Improving Personalized Explanation Generation through Visualization

Shijie Geng[1], **Zuohui Fu**[1], **Yingqiang Ge**[1], **Lei Li**[2], **Gerard de Melo**[3], **Yongfeng Zhang**[1]
[1]Rutgers University [2]Hong Kong Baptist University [3]HPI/University of Potsdam
{sg1309,yongfeng.zhang}@rutgers.edu

Abstract

In modern recommender systems, there are usually comments or reviews from users that justify their ratings for different items. Trained on such textual corpus, explainable recommendation models learn to discover user interests and generate personalized explanations. Though able to provide plausible explanations, existing models tend to generate repeated sentences for different items or empty sentences with insufficient details. This begs an interesting question: can we immerse the models in a multimodal environment to gain proper awareness of real-world concepts and alleviate above shortcomings? To this end, we propose a visually-enhanced approach named METER with the help of visualization generation and text–image matching discrimination: the explainable recommendation model is encouraged to visualize what it refers to while incurring a penalty if the visualization is incongruent with the textual explanation. Experimental results and a manual assessment demonstrate that our approach can improve not only the text quality but also the diversity and explainability of the generated explanations.

1 Introduction

Explainable recommender systems have recently attracted increasing attention both in industry and in the academic community. Such systems aim to provide high-quality recommendations and simultaneously generate explanations for the recommendations (Zhang et al., 2014; Zhang and Chen, 2020). The explanations not only can bridge the gap between how systems and users perceive the relevance of the recommended items, but also can serve to shed light on the recommendation decision process so as to avoid a black box. To provide appropriate explanations, feature-based (Zhang et al., 2014), graph-based (Xian et al., 2019, 2020; Geng et al., 2022; Fu et al., 2020), sentence-based (Chen et al., 2019a; Li et al., 2020, 2021a, 2022), causality-based (Tan et al., 2021, 2022; Xu et al., 2021a,b)

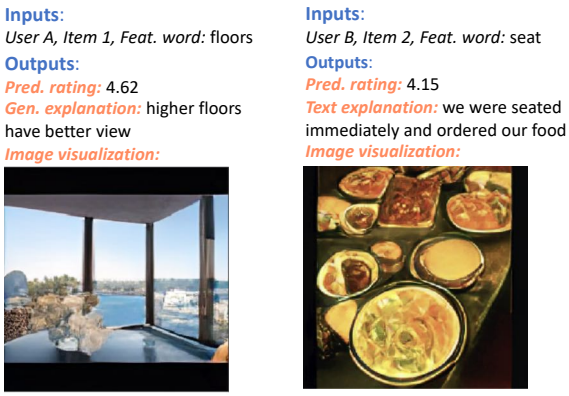

Figure 1: Example cases by the proposed model on TripAdvisor and Yelp datasets respectively.

and neural-symbolic (Shi et al., 2020; Chen et al., 2021, 2022) approaches have been explored in recent years. Among them, PETER (Li et al., 2021a) is a representative sentence-based method that directly generates explanation sentences for given user–item pairs based on Personalized Transformer.

While PETER outperforms previous methods in terms of both explainability and text quality metrics, it also suffers from several shortcomings: PETER tends to repeat certain universally applicable "safe" sentences as explanations (e.g., "the hotel is very nice"). For the 32,003 records in the test split of the TripAdvisor dataset by Li et al. (2020), PETER only generates around 8,100 unique sentences. The duplicate rate is close to 75%, while in reality, the duplicate rate of the TripAdvisor ground truth explanations is only 5.4%. In addition, such models are trained solely on a textual corpus, lacking real-world experiences to generate more authentic explanations, which may lead to empty sentences with insufficient details. Recently, Vokenization (Tan and Bansal, 2020) demonstrates that language understanding can be improved with token-level visual supervisions. This motivates us to consider enhancing text explanation generation with the aid of real-world images.

In this paper, we present an entirely new form

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 244 - 255

of explanation generation model that is immersed in a multimodal environment. The goal is to encourage it to perceive real-world signals and generate visually-enhanced explanations to better assist a user's decision. Specifically, we propose the *Multimodally-Enhanced Transformer for Explainable Recommendation* (**METER**) approach for improved text explanations based on conditional image generation and text–image matching. Unlike traditional caption-to-image generation, our training sentences are explanations that are more comprehensive reviews based on user experiences rather than simple abstract descriptions of the image content. We adopt the generation order "rating $\rightarrow$ text $\rightarrow$ image" based on the consideration that the generation difficulty should gradually increase. With this approach, we seek to guide the model to understand real-world concepts regarding both item attributes and user interests (e.g., a spacious room or modern decoration). Furthermore, METER is encouraged to visualize what it is talking about for the given user–item pair and is penalized in case of a mismatch between the generated visualization and the textual explanation. This is in line with the spirit of the context token prediction module in Li et al. (2021a). While PETER only predicts text tokens as contextual information, our METER additionally generates visual tokens as a supplement. We claim that if a sentence contains more real-world concepts, it is easier to visualize it as an image with higher fidelity. To this end, we introduce a text–image matching discriminator based on contrastive learning which helps to improve both the diversity and faithfulness of the textual explanations. Beyond an auxiliary task for text generation, another advantage of METER is that the generated image visualizations may provide intuitive visual explanations in addition to rating scores and textual explanations.

To empirically evaluate our framework, we conduct experiments and user studies on two real-world datasets in terms of diversity and faithfulness of text explanations, as well as consistency and quality of image visualizations. Our results reveal that using the proposed METER leads to improvements on text diversity and faithfulness, and that the generated image visualizations show high fidelity and good consistency. Overall, we make the following key contributions:

- To the best of our knowledge, this is the first exploration of a multimodal explainable recommender system that jointly generates rating scores, textual explanations, and images. The system will also be promising in creative advertising applications.
- By immersing the model into a multimodal environment, we help it explore the real-world concepts mentioned in the text explanations and in turn enable it to generate more diverse and faithful natural language rationales that are consistent with visual grounding.
- Experiments and a user study on real-world datasets demonstrate the superiority of our approach over several strong baselines.

2 Related Work

Visually-Guided Language Learning There have been numerous efforts on utilizing visual information to facilitate language tasks. The general strategy they typically pursue is to obtain cross-modally aligned semantics through visual grounding. Gella et al. (2017); Zhang et al. (2020); Sigurdsson et al. (2020) draw on the visual modality to bridge the gap between languages and conduct visual grounding to improve unsupervised cross-lingual word mapping or machine translation. Vokenization (Tan and Bansal, 2020) assigns each text token with a corresponding voken and improves text-based pretraining with contextualized, visual-grounded supervisions. VidLanKD (Tang et al., 2021) further solves the shortcomings of Tan and Bansal (2020) by first learning a multimodal teacher model on video-language dataset and then transferring knowledge to the student language model through distillation. Shen et al. (2021) discovers visual impressions from text-only corpus to improve open-domain dialog generation. Li et al. (2021b) learns vision–language representations with cross-modal contrastive learning on a combination of pure text corpus and image–text pairs to advance both single modal and multi-modal downstream tasks. Recently, DALL-E (Ramesh et al., 2021) merges text and visual tokens as a single stream of data and employs a universal Transformer to autoregressively model the multimodal stream. The astonishing success of these methods inspires us to guide personalized explanation generation with visual signals.

Generate Explanations for Recommendation Explainable recommendation has been an important task in both research and industry (Zhang and Chen, 2020). Early approaches mainly attempt to

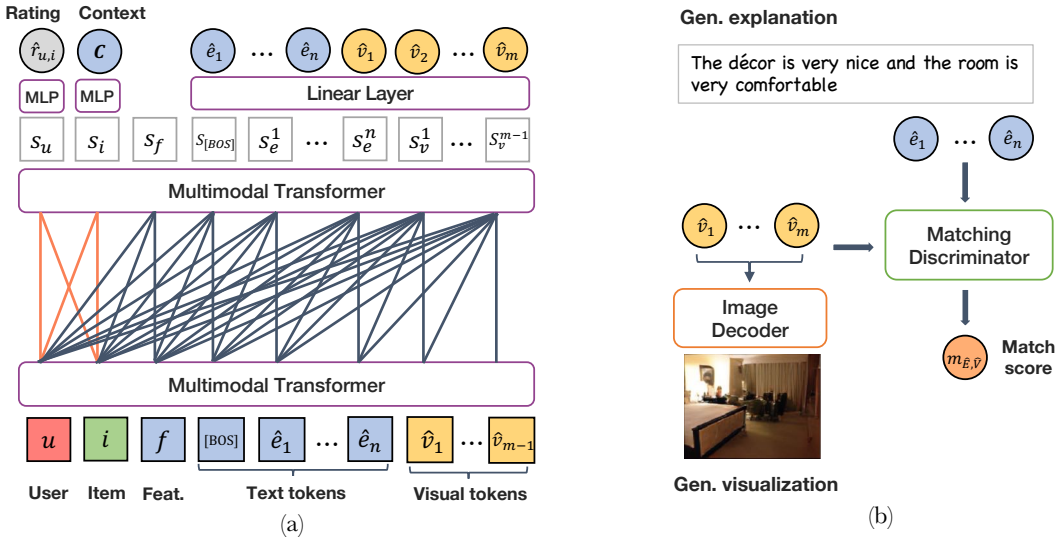

Figure 2: Architecture of METER framework: (a) Multimodally-Enhanced Transformer, which takes user ID u, item ID i, feature word f as initial condition tokens. Text tokens $\{\hat{e}_t\}_{t=1}^{n}$ are first generated triggered by the [BOS] token, next visual tokens $\{\hat{v}_t\}_{t=1}^{m}$ can be generated conditioned on (u,i,f) triplet and text sequence. (b) Text–image matching discriminator that estimates the match score between the generated text explanation and visualization.

make latent factor models interpretable by aligning each latent dimension with the explicit meaning (Zhang et al., 2014; Chen et al., 2016). In recent years, numerous neural models have been proposed to explain recommendations based on user reviews (Chen et al., 2019c,a). There have also been attempts to generate purely visual explanations (Chen et al., 2019b; Tangseng and Okatani, 2020). Compared with other explanation styles for recommendation, sentence-based methods are more straightforward and have been at the center of attention in recent times. Explanation sentences can either be generated by filling predefined templates (Zhang et al., 2014; Wang et al., 2018) or through flexible natural language approaches such as Attn2Seq (Dong et al., 2017), based on recurrent neural networks, and PETER (Li et al., 2021a), which is powered by a personalized Transformer. NETE (Li et al., 2020) combines the advantage of the two styles and produces template-controlled explanations by learning from sentence templates, which is an early form of prompt-based generation. However, none of the previous work has integrated textual and visual features and provided multimodal explanations. To the best of our knowledge, METER is also the first approach to draw on vision for improved textual explanation generation.

3 Methodology

3.1 Overview and Problem Formulation

The goal of our METER framework is to give an estimated rating score $\hat{r}_{u,i}$ that reflects a user u's preference towards item i and generate a multimodal explanation to justify the estimated rating. The generated multi-modal explanation consists of a text sentence $\hat{E}_{u,i}$ and an image visualization $\hat{V}_{u,i}$. The latter may serve as a supplement to the textual explanation for better explainability when text alone provides insufficient information. Moreover, the METER recommendation explanation model is encouraged to visualize what it is talking about for the user–item pairs and will be punished if the generated visualization does not match its textual explanation. By doing so, we aim to improve the quality, diversity, as well as faithfulness of the generated text explanations through visual grounding.

In the following, we shall first elaborate how to represent visual information into visual tokens and how to encode the positional embeddings for different types of tokens used in METER. Subsequently, we describe the Multimodal Enhanced Transformer for autoregressive multimodal explanation generation. Moreover, we will introduce the text–image matching discriminator, which guides the multimodal Transformer to generate better and more diversified text explanations. Finally, we summarize the training objectives of our framework for rating prediction and explanation generation.

3.2 Visual Encoder

To introduce visual signals into the Transformer structure, we follow the idea of VQ-VAEs (van den Oord et al., 2017) to encode an image $I \in \mathbb{R}^{H\times W\times 3}$ into a sequence of discrete patch-level

visual tokens $z_q \in \mathbb{R}^{h\times w\times d}$, where H and W is the original size of the input image, $h \cdot w$ is the number of visual patches, and d is the patch-level feature dimensionality. The visual tokens are constructed by vector-quantization through a learned discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \in \mathbb{R}^d$ of visual representations. To balance efficiency and perceptual quality, we adopt VQ-GAN (Esser et al., 2021) as the visual encoder and decoder in our framework. We first pre-train the vector-quantized visual patch encoder $\mathcal{E}$, decoder $\mathcal{G}$, and the discrete codebook $\mathcal{Z}$ on our collected images. With these pretrained components, we can encode an input image I with the encoder $\mathcal{E}$ as $\hat{z} = \mathcal{E}(I) \in \mathbb{R}^{h\times w\times d}$. Next, we serialize $\hat{z}$ and conduct element-wise quantization for individual encoding $\hat{z}_j$ of $\hat{z}$ onto its closest codebook entry z_k:

$$z_q = \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_j - z_k\| \right) \in \mathbb{R}^{h\times w\times d}$$

The resulting z_q are served as the encoded visual tokens $\{v_j\}_{j=1}^{m}$ of the input image. As for the sequence of visual tokens $\hat{z}_q = \{\hat{v}_j\}_{j=1}^{m}$ produced by METER autoregressively, we can utilize the decoder $\mathcal{G}$ to transform it back to a generated original size image $\hat{I}$: $\hat{I} = \mathcal{G}(\hat{z}_q) \in \mathbb{R}^{H\times W\times 3}$.

3.3 Input Representation

Five distinct types of input tokens can be distinguished: user ID, item ID, feature word, text tokens for explanation, and visual tokens. With the aforementioned vector-quantized visual patch encoder, we obtain a visual token representation for a given image. For text explanations, we directly tokenize them into text token sequences. Intuitively, the generated explanation should reflect both the user's interest preferences and the item attributes. Hence, we have user IDs and item IDs as two special types of tokens to guide the model to talk about the correct topics. Finally, the feature words can serve as conditional inputs to specialize the topic of explanation.

To represent tokens as embeddings, we prepare four embedding codebooks: $\mathcal{U}$ for user IDs, $\mathcal{I}$ for item IDs, $\mathcal{V}$ for text tokens and feature words, and $\mathcal{Z}$ for visual tokens. We set a fixed length m for visual tokens and a maximum length n for text tokens. Thus, the input sequence $\mathbf{S}_0$ can be represented as $\mathbf{S}_0 = [u, i, f, e_1, \cdots, e_n, v_1, \cdots, v_m]$. Before feeding the token sequence into METER, we provide positional embeddings for non-visual tokens and visual tokens separately. As the visual information has a spatial prior and is organized in a 2-D grid, we adopt an axial positional embedding (Ho et al., 2019) for visual tokens. In addition, we prepare an embedding codebook $\mathcal{P}$ for non-visual tokens. The final input sequence representation is the addition of token embeddings and the corresponding positional embeddings.

3.4 Multimodally-Enhanced Transformer

Given a input sequence, we use a Multimodally-Enhanced Transformer to encode it and predict the next token, which can be either a text or visual token. When the input sequence starts with the special token [BOS] alone, the model also predicts the rating score for the candidate user–item pair and contextual words that could reflect the user's preference and the item's attributes. Suppose our multimodal Transformer has L layers, each with h-head multi-head self-attention, and d is the input embedding dimensionality. Then, for input sequence $\mathbf{S}_l$ at layer $l \in [0, L-1]$, the encoded sequence $\mathbf{S}_{l+1}$ can be computed as follows (specifically $\mathbf{S}_L$ denotes the final-layer output):

$$\mathbf{S}_{l+1} = \mathrm{FFN}_l\left(\mathrm{Attention}\left(\mathbf{S}_l\mathbf{W_Q}, \mathbf{S}_l\mathbf{W_K}, \mathbf{S}_l\mathbf{W_V}\right)\right)$$

Here, $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{d\times d_h}$ are weight matrices for projecting query, key, and value respectively (Vaswani et al., 2017), $d_h = d/h$ is the dimensionality for each head. FFN_l is a feed-forward module consisting of two fully-connected layers with ReLU in between for the l-th Transformer layer. The Attention function is defined as

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}$$

with a scaling factor $\sqrt{d_h}$ that maintains the order of magnitude in features. We adopt a similar masking strategy as Li et al. (2021a): the user & item IDs both can attend to all tokens in the sequence, while other non-ID tokens (including feature words, text tokens, and visual tokens) all retain the traditional causal attention masking in order to avoid any leakage of future information. Figure 2 (a) provides an illustration of our masking strategy.

Assuming the final-layer output from the Transformer is $\mathbf{S}_L = [s_u, s_i, s_f, \{s_e\}, \{s_v\}]$, this also serves as a representation of the input sequence for next generation iteration. We can use these vector representations to enable the following four tasks:

Rating prediction The first representation s_u is used to conduct rating score prediction. We regard the score prediction as a regression problem and the goal is to predict the score $\hat{r}_{u,i}$ for the given pair of user/item IDs. Due to the adopted masking strategy, u and i can both attend to each other and capture the correlation between them. Here we make use of a two-layer fully-connected network with sigmoid activation σ to map s_u to a scalar score value: $\hat{r}_{u,i} = \sigma\left(s_u W_1 + b_1\right) W_2 + b_2$, where the dimensionality of input, hidden layer, and output are d, d, and 1 respectively. Mean Squared Error loss (MSE) is used for rating score regression:

$$\mathcal{L}_r = \mathbb{E}_{(u,i)\in\mathcal{T}} \left(r_{u,i} - \hat{r}_{u,i}\right)^2$$

where $r_{u,i}$ is the ground-truth rating score and $\mathcal{T}$ represents the training corpus.

Context token prediction The second representation s_i is designed to predict the context words for a given user–item pair. Similar to s_u, s_i also absorbs the words that are related to a certain user's preference and an item's attributes. Thus, this auxiliary task is able to force the Transformer to exploit the information hidden in the user ID and item ID. Such design can mitigate the problem of identical explanations being generated. By passing s_i into a single fully-connected layer with $\mathrm{Softmax}$ activation, we can obtain a probability distribution over the vocabulary $\mathcal{V}$ for the context word: $P_c = \mathrm{Softmax}\left(s_i W_c + b_c\right)$, where the dimensionality of input and output are d and $|\mathcal{V}|$, respectively. The predicted context tokens are the top-n words with the highest probability. If we represent the probabilities of these context words $\mathbf{C}$ as $\{p_c^t\}_{t=1}^n$, then the negative log likelihood (NLL) loss can be computed as:

$$\mathcal{L}_c = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} -\log p_c^t\right]$$

Explanation/visualization generation The generation of explanation words and visual codes follows the autoregressive style, i.e., decoding one token at a time from left to right. Text generation is triggered by the special [BOS] token, upon which we repeatedly decode words until [EOS] is sampled. If the number of generated text tokens before [EOS] is less than n, we pad the sequence with [PAD]. If the text sequence length is greater than n, we cut it off at length n. To obtain the visual code sequence $\hat{\mathbf{V}}$, we iterate METER for a fixed number of m steps conditioned on the text explanation $\hat{\mathbf{E}}$ and the previously generated visual code sequence. Similar to context word prediction, we adopt a single fully-connected layer for text representations $\{s_e\}$ to produce probability distributions over the text vocabulary $\mathcal{V}$. As for visual representations $\{s_v\}$, we employ another fully-connected layer to produce probability distributions over the discrete visual codebook $\mathcal{Z}$. We can then sample words and visual codes from the obtained probability distributions. For simplicity, we employ greedy decoding as the sampling method to select the word/code with the highest probability. If we denote the probabilities of the sampled words and visual codes as $\{p_e^t\}_{t=1}^n$ and $\{p_v^t\}_{t=1}^m$, respectively, then the token-level language modeling loss for text and visual code generation can be expressed as:

$$\mathcal{L}_e = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} -\log p_e^t\right] + \alpha \cdot \mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m} -\log p_v^t\right]$$

where α is a hyperparameter used to balance the training of textual and visual token generation.

Text–image matching METER is capable of generating text–image explanation pairs. However, we still need to know whether and to what degree the generated image visualization matches the text explanation from a global perspective. Hence we adopt a text–image matching discriminator D to measure the degree of congruency. From another aspect, if a generated sentence contains more real-world concepts, it is easier to ground the sentence to corresponding visual tokens and obtain an image visualization with higher fidelity. With contrastive training, we in turn push METER to generate text explanations with more grounded details. Our discriminator is equipped with two separate encoders for the visual token sequence and the text sequence. Assuming the outputs of the two encoders to be $\hat{\mathbf{E}}$ and $\hat{\mathbf{V}}$, we can construct positive training text–image pairs from the ground truth, as well as negative ones through alternate pairings. Thus, the discriminator loss can be written as:

$$\begin{aligned}\mathcal{L}_{\mathrm{d}} = {} & \mathbb{E}\left[\log\left(D(\mathbf{E}, \mathbf{V})\right)\right] + \mathbb{E}\left[\log\left(1 - D\left(\mathbf{E}, \hat{\mathbf{V}}\right)\right)\right] \\ & + \mathbb{E}\left[\log\left(1 - D\left(\hat{\mathbf{E}}, \mathbf{V}\right)\right)\right]\end{aligned}$$

In summary, the overall training objective function $\mathcal{J}$ consists of the aforementioned four losses:

$$\mathcal{J} = \min_{\Theta}\left(\lambda_e \mathcal{L}_e + \lambda_d \mathcal{L}_d + \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c\right)$$

Here, Θ denotes all trainable parameters, while λ_e, λ_d, λ_r, λ_c are regularization weights to help

Figure 3: t-SNE visualization for the top 88 clusters of sentence semantics when threshold is 0.95. For clarity, we only show a subset of centric explanation sentences.

balance the learning of different tasks. METER is then trained on $\mathcal{J}$ in an end-to-end manner.

4 Experiments and Discussions

4.1 Building Datasets

To conduct experiments, we adopt two publicly available explainable recommendation datasets proposed in Li et al. (2020). For each dataset, training/validation/testing splits are created following the ratio of $8:1:1$.

To enable the visually-enhanced model proposed in this paper, we compile a collection of images portraying real-world concepts. The real-world concepts are obtained by clustering sentence semantics with the help of Sentence-BERT (Reimers and Gurevych, 2019). At first, we use Sentence-BERT to compute the embeddings of all text explanation sentences. Since many ground-truth explanations have similar semantic meanings, we conduct fast clustering to aggregate these explanation sentences into different groups representing similar concepts and topics. Figure 3 gives a t-SNE visualization (Van der Maaten and Hinton, 2008) of the top 88 clusters if setting the similarity threshold to 0.95. From the figure, we can have a glimpse of what kinds of topics these explanation typically show. To ensure a proper amount of clusters, we set the threshold to 0.85. Thus we obtain 16,577 clusters consists of the most common 99,066 explanations for TripAdvisor and 64,937 clusters which cover 283,895 explanations for Yelp.

The explanation sentences at cluster centers are then used as query input to search relevant images through Google Images API. For TripAdvisor and Yelp, we retrieve the top 20 and top 10 images for each centric explanation sentence. As a result, we have a visual concept pool of 331,540 and 649,370 images for TripAdvisor and Yelp, respectively.

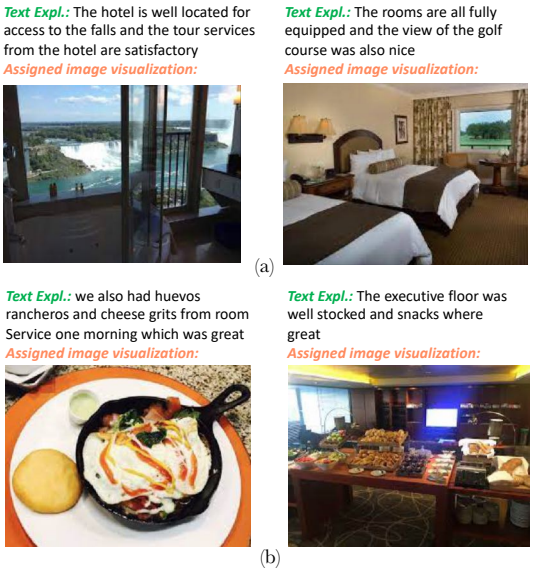

Figure 4: Example ground-truth text explanation–image visualization pairs on (a) TripAdvisor and (b) Yelp datasets.

After collecting enough images about dataset-aware topics, we assign each text explanation the most suitable image visualization by calculating the similarity between the two modalities with CLIP model (Radford et al., 2021). In this way, we build the textual recommendation explanation–image visualization pairs for both datasets and then train METER on the constructed multimodal pairs. In Figure 4, we provide several text explanations with their corresponding assigned image visualizations.

Table 1 shows the statistics of the established multimodal explainable recommendation datasets. Note that the TripAdvisor dataset mainly focuses on the hotel and travel domain, while the majority of the Yelp data is about restaurants. Records in the two datasets consist of: user ID, item ID, rating score (from 1 to 5), feature word, text explanation, and image visualization aligned with the text explanation.

Dataset	Yelp	TripAdvisor
#users	27,147	9,765
#items	20,266	6,280
#explanations	1,293,247	320,023
#features	7,340	5,069
#images	649,370	331,540

Table 1: Statistics of the experimental datasets.

4.2 Implementation Details

To ensure better representative ability of the visual encoder used in METER, the three components

Methods	Text Explainability		Text Diversity		Text Quality				Image Consistency
	FMR↑	FCR↑	DIV↓	USR↑	BLUE-1↑	BLUE-4↑	ROUGE-1↑	ROUGE-2↑	CLIPScore↑
TripAdvisor									
Att2Seq	0.06	0.15	4.32	0.17	15.27	1.03	15.92	2.09	-
Transformer	0.04	0.00	10.00	0.00	12.79	0.71	15.88	2.34	-
NETE	0.78	0.27	2.22	**0.57**	22.39	3.66	27.71	7.66	-
PETER	0.89	0.35	1.61	0.25	24.32	4.55	30.49	9.24	-
METER	**0.90**	**0.39**	**1.42**	0.56	**24.57**	**4.76**	**30.77**	**9.41**	0.62
Yelp									
Att2Seq	0.07	0.12	2.41	0.13	10.29	0.58	13.29	1.31	-
Transformer	0.06	0.06	2.46	0.01	7.39	0.42	12.56	1.09	-
NETE	0.80	0.27	1.48	**0.52**	19.31	2.69	25.56	6.63	-
PETER	0.86	**0.38**	1.08	0.34	20.80	3.43	27.95	7.94	-
METER	**0.88**	0.35	**1.02**	0.42	**21.30**	**3.61**	**28.32**	**8.09**	0.59

Table 2: Performance comparison on the TEST splits of TripAdvisor & Yelp datasets among several explanation generation methods. The metrics are organized into four groups – text explainability (FMR, FCR), text diversity (DIV, USR), text quality (BLUE, ROUGE), and image consistency (CS). Note that here BLEU and ROUGE scores are percentage values, while the other metrics are absolute values.

(i.e., encoder, decoder, and visual codebook) of VQ-GAN are first pre-trained on the collected images of the two datasets. For image visualization generation, we first sample 32 candidate images conditioned on the corresponding explanations, and then use the trained text–image discriminator to produce match scores. The image with the highest match score is finally selected as output. The embedding size d of METER is set to 256, the dimensionality of the feed-forward network's hidden layer is $1,024$. The maximum text length n of the explanation sequence is set to 15, while the length of the visual token sequence m is set to 256, and the standard image size for VQ-GAN is set to 256×256. We keep the most frequent $20,000$ words as the text vocabulary, while the size of the discrete visual codebook is $1,024$. The Multimodally-Enhanced Transformer uses $L = 8$ layers, each endowed with a multi-head attention with $h = 8$ heads. We set the regularization weights λ_e, λ_d, λ_r, and λ_c to 1.0, 1.0, 0.1, and 1.0, respectively. And we choose 7.0 as the value of the balancing hyperparameter α. The METER model is trained with Adam optimization (Kingma and Ba, 2015) under a batch size of 32, and the learning rate is set to 5×10^{-4}. We conduct all experiments on NVIDIA Quadro RTX 6000 GPUs.

4.3 Evaluation Metrics

We conduct our evaluation from three perspectives – explanation generation performance, text–image matching performance, and rating prediction performance. For each of the three aspects, we adopt both automatic and manual forms of evaluation (see Sec. 4.6). For explanation performance, we measure the text quality, diversity, and explainability of the generated explanations. For the text quality, we adopt BLEU-1 and BLEU-4, as well as ROUGE-1 and ROUGE-2. To overcome the drawbacks of the two traditional metrics, we also employ Unique Sentence Ratio (USR) proposed by Li et al. (2020) to quantify the diversity of the generated sentences. For the diversity in feature word level, we adopt Feature Diversity (DIV) proposed in Li et al. (2020), which measures the intersection of features between any two generated explanations. In explainable recommendation, an explanation will normally be valued more by users if it justifies a recommendation's advantage using certain feature words as specified in the datasets. Thus, we adopt two more metrics tailored for explainability evaluation proposed by Li et al. (2020) – Feature Matching Ratio (FMR) and Feature Coverage Ratio (FCR). Specifically, FMR measures whether a generated explanation contains the feature in the ground-truth, while FCR is computed as the number of distinct features contained in the generated explanations, divided by the total number of features in the whole dataset. To assess the text–image matching, we adopt CLIPScore (CS) proposed by Hessel et al. (2021) as an objective metric to measure the degree of correspondence for cross-modality pairs. For the rating prediction performance, we rely on two standard metrics – Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). By including the recommendation experiment, we merely seek to prove that the rating scores predicted by our method are sufficiently

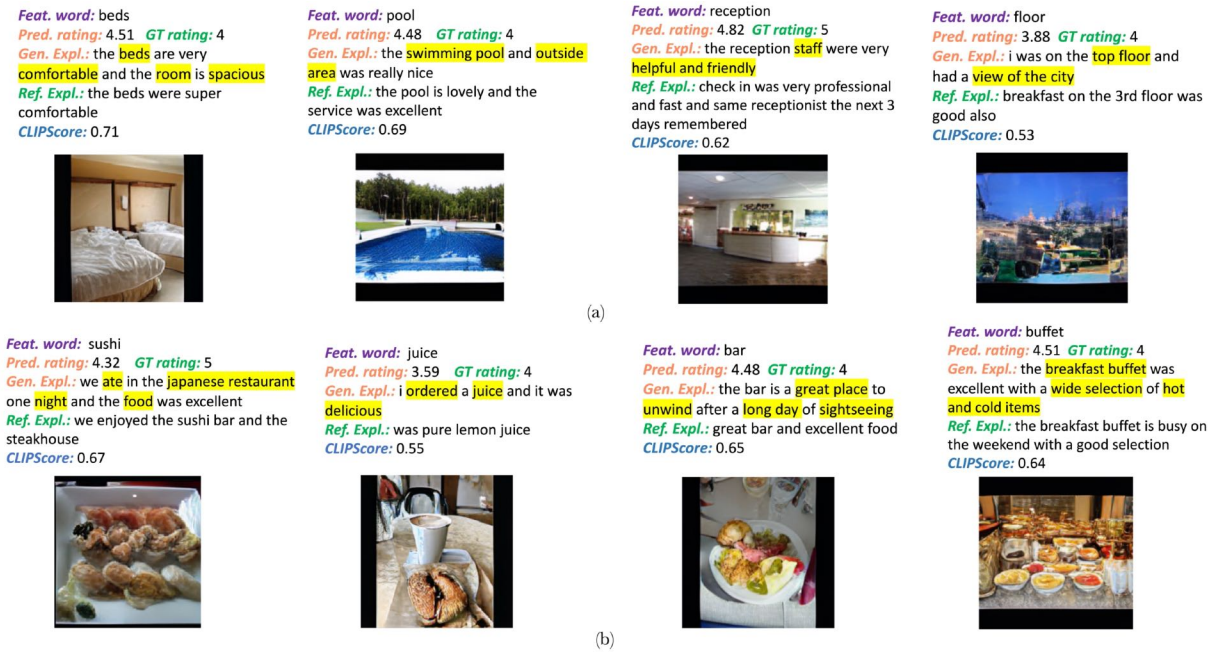

Figure 5: Qualitative results generated by METER with a conditional feature word as input: (a) is from TripAdvisor, while (b) is from Yelp. The real-world concepts in the generated explanations are highlighted.

strong to merit explanation generation, because if a rating prediction is inaccurate, the generated explanation will be less meaningful.

4.4 Comparison Baselines

For the performance comparison, we consider several baselines with regard to the task of explanation generation: **Attn2Seq** (Dong et al., 2017) learns to encode attributes into vectors, and then invokes an attention mechanism to generate reviews conditioned on the attribute vector. **Transformer** (Vaswani et al., 2017) treats user and item IDs as words and trains on the explanation generation task with a vanilla Transformer structure through language modeling. **NETE** (Li et al., 2020) designed a tailored GRU module to incorporate the given feature into the decoding stage. The system can generate template-like explanations while also making recommendations. **PETER** (Li et al., 2021a) is a simple and effective framework that attempts to use the IDs to predict the words in the target explanation. It is built upon a modified attention mask of the Transformer model. With regard to mere recommendation, we compare with two traditional methods in addition to NETE and PETER: **PMF** (Salakhutdinov and Mnih, 2007) conducts probabilistic matrix factorization in latent space. **SVD++** (Koren, 2008) combines factor and neighborhood models to enhance the accuracy.

4.5 Results and Analysis

In this section, we evaluate the performance of the proposed METER approach on two real-world datasets and compare with several representative explanation generation methods in Table 2 and recommendation models in Table 3. From Table 2, we can see that METER achieves the best FMR and DIV against all other methods, showing that METER can cover more diverse feature words during generation while maintaining good explainability. METER notably improves the USR over PETER but is slightly lower than NETE. Note that NETE is a template-based approach so it naturally achieves high USR scores. Among all methods, METER exhibits the best balance between text quality and text diversity, while being the only method that can produce both text and images, with reasonably high Image Consistency. Since automatic metrics cannot completely reflect the quality and faithfulness of generated text explanations, we also conduct a user study in the next subsection for further verification. Moreover, Table 3 indicates that METER can achieve comparable rating performance to other approaches. In Figure 5, we present several real examples illustrating how METER is able to jointly generate not only high-quality rating scores and text explanations but also image visualizations. Taking the first case in (b) as an example, we observe how METER creates coherent explanations

Methods	Yelp		TripAdvisor	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
PMF	1.09	0.88	0.87	0.70
SVD++	**1.01**	**0.78**	0.80	0.61
NETE	**1.01**	0.79	**0.79**	**0.60**
PETER	**1.01**	**0.78**	0.81	0.63
METER	**1.01**	0.79	0.80	0.61

Table 3: Recommendation performance comparison in terms of RMSE and MAE among several methods.

rather than directly copying the feature word into the generated sentence, leading to greater diversity.

4.6 User Study

To genuinely assess the quality of text explanations generated by METER and whether the image visualization matches the text explanation, we conduct a user study on the faithfulness of the generated text explanations with associated visual grounding. We randomly sampled 500 generated explainable contextual sentences as well as corresponding image visualizations. For comparison, we also randomly pick 500 samples from the baselines and randomly mixed them with the samples from our method. We asked 30 human subjects to provide a rating range from $1-5$, where larger scores represent better faithfulness and diversity. For better evaluation, we also provide the original user/item information and ground-truth explanation sentence for their reference. We consider **Faithfulness** as a criterion to assess the degree of explainability of the text, which encompasses both its readability and its cogency to the human participants. A higher **Diversity** represents more lexically varied generated context. We further consider **Consistency** representing to what extent the generated images match the associated generated sentence, while higher **Quality** scores indicate the generated image contains clearer details and better fidelity. We then calculate the overall scores by averaging the ratings given by each human participant across 500 samples each from both the baseline and from our method. The results are reported in Table 4 and show that our method can generate diverse and faithful explanation sentences of a higher quality than PETER, while also attaining a high image quality and good cross-modal consistency.

4.7 Ablation Study

We also provide an ablation study of the training tasks on TripAdvisor dataset. According to Table 5, the context prediction task has a big influence on the explainability and diversity of the generated explanations. The feature word has a vital role in deciding the topic for the model to consider. Obviously the rating prediction task is important for recommendation performance, while the visual generation task is decisive for the image consistency score. As we expected, the discriminator loss can assist the model to generate both diverse explanations and better image visualizations.

	Sentence		Image	
	Faithfulness	Diversity	Consistency	Quality
Baselines	3.41	2.96	2.54	3.04
Ours	**4.57**	**3.70**	**3.06**	**4.19**

Table 4: Manual evaluation performance between METER and baselines. Note that the baseline for sentence generation is PETER, while for image generation it is METER without VQ-GAN tokenizer pretrained on images of certain dataset. Results are not comparable across two domains (Sentence & Image).

	Expl.		Div.		Qual.	Rec.	Cons.
	FMR	FCR	DIV	USR	B4	RMSE	CS
w/o $\mathcal{L}_c$	0.82	0.20	1.73	0.33	4.22	0.80	0.57
w/o $\mathcal{L}_r$	0.85	0.38	1.45	0.54	4.71	3.25	0.60
w/o $\mathcal{L}_v$	0.87	0.37	1.49	0.45	4.58	0.80	0.13
w/o $\mathcal{L}_d$	0.83	0.34	1.58	0.39	4.35	0.80	0.54
w/o f	0.07	0.17	2.51	0.15	1.09	0.81	0.59
METER	**0.90**	**0.39**	**1.42**	**0.56**	**4.76**	**0.80**	**0.62**

Table 5: Ablation study of different training loss components of METER on the TripAdvisor dataset.

5 Conclusion

In this paper, we propose METER, the first attempt to jointly generate rating scores, text explanations, and corresponding image visualizations. We immerse our model in a multimodal environment by putting all modalities to one shared Transformer decoder structure. A text–image matching discriminator is further introduced to encourage sentences with more groundable and fine-grained concepts. Experimental results demonstrate that our framework can provide diverse and faithful text explanations, together with image visualizations as additional intuitive explanations. This proves that visual information offers auxiliary knowledge for the explanation generation model to gain awareness of real-world semantics. Our dataset and code are available at https://github.com/jeykigung/METER. In the future, we plan to investigate generating visually-enhanced explanations for more domains such as fashion and movie.

Acknowledgements

We appreciate the valuable feedback and suggestions of the reviewers. This work was supported in part by NSF IIS 1910154, 2007907, and 2046457. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2019a. Generate natural language explanations for recommendation. In *SIGIR 2019 Workshop on ExplainAble Recommendation and Search*.

Hanxiong Chen, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu, and Yongfeng Zhang. 2022. Graph collaborative reasoning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 75–84, New York, NY, USA. Association for Computing Machinery.

Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural collaborative reasoning. In *Proceedings of the Web Conference 2021*, page 1516–1527, New York, NY, USA. Association for Computing Machinery.

Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019b. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 765–774. ACM.

Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 305–314. ACM.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019c. Co-attentive multi-task learning for explainable recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2137–2143. ijcai.org.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883.

Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 69–78. ACM.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard de Melo, and Yongfeng Zhang. 2022. Path language modeling over knowledge graphs for explainable recommendation. In *Proceedings of the ACM Web Conference 2022*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2019. Axial attention in multidimensional transformers. *ArXiv preprint*, abs/1912.12180.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 426–434, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.

Lei Li, Yongfeng Zhang, and Li Chen. 2021a. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4947–4957, Online. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *ArXiv preprint*, abs/2202.07371.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1257–1264. Curran Associates, Inc.

Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. 2021. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4287–4296, New York, NY, USA. Association for Computing Machinery.

Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. 2020. Neural logic reasoning. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1365–1374. ACM.

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10847–10856. IEEE.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*.

Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 1784–1793, New York, NY, USA. Association for Computing Machinery.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34.

Pongsate Tangseng and Takayuki Okatani. 2020. Toward explainable fashion recommendation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2153–2162.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 165–174. ACM.

Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 285–294. ACM.

Yikun Xian, Zuohui Fu, Handong Zhao, Yingqiang Ge, Xu Chen, Qiaoying Huang, Shijie Geng, Zhou Qin, Gerard de Melo, S. Muthukrishnan, and Yongfeng Zhang. 2020. CAFE: coarse-to-fine neural symbolic reasoning for explainable recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1645–1654. ACM.

Shuyuan Xu, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, and Yongfeng Zhang. 2021a. Causal collaborative filtering. *ArXiv preprint*, abs/2102.01868.

Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Xu Chen, and Yongfeng Zhang. 2021b. Learning causal explanations for recommendation. In *the 1st International Workshop on Causality in Search and Recommendation (CSR'21)*.

Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 83–92. ACM.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.