electronics

MDPI

Article

Transformer-Based Multimodal Infusion Dialogue Systems

Bo Liu [1,2,3,4,†], Lejian He [5,†], Yafei Liu [6], Tianyao Yu [7], Yuejia Xiang [6], Li Zhu [1,*] and Weijian Ruan [2,3,4,8,*]

1 School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
2 The CETC Key Laboratory of Smart City Model Simulation and Intelligent Technology, Shenzhen 518038, China
3 National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen 518038, China
4 The Smart City Research Institute of CETC, Shenzhen 518038, China
5 College of Engineering, Cornell University, Ithaca, NY 14850, USA
6 Jarvis Laboratory, Tencent, Shenzhen 518054, China
7 School of Mathematics and Statistics, Yunnan University, Kunming 650500, China
8 Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518038, China
* Correspondence: zhuli@xjtu.edu.cn (L.Z.); ruanweijian@cetc.com.cn (W.R.)
† These authors contributed equally to this work as co-first authors.

Abstract: The recent advancements in multimodal dialogue systems have been gaining importance in several domains such as retail, travel, fashion, among others. Several existing works have improved the understanding and generation of multimodal dialogues. However, there still exists considerable space to improve the quality of output textual responses due to insufficient information infusion between the visual and textual semantics. Moreover, the existing dialogue systems often generate defective knowledge-aware responses for tasks such as providing product attributes and celebrity endorsements. To address the aforementioned issues, we present a Transformer-based Multimodal Infusion Dialogue (TMID) system that extracts the visual and textual information from dialogues via a transformer-based multimodal context encoder and employs a cross-attention mechanism to achieve information infusion between images and texts for each utterance. Furthermore, TMID uses adaptive decoders to generate appropriate multimodal responses based on the user intentions it has determined using a state classifier and enriches the output responses by incorporating domain knowledge into the decoders. The results of extensive experiments on a multimodal dialogue dataset demonstrate that TMID has achieved a state-of-the-art performance by improving the BLUE-4 score by 13.03, NIST by 2.77, image selection Recall@1 by 1.84%.

Keywords: multimodal; intelligent dialogue system; transformer; conversation understanding; chat bots

Citation: Liu, B.; He, L.; Liu, Y.; Yu, T.; Xiang, Y.; Zhu, L.; Ruan, W. Transformer-Based Multimodal Infusion Dialogue Systems. *Electronics* **2022**, *11*, 3409. https://doi.org/10.3390/electronics11203409

Academic Editor: Rui Pedro Lopes

Received: 29 September 2022
Accepted: 18 October 2022
Published: 20 October 2022

1. Introduction

The recent introduction of conversational agents into the market has elevated our life quality in multiple dimensions (e.g., Apple Siri, Amazon Alexa, Google Assistant and Microsoft Cortana). These intelligent conversational agents are dependent on the dialogue systems behind them, which could generally fall into two categories: task-oriented dialogues systems designed to accomplish a particular task [1,2] and open-domain conversations with casual chi-chat [3,4]. However, most of the agents on the market converse only in unimodal format such as text or voice. There emerge increasing demands for multimodal conversational agents that could understand the visual information from the dialogues and generate corresponding image responses, especially in domains such as fashion, e-commerce retail, entertainment, travel, etc. For example, Figure 1 illustrates a scenario where a customer is looking for travel bags of certain material, brands and style. Providing product images can greatly facilitate the filtering process and help agents find the target product.

Besides capturing the semantics in images, it is worth noting that plenty of domain knowledge might be only presented visually or multimodally, especially in domains such as travel and fashion. For instance, users might query about the features of statues from a specific era in a museum. These statues might have a particular style or material, which can be best represented by images instead of texts. Thus, it is essential to incorporate the multimodality into dialogue systems to help the intelligent conversational agents capture the important visual information and domain knowledge.

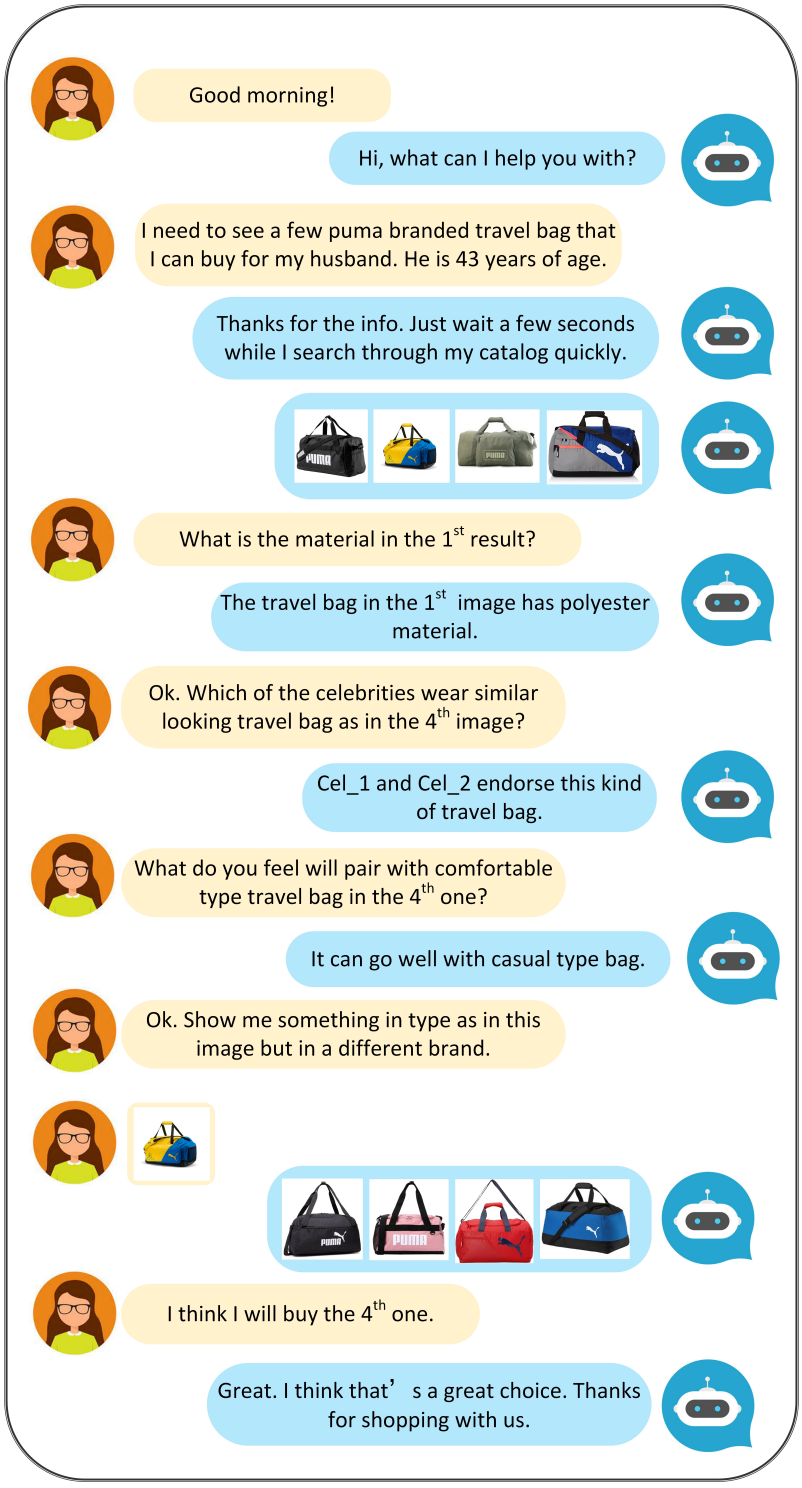

Figure 1. An example of a multimodal dialogue between a client and an agent.

Several studies have paved the way for research in multimodal task-oriented dialogue systems. Saha et al. [5] first released a multimodal dialogue dataset (MMD) and two benchmark models, MHRED and AMHRED, to address two major tasks: the textual response generation task and the image response selection task. Cui et al. [6] designed a user attention-guided multimodal dialogue model (UMD) focusing on the user requirements on the attribute level. An ordinal and attribute-aware response generation model (OAM) [7] was presented to learn enhanced image representation conditioned on the user utterance. MAGIC [8] designed adaptive decoders to separately generate general and knowledge-aware textual responses based on users' intentions. MATE [9] was the first to utilize transformers in the multimodal dialogue systems.

Although there has been promising progress made by the aforementioned studies in multimodal dialogue systems, there are still a few challenges to address. First, existing models usually perform well on generating short responses such as "Yes" or "No" but their performance on longer responses is still limited. Second, longer responses often require additional information from images to address users' requests. It is still challenging to infuse the textual and visual semantics in multimodal dialogue systems. Third, the incorporation of a knowledge vector is essential to certain types of questions such as asking for style tips or product attributes. It remains for challenges to efficiently incorporate the domain knowledge into the multimodal dialogue systems. To address these issues, we present a Transformer-based Multimodal Infusion Dialogue (TMID) model as illustrated in Figure 2. In particular, TMID first embeds the textual dialogues through a transformer encoder and extracts visual features via ResNet. Then, it employs a cross-attention transformer layer to infuse the textual and visual information and creates a multimodal representation for each utterance. The Bi-LSTM context encoder will encode all utterance representations into a context vector. Later, TMID multimodal decoder utilizes the context vector and domain knowledge to generate appropriate responses to users' requests based on the state type of the query.

To sum up, the main contributions of this work are as follows:

- We present a novel efficient Transformer-based multimodal dialogue system that performs considerably better on generating relatively long textual responses compared to previous studies.
- We apply a cross-attention mechanism to achieve better information infusion between texts and images.
- We conduct extensive experiments to evaluate TMID and achieve a stunning improvement of 13.53, 2.77, 1.84% on BLEU-4 (51.59), NIST (8.8317), and Recall@1 (99.99%) compared to the state-of-the-art method.

2. Related Work

The prominent related work of this study will be introduced in this section, which generally falls into three categories: unimodal dialogue systems, multimodal dialogue systems, and transformer-based multimodal dialogue systems.

2.1. Unimodal Dialogue Systems

With the increasing demand for intelligent conversational agents, many researchers have dedicated considerable efforts in building robust textual dialogue systems. These dialogue systems can be generally grouped into two categories based on their applications: open-domain and task-oriented dialogue systems. Though our work focuses on the task-oriented dialogue systems, some open-domain methods have put forward much progress in multmodal dialogue systems such as HRED [10] that extends the dialogue capacity by encoding multiple turns of textual context into the dialogue system and generating responses in a hierarchical encoder–decoder framework.

Task-oriented dialogue systems, by contrast, are designed to assist users to accomplish specific tasks in vertical domains [1,2]. They usually follow a typical pipeline. First, task-oriented dialogue systems encode utterances to classify the user's intentions. Then,

they employ a policy network to determine the next action. In the end, given the current user intention, they generate responses by either the predefined templates or using generation-based methods. The pipeline approach has performed decently well in task-oriented dialogue systems. However, it is still worth noting that problems such as error propagation and heavy interdependence among components might emerge in this structure [11–13]. There have been several studies tackling these problems, including end-to-end dialogue systems incorporating supervised learning and reinforcement learning [14,15], and knowledge-aware dialogue systems generating more informative responses [2,16]. Although the existing studies have pushed forward the progress of intelligent dialogue systems, they all only consider one modality in their works, which neglects the important visual semantics and knowledge in human conversations.

2.2. Multimodal Dialogue Systems

Since images carry important visual information, there have been several efforts to build multimodal dialogue systems. As a leading study, Saha et al. [5] constructed a Multimodal Dialogue (MMD) benchmark dataset consisting of more than 150k conversation sessions with domain knowledge curations. They also proposed two baseline models for the text response generation task and the image response selection task: MHRED and AMHRED, which ignored incorporating the domain knowledge. To address that, Liao et al. [12] proposed a knowledge-aware multimodal dialogue (KMD) system to encode the style tips knowledge into hierarchical neural model with attention mechanisms. Nie et al. [8] presented a multimodal dialogue system with adaptive decoders (MAGIC) that could generate general responses, knowledge-aware responses, and multimodal responses dynamically based on user intentions. Moreover, a few studies applied attention mechanisms into their models [17–19]. Chauhan et al. [7] presented an ordinal and attribute-aware response generation model (OAM) to learn enhanced image representation conditioned on the user utterance. Cui et al. [6] presented a user attention-guided multimodal dialogue model (UMD) that paid more attention to the user requirements explicitly in the attribute level.

2.3. Transformer-Based Multimodal Dialogue Systems

The extraordinary success of Transformer [20] in the field of natural language processing has also drawn researchers' attention on applying transformers to multimodal tasks [21–23]. However, rare efforts have been dedicated to applying transformers into multimodal dialogue systems. He et al. [9] proposed MATE that first utilized transformers in capturing context-aware dependencies of semantic elements. Although MATE achieved state-of-the-art performance on the MMD dataset for the textual response generation task, there still exists a large space for improvement on the textual response quality. Furthermore, MATE only focuses on the textual response generation task while ignoring the image selection task. To address these issues, we propose a novel transformer-based multimodal dialogue system that can encode the multimodal semantics using cross-attention mechanisms to generate appropriate multimodal responses.

3. Method

In this paper, we propose a Transformer-based Mutlimodal Infusion Dialogue (TMID) system illustrated in Figure 2. The overall architecture of our model can be split into two major components: multimodal context encoder and multimodal response decoder. In this section, we formalize our problem first and then elaborate the details of our method.

3.1. Problem Definition

A complete multimodal dialogue system can understand and generate multimodal information. Therefore, in this work, we address both the textual response generation task and the image selection task. Precisely, given a user query $Q = \{(U_k, I_k)\}$ and a multimodal conversational history $H_k = \{(U_t, I_t)\}_{t=1}^{k-1}$, the task is to generate a multimodal system

response $R = (U_r, I_r)$. Here, each turn of the dialogue history or query (U_t, I_t) consists of two parts: the textual utterance $U_t = \{w_i^t\}_{i=1}^{n_t}$ that contains n_t words and the image utterance $I_t = \{img_j^t\}_{j=1}^{n_t'}$ that contains n_t' images. It is worth noting that at some turns there might only exit one modality. The system response to generate can also be unimodal depending on the intention of of the user query.

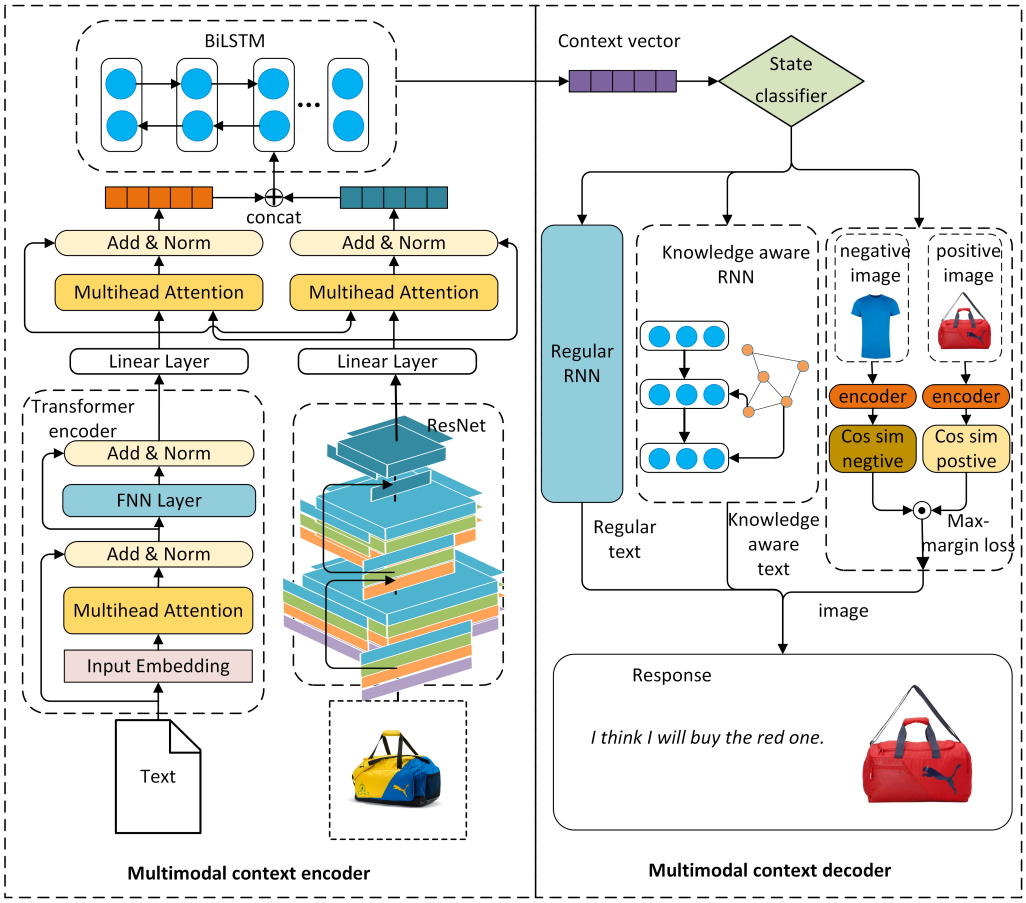

Figure 2. Framework of Transformer-based Mutlimodal Infusion Dialogue (TMID).

3.2. Multimodal Context Encoder

In order to learn the multimodal semantics of the context, we construct a multimodal context encoder as illustrated in Figure 2.

3.2.1. Transformer-Based Text Encoder

We employ a Transformer encoder [20] to extract the textual embeddings. The encoder is composed of a stack of six identical layers, which have two sub-layers each: a multi-head self-attention layer and and a position-wise fully connected feed-forward layer. We also use a residual connection [24] and layer normalization [25] after each sub-layer. More precisely, the final output of each sub-layer is $LayerNorm(x + SubLayer(x))$ where $SubLayer(x)$ is either the function of the multi-head attention layer or the fully connected feed-forward layer.

To be more specific on the multi-head attention sub-layer, it contains h single-head attentions as illustrated below:

$$head_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}} V_i) \quad (1)$$

where $Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V$. Here, Q, K, V denote the query, key, value matrix, respectively, and W_i^Q, W_i^K, W_i^V are corresponding projection parameter matrices.

Note that $d_k = d_v = d_{model}/h$ to perform as a scaling factor to reduce the dimension for each head and the total computational cost. Multi-head attention allows the model to jointly attend to information from different representation subspace at different positions as presented below:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (2)$$

where W^O is the parameter matrix for the multi-attention mechanism.

3.2.2. Cross-Attention Transformer Layer for Multimodal Infusion

We use a ResNet [24] to extract the visual features from product images. To understand and generate multimodal utterances, it is essential to infuse the textual and visual semantics. Most of the previous works directly concatenate the textual and image embeddings together for each utterance. We believe that this practice would limit the model's ability to extract useful multimodal semantics from the utterances. For example, if a user asks for a specific style of backpack in one particular image yet another color presented in another image, then a good multimodal representation should be able to pay attention to the style and color in those two images respectively given the text, and also pay attention to the key words "style", "color" and the image number. Therefore, we implement a cross-attention mechanism after we extract the textual and visual embeddings.

As shown in Figure 2, we construct a co-attentional transformer layer TRM to obtain the multimodal utterance representations. Intermediate visual and textual representations E_i^V and E_i^W from the previous text and image encoders are fed into a cross-attention transformer layer. Similar to the textual transformer encoder, we need to compute query, key, and value matrices. However, instead of inputting the corresponding key and value matrices to that modality's multi-headed attention layer, we feed them into the other modality's multi-headed attention layer, which will generate attention-pooled features for each modality conditioned on the other. Specifically, we treat the visual embeddings as key and value matrices and input them into the textual transformer encoder layer with the textual embeddings as queries $TRM_{text}(E^W, E^V, E^V)$, and vice versa for the visual transformer $TRM_{visual}(E^V, E^W, E^W)$. Through the cross-attention transformer layer, we apply image-conditioned language attentions on the visual representations H^V and text-conditioned image attentions on the textual representations H^W.

3.2.3. Bi-LSTM Utterance Encoder

After we obtain the textual H^W and visual representations H^V from the cross-attention transformer layer, we first concatenate them and then feed them into a high-level Bidirectional LSTM. The final hidden state of the high-level utterance encoder is treated as our context representation c, which will later be used to identify the intention of the query and to generate corresponding responses in either textual, visual or multimodal format .

3.3. Multimodal Response Decoder

In a multimodal dialogue system, the ability to generate multimodal system responses to users' queries is essential. However, many existing models only focus on the textual response generation task and ignore the image response at all, making them less robust regarding the nature of multi-modality. In this work, we design a multimodal response decoder that can generate either unimodal or multimodal responses based on the requirements of queries. Before we input the context vector to the multimodal response decoder, we need to identify the state of the query first. The domain experts who have constructed the MMD dataset group the states of queries into 15 types such as greeting, show-image, ask-attribute [5]. Given a specific state type of the query, the required system response is in a different format. For example, the greeting state type requires a textual response while show-orientations require multimodal responses. Therefore, we construct a simple multilayer perceptron with a cross-entropy loss to classify the state type of query using the

context vector from the multimodal utterance encoder. This simple neural network can classify the state type at a very high accuracy of 99.4% in our model.

Once we have identified the state type of the query, we pass the context vector into the multimodal response decoder, which is composed of two components: the image response decoder and the textual response decoder. Since the multimodal dialogue system only involves two modalities, it is intuitive to construct decoders for both image and text. This design of state type classification and multimodal response decoder boosts the efficiency of generating proper responses in an appropriate format. Among the 15 state types, there are 3 special ones that require the incorporation of domain knowledge: goes-with, ask-attribute and celebrity. In order to integrate the domain knowledge into decoder, we extract the domain knowledge from the corresponding preprocessed information and create a knowledge embedding vector. Once our model identifies the state type of query that falls into any of these three, we incorporate the knowledge embedding into the textual response decoder.

3.3.1. Image Response Decoder

In a multimodal dialogue setting, users might ask for specific products they want. The system is expected to search through their catalog of products and respond to the users with corresponding images. Therefore, the task of image response is indeed to rank a given set of images depending on their relevance to the context of dialogue. Therefore, we construct an image response decoder to select the best response to users. The image response decoder has two steps: constructing a multimodal product representation p from the visual semantics and the product attributes, and then computing the similarity between the context vector c and the product representation.

As shown in Figure 2, we first encode the visual semantics and product information into a combined representation. Similar to the image encoder in the multimodal context encoder, a ResNet is used to extract the visual semantics from the images, creating visual representations of the products. In addition to visual representations, we also encode textual product attributes into the final product representations. We input the embeddings of the key-value pairs of product attributes into an RNN layer sequentially and then obtain the final hidden state as the product attribute representation. The visual representation and the product attribute representation are concatenated together as a multimodal product representation p, which is then used to compute the cosine similarity with the context vector. Once we have obtained the cosine similarity for positive and negative samples, we calculate the Euclidean distance between them as follows:

$$d = ||CosSim(c, p_{pos}) - CosSim(c, p_{neg})||_2 \tag{3}$$

Then, we employ a contrastive loss function [26] to optimize the parameters when training our model:

$$L = \frac{1}{2N} \sum_{n=1}^{N} yd^2 + (1-y)\max(margin - d, 0)^2 \tag{4}$$

where d is the Euclidean distance between the positive and negative product similarity with the context vector, $y = 1$ for positive samples and $y = 0$ for negative samples, N denotes the number of positive and negative pairs, $margin$ is a predefined limit. The contrastive loss [26] function enables the discrepancy between the positive and negative samples to be as large as possible while the discrepancy between positive samples is as small as possible.

3.3.2. Textual Response Decoder

The quality of textual responses is of great significance to our final evaluation on the performance of a multimodal dialogue system. As shown in Figure 2, a textual response decoder is constructed to generate textual responses based on the context vector. The textual response decoder is an RNN. We first initialize the first hidden state of the RNN

decoder h_0 using the context vector c, which carries rich multimodal information from the dialogues. Then, we iteratively update the hidden state at time t h_t by its previous hidden state h_{t-1} and the embedding of the previous word E_{t-1} in the target response. Later, we pass the hidden state at each step into a linear layer and get a probability distribution over the vocabulary, which is optimized by minimizing the cross-entropy loss.

Recall that for several state types of queries, the incorporation of domain knowledge is required. These state types include style tips, product attributes and celebrity endorsements. To determine whether to incorporate domain knowledge or not, we design the model to automatically incorporate the corresponding knowledge vector for each of the three state types once they are classified by the state classifier. The incorporation of domain knowledge is achieved by adding the knowledge vector E_k into the word embedding of the previous token in the target response E_{t-1} when updating the hidden state at time step h_t. Specifically, we concatenate E_{t-1}, E_k and an attentive context vector c_a, which is obtained by combining the hidden states of the previous sentence through the attention mechanism. Then, we iteratively update the hidden state at time t h_t by its previous hidden state h_{t-1} and the concatenated embedding.

With respect to the knowledge vector, we obtain it from the knowledge base by using the previous hidden state h_{t-1} as the query q. Since the hidden state at the first step h_0 is actually the context vector c, we use it as our query at the beginning of the decoding and a special token 'st' is fed into our textual response decoder. The processes of generating knowledge vectors using queries are similar for all three types of knowledge. This is due to the similarity of the data format among styletips, product attributes and celebrity endorsements. We can create pairwise entries for all knowledge entries. For example, for knowledge of style tips, we construct a pairwise entry of two products that go well with each other such as (jeans, T-shirts). Then, we embed each item of the pair into a vector and concatenate them as a knowledge entry. We store all knowledge entries into a Memory Network and use the query to obtain the knowledge vector E_k. Similarly, the knowledge entries to be stored in the Memory Network for the knowledge of product attributes is the concatenated embeddings of the key-value pairs such as (color, black). For celebrity endorsements, the knowledge entries are the preference distributions over all products of different celebrities. For instance, for celebrity x, their preferences on N_p products can be represented as a one-dimensional vector of N_p. Then, we store all those preference vectors as knowledge entries in the Memory Network.

3.4. MMD Dataset

In this study, we adopt the MMD dataset as our main dataset for experiments and comparison [5]. The MMD dataset is the very first large-scale multimodal conversation dataset that provides a strong foundation for training and evaluating Multimodal Dialogue Systems. It consists of over 150k converstaions where users state their preferences and the agent tries to find the products that satisfy users' needs and requirements. The dialogues between users and agents represent shopping experiences that would usually involve both texts and images, which require multimodal understanding and feature extraction. Each conversation consists of approximately 40 utterances and every user's utterance has been associated with a state type, in other words, their intention of that utterance. The average number of words in a shopper's question is 12, while it is 14 in the agent's response. The average number of positive and negative image responses is 4. Within the dataset, over 1 million fashion products along with relevant domain knowledge in different forms are collected from several popular online retailing websites, including Amazon, Jalong and Abof and curated by domain experts.

Note that instead of utilizing the fixed visual features of products extracted from the FC6 layer of the VCGNet-16 [27] as MHRED and KMD did, we have followed the practice of MAGIC to use the original images of the products from the websites to achieve better representation construction and information infusion with textual semantics. We utilized the images crawled by MAGIC from those websites. Our work focuses on two

major research tasks proposed by [5] on the MMD dataset: the textual response generation task and the image response selection task. Many previous studies only focus on the textual response generation task, which weakens the robustness and flexibility of their multimodal dialogue systems. In regard to the training data, we treat each utterance in the conversations as the target response, and the utterances before as the context history for multimodal understanding.

4. Results

4.1. Experiment Setup

We chose the MMD dataset in our model training and testing [5]. We followed the same split of training–validation–testing (75%, 15%, 15%) in the original MMD dataset for the following two reasons: first, most studies adopting the MMD dataset followed their split, and second, the same split should make model performance comparisons fair and convincing. We adopted PyTorch as the deep learning framework in our model. During the training, we followed the practice in previous studies [8,12] to use two-turn utterances before the responses as the context for the multimodal transformer encoder. The vocabulary size was 24,622. The dimension of word embeddings was set as 300 for both the utterances and the generated textual responses. We set the dimension of the Bi-LSTM as 512 both in the textual encoder and the context encoder empirically. The dimension of the knowledge vector to be added during the decoding stage was also set as 512. With respect to the image decoder, one positive product image and four negative product images were utilized to compute the similarity. In the calculation of the contrastive loss, we set the margin as 1. Adam [28] was used for the optimization and the learning rate was initialized as 0.0001.

4.2. Baseline Models

To demonstrate the performance of TMID, we compared it with several representative methods from different perspectives: (1) Text-only methods such as Seq2seq and HRED, which only encode textual semantics into their model and ignore the visual information. (2) MMD benchmark methods by Saha et al. [5], including MHRED and AMHRED. (3) Methods that adopt attention mechanisms, including UMD and OAM. (4) Models incorporating domain knowledge such as KMD and MAGIC. (5) The method MATE utilizes transformers in this field. Details about these baselines have been introduced previously in Related Work.

4.3. Evaluation Metrics

We adopted different metrics to evaluate TMID on the textual response generation task and the image response selection task. For the textual response generation task, we utilized the BLEU-N [29] and NIST [30] as our evaluation metrics, which measure the similarity between the generated responses and the target responses. BLEU-N indicates the number of n-gram overlaps between the target and the generated responses. Higher BLEU-N scores represent more overlaps and higher similarity. Since the length of around 20% target responses is less than 4, we calculated BLEU-1 to BLEU-4 for the textual response evaluation. Based on BLEU, NIST considers the weights of n-grams dynamically where the weight of an n-gram is proportional to its rareness. Note that all textual responses, including those knowledge-aware textual responses, were evaluated together. As for the image response selection task, we followed the practice in [5,12] and used Recall-m (1 to 3 in TMID) where the selection is correct only if the positive product is ranked in the top-m ones. We used the same evaluation scripts as MAGIC [8].

4.4. Experimental Results

4.4.1. Textual Response Generation Task

Table 1 summarizes the performance of TMID and baselines on the task of textual response generation. We have the following observations from Table 1. First, TMID has superior performance compared to the baselines on both BLEU and NIST scores, which

demonstrates the effectiveness of the TMID architecture. Specifically, TMID outpeforms the state-of-the-art model MATE on BLEU-1, BLEU-2, BLEU-3, BLEU-4 and NIST by 8.26, 10.20, 11.76, 13.53 and 2.77, respectively. The improvements by TMID are the largest compared to previous studies in the field of multimodal task-oriented dialogue systems [5–9,17]. In addition to the great improvements on all BLEU scores, we find that TMID has particularly improved the performance on generating longer textual responses compared to the baslines. We can see that the improvement on BLEU-N compared to MATE increased as the length of the responses N increased with BLEU-4, achieving a stunning 13.53 improvement. This demonstrates that the transformer-based multimodal encoder with information infusion can extract textual and visual semantics much better. Furthermore, similar to previous studies, TMID has achieved a relatively high score of 64.81 on BLEU-1 since it can generate more accurate short responses such as "Yes" and "No" to many knowledge-aware questions, e.g., "Does this bag go well with the T-shirts?", which expect short responses. Moreover, TMID without the cross-attention mechanism has also achieved the same level of performance as TMID with it, though it has slightly lower BLEU and NIST scores.

Table 1. Performance comparison between TMID and baseline models on textual response generation.

Method	Framework	BLEU1	BLEU2	BLEU3	BLEU4	NIST
Text-only	Seq2seq	35.39	28.15	23.81	20.65	3.3261
Text-only	Seq2seq	35.44	26.09	20.81	17.27	3.1007
MMD Benchmarks	MHRED	32.6	25.14	23.21	20.52	3.0901
MMD Benchmarks	AMHRED	33.56	28.74	25.23	21.68	2.4600
Attention	UMD	44.97	35.06	29.22	25.03	3.9831
Attention	OAM	48.3	38.24	32.03	27.42	4.3236
Domain Knowledge	MAGIC	50.71	39.57	33.15	28.57	4.2135
Transformer	MATE	56.55	47.89	42.48	38.06	6.0604
Transformer †	**TMID** †	**64.69**	**57.92**	**54.09**	**49.45**	**8.7131**
Cross attention ‡	**TMID** ‡	**64.81**	**58.09**	**54.24**	**51.59**	**8.8317**

‡ The best performance, † second best performance.

Additionally, the improvement by the cross-attention mechanism was greater for longer responses, as BLEU-4 has the largest improvement among all BLEU scores. Improvement on NIST was also relatively considerable for longer responses.

4.4.2. Image Response Selection Task

The performance comparison of TMID and baselines on the image response selection task is summarized in Table 2. We have the following observations from Table 2. First, TMID outperformed all baselines on Recall@m of the best image selection task achieving a stunning 100% of Recall@2 and Recall@3. Recall@1 is also very close to 100%, which indicates TMID's excellent capability of selecting the best image response to user queries. We also see that TMID—using the Contrastive Loss for optimization—achieves higher Recall@1.

Table 2. Performance comparison between TMID and baseline models on image response selection.

Method	Framework@2	Recall@3		
Text-only	Seq2seq	0.5926	0.7395	0.8401
Text-only	Seq2seq	0.4600	0.6400	0.7500
MMD Benchmarks	MHRED	0.7200	0.8600	0.9200
MMD Benchmarks	AMHRED	0.7980	0.8859	0.9345
Domain Knowledge	KMD	0.9198	0.9552	0.9755
Domain Knowledge	MAGIC	0.9813	0.9927	0.9965
Cross attention †	**TMID** †	**0.999858**	**1**	**1**
Cross attention ‡	**TMID (Contrastive Loss)** ‡	**0.9999717**	**1**	**1**

‡ The best performance, † second best performance.

5. Discussion

TMID has achieved superior performance in both textual response generation and image response selection tasks. Particularly, TMID with cross attention mechanism has even achieved better performance. We argue the following conclusions from this observation: (1) The majority of TMID's improvement comes from the elevated multimodal, especially the textual understanding of the dialogue context by transformers. Transformer is verified to be considerably effective in the field of multimodal dialogue systems. (2) The cross-attention mechanism can better infuse information from texts and images to increase TMID's mutlimodal understanding. (3) We claim that this superior performance on the image selection task is because TMID uses the ResNet to directly extract the visual features of the original product images and incorporates the product knowledge into the product representations as MAGIC. (4) TMID with the Contrastive Loss for optimization achieving higher Recall@ demonstrates that the Contrastive loss function can strengthen TMID's ability to distinguish positive and negative products.

We also found that TMID with cross-attention mechanism performs better on longer responses with higher BLEU4 and NIST. Recall that NIST weighs more on the rare n-gram responses. Since the MMD dataset has less long responses, higher NIST scores indicate higher performance on long response generation. It is reasonable that the cross-attention performs well for longer responses since short responses such as "Yes" or "No" do not need much information from images. However, in long sentences, there exit certain words that need attention to be macthed with the visual information. For example, if the user asks "What is the material of the T-shirts?" and refers to a product image, then the model needs to pay attention not only to the "material" and "T-shirts" in the question but also to the product image the user refers to. Here, the cross-attention mechanism enables TMID to infuse the textual and visual information appropriately so that its performance on longer responses is improved.

5.1. Model Ablation Analysis

Through extensive experiments, TMID has demonstrated extraordinary performance in multimodal dialogue understanding and generation, especially in the case with a cross-attention mechanism applied. However, we are still uncertain if the boosted performance is substantially caused by cross attention or other techniques we adopted in our framework. Therefore, we have conducted granular ablation analysis to identify the major contributor of improved performance. Table 3 has summarized the results of our ablation study. In addition to utilizing cross attention in TMID, we have tested the the effect of text encoder in our dialogue system. We replaced Transformer-based text encoder with a less effective Bi-LSTM encoder of which the results are as expected. The BLEU scores of TMID with Bi-LSTM text encoder have all dropped by more than 1 point compared to the original TMID, especially for BLEU4, which dropped by 3.33. This indicates that Transformer-based text encoder plays an important role in understanding the dialogues and generating proper responses accordingly, particularly for longer responses. To investigate the impact of utterance encoder on TMID performance, we have trained TMID with a Bi-GRU utterance encoder, which has achieved comparable but slightly worse results compared to Bi-LSTM. We further investigated the effect of domain knowledge on TMID performance. Without domain knowledge in the decoder, we find out that the BLEU and NIST scores all slightly decreased. To sum up, Transformer-text encoder and cross-attention mechanism have major contributions to the superior performance of TMID, while domain knowledge has a slight positive impact on TMID decoding.

Table 3. Ablation study of TMID.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	NIST
Bi-LSTM text encoder	63.12	56.11	52.28	46.12	8.5816
No domain knowledge	64.31	57.57	53.92	49.04	8.6839
Bi-GRU utterance encoder	64.58	57.68	54.03	49.51	8.7082
TMID †	**64.69**	**57.92**	**54.09**	**49.45**	**8.7131**
TMID (Cross attention) ‡	**64.81**	**58.09**	**54.24**	**51.59**	**8.8317**

‡ The best performance, † second best performance, also as baseline for ablation study.

5.2. Case Study

We extracted a few sample responses of all types generated by MATE and TMID, and compared them against the ground truth summarized in Table 4. Note that these examples are carefully selected to best represent the overall response generation ability. We can see that in cases of general greetings, TMID's responses are more accurate and have less grammar mistakes than MATE. When asked on the different orientations of the product image, both frameworks can correctly provide the answer but TMID is still more accurate and close to the ground truth. However, when it comes to knowledge-ware response generation, TMID has better performance. Regarding style tips, TMID can mention more details than MATE. For instance, MATE only responded "it can go well with mocas style footwear" but TMID additionally matches it with flexible style shorts, which is closer to the ground truth. We found that TMID did particularly well in describing product attributes compared to MATE and generated responses very similar to the ground truth. TMID can also capture more accurate celebrity endorsement than MATE in most cases.

Table 4. Case study of TMID.

Methods	Text Responses
MATE	hi, please i help i with can help you
TMID	hello, please tell me how can i help you
Ground Truth	hello, please tell me what can i help you with?
MATE	the similar looking ones are image from the front, right, back and left
TMID	the similar looking ones are image from the front, right, back and left orientations respectively
Ground Truth	the similar looking ones are image from the front, right, back and left viewpoints respectively
MATE	absolutely . thats a thats a great choice
TMID	absolutely . i think thats a great choice
Ground Truth	absolutely, i think thats a great training shoes thank you for shopping with us
MATE	it can go well with mocas style footwear
TMID	it can go well with mocas style footwear and with flexible style, flexible with shorts
Ground Truth	it can go well with casual fitted, casual type trousers and with flexible style, flexible fit shorts
MATE	2nd product will go well with it
TMID	1st product will go well with it
Ground Truth	1st product will go well with it
MATE	regarding the second item, crocs, is a designer, of, for the, under the,
TMID	for the second item, crocs, inc. is a rapidly growing designer, manufacturer and retailer of footwear for men, women and children under the crocs brand
Ground Truth	regarding the second item, crocs, inc. is a rapidly growing designer, manufacturer and retailer of footwear for men, women and children under the crocs brand
MATE	celebrities cel_2644 endorses this kind of slippers
TMID	celebrities cel_193 endorse this kind of slippers
Ground Truth	celebrities cel_578 and cel_193 endorse this kind of slippers

6. Conclusions

This paper proposes a Transformer-based multimodal infusion dialogue system that encodes and generates knowledge-aware multimodal responses to users' queries. Specifically, we first adopt a transformer encoder to encode textual information and ResNet to extract visual features. Then we employ a cross-attention mechanism to infuse the textual and visual information to create multimodal representations, which will later be passed into a sentence-level context encoder to obtain the context vector. We incorporate the knowledge vectors into the textual decoder to generate knowledge-aware textual responses and utilize an image decoder with a constrative loss function to select the best image response based on the context vector and the product representations.

Although the proposed model has achieved excellent performance on both the textual response generation task and the best image selection task, there is still a gap between the study and the industrial application. First, the MMD dataset only covers a tiny portion of the products in the retail domain and is restricted to retailing. Hence, the model might have limitations in other domains or new products. Second, the model only focuses on two tasks in the field of multimodal dialogue systems. However, there are other tasks such as image response generation in cases where the system does not have relevant products but wants to generate an image to see if it meets the user's need. Reinforcement learning for Multimodal Image Retrieval could be useful in this case [31]. Moreover, it is more comprehensive to include product images from real-life settings where models actually wear the products in different scenarios such as in the street or doing sports. To well capture the visual information, pose tracking and image boundary should be tackled in the embedding process [32–34]. Last, there are still a lot of noise in the image selections in dialogues. We need to pay close attention to the impact of noise labels in both text [35] and visual embedding of Multimodal dialogue systems. In the future, we will further explore the aforementioned issues and extend our current model.

Author Contributions: Conceptualization, B.L., L.H. and Y.L.; methodology, Y.L., Y.X. and L.H.; software, L.H.; validation, B.L. and T.Y.; formal analysis, B.L., Y.L., L.H. and Y.X.; investigation, B.L., L.Z. and W.R.; resources, B.L., L.Z. and W.R.; data curation, T.Y. and L.H.; writing—original draft preparation, L.H.; writing—review and editing, B.L., Y.L., Y.X., T.Y. and L.H.; visualization, T.Y. and L.H.; supervision, B.L., L.Z. and W.R.; project administration, B.L., L.Z. and W.R.; funding acquisition, L.Z. and W.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the national key research and development project: 2019YFB2102500.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study has used a public dataset called MMD. Details about MMD can be found here in this link: https://amritasaha1812.github.io/MMD/, accessed on 28 September 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mrkšić, N.; Séaghdha, D.O.; Wen, T.-H.; Thomson, B.; Young, S. Neural belief tracker: Data-driven dialogue state tracking. *arXiv* **2016**, arXiv:1606.03777.
2. Wen, T.-H.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L.M.; Su, P.-H.; Ultes, S.; Young, S. A network-based end-to-end trainable task-oriented dialogue system. *arXiv* **2016**, arXiv:1604.04562.
3. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A diversity-promoting objective function for neural conversation models. *arXiv* **2015**, arXiv:1510.03055.
4. Shang, L.; Lu, Z.; Li, H. Neural responding machine for short-text conversation. *arXiv* **2015**, arXiv:1503.02364.
5. Saha, A.; Khapra, M.; Sankaranarayanan, K. Towards building large scale multimodal domain-aware conversation systems. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
6. Cui, C.; Wang, W.; Song, X.; Huang, M.; Xu, X.-S.; Nie, L. User attention-guided multimodal dialog systems. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 445–454.

7. Chauhan, H.; Firdaus, M.; Ekbal, A.; Bhattacharyya, P. Ordinal and attribute aware response generation in a multimodal dialogue system. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5437–5447.
8. Nie, L.; Wang, W.; Hong, R.; Wang, M.; Tian, Q. Multimodal dialog system: Generating responses via adaptive decoders. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1098–1106.
9. He, W.; Li, Z.; Lu, D.; Chen, E.; Xu, T.; Huai, B.; Yuan, J. Multimodal dialogue systems via capturing context-aware dependencies of semantic elements. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2755–2764.
10. Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
11. Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; Yin, D. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1437–1447.
12. Liao, L.; Ma, Y.; He, X.; Hong, R.; Chua, T.-S. Knowledge-aware multimodal dialogue systems. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 801–809.
13. Nie, L.; Song, X.; Chua, T.-S. Learning from multiple social networks. *Synth. Lect. Inf. Concepts Retr. Serv.* **2016**, *8*, 1–118.
14. Bordes, A.; Boureau, Y.-L.; Weston, J. Learning end-to-end goal-oriented dialog. *arXiv* **2016**, arXiv:1605.07683.
15. Li, X.; Chen, Y.-N.; Li, L.; Gao, J.; Celikyilmaz, A. End-to-end task-completion neural dialogue systems. *arXiv* **2017**, arXiv:1703.01008.
16. Williams, J.D.; Zweig, G. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv* **2016**, arXiv:1606.01269.
17. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
18. Mei, H.; Bansal, M.; Walter, M.R. Coherent dialogue with attention-based language models. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
19. Yao, K.; Zweig, G.; Peng, B. Attention with intention for a neural network conversation model. *arXiv* **2015**, arXiv:1510.08565.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
21. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv* **2019**, arXiv:1908.02265.
22. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
23. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for Computational Linguistics. Meeting, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
26. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Las Vegas, NV, USA, 27–30 June 2016; Volume 2, pp. 1735–1742.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
29. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 June 2002; pp. 311–318.
30. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research, San Diego, CA, USA, 24–27 March 2002; pp. 138–145.
31. Kaushik, A.; Jacob, B.; Velavan, P. An Exploratory Study on a Reinforcement Learning Prototype for Multimodal Image Retrieval Using a Conversational Search Interface. *Knowledge* **2022**, *2*, 7. [CrossRef]
32. Ruan, W.; Ye, M.; Wu, Y.; Liu, W.; Chen, J.; Liang, C.; Li, G.; Lin, C.W. TICNet: A Target-Insight Correlation Network for Object Tracking. *IEEE Trans. Cybern.* **2022**, *52*, 12150–12162. [CrossRef] [PubMed]
33. Ruan, W.; Chen, J.; Wu, Y.; Wang, J.; Liang, C.; Hu, R.; Jiang, J. Multi-Correlation Filters With Triangle-Structure Constraints for Object Tracking. *IEEE Trans. Multimed.* **2019**, *21*, 1122–1134. [CrossRef]
34. Ruan, W.; Liu, W.; Bao, Q.; Chen, J.; Cheng, Y.; Mei, T. POINet: Pose-Guided Ovonic Insight Network for Multi-Person Pose Tracking. In Proceedings of the 27th ACM International Conference on Multimedia (ACM MM), Nice, France, 21–25 October 2019; pp. 284–292.
35. Liu, B.; Xu, W.; Xiang, Y.; Wu, X.; He, L.; Zhang, B.; Zhu, L. Noise Learning for Text Classification: A Benchmark. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Korea, 12–17 October 2022; pp. 4557–4567.