ZRIGF: An Innovative Multimodal Framework for Zero-Resource Image-Grounded Dialogue Generation

Bo Zhang
Dalian University of Technology
Dalian, China
zhangbo1998@mail.dlut.edu.cn

Jian Wang*
Dalian University of Technology
Dalian, China
wangjian@dlut.edu.cn

Hui Ma
Dalian University of Technology
Dalian, China
huima@mail.dlut.edu.cn

Bo Xu
Dalian University of Technology
Dalian, China
xubo@dlut.edu.cn

Hongfei Lin
Dalian University of Technology
Dalian, China
hflin@dlut.edu.cn

ABSTRACT

Image-grounded dialogue systems benefit greatly from integrating visual information, resulting in high-quality response generation. However, current models struggle to effectively utilize such information in zero-resource scenarios, mainly due to the disparity between image and text modalities. To overcome this challenge, we propose an innovative multimodal framework, called ZRIGF, which assimilates image-grounded information for dialogue generation in zero-resource situations. ZRIGF implements a two-stage learning strategy, comprising contrastive pre-training and generative pre-training. Contrastive pre-training includes a text-image matching module that maps images and texts into a unified encoded vector space, along with a text-assisted masked image modeling module that preserves pre-training visual features and fosters further multimodal feature alignment. Generative pre-training employs a multimodal fusion module and an information transfer module to produce insightful responses based on harmonized multimodal representations. Comprehensive experiments conducted on both text-based and image-grounded dialogue datasets demonstrate ZRIGF's efficacy in generating contextually pertinent and informative responses. Furthermore, we adopt a fully zero-resource scenario in the image-grounded dialogue dataset to demonstrate our framework's robust generalization capabilities in novel domains.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Intelligent agents**.

KEYWORDS

Image-grounded dialogue, multimodal fusion, zero resource, contrastive learning

---

*Corresponding author.

1 INTRODUCTION

Dialogue systems have attracted considerable interest in AI research due to their promising applications in virtual assistants and social chatbots [27, 39]. Although existing models generate fluent responses thanks to advanced neural architectures like Transformer [33] and pre-training techniques such as DialoGPT [37], a noticeable gap remains when comparing machine-generated dialogues to human-to-human conversations. A primary factor contributing to this disparity is that, unlike humans who can associate dialogue with background knowledge such as visual and auditory perceptions, machines can only access limited information from superficial text.

Bridging the gap between human and machine communication requires incorporating visual scene information into dialogue generation. Recent research has explored various vision-language dialogue tasks, including visual dialogue [1, 3] and image-grounded dialogue [16, 28, 31, 35]. Visual dialogue entails answering questions about the factual content of an image through multi-turn conversations, while image-grounded dialogue aims to generate fitting responses to given images and dialogue contexts through casual conversation. The remarkable success of ChatGPT [24] has spurred increased research focus on image-grounded dialogue due to its lack of multimodal capabilities. However, there are still some challenges remaining unresolved.

One of the major challenges in image-grounded dialogue research is the scarcity of large-scale dialogue datasets that are naturally related to images, as obtaining such data from human annotators is costly. Although some crowdsourced datasets have been proposed [17, 23, 30], their limited training size hinders the development of dialogue generation models applicable to diverse domains. Consequently, there is a growing interest in improving image-grounded dialogue performance within zero-resource scenarios, where context-image-response triples obtained from crowdsourcing remain unused for training purposes. Previous endeavors, such as [16] and [28], have explored zero-resource image-grounded
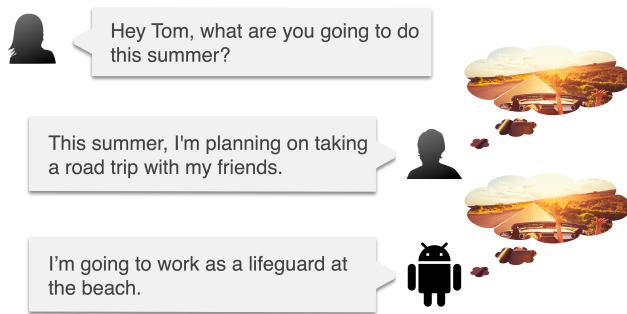
Bo Zhang, Jian Wang, Hui Ma, Bo Xu, & Hongfei Lin

Figure 1: An example of image-grounded dialogue, illustrating a comparison between human and robot responses to a given image.

dialogue generation; however, these efforts only focused on utilizing retrieved images to enhance response quality, neglecting the crucial aspect of generalization ability in zero-resource scenarios. To address this issue, we propose a novel approach with strong generalization ability that can effectively adapt and perform well in new domains, even devoid of labeled data for fine-tuning.

Moreover, a significant modality gap exists between images and texts, particularly between dialogue context and corresponding images, which can result in suboptimal use of visual information. A common approach [23] for image-grounded dialogue generation involves encoding images and texts separately and then generating responses during the decoding process; however, this method completely disregards the modality gap. Recent models, such as Maria [16] and VisAD [28], have attempted to incorporate visual information into dialogue generation before or during the decoding process. These approaches share similarities with text-only knowledge-grounded dialogue generation [22, 36], but they fail to completely bridge the modality gap due to the persistent lack of correlation among different modality vectors. As shown in Figure 1, a person asks about summer vacation plans, and the dialogue partner (either a human or a robot) provides a response based on the given image. The human response is coherent and contextually appropriate, while the robot response is mismatched and irrelevant to the image content, which shows a highway and not a beach or any sand. This example shows the importance of aligning different modality vectors for effective image-grounded dialogue generation.

To address the challenges inherent in a fully zero-resource scenario for image-grounded dialogue generation, we present an innovative multimodal encoders-decoder architecture, termed ZRIGF. The encoders and decoder are functionally segregated into four modules: a text-image matching module, a text-assisted masked image modeling module, a multimodal fusion module, and an information transfer module. To combine these modules, we devise a learning strategy consisting of two stages: contrastive pre-training and generative pre-training. **First**, during the contrastive pre-training stage, the encoders serve as a text-image matching module that maps image and text inputs into a unified encoded vector space, thereby aligning different modality vectors. This matching module trains on a large corpus of annotated image-text pairs using pre-trained image and text encoders. Additionally, we implement

a text-assisted masked image modeling module that preserves the original pre-training visual features and realigns the multimodal features. **Second**, during the generative pre-training stage, a multimodal fusion module and an information transfer module are employed to generate insightful responses based on aligned multimodal representations. These modules further facilitate alignment of representations. The entire model in this stage is trained using a corpus of annotated context-response pairs augmented with relevant images obtained through the text-image matching module.

To assess the efficacy of our proposed ZRIGF, we perform experiments on two publicly available dialogue datasets, namely the text-based Reddit Conversation Corpus [8] and the image-grounded Image-Chat dataset [30]. The experimental results demonstrate that ZRIGF outperforms competitive baselines on most automatic evaluation metrics and achieves comparable performance with ChatGPT on human evaluation metrics. Moreover, ZRIGF exhibits remarkable robustness and generalization ability under different data settings, especially in the fully zero-resource scenario.

The contributions of this paper can be summarized as follows:

- We propose a novel framework for image-grounded dialogue generation within a fully zero-resource scenario, proficiently addressing the modality gap between images and texts through a two-stage learning strategy: contrastive pre-training and generative pre-training.
- We introduce pioneering components such as a text-assisted masked image modeling module and an information transfer decoder, which further align multimodal representations and generate informative responses.
- We conduct comprehensive experiments on two publicly available dialogue datasets, showcasing the superior performance and robust generalization ability of our proposed ZRIGF in comparison to state-of-the-art models.

2 RELATED WORK

2.1 Image-Grounded Dialogue Generation

Image-grounded dialogue generation is a task that seeks to generate natural and engaging responses based on provided images and dialogue context. This task demands the optimal utilization of both textual and visual information. Early attempts at image-grounded dialogue generation [23, 30] were similar to visual dialogue, wherein the conversation topic revolves around a given image. However, recent research has shifted towards open-domain chitchat, utilizing the image to supplement conversation rather than dictate it.

The most elementary method, as delineated in [11], concatenates encoded image and text vectors before feeding them into the decoder to generate a response. Despite its simplicity, this approach neglects to capture the intricate relationship between visual and textual modalities, resulting in subpar responses. To overcome this limitation, researchers devised more sophisticated methods to more effectively incorporate visual information and bridge the modality gap. For example, Liang *et al.* [16] fed the encoded image information and embedded text into an autoregressive model [4] to generate responses. During the decoding stage, the masked concept prediction objective was employed to assist the model in aligning multimodal representations. Similarly, Shen *et al.* [28] used a more complex model structure, passing the encoded image vector and

text vector through a co-attention encoder and a cascade decoder to bridge the modality gap. Nevertheless, these methods still struggle with the modality gap problem due to the lack of alignment between different modality vectors.

Our proposed framework advances these existing methods and more effectively addresses the modality gap problem by incorporating a two-stage learning strategy. This design ensures a more thorough integration of visual and textual information, yielding more coherent and contextually appropriate responses in image-grounded dialogue generation.

2.2 Zero-Resource Learning

Zero-resource learning refers to scenarios where no training data exists for a specific task or domain. This scenario has been explored in various works, including Zero Resource Neural Machine Translation [2, 9], Zero Resource Speech Challenge [6, 7], and Zero Resource Dialogue Generation [15, 38].

Li *et al.* [15] introduced a double latent variable model and a variational learning approach to investigate knowledge-grounded dialogue generation within a zero-resource scenario. Since then, the majority of dialogue generation work has been founded on this method, encompassing zero-resource image-grounded dialogue generation. For example, Yang *et al.* [35] hypothesized the presence of a latent image variable underpinning the textual dialogues and endeavored to recover it using text-to-image generation techniques. However, the text-to-image generation technique requires some annotated context-image-response data, which remains unavailable in zero-resource scenarios. Liang *et al.* [16] and Shen *et al.* [28] employed retrieval methods to obtain images corresponding to dialogue context, excluding the availability of context-image-response triples obtained from crowdsourcing for training purposes. The primary objective of these methods is to enhance dialogue generation by incorporating retrieved images. Although these methods have made significant progress, they still require annotated data for fine-tuning before effectively applying them to new domains.

Our proposed framework addresses the challenges inherent in a fully zero-resource scenario for image-grounded dialogue generation. By incorporating the learning strategy, our model effectively aligns multimodal vectors without relying on the annotated triples. This approach enables our model to generalize well to new domains where no labeled data is available for fine-tuning, making it a more practical solution for real-world applications.

3 METHODOLOGY

3.1 Task Formalization and Model Overview

The task of image-grounded dialogue generation can be defined as: given a dialogue context C and an associated set of the top k most relevant images $I_{1:k}$, the goal is to generate a coherent and contextually appropriate response R. The dialogue context C consists of a sequence of tokens $c_1, c_2, \ldots, c_n$, where c_i represents the ith token in the conversation.

In a zero-resource scenario, the objective is to learn a text-image matching module $P(I_{1:k}|C)$, which provides a probability distribution over image sets, and a generative model $P(R|C, I_{1:k})$ capable of handling the set of top k images collectively. By combining these two components, we obtain the following equation for a zero-resource image-grounded dialogue generation model $P(R|C)$:

$$P(R|C) = P(R, I_{1:k}|C) = P(R|C, I_{1:k})P(I_{1:k}|C) \quad (1)$$

The equality holds because the generative model can effectively utilize the information derived from the set of top k images, and the text-image matching module is capable of providing a probability distribution over image sets.

Our proposed ZRIGF consists of four main components: a text-image matching module, a text-assisted masked image modeling module, a multimodal fusion module, and an information transfer module. These components are merged into our two-stage learning strategy: contrastive pre-training and generative pre-training.

3.2 Contrastive Pre-training

As shown in Figure 2, the aim of contrastive pre-training is to learn a text-image matching module $P(I_{1:k}|C)$ capable of providing a set of relevant images when given a dialogue context C. To achieve this, we utilize two modules: a text-image matching module and a text-assisted masked image modeling module. These modules are trained on a large corpus of annotated images and text pairs.

3.2.1 Text-Image Matching Module. The text-image matching (TIM) module takes an image-text pair as input and produces a similarity score indicating the extent of their correspondence. The input image is first processed by an image encoder E_I, initialized with a pre-trained ViT [5], to obtain the image hidden state $\mathbf{h}_I$. Then, we pool the model to acquire its image representation $\bar{\mathbf{h}}_I$ by simply taking $\mathbf{h}_I$ corresponding to the first token. Similarly, the input text is processed by a text encoder E_T, which is initialized with a pre-trained BART Encoder [13], to obtain its text representation $\bar{\mathbf{h}}_T$. We then compute the cosine similarity between $\bar{\mathbf{h}}_I$ and $\bar{\mathbf{h}}_T$ as follows:

$$s_{IT} = \cos(\bar{\mathbf{h}}_I, \bar{\mathbf{h}}_T) = \frac{\bar{\mathbf{h}}_I^T \bar{\mathbf{h}}_T}{\|\bar{\mathbf{h}}_I\|\|\bar{\mathbf{h}}_T\|} \quad (2)$$

We train this module using the CLIP loss [26], which aims to learn a joint embedding space for images and texts by maximizing the similarity between corresponding image-text pairs (i.e., positive pairs) and minimizing the similarity between non-matching image-text pairs (i.e., negative pairs). The loss function is defined as follows:

$$\mathcal{L}_{match} = -\mathbb{E}_{(I,T)\sim P(I,T)}\left[\log \frac{\exp(s_{IT}/\tau)}{\sum_{T'\sim P(T)} \exp(s_{IT'}/\tau)} \right.$$
$$\left. + \log \frac{\exp(s_{IT}/\tau)}{\sum_{I'\sim P(I)} \exp(s_{I'T}/\tau)} \right] \quad (3)$$

where τ represents a learnable temperature parameter controlling the distribution's sharpness, $P(I, T)$ is the joint distribution of images and texts, $P(I)$ is the marginal distribution of images, and $P(T)$ is the marginal distribution of texts. By minimizing this loss function, we can align different modality vectors in the unified encoded vector space during the encoding process. This alignment can help bridge the modality gap and facilitate multimodal fusion in subsequent stages.

3.2.2 Text-Assisted Masked Image Modeling Module. While the alignment of different modality vectors is advantageous for multimodal fusion, it may also result in a loss of pre-training information due to projection. To address this issue, we employ a text-assisted
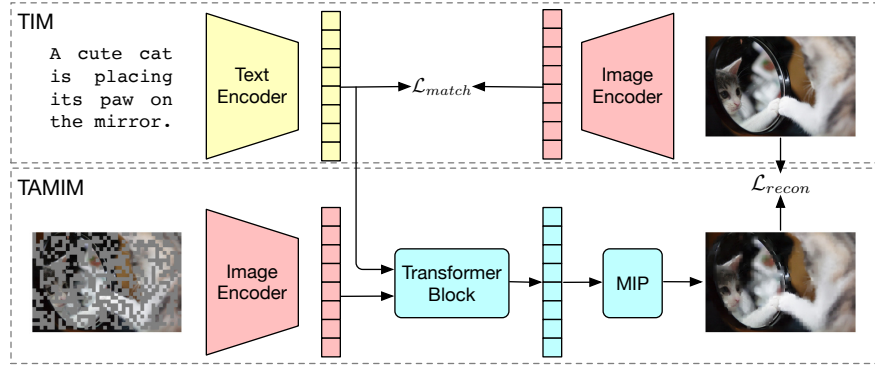
Figure 2: Overview of the contrastive pre-training stage.

masked image modeling (TAMIM) module during the contrastive pre-training stage. The TAMIM module takes a masked image and its corresponding text as input, in which a random portion of the image has been replaced with a mask token. The module's objective is to reconstruct the masked portion of the image using the unmasked regions, supplemented by the text representation.

To achieve this, we first obtain the masked image vector $\mathbf{h}_{I'}$ using the image encoder E_I. We then integrate the text vector $\mathbf{h}_T$ and the masked image vector $\mathbf{h}_{I'}$ through a transformer block TB [33], consisting of multi-head attention and a position-wise feed-forward network. Next, we feed the output vector into a masked image prediction head MIP, composed of a convolution layer followed by a pixel-shuffle operation [29]. This process can be described as follows:

$$\mathbf{h}_{I'} = E_I(\mathbf{v}_I \odot \mathbf{M}) \tag{4}$$

$$\hat{\mathbf{v}}_I = MIP(TB_I(\mathbf{h}_{I'}, \mathbf{h}_T, \mathbf{h}_T)) \tag{5}$$

where $\mathbf{v}_I$ denotes the original patch vector, $\hat{\mathbf{v}}_I$ represents the reconstructed patch vector, $\mathbf{M}$ is a binary mask with dimensions identical to $\mathbf{v}_I$, and $\odot$ indicates element-wise multiplication.

Afterward, we compute the reconstruction loss between the original patch vector $\mathbf{v}_I$ and the reconstructed patch vector $\hat{\mathbf{v}}_I$ using the mean absolute error (MAE) loss:

$$\mathcal{L}_{recon} = \frac{1}{N_M} \sum_{i=1}^{N_M} |\mathbf{v}_I^{(i)} - \hat{\mathbf{v}}_I^{(i)}| \tag{6}$$

where N_M is the total number of masked image patches. By minimizing this loss function, the TAMIM module can accurately reconstruct the masked portions of the image using the unmasked regions and the text representation. This can help preserve more pre-training visual information and further enhance the alignment between different modality vectors.

The overall loss function for the contrastive pre-training stage is a linear combination of the text-image matching loss and the text-assisted masked image modeling loss:

$$\mathcal{L}_{con\_pre} = \mathcal{L}_{match} + \lambda_1 \mathcal{L}_{recon} \tag{7}$$

where λ represents a hyperparameter controlling the trade-off between the two objectives.

3.3 Generative Pre-Training

As depicted in Figure 3, the generative pre-training stage's goal is to learn a generative model $P(R|C, I_{1:k})$ skilled in generating coherent and contextually appropriate responses when given a dialogue context C and an associated set of images $I_{1:k}$. To accomplish this, we use a multimodal fusion module and an information transfer module. These modules are trained on a corpus of annotated context-response pairs, where the relevant images are obtained utilizing the TIM module. During training, the context and response are concatenated to obtain the relevant images, whereas, during testing, only the context is used.

3.3.1 Multimodal Fusion Module. The multimodal fusion (MF) module takes the dialogue context hidden state $\mathbf{h}_C$ and the top k image hidden states $\mathbf{h}_{I_1}, \ldots, \mathbf{h}_{I_k}$ as input, striving to effectively fuse multimodal information. Before achieving this, we initially apply an objective to ensure the encoders preserve the information contained in the TIM module. This objective aligns $\mathbf{h}_C$ and $\mathbf{h}_{I_1}, \ldots, \mathbf{h}_{I_k}$ using binary cross-entropy loss:

$$\mathcal{L}_{preser} = -\frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[y_{ij} \log \sigma(s_{ij}) + (1 - y_{ij}) \log(1 - \sigma(s_{ij})) \right] \tag{8}$$

where n represents the number of dialogue contexts, m represents the number of images, y_{ij} is an element with $y_{ij} = 1$ if the ith context and the jth image form a matched pair, and $y_{ij} = 0$ otherwise, $\sigma(\cdot)$ is a logistic sigmoid function.

Then, we employ an attention mechanism to weigh the significance of each image representation concerning the dialogue context and perform the weighted sum over the image hidden state:

$$\alpha_i = \frac{\exp(s_{I_iC})}{\sum_{j=1}^{k} \exp(s_{I_jC})}, \quad \mathbf{h}_{I_{att}} = \sum_{i=1}^{k} \alpha_i \mathbf{h}_{I_i} \tag{9}$$

This makes the fused image hidden state contain the beneficial information of k images, while ignoring their noise information.

Next, we use two transformer blocks to fuse text and image hidden states separately:

$$\mathbf{h}_{I_C} = TB_I(\mathbf{h}_{I_{att}}, \mathbf{h}_C, \mathbf{h}_C) \tag{10}$$

$$\mathbf{h}_{C_I} = TB_C(\mathbf{h}_C, \mathbf{h}_{I_{att}}, \mathbf{h}_{I_{att}}) \tag{11}$$
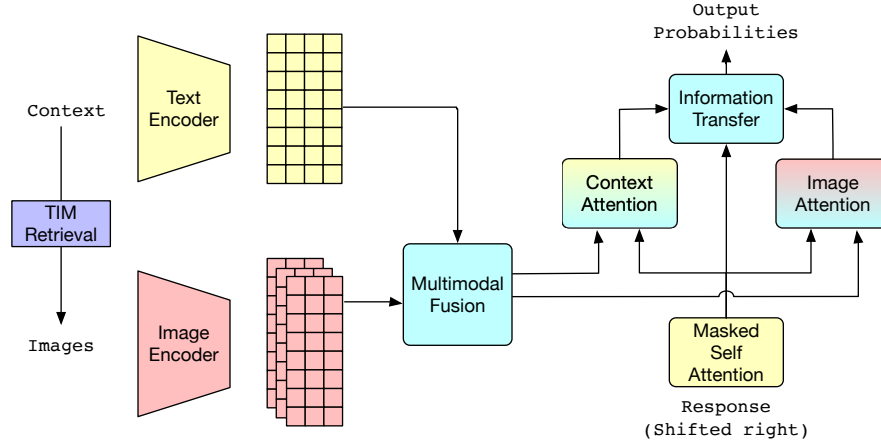
Figure 3: Overview of the generative pre-training stage.

3.3.2 Information Transfer Module. When given the first $t-1$ response tokens $R_{1:t-1}$, the information transfer (IT) module predicts the probability of the tth token based on the fused dialogue context hidden state $\mathbf{h}_{C_I}$ and the fused image hidden state $\mathbf{h}_{I_C}$. This module is initialized with a pre-trained BART decoder, and the response hidden state $\mathbf{h}_R$ can be obtained from $R_{1:t-1}$ through its masked self-attention module. We then incorporate $\mathbf{h}_{C_I}$ and $\mathbf{h}_{I_C}$ into $R_{1:t-1}$ by the cross-attention module of the BART decoder to obtain $\mathbf{h}_{CR}$ and $\mathbf{h}_{IR}$, respectively. Next, an information transfer mechanism is employed to integrate all the information to obtain the predicted representation of $\hat{\mathbf{h}}_R$:

$$\gamma_* = \sigma(\mathbf{W}_*[\mathbf{h}_R; \mathbf{h}_{*R}] + b_*) \tag{12}$$

$$\hat{\mathbf{h}}_R = FFN(\mathbf{W}_R[\mathbf{h}_R; \gamma_C \mathbf{h}_{CR}; \gamma_I \mathbf{h}_{IR}] + b_R) \tag{13}$$

where $*$ here is denoted as C or I, $\mathbf{W}_*$ and b_* are learnable parameters, and FFN is a position-wise feed-forward network. The information transfer mechanism selectively incorporates the fused dialogue context and image information into the response representation based on the context and image attentions γ_C and γ_I, respectively.

Finally, the probability distribution of the tth token is obtained by applying a linear transformation followed by a softmax activation function $\eta(\cdot)$ to the corresponding $\hat{\mathbf{h}}_R$ at position $t-1$:

$$P(R_t|R_{1:t-1}, C, I_{1:k}) = \eta(\mathbf{W}_P \hat{\mathbf{h}}_{R_{t-1}} + b_P) \tag{14}$$

where the weight of $\mathbf{W}_P$ comes from the word embedding. The primary training objective of $P(R|C, I_{1:k})$ is to minimize the negative log-likelihood between the predicted probability distribution and the ground-truth response tokens:

$$\begin{aligned} \mathcal{L}_{gen} &= -\log P(R|C, I_{1:k}) \\ &= -\sum_t \log P(R_t|R_{1:t-1}, C, I_{1:k}) \end{aligned} \tag{15}$$

The overall loss function for the generative pre-training stage is a linear combination of the generative loss and the information preservation loss:

$$\mathcal{L}_{gen\_pre} = \mathcal{L}_{gen} + \lambda_2 \mathcal{L}_{preser} \tag{16}$$

3.4 Zero-Resource Learning Detail

3.4.1 Training. During the zero-resource training stage, we initially train the contrastive pre-trained text-image matching module according to Equation (7), and subsequently train the generative pre-trained dialogue model as per with Equation (16). ZRIGF is trained without explicit context-image-response pairs, as it leverages the learned representations of the individual components to enable effective dialogue generation.

Throughout the generative pre-training stage, the encoder undergoes alterations with each training iteration, leading to corresponding modifications in the TIM module. This results in a complex process of image retrieval based on the evolving TIM module. To enhance efficiency, we uniformly retrieve all relevant images after the contrastive pre-training stage.

3.4.2 Inference. For fully zero-resource scenarios in new domains, we simply use the pre-trained ZRIGF without any additional fine-tuning. During the zero-resource inference stage, the model first selects the top k most relevant images $I_{1:k}$ for a given dialogue context C using the TIM module $P(I_{1:k}|C)$ according to Equation (2). Then, the generative model $P(R|C, I_{1:k})$ is utilized to generate a contextually appropriate response R according to Equation (14). As the model has been pre-trained to effectively fuse multimodal information and transfer this information into the generated response, it can generate coherent and contextually appropriate responses that are grounded in the provided images.

4 EXPERIMENTS

4.1 Datasets

The Microsoft COCO 2017 dataset [19] is utilized for contrastive pre-training to learn the text-image matching module and the text-assisted masked image modeling module. This dataset is segmented into 591K training samples and 25K validation samples. It comprises 123,287 images, each accompanied by five captions, providing an abundant source of image-text pairs for training both the TIM module and the TAMIM module.

The Reddit Conversation dataset [8] is leveraged for generative pre-training, in which we train our model to generate contextually appropriate responses given a dialogue context. This dataset comprises over 15 million dialogues extracted from the Reddit platform, providing a large-scale and diverse collection of context-response pairs. The version used in this study, as released by [35], has been preprocessed to retain 1M dialogues for training, 20K dialogues for validation, and 20K dialogues for testing. During the generative pre-training, we utilize the TIM module learned from the contrastive pre-training stage to obtain the relevant images for the given dialogue contexts and responses. In addition to the images found in the COCO dataset, an extra 500K images from the Open Images dataset [12] are employed as additional image indices.

The Image-Chat dataset [30] is used for evaluating the performance of ZRIGF in zero-resource scenarios. This dataset is divided into 186,782 training examples, 5,000 validation examples, and 9,997 testing examples. It is specifically constructed for image-grounded dialogue generation tasks and comprises over 100,000 human-human dialogues grounded in images, with each dialogue consisting of a context, a relevant image, and a response. In the fully zero-resource scenario, we use only its test set for validation.

4.2 Implementation Details

We implement our ZRIGF [1] based on the Hugging Face Transformers library [34]. For the image encoder E_I, we initialize it with the ViT-Base model and use a patch size of 16. The text encoder E_T and decoder are initialized with the BART-Large model. The image and text dimensions are transformed by linear transformations. We employ the AdamW optimizer [21] with a learning rate of $2e^{-5}$ for both contrastive and generative pre-training stages.

During the contrastive pre-training stage, the text encoder and decoder are frozen, and a batch size of 128 is employed. We train the model for 20 epochs, with a weight decay of 0.05 and a linear learning rate warm-up over the first 10% of the total steps. We set the temperature parameter τ in Equation (3) to 0.07, and the trade-off hyperparameter λ_1 in Equation (7) to 0.2. For the TAMIM module, we use a random masking strategy with a patch size of 32×32 and a mask ratio of 0.4.

For the generative pre-training stage, we use a batch size of 32 and train the model for 10 epochs with a label smoothing factor of 0.1. In Equation (16), the hyperparameter λ_2 is set to 0.1. The number of top relevant images k is set to 3. Response tokens are generated using beam search with a beam size of 3.

4.3 Baseline Models

To evaluate the performance of our ZRIGF, we compare it against the following baseline models: (1) **Seq2Seq** [32], which is a standard sequence-to-sequence architecture consisting of an encoder and a decoder with long short-term memory (LSTM) cells; (2) **BART** [13], specifically the large BART model, which is a pre-training sequence-to-sequence model that employs a Transformer architecture and has demonstrated strong performance in various natural language generation tasks; (3) **ImgVAE** [35], which is an image-grounded dialogue generation model that leverages a variational autoencoder to incorporate visual information; (4) **Maria** [16], which is a neural

[1]The code is available at https://github.com/zhangbo-nlp/ZRIGF.

conversation agent that can generate responses based on visual world experiences retrieved from a large-scale image index; and (5) **ChatGPT** [24], which is a state-of-the-art conversational AI model based on the GPT-3.5 and GPT-4 series of Large Language Models (LLMs), trained to follow an instruction in a prompt and provide a detailed response. Among them, ImgVAE and Maria are multimodal models, while all others are unimodal text models.

We have implemented all methods with the exception of ImgVAE, for which we directly utilized the results from [35]. In the case of ChatGPT, we used GPT-3.5 and employed the following prompt to predict the conversation:

> As a dialogue model, I would like you to predict the conversation, utilizing </s> as the conversation separator. Please provide the response only.

4.4 Evaluation Metrics

4.4.1 Automatic Metrics. To evaluate the performance of our proposed ZRIGF, we use several widely adopted automatic evaluation metrics. **Perplexity** (PPL) is used to assess the high-level general quality of the generation model, with a relatively lower Perplexity value indicating a more fluent response. Traditional metrics such as **BLEU-1** [25] and **ROUGE-L** [18] are used to compare the generated responses with reference human-generated responses, considering aspects such as 1-gram overlap and the longest common subsequence. Embedding-based metrics [20] such as **Average**, **Extrema** and **Greedy** are used to compute the word embedding similarity between the generated response and the ground truth, based on cosine similarities. Diversity metrics [14] such as **Dis-1** and **Dis-2** are used to report the degree of diversity in generated responses, measuring the ratio of unique uni/bi-grams over all uni/bi-grams.

4.4.2 Human Evaluation. For a more comprehensive evaluation, we conduct a human evaluation to assess the quality of the generated responses from different aspects. Specifically, we invite three human evaluators to rate 100 randomly generated responses in the test set on a scale from 1 to 5 in terms of **relevance**, **informativeness**, and **fluency**. Relevance evaluates how well the generated response relates to the given context and image, informativeness measures the amount of new and useful information provided by the response, and fluency assesses the grammaticality and readability of the generated text. The mean ratings are adopted as measures, and Fleiss' **Kappa** [10] statistic is applied to evaluate the concordance among evaluators.

5 RESULTS AND DISCUSSION

5.1 Automatic Evaluation Results

Table 1 presents the results of the automatic evaluation metrics for the proposed ZRIGF and the baseline models on the Reddit Conversation and Image-Chat datasets. For the Reddit Conversation dataset, we observe that our ZRIGF outperforms all the baseline models in terms of PPL, BLEU-1, ROUGE-L, Extrema, Greedy, and Dis-1 metrics, demonstrating the effectiveness of the proposed multimodal architecture and the learning strategy in generating contextually appropriate responses. ChatGPT outperforms ZRIGF on the Average and Dis-2 metrics. This is because ChatGPT is based on LLM, which enables it to generate more diverse responses.

Table 1: Assessment of automated metrics: † denotes a zero-resource scenario without annotated images, while ‡ indicates a fully zero-resource scenario. Bold font highlights the best performance in each column.

Task	Methods	PPL	BLEU-1	ROUGE-L	Average	Extrema	Greedy	Dis-1	Dis-2
Reddit Conversation	Seq2Seq	77.27	12.21	10.81	78.38	40.06	62.64	0.53	1.96
	BART	44.73	13.51	12.50	80.21	41.63	63.72	4.17	16.98
	ChatGPT‡	-	11.62	11.29	**82.39**	37.48	62.05	5.28	**38.63**
	ImgVAE	72.06	12.58	12.05	79.95	42.38	63.55	1.52	6.34
	Maria	56.23	14.10	12.66	81.76	43.04	63.98	4.83	22.87
	ZRIGF	**36.21**	**16.06**	**14.51**	82.27	**43.79**	**64.53**	5.79	26.57
Image-Chat	Seq2Seq†	50.82	11.34	13.65	82.95	47.45	65.67	1.28	7.80
	BART†	37.26	13.41	14.24	84.48	48.57	66.49	2.44	15.79
	ChatGPT‡	-	10.77	11.62	86.17	43.02	64.66	**5.32**	37.77
	ImgVAE	41.94	16.07	15.98	85.81	49.59	67.44	1.68	7.22
	Maria†	37.49	14.74	14.59	85.72	50.58	66.89	2.57	11.99
	Maria‡	135.49	11.75	12.13	83.51	45.57	64.48	1.89	7.32
	ZRIGF	**29.82**	16.86	**17.21**	**86.30**	**51.41**	**68.56**	2.59	10.62
	ZRIGF†	30.58	**17.29**	17.03	86.00	50.99	68.05	4.28	24.10
	1/4 Data†	35.41	16.35	16.59	85.75	49.95	67.20	4.61	22.66
	1/8 Data†	40.07	16.01	16.08	84.91	48.74	66.32	4.31	19.17
	Zero Data‡	105.12	15.17	15.13	84.52	45.95	65.70	5.25	29.38

For the Image-Chat dataset, we evaluate our ZRIGF in three scenarios: a scenario with annotated images, a zero-resource scenario without annotated images (†), and a fully zero-resource scenario without any training data (‡). We also report the results of our ZRIGF when trained with 1/4 and 1/8 of the training data to examine its robustness under data scarcity. We can see that ZRIGF† achieves the best performance on most metrics compared to baseline models, indicating its effectiveness in generating informative and coherent responses that are grounded on images. The results also show that ZRIGF can handle the zero-resource scenario well, as ZRIGF† performs competitively with ZRIGF and its diversity has been significantly enhanced. Notably, even when trained with only 1/4 of the available training data in the second scenario, ZRIGF still surpasses ImgVAE, which used all available training data. Moreover, ZRIGF shows remarkable robustness and generalization ability in different data settings, as its performance does not degrade significantly when reducing the amount of training data or using zero data. In contrast, Maria suffers from a severe performance drop when switching from full training data to zero training data. ChatGPT performs well in terms of Dis-1 and Dis-2 metrics, but it lags behind ZRIGF‡ in terms of BLEU-1 and ROUGE-L metrics, suggesting that it may generate responses that are diverse but not specific or relevant to the given context.

5.2 Human Evaluation Results

Table 2 shows the results of the human evaluation for the Image-Chat task in a fully zero-resource scenario. We compare ZRIGF with Maria and ChatGPT models. We can see that ZRIGF obtains the highest rating in terms of relevance, indicating that it can generate responses that are closely related to the given context and image.

Table 2: Human evaluation outcomes for the Image-Chat dataset in a fully zero-resource scenario.

Methods	Relevance	Informativeness	Fluency	Kappa
Maria	2.26	1.56	3.37	0.41
ChatGPT	3.38	**3.02**	**4.46**	0.48
ZRIGF	**3.74**	2.97	4.23	0.45

ChatGPT achieves the highest rating in terms of informativeness and fluency, suggesting that it can generate responses that are rich in content and grammatically correct. However, ChatGPT also receives a lower rating in terms of relevance than ZRIGF, implying that it may generate responses that are informative but not specific or consistent with the given context and image. It is possible that ChatGPT is based on a text-only LLM, which does not incorporate multimodal information. ZRIGF achieves results similar to ChatGPT in terms of informativeness, thanks to the incorporation of multimodal information. Maria obtains the lowest rating in all aspects, reflecting its poor performance in a fully zero-resource scenario.

5.3 Ablation Study

To analyze the impact of each module in ZRIGF, we conduct an ablation study by removing one module at a time and evaluating the performance on both Reddit Conversation and Image-Chat datasets. Table 3 shows the results of the ablation study in terms of BLEU-1 and ROUGE-L metrics. We observe that removing any module leads to a decrease in performance, indicating that each module contributes to the overall quality of the generated responses. Among the four modules, we find that removing TAMIM causes the

Context	A: The sky has my attention. B: It's beautiful but terrifying to imagine what is in the sky.

Human — This is a cool looking sky.

BART — I think it 's a beautiful sky.

ChatGPT — What's making it so terrifying for you?

Maria — It's beautiful, but the sky is nt really hot.

ZRIGF

I think it's a meteor coming to take me to the heavens!

Figure 4: Case study on Image-chat test set in a fully zero-resource scenario. Maria and ZRIGF retrieve relevant images based on context and generate responses accordingly.

Table 3: Ablation study.

Methods	Reddit Conversation		Image-Chat (Zero Data)	
	BLEU-1	ROUGE-L	BLEU-1	ROUGE-L
ZRIGF	**16.06**	14.51	**15.17**	**15.13**
-TIM	15.19	13.95	14.77	15.02
-TAMIM	13.68	11.33	12.33	13.31
-MF	15.46	**14.56**	14.90	14.89
-IT	15.30	13.89	14.17	14.50

most significant drop in performance, suggesting that TAMIM plays a vital role in preserving and realigning the multimodal features for dialogue generation. Removing IT also results in a noticeable decline in performance, especially in the fully zero-resource scenario, which implies that IT is critical to the generalization ability of ZRIGF. TIM and MF are relatively less essential but still beneficial for generating insightful responses based on multimodal representations.

We further explored the impact of the relevance of retrieved images on the generation performance, detailed in Appendix A.

5.4 Case Study

To further illustrate the effectiveness of the proposed ZRIGF, we conduct a case study by analyzing the responses generated in Figure 4. We compare our model to baseline models on the Image-Chat test set in a fully zero-resource scenario. As can be seen from the figure, ZRIGF effectively utilizes the visual information from retrieved images, generating more detailed and engaging responses. ChatGPT,

benefiting from a powerful LLM, produces fluent and contextually relevant responses; however, its inability to incorporate visual information results in less engaging responses compared to ZRIGF. Maria generates a response somewhat related to the retrieved image, but it lacks relevance and accuracy. This case study demonstrates that ZRIGF successfully leverages the integrated multimodal information to generate more image-grounded and contextually appropriate responses.

6 CONCLUSION

In this paper, we have proposed ZRIGF, an innovative multimodal framework for zero-resource image-grounded dialogue generation. ZRIGF consists of four modules: a text-image matching module, a text-assisted masked image modeling module, a multimodal fusion module, and an information transfer module. ZRIGF implements a two-stage learning strategy that combines contrastive pre-training and generative pre-training to align multimodal features and generate insightful responses. We have evaluated ZRIGF on two publicly available dialogue datasets and demonstrated its effectiveness and robustness in generating cohesive and contextually pertinent responses that are grounded on images. We have also shown that ZRIGF can generalize well in the fully zero-resource scenario without requiring any training data. Our work contributes to the advancement of image-grounded dialogue systems and opens up new possibilities for multimodal dialogue generation in zero-resource scenarios.

ACKNOWLEDGMENTS

REFERENCES

[1] Feilong Chen, Duzhen Zhang, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo XU. 2022. Unsupervised and Pseudo-Supervised Vision-Language Alignment in Visual Dialog. In *Proceedings of the 30th ACM International Conference on Multimedia.* 4142–4153.

[2] Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1925–1935.

[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[4] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Advances in Neural Information Processing Systems*, Vol. 32.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations.*

[6] Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The Zero Resource Speech Challenge 2021: Spoken language modelling. In *Interspeech 2021-Conference of the International Speech Communication Association.*

[7] Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units. (2020).

[8] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting Neural Response Generation with Context-Aware Topical Attention. In *Proceedings of the First Workshop on NLP for Conversational AI.* 18–31.

[9] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 268–277.

[10] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[11] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional Dialogue Generation Using Image-Grounded Language Models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–12.

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 7871–7880.

[14] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 110–119.

[15] Linxiao Li, Can Xu, Wei Wu, YUFAN ZHAO, Xueliang Zhao, and Chongyang Tao. 2020. Zero-Resource Knowledge-Grounded Dialogue Generation. In *Advances in Neural Information Processing Systems*, Vol. 33. 8475–8485.

[16] Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A Visual Experience Powered Conversational Agent. 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).* 5596–5611.

[17] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 675–684.

[18] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out.* 74–81.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision − ECCV 2014.* 740–755.

[20] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2122–2132.

[21] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations.*

[22] Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 522–532.

[23] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 462–472.

[24] OpenAI. 2022. *ChatGPT: Optimizing Language Models for Dialogue.* Retrieved November 30, 2022 from https://openai.com/blog/chatgpt/

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* 311–318.

[26] 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*, Vol. 139. 8748–8763.

[27] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. (2018). arXiv:1801.03604

[28] 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.* 4287–4296.

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1874–1883.

[30] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-Chat: Engaging Grounded Conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 2414–2429.

[31] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. [n. d.]. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* 38–45.

[35] Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open Domain Dialogue Generation with Latent Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14239–14247.

[36] Bo Zhang, Jian Wang, Hongfei Lin, Hui Ma, and Bo Xu. 2022. Exploiting Pairwise Mutual Information for Knowledge-Grounded Dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2231–2240.

[37] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* 270–278.

[38] Xueliang Zhao, Tingchen Fu, Chongyang Tao, Wei Wu, Dongyan Zhao, and Rui Yan. 2022. Learning to Express in Knowledge-Grounded Conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2258–2273.

[39] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

A ANALYSIS OF IMAGE RELEVANCE

A.1 Impact of irrelevant Images

Table 4: Performance comparison with the incorporation of irrelevant images.

Methods	Reddit Conversation		Image-Chat (Zero Data)	
	BLEU-1	ROUGE-L	BLEU-1	ROUGE-L
ZRIGF	**16.06**	**14.51**	**15.17**	**15.13**
BART	13.51	12.50	13.41	14.24
Random	12.11	11.47	12.32	13.60

The impact of irrelevant images on generation performance is a critical factor in understanding the overall behavior and dependency of the ZRIGF model on visual information. To investigate this, we randomly replaced the top k relevant images with irrelevant ones taken from the Open Images dataset. Table 4 demonstrates the comparison between the performance of the ZRIGF model, BART, and a Random selection approach. The BART model can be considered as the case where no images are used. It can be observed that the BLEU-1 and ROUGE-L scores drop markedly for ZRIGF when irrelevant images are used, compared to the BART model, where no images are incorporated.

The results underline the ZRIGF's strong reliance on the quality of retrieved images, showing that the presence of irrelevant images has negative effects on response generation. This can be attributed to the model's design, which presumably leverages the visual context in generating responses. Therefore, the incorporation of irrelevant images likely confuses the model, leading to less

coherent and contextually accurate output. This also demonstrates the model's ability to retrieve relevant images.

A.2 Impact of Image Quantity

Table 5: Performance comparison with various numbers of images.

Methods	Reddit Conversation		Image-Chat (Zero Data)	
	BLEU-1	ROUGE-L	BLEU-1	ROUGE-L
k=1	15.54	13.72	13.74	14.99
k=3	**16.06**	**14.51**	**15.17**	**15.13**
k=5	15.77	14.49	15.04	15.02

The role of hyper-parameter k, which controls the number of retrieved images, is a key aspect to consider in understanding the ZRIGF model's behavior. We conducted a series of experiments to study how varying the value of k affects the generation performance. Specifically, we set k to 1, 3, and 5 and analyzed the model's output with different quantities of retrieved images. Table 5 illustrates the performance of ZRIGF for different values of k. The best results were achieved when $k = 3$, indicating an optimal trade-off between retrieving more images and retrieving more relevant images. When k is too small, ZRIGF may miss valuable images that could provide essential information for response generation. On the other hand, if k is too large, the model might retrieve irrelevant or redundant images, leading to ambiguity and potential noise in the response generation process.