Triplet-Free Knowledge-Guided Response Generation

Dongming Li[1,2]*, **Jianfeng Liu**[2], **Baoyuan Wang**[2]†
[1] The Chinese University of Hong Kong, Shenzhen
[2] Xiaobing.AI
dongmingli@link.cuhk.edu.cn
{liujianfeng, wangbaoyuan}@xiaobing.ai

Abstract

Generating vivid and informative responses (e.g., comments for social posts and utterances for dialogues) is challenging without giving relevant knowledge. Prior works focus on constructing the "latent" knowledge first and then learning how to "ground" it based on pseudo (context, knowledge, response) triplets. However, the retrieval between real responses and their latent knowledge is difficult in nature. In this paper, instead of focusing on how to ground knowledge given the responses, we take a different perspective to optimize the final responses for given guided knowledge directly. This allows us to re-formulate the entire problem in a simplified yet more scalable way. Specifically, we pretrain a response language model (LM) to measure the relevance and consistency between any context and response, then use search engines to collect the top-ranked passages to serve as the guiding knowledge without explicitly optimizing the "best" latent knowledge that corresponds to a given response. The final response generation model is trained through reinforcement learning by taking both the response LM prior and knowledge-injection rate as rewards. For better evaluations, we construct a new Chinese benchmark, "IceKC", using fresh multimodal online social posts. Both automatic evaluations and human evaluations show our zero-resource approach performs significantly better than prior works.[1]

1 Introduction

Response generation, including dialogue utterances and post comments, is a testbed for machine intelligence and has many applications. However, previous AI models tend to output generic and bland responses as shown in (Li et al., 2016; Shao et al., 2017), which led to a few recent works that leverage external knowledge to improve the generation quality from both diversity and informativeness perspectives (Zhou et al., 2018; Dziri et al., 2019; Dinan et al., 2018; Moon et al., 2019; Wu et al., 2019; Hayati et al., 2020; Liu et al., 2021b; Komeili et al., 2022). Although impressive progress has been made, major hurdles still exist. Constructing a context-knowledge-response triplet dataset via crowd-sourcing (i.e., Dinan et al. (2018); Komeili et al. (2022)) for training purposes might be too expensive to scale up, which also violates the mainstream paradigm of large-scale self-supervised pretraining (Radford et al., 2021; Jia et al., 2021). As expected, prior works (Li et al., 2019b; Lian et al., 2019; Zhao et al., 2020a; Lin et al., 2020) show that models trained on such manually-constructed triplets cannot be generalized to other languages and unseen domains. More recently, zero-resource methods (Li et al., 2020; Chen et al., 2021b; Liu et al., 2021a) are proposed where knowledge-aware generation is learned on such triplets with either matched or inferred knowledge. However, one critical challenge is that given the response, their corresponding "knowledge" is extremely difficult to

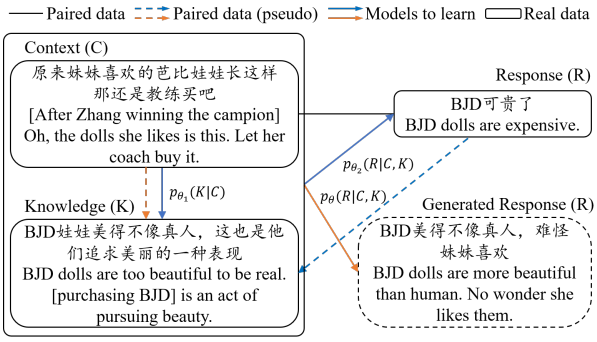


Figure 1: A graphical illustration of different zero-resource approaches for generation with knowledge. Triplet-based approach is highlighted in blue, which retrieves the most likely knowledge for training. Our approach is highlighted in orange, which aims to generate knowledge-guided responses directly.

*Work done during an internship at Xiaobing.AI
†Corresponding Author
[1]Our code and IceKC dataset are publicly available at https://github.com/dongmingli-Ben/triplet-free.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12881–12899
July 9-14, 2023

retrieve from a vast knowledge space, especially when retrieving from Internet search results. As shown in Figure 1, such constructed triplets are very unreliable because the corresponding "knowledge" is scattered over the Internet and the retrieval results are noisy where irrelevant sentences can have high overlap with the response. Forcing the generator to associate the response with such pseudo knowledge is insensible. This raises the question: can we relax the requirement of constructing such triplets for training purposes?

To mitigate those challenges, we introduce a new training methodology that does not require constructing any triplet. Our key idea is inspired by a relaxed casual graphical model where the conditional probability distribution ($p(R|C,K)$) of response (R) given both context (C) and knowledge (K) can be approximated by its lower-bound which only contains two prior, $p(R|C)$ and $p(R|K)$. More specifically, given a context, we use search engines to sample a knowledge passage without explicitly inferring or optimizing, then we use reinforcement learning to optimize the response generation model by jointly considering both prior models as critics to provide reward signals. This allows us to steer the focus from knowledge selection to knowledge-guided generation, which therefore pivots the optimization to explore how to generate informative and engaging responses given both context and knowledge. Without loss of generality, to validate our idea, we leverage a pretrained response language model for $p(R|C)$ and a non-parametric LM for $p(R|K)$ in the current experiments, leaving more advanced prior models for future work. To further encourage more investigation from the community, we construct a benchmark on Chinese multi-modal knowledge-grounded social post commenting, called "IceKC", that facilitates more faithful evaluations. Extensive experiments show that our approach performs significantly better than previous work under a zero-resource setting.

Our contributions are three-fold: (1) We propose a novel zero-resource training strategy for knowledge-guided response generation without the need to build triplets, which is model-agnostic as well as more flexible and easy to scale. (2) We construct a new benchmark from social media posts, which can be used to more faithfully evaluate knowledge-aware multimodal commenting systems. (3) Experiments show that our approach generates significantly better responses than other strong baselines.

2 Approach

Prior works all focus on constructing the "latent" knowledge (K) first given both the context (C) and response (R). However, as we argued before, such triplet-based approaches are challenging due to the difficulty in knowledge retrieval, which results in noisy triplets (C, K, R) and thus hurts the training. In this paper, we take a triplet-free approach by directly feeding sampled knowledge (i.e., search engine) into the response generation model $p(R|C,K;\theta)$, for any given context. As for a fixed context and sampled knowledge pair (C,K), there is no corresponding "ground-truth" response (R) available to apply any supervised learning method. We resolve such unpaired training by leveraging a pretrained response language model that provides distribution-level reward signals.

2.1 Triplet-free Knowledge-guided Learning

Mathematically, we achieve our triplet-free knowledge-guided learning by optimizing the lower-bound of $p(R|C,K)$ inspired by the causal graphical model (Hlaváčková-Schindler et al., 2007; Tuan et al., 2020), where one can derive the following inequalities:

$$p(R|C,K) > p(R|C); p(R|C,K) > p(R|K)$$

hence,

$$\begin{aligned} p(R|C,K) &= p(R|C,K)^{\alpha+1-\alpha} \\ &> p(R|C)^{\alpha} p(R|K)^{1-\alpha}, \alpha \in (0,1) \end{aligned} \tag{1}$$

By further taking the logarithm on both sides of Equation 1, we can get

$$\begin{aligned} \log p(R|C,K) &> \alpha \log p(R|C) \\ &\quad + (1-\alpha) \log p(R|K) \\ &= r_c(C,R) + r_k(K,R) \end{aligned} \tag{2}$$

where $r_c(C,R)$ defines as the reward that measures how consistent and sensible R is for given C; while $r_k(K,R)$ measures how much knowledge is injected from K to R. In principle, they both can be flexibly defined, such as pretrained language models (LM) or adversarial discriminative networks. In our current implementation, we use a pretrained LM to model $r_c(C,R)$ while using a simple non-parametric LM to model $r_k(K,R)$, which will be

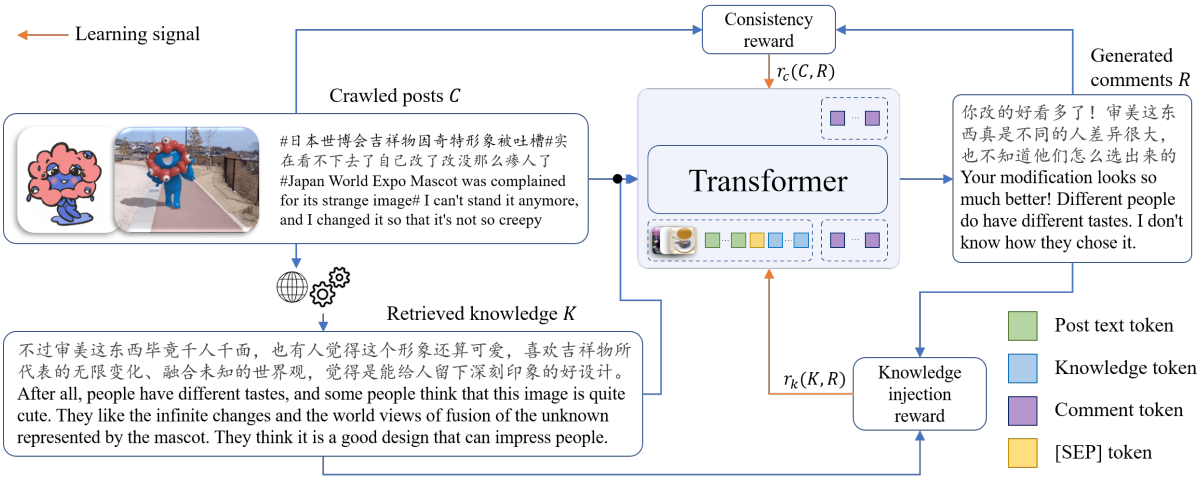


Figure 2: An overview of our approach. The learning objective is to maximize the total reward obtained from two critics: consistency reward from pretrained LM and knowledge injection reward from non-parametric knowledge injection model

discussed in more detail later. Our entire learning objective is simplified as:

$$\max_{\theta} \mathbb{E}_{p(R|C,K;\theta)}[r_c(C,R) + r_k(K,R)] \quad (3)$$

which is trained using two separately constructed corpus: context-response, denoted as $\mathcal{D}^{cr}(C_i, R_i)_{i=1}^{n}$ and context-knowledge, denoted as $\mathcal{D}^{ck}(C_j, K_j)_{j=1}^{m}$, where both n and m represent the number of training samples.

Consistency Reward For the sake of optimization efficiency, we first pre-train an LM on context-response corpus $\mathcal{D}^{cr}$ with unlikelihood objective (Welleck et al., 2019), i.e.

$$\min_{\phi} \mathbb{E}[-\log p(R|C,\phi) + \mathcal{L}_{UL}(R|C,\phi)] \quad (4)$$

where ϕ denotes the parameters of the LM, and $\mathcal{L}_{UL}$ denotes the token-level unlikelihood loss. Once it is pretrained, one can easily use this prior model to estimate $p(R|C,\phi)$. Because language models generally favor short sentences over longer sentences, we further add a length adjustment to the LM to encourage the generation of relatively long responses. Therefore,

$$r_c(C,R) = \log p(R|C,\phi) + r_l(R) \quad (5)$$

$r_l(R)$ is defined as the length of the response under some margin m_l to control the extent of encouragement for longer responses, i.e.

$$r_l(R) = \min(len(R), m_l) \quad (6)$$

Non-parametric Knowledge Injection Reward Different from $p(R|C)$, where context-response corpus can be obtained relatively easily, natural knowledge-response corpus is hard to obtain even manually. One workaround is to design $r_k(K,R)$ following simple heuristics, which is feasible because $p(R|K)$ measures how likely the response R is based on knowledge K, in other words, how much of K is injected in R. In our current implementation, following Li et al. (2020), we use precision-based n-gram matching score BLEU-n (Papineni et al., 2002) with knowledge margin m_k as $r_k(K,R)$, since it may not be wise to incorporate all available knowledge in K. Hence,

$$r_k(K,R) = \min(\text{BLEU\_2}(K,R), m_k) \cdot \alpha \quad (7)$$

where α is a hyper-parameter to balance consistency reward and knowledge injection reward. We choose $n = 2$ based on the heuristics that phrases usually consist of two Chinese characters or English words. Note that, although $p(R|K)$ is not restricted to n-gram-based methods, our empirical experiments show that n-gram based score gives surprisingly better results than embedding-based methods such as BLEURT (Sellam et al., 2020), possibly due to the inherent flaws in the soft alignment of embeddings. We leave further investigations in this direction as future work.

Soft Q-Learning Since Equation 3 is challenging to optimize due to the inherent non-differentiable sampling in the process, we resort to applying soft q-learning (Guo et al., 2021), which turns out to work well compared with other RL optimization methods. Specifically, we optimize

the final knowledge-guided response generation model $p(R|C, K; \theta)$ through Equation 3 over the context-knowledge corpus $\mathcal{D}^{ck}$ after we pretrained $p(R|C, \phi)$. Figure 2 illustrates our approach.

2.2 Training Corpus Construction

We construct both $\mathcal{D}^{cr}$ and $\mathcal{D}^{ck}$ by crawling from the Chinese social post platform, Weibo; while all the knowledge is obtained from search engine. Raw post-comment (i.e., context-response) pairs are cleaned under strict rules to remove noisy contents for LM training. $\mathcal{D}^{ck}$ is then constructed by measuring the confidence of the retrieved results and filtering noisy contents in retrieved passages. Details on constructing post-comment corpus and post-knowledge (i.e., context-knowledge) corpus can be found in Appendix C and D, respectively. We also construct a text-only English dataset from the training set of Wizard of Wikipedia (Dinan et al., 2018). We use a pretrained DialoGPT model for the consistency reward and D^{cr} is not used. For D^{ck}, we use the dialogue and the candidate knowledge sentences of the Wikipedia training set, as the candidate knowledge sentences are retrieved according to the dialogue history. Statistics of the training corpus are in Table 9 in Appendix E.

3 IceKC

To evaluate models on Chinese multimodal social post commenting, we construct a new benchmark testset for evaluation purposes only. To ensure that external knowledge is necessary to generate the comment, we collect testcases from Weibo trending posts[2] starting from February 2022 so that the topics are fresh and not likely to overlap with training data collected previously. Annotators are instructed to choose a post in one of the trending topics, use search engines to search for an appropriate knowledge sentence, and write a sensible comment to the original post based on the information in the knowledge sentence. Although we focus on knowledge-aware comments generation in this work, this benchmark testset can also be used to evaluate other tasks, including query generation and knowledge selection, as shown in Appendix A.

[2]Sina Weibo is the largest Chinese social platform for individuals and trending posts are the posts that are most frequently viewed and discussed by users on Weibo at a specific time. History trending topics are obtained from https://www.weibotop.cn/2.0/.

IceKC	Utt.	Query		Domain	Len.
		Text	Image		
Overall #	997	489	508	81	22.39
Factual #	556	294	262	58	22.08
Opinion #	441	195	246	44	22.77
Text-only #	347	347	0	52	23.72

Table 1: Statistics of our IceKC benchmark. Utt. denotes the number of comments; Query denotes the number of queries, with Text and Image indicating text query or image query; Domain denotes the number of unique domains; and Len. denotes the average length of the comments."Factual", "Opinion", and "Text-only" denotes cases with factual knowledge, with opinion-like knowledge, and without images respectively.

3.1 Post Selection

Annotators are shown a list of Weibo trending topics, which are fresh and updated every second. To stress the role of external knowledge, annotators are instructed to select individuals' posts where the background information is not fully presented in the post. Considering that most posts contain some images along with text content, we instruct annotators to select posts with/without images at a balanced ratio so that the performance of models on text-only posts and multimodal posts can both be evaluated throughout.

3.2 Search Query

After choosing a post, annotators are instructed to use search engines to obtain some background knowledge. Specifically, different from previous work (Zhou et al., 2018; Moghe et al., 2018; Dinan et al., 2018; Komeili et al., 2022), we also consider the rich information conveyed in images. Annotators can either use a text query or choose an image as a query to retrieve relevant documents via baidu search engine[3] or baidu image search engine[4]. To explore the effect of images, annotators are encouraged to use images as queries to find relevant documents. Annotators can try another query if they do not find the retrieved result satisfactory.

3.3 Knowledge Selection and Comment Generation

During benchmark construction, annotators can click on any of the retrieved documents on the front page to expand the document and choose a sentence that they decide to be appropriate to write a comment based on the sentence. Unlike previous work

[3]https://www.baidu.com/

[4]https://image.baidu.com/

focusing on factual knowledge only, such as using Wikipedia as the only knowledge source (Dinan et al., 2018; Li et al., 2020, 2022), we do not restrict the scope of knowledge. However, following Moghe et al. (2018), we ask annotators to distinguish different types of knowledge by either factual (such as encyclopedia) or opinion-like (such as movie reviews and comments) because we find that different strategies work differently for different types of knowledge. For instance, a retrieve-rerank-rewrite (Cao et al., 2018; Wang et al., 2019) approach may perform better on opinion-like knowledge. Annotators are encouraged to balance factual and opinion-like knowledge so that models' performance can be evaluated on different types of knowledge. To stress the role of images in commenting, annotators are encouraged to write comments that echo the post's images whenever appropriate.

3.4 Statistics

The statistics of IceKC can be found in Table 1. As far as we know, this is the first Chinese multimodal benchmark testset that is designed to evaluate knowledge-aware commenting systems. Again, note that this benchmark is only used for testing, regardless of how the training set is constructed.

4 Experiment

4.1 Dataset

We evaluate our approach on both our Chinese IceKC benchmark and two public knowledge-grounded dialogue benchmarks, Wizard of Wikipedia (WoW) (Dinan et al., 2018) and Wizard of Internet (WizInt) (Komeili et al., 2022). Models are provided with the complete dialogue history and the golden knowledge sentence in evaluation.

WoW The WoW dataset contains dialogues between two participants where one of them, the wizard, is provided potentially relevant knowledge sentences retrieved from Wikipedia passages. The testset for evaluation is further divided into seen and unseen according to whether the topic of the testing dialogue has appeared in the training dialogues. We further remove a few turns where the wizard chooses not to use any retrieved knowledge sentences for response. After the removal, there are 4087 turns and 4125 turns for model evaluation in Test Seen and Test Unseen, respectively.

WizInt Similar to WoW, WizInt is constructed with dialogues of a wizard and an apprentice. Different from WoW, the wizard produces internet search queries, selects knowledge sentences across the Internet which are appropriate, and gives a response based on the knowledge sentence to the apprentice. In evaluation, we use solely the test set of the WizInt dataset, which has 1957 turns for evaluation. However, because Internet-retrieved contents are noisy, the golden knowledge sentences are sometimes not accurately cut into sentences. To avoid cases where useful information is scattered among the long paragraph, we remove any turns whose golden knowledge is more than 50 words long. 1041 turns are available for evaluation after all preprocessing.

4.2 Baselines

We report the performance of the following methods for comparison:

- BASE: The context-response LM trained for consistency reward without using any knowledge.
- ZRKGC (Li et al., 2020)[5]: A variational method to learn the relation between response, context, and knowledge from pseudo triplets where pseudo knowledge is inferred from pseudo knowledge pool constructed by n-gram matching, with Unified pre-trained Language Model (UniLM) (Dong et al., 2019).
- UKSDG (Chen et al., 2021b)[6]: An unsupervised method that retrieves the most likely knowledge from candidates first and leverages knowledge distillation to alleviate the noisy labeling problem.
- KAT-TSLF (Liu et al., 2021a)[7]: A three-stage learning framework that retrieves pseudo knowledge from an unlabeled knowledge base and trains the model on such weakly constructed triplets.
- OURS: Our proposed full method.

To shed light on whether zero-shot inference with large-scale pretrained models is sufficient for the task, we evaluated two large Chinese language models, PANGU-α (Zeng et al., 2021) and EVA (Zhou et al., 2021). Since PANGU-α and EVA only support text input, we use the following prompt to perform zero-shot inferences: "{post text}；根

[5] https://github.com/nlpxucan/ZRKGC
[6] https://github.com/ErenChan/UKSDG.
[7] https://github.com/neukg/KAT-TSLF.

Model	PARAM	BL-1	BL-2	BL-3	BL-4	R-1	R-2	R-L	F1	KF1
DIALOGPT$_b$	335M	0.035	0.006	0.001	0.000	0.094	0.014	0.053	0.079	0.035
ZRKGC	154M	0.164	0.046	0.014	0.004	0.203	0.034	0.167	0.167	0.152
UKSDG	174M	0.164	0.090	0.056	0.038	0.327	0.155	0.184	0.299	0.425
KAT-TSLF	198M	0.156	0.073	0.040	0.025	0.294	0.114	0.167	0.264	0.416
OURS	220M	**0.211**	**0.125**	**0.083**	**0.061**	**0.341**	**0.162**	**0.232**	**0.309**	0.505
DIALOGPT$_{ft}$	335M	0.165	0.103	0.070	0.052	0.303	0.151	0.208	0.281	0.561
T5$_{ft}$	220M	0.196	0.132	0.094	0.073	0.341	0.183	0.234	0.317	0.614
HUMAN	-	-	-	-	-	-	-	-	-	0.384

Table 2: Evaluation results on WoW test seen. R-n denotes the ROUGE score using up to n-gram. DIALOGPT$_b$ denotes the pretrained DIALOGPT model used for the consistency reward. $_{ft}$ denotes models finetuned on the WoW train set.

Model	PARAM	BL-1	BL-2	BL-3	BL-4	R-L	F1	KF1
BASE	119M	0.00	0.00	0.00	0.00	0.01	0.01	0
PANGU-α	2.6B	0.11	0.07	0.05	0.03	0.15	**0.32**	0.47
EVA	2.8B	0.04	0.01	0.00	0.00	0.06	0.09	0.00
ZRKGC	147M	0.15	0.06	0.02	0.01	0.14	0.17	0.14
OURS	119M	**0.23**	**0.15**	**0.10**	**0.08**	**0.23**	0.30	0.48
HUMAN	-	-	-	-	-	-	-	0.40

Table 3: The performance on the text-only portion of our IceKC benchmark. BL-n denotes the BLEU score using up to n-gram and R-L denotes the ROUGE-L score. The sizes of models are reported in PARAM.

据下面信息进行评论：{knowledge}；评论：” (“{post text}; comment based on the following information: {knowledge}; comment:”), where {post text} and {knowledge} denotes the text component of posts and knowledge to be grounded on respectively.

4.3 Metrics

For evaluation, following Dinan et al. (2018); Li et al. (2020); Komeili et al. (2022), we adopt F1 score[8]. We also report the BLEU scores (Papineni et al., 2002) up to 4 grams and ROUGE (Lin, 2004) scores. To evaluate models' ability to incorporate knowledge into responses, following previous work (Li et al., 2020; Komeili et al., 2022), we also report the KF1 (knowledge F1) score, i.e., the uni-gram F1 score using knowledge as reference. Although in the ideal case, a high KF1 score indicates that the model is capable of integrating knowledge into its responses, we also report the KF1 score of humans to serve as a base for comparison of the KF1 score.

4.4 Implementation

To utilize the generalizability of pretrained language models and make the reinforcement learning process smooth, we use BART base[9] (Lewis et al., 2020) for Chinese and T5 base[10] (Raffel et al., 2020) for English as the backbone of our models. We set α to be 100 and 200 to balance the scale of $r(C, R)$ and $r(K, R)$ for Chinese corpus and English corpus, respectively. Following Guo et al. (2021), we warm up the training of the generator using off-policy updates on real responses for the first 20k steps and train the model using both off-policy (and on-policy) updates for further steps. To align the reward, we linearly transform it to be centered around 0 and approximately bounded by ± 50, employing two additional hyper-parameters. We use 4 as the batch size and a learning rate of 1e-5 in our experiments. The experiments are conducted on Nvidia V100 and A100 GPUs. For more information on the model architecture for multi-modal context and implementation, please refer to Appendix B.

4.5 Results

We show the performance of different models on our IceKC benchmark in Table 3. Since PANGU-α, EVA, and ZRKGC are proposed for text-to-text generation, for a fair comparison, we compare the performance of the models on a subset of IceKC whose posts contains only text. Evaluation results on Wow and WizInt are in Table 2, 4, and 5. Results on the full WizInt are provided in Appendix F for reference. Human evaluation results are in Section 4.6. Here are some observations.

First, our higher score compared to the low

[8]We implement F1 score based on https://github.com/facebookresearch/ParlAI/blob/main/parlai/core/metrics.py

[9]https://huggingface.co/fnlp/bart-base-chinese

[10]https://huggingface.co/t5-base

Model	PARAM	BL-1	BL-2	BL-3	BL-4	R-1	R-2	R-L	F1	KF1
DIALOGPT$_b$	335M	0.035	0.005	0.001	0.000	0.099	0.014	0.055	0.083	0.039
ZRKGC	154M	0.166	0.044	0.014	0.004	0.204	0.032	0.167	0.166	0.153
UKSDG	174M	0.145	0.074	0.043	0.028	0.293	0.126	0.163	0.267	0.345
KAT-TSLF	198M	0.163	0.078	0.043	0.026	0.302	0.120	0.174	0.272	0.421
OURS	220M	**0.213**	**0.124**	**0.083**	**0.061**	**0.341**	**0.160**	**0.232**	**0.310**	0.494
DIALOGPT$_{ft}$	335M	0.177	0.103	0.070	0.051	0.301	0.149	0.209	0.283	0.568
T5$_{ft}$	220M	0.199	0.134	0.098	0.074	0.341	0.184	0.237	0.318	0.610
HUMAN	-	-	-	-	-	-	-	-	-	0.385

Table 4: Evaluation results of the WoW test unseen.

Model	PARAM	BL-1	BL-2	BL-3	BL-4	R-1	R-2	R-L	F1	KF1
DIALOGPT	335M	0.025	0.004	0.001	0.000	0.069	0.009	0.038	0.057	0.021
ZRKGC	154M	0.116	0.028	0.008	0.002	0.142	0.020	**0.125**	0.117	0.105
UKSDG	174M	0.069	0.015	0.004	0.001	0.165	0.036	0.078	0.139	0.113
KAT-TSLF	198M	0.090	0.023	0.007	0.002	0.202	0.051	0.109	0.176	0.263
OURS	220M	**0.116**	**0.036**	**0.013**	**0.005**	**0.219**	**0.059**	0.124	**0.191**	0.275
DIALOGPT$_{ft}$	335M	0.105	0.039	0.015	0.006	0.226	0.069	0.121	0.203	0.507
T5$_{ft}$	220M	0.099	0.045	0.018	0.009	0.231	0.086	0.128	0.210	0.353
HUMAN	-	-	-	-	-	-	-	-	-	0.243

Table 5: Evaluation results on WizInt. $_{ft}$ denotes models finetuned on the WizInt train set.

zero-shot performance of large pretrained language models indicates the necessity of explicitly injecting knowledge into generation. While PANGU-α achieves a slightly higher KF1 score, our approach considers the consistency between context and response, leading to higher scores when incorporating the same level of knowledge.

Second, our approach outperforms triplet-based models in incorporating knowledge into responses, as shown by the higher KF1 score of our model. Triplet-based models struggle with knowledge incorporation due to loosely associated pseudo knowledge selected based on n-gram overlapping. This often results in overlaps on non-important characters rather than keywords or topic words, even after strict processing. By optimizing the lower bound instead of directly learning $p(R|C, K)$ using pseudo triplets, our approach overcomes this issue and achieves higher scores in BLEU, ROUGE, and F1. Furthermore, an interesting observation from Table 5 is that finetuned DIALOGPT has a much higher KF1 score but a lower F1 score compared to finetuned T5. This suggests that simply copying knowledge does not guarantee better performance. Our approach achieves higher scores by selectively incorporating knowledge based on contextual constraints.

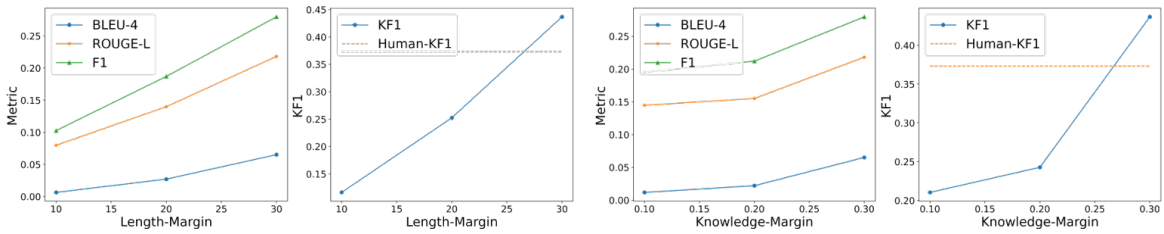
Third, the significant improvement of OURS over the base DIALOGPT model with a smaller parameter size indicates that our training method, rather than model size and initialization, is primarily responsible for the performance gain.

Model	IceKC	BL-1	R-L	F1	KF1
OURS	Overall	**0.213**	**0.218**	**0.280**	0.437
$-\mathcal{I}$		0.209	0.216	0.277	0.439
HU.		-	-	-	0.373
OURS	Factual	**0.207**	**0.209**	**0.269**	0.428
$-\mathcal{I}$		0.203	0.208	0.266	0.429
HU.		-	-	-	0.339
OURS	Opinion	**0.221**	**0.230**	**0.293**	0.448
$-\mathcal{I}$		0.218	0.227	0.292	0.451
HU.		-	-	-	0.417

Table 6: The performance of our model on our IceKC benchmark under different settings. $-\mathcal{I}$ denotes the scores of the model when the images in posts are not fed to the model in inference. The KF1 score of human is reported as HU. in the table.

4.6 Discussions

Impact of Knowledge Type and Post Modality To investigate the impact of different knowledge types and the inclusion of images in the context, we conducted ablation studies, as presented in Table 6. Our findings indicate that our model effectively incorporates more knowledge into generated responses for opinion-like knowledge, as evidenced



(a) Length margin (b) Knowledge margin

Figure 3: (a) Performance of models with different length margins; (b) Performance of models with different knowledge margins. KF1 of human is shown in dotted line for reference.

by the higher KF1 score in that category. This outcome is expected since opinion-like knowledge consists of users' opinions on events, which can be relatively easily transformed into responses. The effectiveness of knowledge incorporation is further supported by the higher BLEU, ROUGE, and F1 scores obtained for opinion-like knowledge. Furthermore, when comparing the scores of models with and without using images in the posts, the higher scores achieved by the model incorporating images suggest that the visual modality enhances the ability to provide comments. Although the difference in scores is not substantial, this could be attributed to the fact that the language model trained for consistency reward may not be highly sensitive to images.

Consistency Reward To examine the impact of consistency reward, we vary the length adjustment margin, where a larger margin implies a greater influence of consistency reward on the model. The results, depicted in Figure 3(a), demonstrate the effectiveness of consistency reward. During our experiments, we observed that a low length margin led to an unstable learning process and resulted in the generation of incoherent and nonsensical comments. Conversely, a higher length margin, associated with higher BLEU, ROUGE, and KF1 scores, indicates that an appropriate margin ensures a more stable learning process. Under such conditions, the model learns to selectively incorporate knowledge more effectively.

Knowledge Reward To study the effect of knowledge injection reward in learning, we vary m_k and report the performance of different models in Figure 3(b). The increasing BLEU and ROUGE scores as the knowledge injection reward margin increases indicate that encouraging the model to copy knowledge, at an appropriate level, can enhance performance. This observation suggests that our models possess the ability to selectively incorporate relevant and appropriate knowledge.

Model	CR			KR			FL		
	μ	κ	p	μ	κ	p	μ	κ	p
ZRKGC	1.1	0.7	0.00	0.6	0.5	0.00	1.3	0.8	0.00
UKSDG	1.4	0.6	0.67	0.8	0.4	0.00	1.8	0.7	0.01
KAT-TSLF	1.4	0.6	0.64	1.3	0.5	0.08	1.6	0.5	0.00
OURS	1.4	0.6	-	1.4	0.6	-	1.9	0.8	-

Table 7: Human evaluation results of WizInt samples. The average score, Cohen's Kappa statistic, and the paired t-test p-values are denoted as "μ", "κ", and "p". The null hypothesis of each p-value is that the mean scores of the corresponding model and OURS are the same.

Human Evaluation We conducted a human evaluation to qualitatively analyze the performances. To this end, we followed the methodology of Li et al. (2020); Liu et al. (2021a) and randomly sampled 100 turns of utterances from WizInt, including the full dialogue history and golden knowledge. The generated responses were then evaluated by human judges on contextual relevance (CR), knowledge relevance (KR), and fluency (FL). The evaluation process was double-blind, and we recruited two well-educated human evaluators who assigned scores in the range of $0, 1, 2$, with 0 representing "bad," 1 representing "mediocre," and 2 representing "good." The results of the human evaluation are presented in Table 7. We observed that responses generated by our triplet-free approach were more knowledge-rich while maintaining a similar level of contextual relevance to the dialogue context. This suggests that our approach enables selective integration of knowledge that aligns with contextual constraints. For more detailed analysis, we provide specific cases in Appendix G for reference.

5 Related Work

Generating vivid responses, such as comments (Zheng et al., 2017; Qin et al., 2018; Li et al.,

2019a; Yang et al., 2019) and dialogue utterances (Huang et al., 2018), is known to be difficult for neural models. To address the problem, previous work attempts to use additional key phrases (Ni and McAuley, 2018), user profiles (Zeng et al., 2019), and images (Chen et al., 2021a) to serve as an external knowledge to be grounded on. More recently, unstructured documents (Moghe et al., 2018; Zhou et al., 2018; Dinan et al., 2018), such as those retrieved from the Internet (Komeili et al., 2022), are prevalent, especially for open-domain contexts. Generally, knowledge-grounded generation is decomposed into two tasks, knowledge selection (Kim et al., 2020; Zhao et al., 2020b) and knowledge-aware generation (Dinan et al., 2018), and additional search query generation for Internet-retrieved knowledge (Komeili et al., 2022). Given knowledge-grounded datasets, supervised approaches (Li et al., 2019b; Lin et al., 2020) have been proposed for knowledge-aware generation.

Several unsupervised approaches (Lian et al., 2019; Li et al., 2020; Chen et al., 2021b; Bai et al., 2021; Liu et al., 2021a) have also been proposed. They rely on retrieving a most likely knowledge sentence for a specific response by either n-gram matching or model inference and leverage such weak triplets to train models. While triplets can be relatively reliable on human-annotated datasets such as WoW (Dinan et al., 2018), it is generally not feasible in Internet-retrieved documents (Komeili et al., 2022) and leads to poor generalization (Chen et al., 2021b) to different domains. Although some approaches (Li et al., 2020; Liu et al., 2021a) build on unreliable automatically constructed triplets and are free from ill generation, they still struggle to incorporate knowledge since knowledge sentences are loosely associated with responses in such triplets.

6 Conclusion

In our study, we examine the effectiveness of previous approaches in utilizing unstructured text for knowledge-grounded response generation in a zero-resource setting. Additionally, we propose a novel approach to tackle this challenging task. Instead of adopting a triplet-based approach, which focuses on identifying the specific knowledge underlying a response, we employ a triplet-free approach that aims to generate coherent and knowledgeable responses given appropriate knowledge. Furthermore, we develop the first benchmark specifically designed for Chinese multimodal knowledge-grounded commenting to evaluate the effectiveness of our approach. Experimental results demonstrate that optimizing the lower bound for $p(R|C, K)$ without relying on triplets as learning signals can achieve superior performance compared to existing triplet-based approaches.

7 Limitations

While we have demonstrated the promising potential of utilizing the lower bound as a substitute for $p(R|C, K)$ in knowledge-grounded generation tasks, there are several limitations that need to be acknowledged. First, the use of the language model (LM) as learning signals can introduce flaws. The model may exploit the LM's weaknesses by generating comments with a high likelihood based on the LM but are nonsensical in reality, resembling adversarial samples. In our experiments, we observed that generating adversarial text samples, unlike vision models, proved challenging, and we did not encounter completely nonsensical comments. However, we did observe the model exploiting the flaws in the LM, indicated by certain common patterns in the generated comments. Second, there are better alternatives to a hard knowledge injection reward, such as an n-gram matching-based BLEU score used in this study. In some cases, a knowledge-grounded comment may not have any word overlaps with the knowledge instances, resulting in a n-gram-based score of 0. Ideally, an embedding-based soft knowledge reward would be more desirable for this reason. However, in our experiments, we found that the soft knowledge reward based on methods like (Kusner et al., 2015; Sellam et al., 2020) was easily exploitable, as the model learned to echo keywords from the context to achieve a high soft knowledge reward. Third, our approach primarily focuses on scenarios where well-constructed triplets are not readily available, such as when retrieving information from the Internet. However, in cases where pseudo knowledge construction is highly accurate, such as applications with more limited scopes, our approach may not outperform triplet-based approaches. Fourth, it is important to note that our method could potentially be used to generate offensive or prejudiced texts. Addressing biases in generative models is a longstanding issue, and it is not the main focus of this work. However, the ethical implications can be partially mitigated by integrating our approach

with other debiasing technologies.

8 Ethical Consideration

In this work, we introduce IceKC, which aims to facilitate Chinese multimodal knowledge-guided evaluation for future research in knowledge-powered generation. To ensure strict adherence to ethical guidelines, we took several measures during the construction of IceKC and the self-constructed training datasets. All data used for training and testing, including user posts, user comments, and Internet search results, are publicly visible. However, user information such as user ID, user name, age, and geographical location is not collected. We also implemented hate comment filtering using swearing word lists.

Data Privacy The construction of the IceKC benchmark underwent approval from an internal review board. We carefully designed the construction and release protocols to avoid any privacy violations of Weibo users. The posts collected from historical trending posts are publicly available, but user information is not included in the dataset. We conducted a rigorous review process to remove any user privacy information found within the posts, such as email addresses and phone numbers. Additionally, toxic and biased texts were eliminated during the review process. To further protect data privacy, IceKC is released under strict terms for academic use only. Users acquiring the data must agree to our requirements for academic use.

Annotators Annotators are Chinese undergraduate students interning at our institution. They are compensated with a monthly salary for their annotation tasks, which include IceKC, as well as other duties assigned by the institution. They are well-informed about the ongoing research and understand that the curated data will be used for research purposes.

References

Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li. 2021. Learning to copy coherent knowledge for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12535–12543.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161.

Kezhen Chen, Qiuyuan Huang, Daniel McDuff, Xiang Gao, Hamid Palangi, Jianfeng Wang, Kenneth Forbus, and Jianfeng Gao. 2021a. Nice: Neural image commenting with empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4456–4472.

Xiuyi Chen, Feilong Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021b. Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1230–1244.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

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. *Advances in Neural Information Processing Systems*, 32.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31.

Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Text generation with efficient (soft) q-learning. *arXiv preprint arXiv:2106.07704*.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. 2007. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46.

Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

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*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B 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*, pages 110–119.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Wei Li, Jingjing Xu, Yancheng He, ShengLi Yan, Yunfang Wu, and Xu Sun. 2019a. Coherent comments generation for chinese articles with a graph-to-sequence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4843–4852.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019b. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021a. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021b. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711.

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*, pages 311–318.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2470–2477.

Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 151–156.

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. *CoRR*, abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Yi-Lin Tuan, Wei Wei, and William Yang Wang. 2020. Knowledge injection into dialogue generation via language models. *arXiv preprint arXiv:2004.14614*.

Kai Wang, Xiaojun Quan, and Rui Wang. 2019. Biset: Bi-directional selective encoding with template for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2153–2162.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5077–5089.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Wenhuan Zeng, Abulikemu Abuduweili, Lei Li, and Pengcheng Yang. 2019. Automatic generation of personalized comment based on user profile. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 229–235.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B 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*, pages 270–278.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2002.10348*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Hai-Tao Zheng, Wei Wang, Wang Chen, and Arun Kumar Sangaiah. 2017. Automatic generation of news comments based on gated attention neural networks. *IEEE Access*, 6:702–710.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

A Additional Information on IceKC Benchmark

A comparison between our IceKC benchmark and other publicly available knowledge-grounded generation benchmarks is shown in Table 8.

B Model Architecture and Implementation Details

To effectively incorporate both text and image in the Chinese multimodal model, we draw inspiration from SimVLM (Wang et al., 2021) and make a

Dataset	Domain	Language	Modal	Session #	Query generation	Knowledge selection	Knowledge-aware generation	Overall performance
Holl-e (Moghe et al., 2018)	movie	en	text	908	×	×	×	✓
CMU_DoG (Zhou et al., 2018)	movie	en	text	630	×	×	×	✓
WoW (Dinan et al., 2018)	open-domain	en	text	1,933	×	✓	✓	✓
WizInt (Komeili et al., 2022)	open-domain	en	text	503	✓	✓	✓	✓
IceKC	open-domain	zh	multi-modal	997	✓	✓	✓	✓

Table 8: Comparison between our proposed benchmark and other existing datasets. The crosses and checkmarks indicate whether the dataset can be used to evaluate a specific aspect of the model.

few modifications to accommodate multiple images in the post. Figure 4 illustrates the architecture of our model. In our approach, we first transform images into patch embeddings and add positional and image ID embeddings to differentiate patch positions and different images. These image embeddings are then concatenated with text embeddings to represent the posts. Based on the finding that 2D spatial positional encoding is not more effective than simple 1D positional encoding (Dosovitskiy et al., 2020), we utilize 1D positional encoding to encode both image and text streams. It's important to note that although the LM and the generator share the same model architecture, they have different inputs during training. The LM takes image features $i_1, ..., i_m$ and text features $t_1, ..., t_n$ as input, while the generator takes $i_1, ..., i_m, t_1, ..., t_n, e_{[SEP]}, k_1, ..., k_l$ as input, where $e_{[SEP]}$ represents the embedding of the [SEP] token.

In the multimodal IceKC test set, images are resized to 224×224 pixels before being fed to the models, and the patch size is set to 16×16. If a post contains more than 9 images, we keep only the first 9 due to the limited capacity of the BART base model. Considering the potentially long sequence of image patches, text tokens, and knowledge sentence tokens, we truncate the sequence to 768 tokens to account for capacity limitations. During generator training, we initialize the generator from the LM to accelerate training speed and facilitate the reinforcement learning process.

C Post Comment Corpus

We crawl a large number of posts and their corresponding comments from Weibo. Due to the inherent noise in online posts and comments, we perform extensive cleaning and filtering on the raw post comment corpus to provide a cleaner corpus for LM training.

Specifically, following DialoGPT (Zhang et al., 2020), EVA (Zhou et al., 2021), and Pchatbot (Qian et al., 2021), we apply the following procedure: (1) Remove posts that has videos, urls or without comments, or is a repost of some other posts; (2) Remove posts that are votes or have other external links; (3) Remove the reply part (e.g. "@xxx") in comments; (4) Remove posts that are too long (i.e. more than 256 Chinese words, using jieba[11] to perform Chinese word segmentation; (5) Normalize duplicated characters to three times, for instance, "哈哈哈哈哈哈" ("hahahahahaha") to "哈哈哈" ("hahaha"); (6) Reduce the times of duplication of the comment under the same post to three; (8) Remove urls in comments; (9) Reduce the times of duplication of a comment in the whole dataset to 10,000; (10) Remove comments that contain 90% of tri-grams that have been seen more than 1,000 times in the whole dataset; (11) Remove non-Chinese comments; (12) Remove comments longer than 50.

D Post Knowledge Corpus

To construct post knowledge pair corpus for model training in the second stage, we first retrieve web results using search engines and then perform extensive cleaning to remove irrelevant results in the noisy online contents.

Knowledge Retriever To gather relevant knowledge for each post, we employed both text and image queries to search the Internet. The entire text content of the post was used as the text query, while all images served as image queries. For text search, we utilized Baidu Search[12], and for image search, we relied on Baidu Image Search[13]. It's worth not-

[11] https://github.com/fxsjy/jieba
[12] https://www.baidu.com/
[13] https://image.baidu.com/

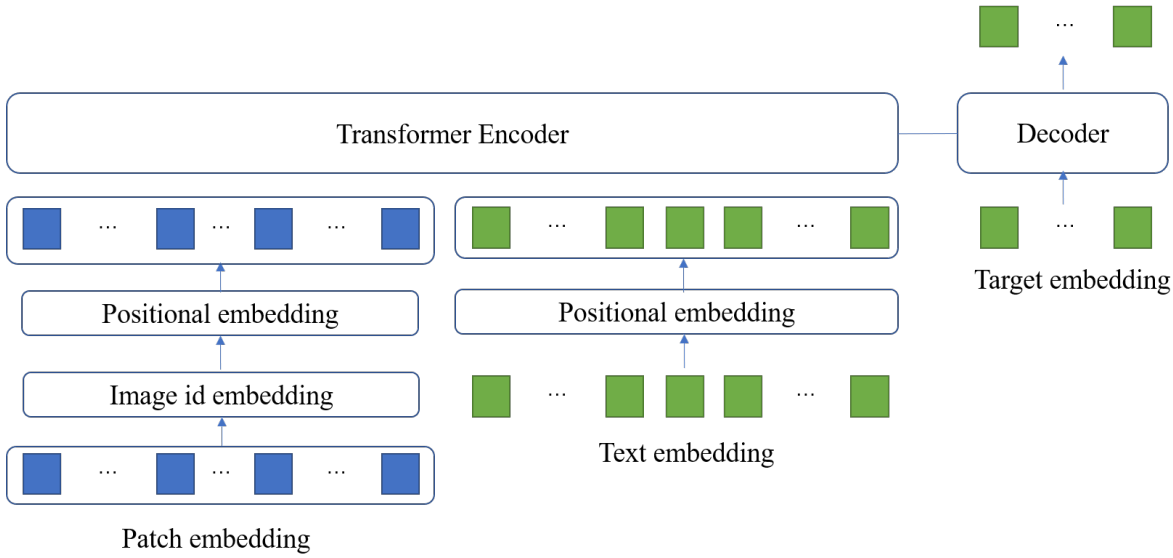


Figure 4: Illustration of the architecture of both our language model and the generator.

ing that Baidu Search truncates text queries longer than 38 Chinese characters, but our dataset's posts typically do not exceed this threshold. To maintain quality, we only retrieved passages from the first page of Baidu Search, considering that search results on the internet tend to be noisy, and subsequent pages usually contain lower-quality content. The retrieved passages were then filtered using specific rules to eliminate merchandise advertisements and other noisy content such as online shopping websites and online novel webpages. Finally, we segmented all the retrieved passages into sentences using HanLP[14].

Semantic Reranking The retrieved sentences from the knowledge retriever remain highly noisy, even after rigorous filtering for advertisements. This is primarily due to the fact that the relevant information often constitutes only a small portion of the original passage. To eliminate irrelevant information within the retrieved sentences, we perform separate reranking for sentences obtained from text and image queries. This reranking process involves calculating the cosine similarity between the TFIDF (term frequency-inverse document frequency) representations of the post text and the retrieved sentences. Reranked knowledge sentences from both text and images are then compared to determine whether the post contains well-known information that can be retrieved. If there is significant semantic overlap between the reranked knowledge sentences obtained from text and images, it suggests that the text and images contain salient information about well-known events. Empirical findings indicate that the results retrieved from image queries are considerably noisier compared to those from text queries. This may be attributed to the fact that ordinary images, such as views of lesser-known mountains in unknown cities, lack relevant knowledge backgrounds. To address this, we assess the confidence of the retrieved results from both text and images and combine them to obtain the final retrieved knowledge instances. Specifically, if both the TFIDF rerank score of knowledge instances from text and the TFIDF rerank score of knowledge instances from images exceed 0.3, it is likely that the post pertains to public events or known content. In such cases, the top 10 knowledge instances based on their scores are retained. Additionally, considering that longer text queries tend to yield more accurate results, we also preserve the retrieved results if the text query exceeds 25 Chinese characters. To eliminate low-relevance knowledge instances and instances with high overlap with the post, possibly because the knowledge instance was retrieved from the same source as the post, we discard all instances with scores lower than 0.2 or higher than 0.6, along with post text shorter than 15 Chinese characters. Empirical observations reveal that, following the aforementioned processing steps, the retrieved knowledge instances are typically related to the context of the post and serve as an external knowledge source.

[14] https://github.com/hankcs/HanLP

Language	Corpus	Context #	Image #	Response #	Knowledge #
Chinese	D^{cr}	875,844	1,658,933	51,305,187	-
	D^{ck}	61,758	145,657	3,025,778	333,991
English	D^{cr}	147,116,725*	-	147,116,725*	-
	D^{ck}	77,518	-	77,518	77,518

Table 9: Statistics for training corpus. Numbers is obtained by counting the turns for English dialogues except that numbers with * represent the number of dialogues. Note that the knowledge and responses in context-response corpus are not paired.

Model	PARAM	BL-1	BL-2	BL-3	BL-4	R-1	R-2	R-L	F1	KF1
DIALOGPT$_b$	335M	0.026	0.004	0.001	0.000	0.069	0.009	0.042	0.058	0.022
ZRKGC	154M	**0.143**	**0.034**	0.009	0.002	0.171	0.022	**0.137**	0.142	0.121
UKSDG	174M	0.071	0.015	0.004	0.001	0.164	0.033	0.080	0.138	0.099
KAT-TSLF	198M	0.090	0.023	0.007	0.002	0.195	0.046	0.095	**0.169**	0.221
OURS	220M	0.109	0.033	**0.012**	**0.005**	**0.197**	**0.048**	0.114	0.167	0.219
DIALOGPT$_{ft}$	335M	0.093	0.037	0.015	0.007	0.216	0.067	0.112	0.185	0.772
T5$_{ft}$	220M	0.096	0.041	0.019	0.009	0.219	0.074	0.126	0.194	0.234
HUMAN	-	-	-	-	-	-	-	-	-	0.194

Table 10: Evaluation results on the full WizInt testset.

E Training Corpus Statistics

Statistics of D^{cr} and D^{ck} is shown in Table 9.

F Additional Evaluation Results

Evaluation results of the models on the full WizInt test set is shown in Table 10.

G Cases

Example cases of different approaches on Chinese multimodal posts are shown in Figure 5. Example cases on English dialogues are at Table 11, 12, 13, and 14.

Figure 5: Cases in the benchmark and the generated comments from different models. The English translation of the post, knowledge, and generated comments are provided under Chinese texts. HUMAN denotes the human written golden comment. Knowledge injected in comments are highlighted.

Dialogue History	A: Flip flops are one of the oldest forms of footwear, we have evidence of them going back to 4000 BCE B: Hey that is pretty cool. I guess I am retro. A: Flip flops became popular in America when soldiers returned home from WWII wearing the Japanese version, zoris. B: Neat I should call them zoris. I like that you can find them in all colors. A: And the materials are usually rubber, foam, vinyl, or webbing, materials that can survive getting wet. B: I like that but they are really hard to walk in in the rain and when they are wet. A: Yes, sometimes they get all squishy and uncomfortable feeling. B: And I think I could turn an ankle yeah you have to start walking very slow its awful.
Knowledge	Foot doctors caution that if you plan on wearing flip-flops, shop for brands that offer arch support, and don t wear them for long periods of time.
Ground Truth	The typically flat bottom doesn't offer much in the way of **arch support**, but some **brands make good arch support flip flops** now.
DIALOGPT$_b$	A:
ZRKGC	A: you don't have to wear flip-flops.
UKSDG	A: i'm not sure, but i know that foot doctors caution that are able to wear a flip flops.
KAT-TLSF	A: and if you're wearing flip-flops, you can turn them.
OURS	A: I don't think you have to wear flip-flops, but I'd suggest you **shop for brands that offer arch support and don' t wear them for long periods**

Table 11: A case from WizInt test set. The response generated by the DIALOGPT$_b$ is empty with top-p decoding. Appropriate knowledge incorporation is bold. Improper incorporated knowledge is highlighted in red.

Dialogue History	A: I trade on the Forex market on a daily basis
Knowledge	Because of the worldwide reach of trade, commerce, and finance, forex markets tend to be the largest and most liquid asset markets in the world.
Ground Truth	**Forex markets are generally have the most liquid assets and are the largest in teh world**.
DIALOGPT$_b$	B: I'm a trader, I'm just not very good at it.
ZRKGC	B: i've been in the us for a year, and it is not sure that you know about this.
UKSDG	B: that's great! i've heard that the forex market is the second most populated major city in the us.
KAT-TLSF	B: forex is a **global market**. the forex is a global market. the forex is a global market.
OURS	B: I'm not a huge trader but I do know that the **forex markets tend to be the largest and most liquid asset markets in the world.**

Table 12: Another case from WizInt test set where the dialogue history is short. Hallucinated knowledge in responses is highlighted in blue.

Dialogue History	A: Tango and cash is an action movie with lot's of comedy
Knowledge	R | 1h 44min | Action, Comedy, Crime | 22 December 1989 (USA)
Ground Truth	It's pretty good, a more mature comedy, definitely deserves it's **R rating**.
DIALOGPT$_b$	B: It's also a very underrated movie.
ZRKGC	B: it's the best action movie of all time.
UKSDG	B: i love tango and cash!
KAT-TLSF	B: tango and cash 1989 imdb r | 1h 44min action, comedy, crime
OURS	B: I'm not sure about that, but I do know that the film was **made in 1989**

Table 13: A case from WizInt test set where the knowledge sentence is not a regular sentence.

Dialogue History	A: Danielle Steel is a great romance novelist. I really enjoyed No Greater Love. B: No Greater Love is a great fictionalized telling of the sinking of the Titanic. A: Yes you are right. It is so good. I also like to read Nora Roberts Irish series. But Danielle is my favorite. B: She has written over 179 books, and is the 4th bestselling author of all time, that's a huge accomplishment. A: Have you read her works? I want to get her latest "Flying Angels" next. B: With over 800 million copies of her books in circulation you think I would've read one, but no, I have not. A: Well you should when you can. LOL. I started reading her books in 1978.
Knowledge	Going Home (1973) Hardcover Paperback Kindle
Ground Truth	**Going Home** was her first book, it was **published in 1973**.
DIALOGPT$_b$	B: .com
ZRKGC	B: i
UKSDG	B: did you read her books?
KAT-TLSF	B: i'm going to start a new kindle in the fall. i'm going to start a new kindle in the fall. i'm going to start a new kindle in the fall.
OURS	B: I'm not sure about you but I do know that I have a **hardcover paperback Kindle Kindle version of Going Home.**

Table 14: Another case from WizInt test set where the knowledge sentence retrieved from Internet search is irregular.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7, Limitations.

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1, Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3, IceKC; section 4.1, 4.2, 4.3

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1, 4.2, 4.3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.1, 4.2, 4.3; section 2.2; appendix A.2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A.2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.4; appendix F.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.4; section 4.1

C ☑ Did you run computational experiments?

Section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Model size and computing infrastructure is reported in tables of section 4 and section 4.4 (Implementation). The total computational budget is not reported because it is difficult to estimate all GPU hours used to do the experiments.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Hyperparameters are reported in section 4.4 (Implementation). With the proposed approach, a simple few experiments give good looking results already. Therefore extensive hyperparameter search is not included.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
All reported scores in the tables are models from one single run. Because training a generative model is expensive and takes days. We do not re-train models.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3, Appendix H

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The instructions are in Chinese for the IceKC annotation. For human evaluation on WizInt, instructions are relatively simple and are reported in text.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.2, H

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A.2, H

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix A.2, H

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A.2, H