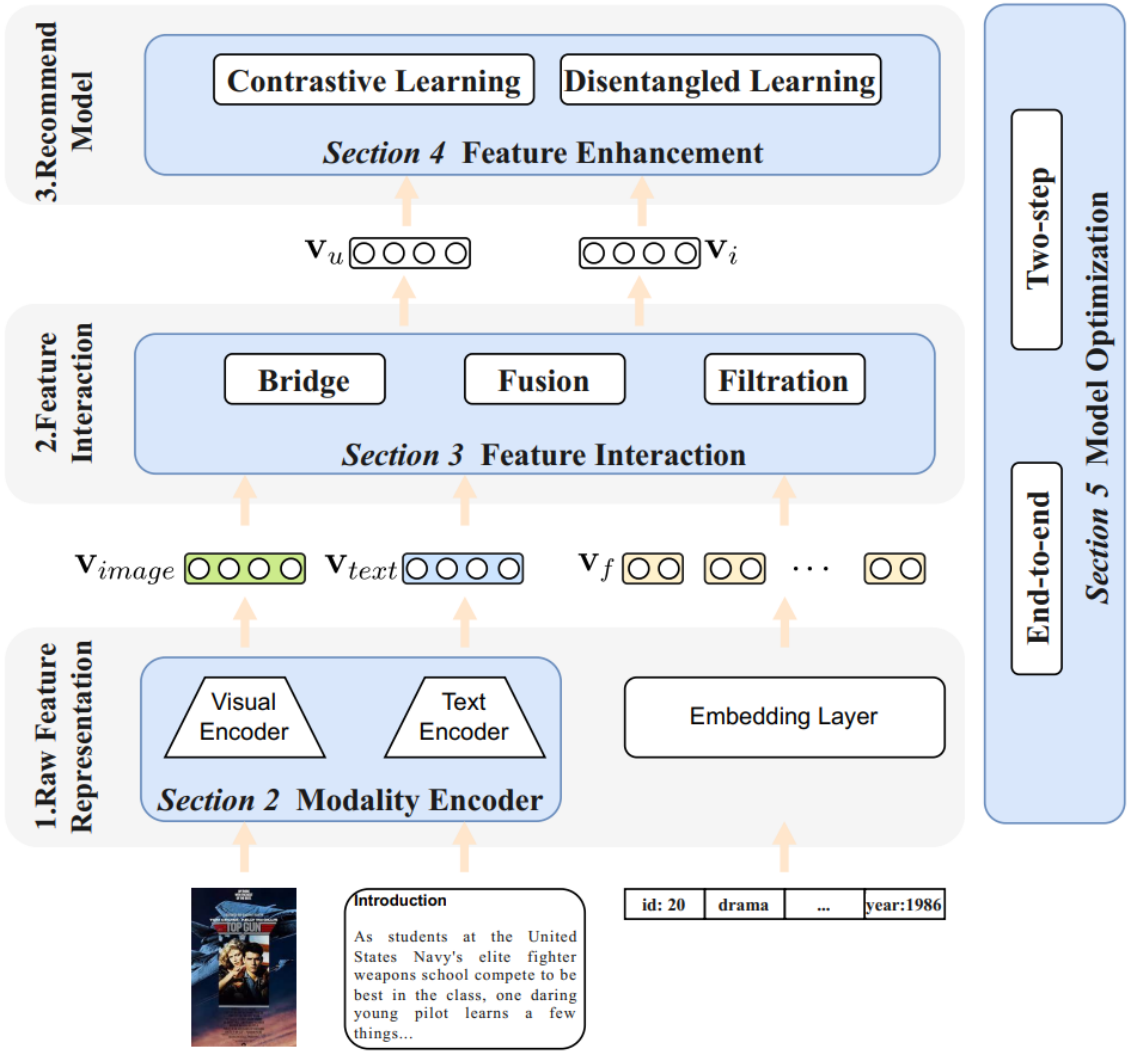
- **两阶段 (Two Step)**
  - 模态编码器与推荐模型分开训练。通常直接使用预训练好的编码器
- PromptMM (WWW'24)
  - Teacher Prompt Training: 训练用于降维模态特征的提取器(看做是编码器一部分)
  - Knowledge Distillation: 把模态表征蒸馏给ID表征



Wei, Wei, et al. "PromptMM: Multi-Modal Knowledge Distillation for Recommendation with Prompt-Tuning." WWW. 2024.

大纲

- 背景和流程
- 模态编码器
- 特征交互
- 特征增强
- 模型优化
- 未来的方向与讨论

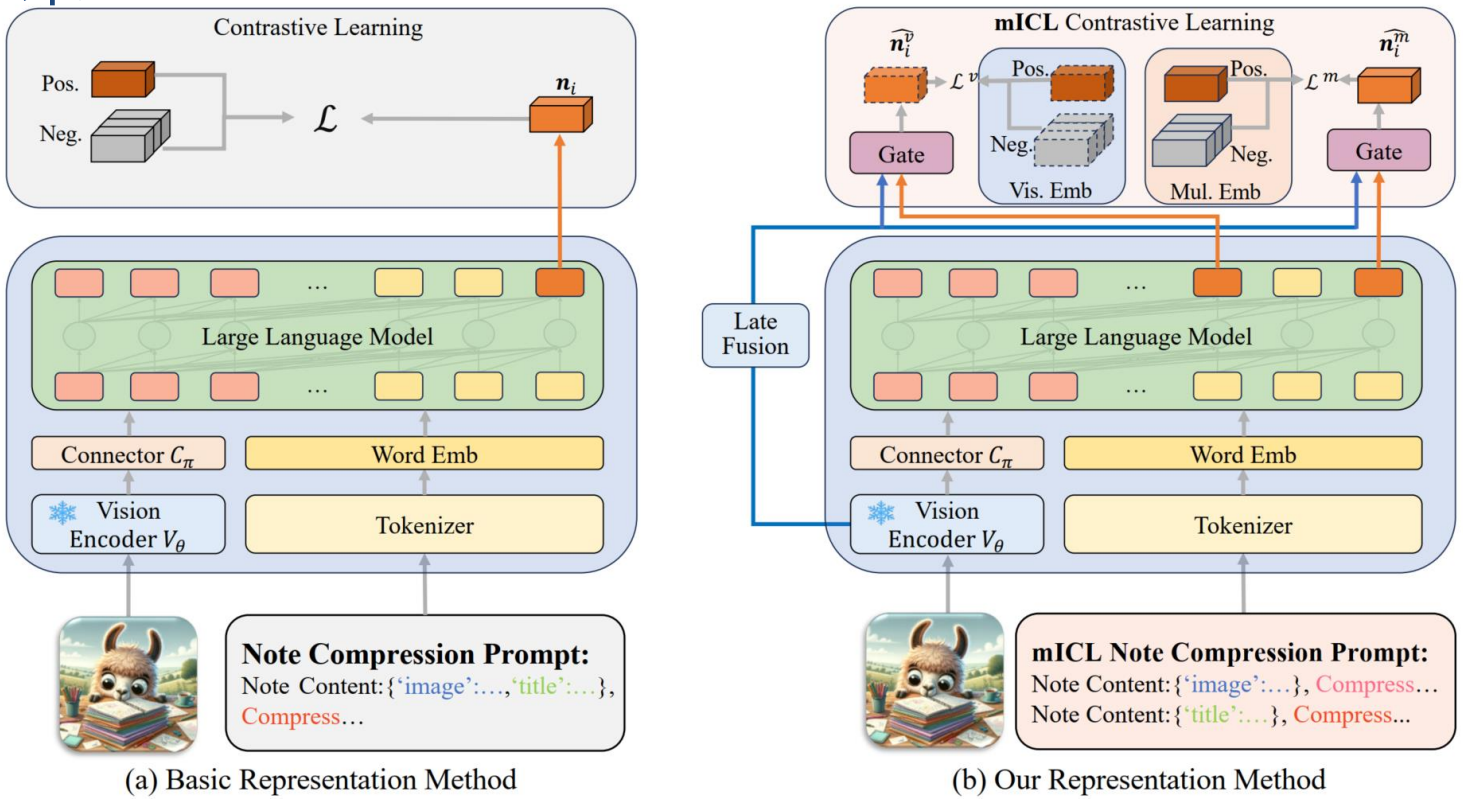


未来的方向与讨论

- **未来方向**
  - **统一的解决框架**: 设计一个可以将之前提到的每个部分的技术都可以囊括进去的整体框架
  - **模型可解释性**: 多模态推荐系统通常非常复杂且黑盒，可解释的MRS可以提升可信度和透明性
  - **计算复杂性**: 模态编码器与RS的适配和端到端训练的高消耗是相互冲突的，如何平衡两者
  - **不完整和有偏的数据**: 在现实世界中，多模态数据通常是不完整或者有偏的，如何解决这两个问题

未来的方向与讨论

- **未来方向**
  - **MLLM的应用**: MLLM展现出了强大的理解和推理能力，如何将MLLM应用到多模态推荐中来



Zhang, Chao, et al. "NoteLLM-2: Multimodal Large Representation Models for Recommendation." arXiv. 2024.

Our Survey Paper

Github Repo

Thank You!

刘启东

西安交通大学 & 香港城市大学

liuqidong@stu.xjtu.edu.cn

2024年6月24日