

Article

Knowledge Interpolated Conditional Variational Auto-Encoder for Knowledge Grounded Dialogues

Xingwei Liang ^{1,2,†} , Jiachen Du ^{2,†}, Taiyu Niu ², Lanjun Zhou ^{2,*} and Ruifeng Xu ^{2,*} 

¹ Konka Group Co., Ltd., Shenzhen 518053, China

² Harbin Institute of Technology, Shenzhen 518055, China

* Correspondence: bluejade.zhou@gmail.com (L.Z.); xurui Feng@hit.edu.cn (R.X.)

† These authors contributed equally to this work.

Abstract: In the Knowledge Grounded Dialogue (KGD) generation, the explicit modeling of instance-variety of knowledge specificity and its seamless fusion with the dialogue context remains challenging. This paper presents an innovative approach, the Knowledge Interpolated conditional Variational auto-encoder (KIV), to address these issues. In particular, KIV introduces a novel interpolation mechanism to fuse two latent variables: independently encoding dialogue context and grounded knowledge. This distinct fusion of context and knowledge in the semantic space enables the interpolated latent variable to guide the decoder toward generating more contextually rich and engaging responses. We further explore deterministic and probabilistic methodologies to ascertain the interpolation weight, capturing the level of knowledge specificity. Comprehensive empirical analysis conducted on the Wizard-of-Wikipedia and Holl-E datasets verifies that the responses generated by our model performs better than strong baselines, with notable performance improvements observed in both automatic metrics and manual evaluation.

Keywords: Knowledge Grounded Dialogue (KGD); Knowledge Interpolated conditional Variational auto-encoder (KIV); interpolation of latent variables; Conditional Variational auto-encoder (CAVE)



Citation: Liang, X.; Du, J.; Niu, T.; Zhou, L.; Xu, R. Knowledge Interpolated Conditional Variational Auto-Encoder for Knowledge Grounded Dialogues. *Appl. Sci.* **2023**, *13*, 8707. <https://doi.org/10.3390/app13158707>

Academic Editor: Rocco Zaccagnino

Received: 2 June 2023

Revised: 20 July 2023

Accepted: 25 July 2023

Published: 28 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

End-to-end neural conversation models have shown significant promise, demonstrating remarkable advancements from both academia and industry perspectives [1–6]. Nonetheless, these conventional conversation models often grapple with generating informative and engaging responses due to their limited capacity to retain and leverage background knowledge [7,8]. To surmount this knowledge-absence issue prevalent in existing conversation models, Knowledge Grounded Dialogue (KGD) generation is recently proposed for generating responses by simultaneously referring to both the background knowledge and the dialogue context [7,9–17]. The objective is to enhance dialogue response generation to facilitate engaging and in-depth conversations, while avoiding the inclusion of non-factual information.

Existing methodologies in KGD can be broadly classified into extractive and generative models. Extractive models typically view KGD as a reading comprehension task [18] or a document question-answering challenge [19], employing established models such as BiDAF [20] to address this issue. Nevertheless, these models, primarily focusing on extracting knowledge snippets, need to provide more engaging and natural responses akin to human conversation. To mitigate these limitations, attention has turned towards integrating external factoid knowledge into generative dialogue models [21–24]. Most of these works focus on selecting appropriate knowledge and combining the chosen knowledge during the response generation process via copy mechanisms [10,25] or memory network [26]. However, it is crucial to note that responses exhibit varying degrees of knowledge specificity. Some responses may rely heavily on external knowledge, while

others may predominantly depend on the query due to differing dialogue contexts. Despite this, existing generative models do not explicitly model the instance-variety of knowledge specificity in KGD.

The Conditional Variational Auto-Encoder (CVAE) has emerged as an effective model for integrating information from multiple sources within a latent space, showing promise in the domain of response generation [27]. Previous research has utilized CVAE by incorporating a latent variable that is conditioned on the concatenation of dialogue context and additional knowledge information such as dialog acts [28], persona sentences [29] or even images [30]. However, existing CVAE models encounter the issue of representation entanglement [31]. This issue arises due to the single latent variable in CVAE, making it challenging for the model to learn informative and interpretable representations for context and knowledge simultaneously. This issue can lead to inconsistent and unnatural knowledge-grounded responses by the CVAE models.

In this work, we build upon the strengths of CVAE to address the challenges of KGD. To enable the explicit capture of knowledge specificity by CVAE and to overcome its representation entanglement problem, we present a novel Knowledge Interpolated conditional Variational auto-encoder (KIV). This innovative approach introduces two separate latent variables to model dialogue context and external knowledge independently. Unlike prior models that concatenate multiple latent information sources, our approach uses linear interpolation to seamlessly integrate the latent variables associated with context and knowledge, where the interpolation weight corresponds to the level of knowledge specificity. In addition to proposing a deterministic method for obtaining the interpolation weight, we explore a probabilistic interpolation method. This approach views the interpolation weight as a latent variable and models it using a Logistic-Normal distribution. Our proposed probabilistic interpolation method empowers the model to robustly emulate the process of adaptively leveraging background knowledge in response generation.

We evaluate the effectiveness of our proposed model on the Wizard-of-Wikipedia [7] and Holl-E [10] datasets. Both evaluations confirm that our model significantly outperforms the vanilla CVAE and other existing KGD models. Additionally, qualitative analyses demonstrate that the interpolated latent variable successfully controls the knowledge specificity of the generated responses, further offering human-interpretable meaning representations.

In summary, our contributions to the field are as follows:

- We introduce a novel Knowledge Interpolated conditional Variational auto-encoder (KIV) for knowledge grounded dialogue generation. This approach utilizes two distinct latent variables for context and knowledge and fuses them by linear interpolation;
- We explore deterministic and probabilistic methodologies for obtaining the interpolation weight that signifies the level of knowledge utilization;
- A series of extensive experiments are conducted to validate the effectiveness of our proposed model. These experiments further illustrate the interpretability of our interpolation methodologies.

2. Related Work

2.1. Knowledge Grounded Dialogue

Prior research on Knowledge Grounded Dialogue shows that extractive models often generate more suitable responses than their generative counterparts, as indicated by [8,10]. Despite this, it has been noted that generative models can produce more captivating responses that resemble natural human dialogue [21,32]. The primary focus of most KGD generative models is to learn external knowledge representations, often through neural memory networks [7,8] or intricate attention mechanisms [23,33–35]. Ghazvininejad et al. [9] took a unique approach, encoding the dialogue history and documents separately to imbue responses with facts from the external world. Other researchers, including Yang et al. [16], Chen et al. [36], Wang et al. [37], Zhou et al. [38], Li et al. [39], have integrated knowledge graph representation into the response generation process. A few works concentrate on seamlessly integrating external knowledge with dialogue

context. For instance, Li et al. [40] introduce a two-pass decoding strategy for document grounded conversations. Wu et al. [41] defined knowledge identification as finding relevant knowledge in an extensive document that aligns with a user's current query within the conversation context. In their work [15], introduce a KGD model for document-grounded dialogue generation. The model leverages both structured and unstructured knowledge sources to enhance its performance. It incorporates a comprehensive common-sense knowledge network enriched with named entities, along with a domain-specific factual knowledge base. These knowledge sources are utilized to improve the understanding of utterances and generate more informed and contextually appropriate responses. Overall, the proposed model effectively combines multi-source heterogeneous knowledge to enhance document-grounded dialogue generation. Ye et al. [32] jointly employed a CVAE model to represent context and knowledge within a unified latent variable.

Contrasting with previous research, our work enhances knowledge fusion methodologies in two distinct ways: (1) By employing interpolating latent variables, we facilitate a fusion of knowledge and context that is both interpretable and controllable. (2) We innovatively manage the fusion ratio of instance-variety knowledge by using deterministic and probabilistic interpolation weight schemes, allowing for dynamic control over the process.

2.2. Latent Space Interpolation

Latent-space interpolation is a widely adopted technique for evaluating generative latent variable models, typically employed to verify the effective generalization of a generative model [42]. This latent space representation encapsulates all critical information necessary to depict the original data's features. The model learns the data features and simplifies its representation to facilitate easier analysis. This process is integral to Representation Learning [43], a collection of methods designed to enable a system to discern the requisite representations for feature detection from raw data.

Interpolation has traditionally been used to bolster the robustness and effectiveness of representation in supervised learning [44], and to improve semi-supervised learning models [45]. Within text generation, interpolation is commonly employed to demonstrate that generative models can effectively learn smooth latent representations [46,47]. Unlike previous work, Gao et al. [48] incorporated interpolation into their model to promote seamless transitions between two sub-modules. However, the interpolation weight in their model was randomly chosen. In contrast, our model utilizes the interpolation weight as a control variable for knowledge specificity, which needs to be estimated in our model.

3. CVAE for Knowledge-Grounded Dialogue

The task of knowledge-grounded dialogue generation can be framed as follows: Given a dialogue context $C = (C_1, C_2, \dots, C_{|c|})$ comprising $|c|$ turns of conversation, and a segment of knowledge text $k = (k_1, k_2, \dots, k_{|k|})$ containing $|k|$ words, the goal is to generate a response $y = (y_1, y_2, \dots, y_{|y|})$ that aligns with the provided context and is informed by the knowledge text. This is achieved by maximizing the probability $p(y|C, k)$. A practical approach for solving this problem involves using the Conditional Variational Autoencoder (CVAE) framework [27]. This framework approximates the distribution of the random variable y (representing the response) conditioned on c (representing the context) and k (representing the knowledge). This is accomplished by introducing a latent variable z .

The training objective of CVAE can be formulated as maximizing the Evidence Lower Bound (ELBO), as detailed below:

$$\begin{aligned} \mathcal{L}_{\text{CVAE}} = & -\text{KL}(q_{\phi}(z|y, c, k) || p_{\theta}(z|y, c, k)) \\ & + \mathbb{E}_{z \sim q_{\phi}(z|c, k)} \log p_{\theta}(y|c, k, z) \end{aligned} \quad (1)$$

In the above formula, KL represents the Kullback–Leibler divergence. $p_{\theta}(y|z, c, k)$ acts as the decoder, reconstructing the response y using the latent variable z , as well as context c and knowledge k ; $q_{\phi}(z|y, c, k)$ serves as the inference model, approximating the true posterior; $p_{\theta}(z|c, k)$ is the prior model, which samples the latent variable from the prior

distribution. In these models, θ, ϕ are parameters that respectively pertain to the inference and decoder models.

The CVAE model mentioned above employs a single latent variable to encode information from context and knowledge sources. A typical implementation of this CVAE involves concatenating c and k as input to the encoder, subsequently sampling the latent variable $z \sim q_\phi(z|[c, k])$ [28,29], as illustrated in Figure 1a. However, it has been noted that, in the absence of explicit supervision, the solitary latent variable generated by the standard CVAE fails to learn disentangled representations that accurately reflect the distinct latent structures of different sources [31]. This limitation significantly impedes the model's performance enhancement and interpretability potential by exploiting variational latent variables.

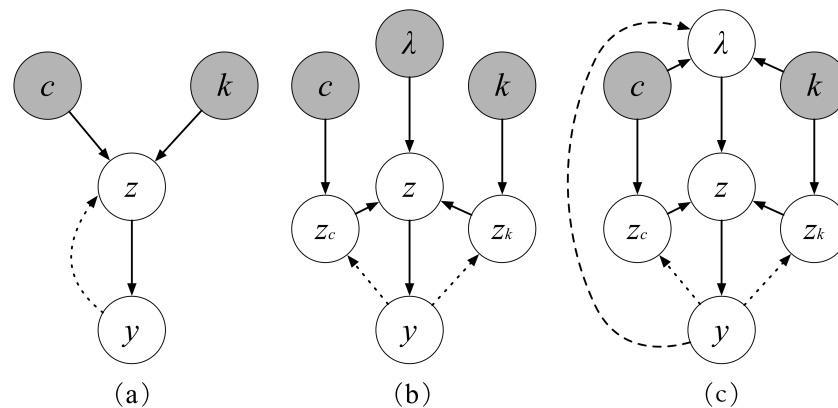


Figure 1. Comparison of graphical models: (a) the Vanilla CVAE; (b) KIV_d , the KIV model with a deterministic interpolation weight; (c) KIV_p , the KIV model with a probabilistic weight, in which the weight λ is a latent variable sampled from the Logistic-Normal distribution conditioned on c and k .

4. Knowledge Interpolated Conditional Variational Auto-Encoder (KIV)

To address the representation entangled issue of the standard CVAE, we introduce our KIV model to explicitly disentangle the latent variable z by incorporating two new variables z_k and z_c . Using two independent variational encoders, KIV initially encodes the given context and knowledge separately. Upon acquiring the disentangled latent variables z_k and z_c , we employ linear interpolation to blend z_k and z_c . This approach allows for smooth transitions between latent variables in an interpretable and controllable way. Specifically, the latent variables z_k and z_c in our model are linearly interpolated using a weight factor $\lambda \in (0, 1)$, thereby constructing a 1-simplex as follows:

$$z = \lambda z_k + (1 - \lambda) z_c. \quad (2)$$

In this study, we denote the interpolation weight corresponding to knowledge as λ , and the interpolation weight of context is given by $1 - \lambda$. The interpolation weight λ signifies the knowledge specificity and relevance of the response, thus facilitating better interpretation and connection of empirical findings. By incorporating the interpolated latent variable into Equation (1) and applying the convex property of the KL-divergence, we can deduce a new ELBO as the training objective:

$$\begin{aligned} \mathcal{L}_{KIV} = & -\lambda \text{KL}(q_\phi(z_k|y, k) || p_\theta(z_k|k)) \\ & -(1 - \lambda) \text{KL}(q_\phi(z_c|y, c) || p_\theta(z_c|c)) \\ & + \mathbb{E}_{z_c \sim q_\phi, z_k \sim q_\phi} \log p_\theta(y|c, k, z) \end{aligned} \quad (3)$$

A detailed derivation showing that $\mathcal{L}_{KIV}$ still serves as a valid lower bound of $\log p(y|c, k)$ is provided in Appendix A.

As illustrated in Figure 2, the proposed KIV model is composed of context/knowledge encoders and a response decoder, the specifics of which are expanded upon in the following section. The interpolation methods will be introduced in the subsequent section.

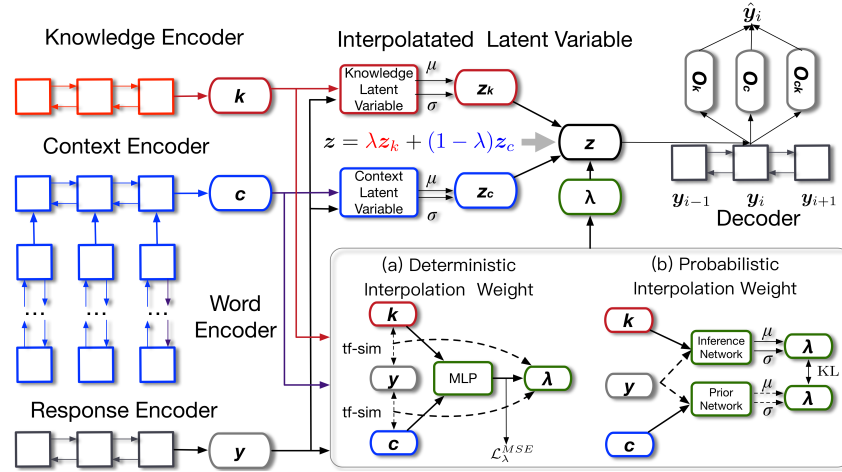


Figure 2. The detailed architecture of our proposed Knowledge Interpolated Conditional Variational Auto-Encoder. Two latent variables z_k, z_c , corresponding to context and knowledge, are smoothly fused by (a) deterministic and (b) probabilistic interpolation methods.

4.1. Context and Knowledge Encoders

The context and knowledge latent variables z_c, z_k are derived from two variational neural modules that share the same architecture but utilize different parameters. Taking the knowledge latent variable z_k as an example, its variational neural module, f^k , is composed of a posterior network f_{pst}^k and a prior network f_{pri}^k . The prior distribution is a factorized normal distribution $p_\theta(z_k|k) \sim \mathcal{N}(\mu_{pri}^k, \sigma_{pri}^k)$, which is parameterized by the prior network f_{pri}^k as follows:

$$[\mu_{pri}^k, \log \sigma_{pri}^k] = f_{pri}^k(k). \quad (4)$$

In the training phase of our model, k and y jointly define the posterior distribution $q_\phi \sim \mathcal{N}(\mu_{pst}^k, \sigma_{pst}^k)$ by

$$[\mu_{pst}^k, \log \sigma_{pst}^k] = f_{pst}^k([k, y]), \quad (5)$$

where f_{pst}^k and f_{pri}^k are multi-layer perceptrons with tanh activation function. The knowledge representation k and response representation y are derived by extracting the final hidden state from two bidirectional GRU encoders.

Likewise, we can obtain the parameterized prior distribution $p_\theta(z_c|c)$ and posterior distribution $q_\phi(z_c|c, y)$ of the context latent variable z_c via its variational neural module f^c . The context representation c is obtained by extracting the final hidden state from a hierarchical GRU encoder [3]. This method uses a word-level GRU network for each utterance and then feeds the outputs of the word-level GRU's last hidden state into an utterance-level GRU network.

4.2. Response Decoder

Upon obtaining z_c and z_k , an interpolation network produces the interpolated latent variable z as described in Equation (2). The decoder, a GRU-based recurrent neural network f^{dec} , maintains a hidden state h_t^{dec} at each step. The initial hidden state h_0^{dec} is configured by the latent variable z through a MLP: $h_0^{dec} = \text{MLP}(z)$. To fully harness the context c , knowledge k , and latent information inherent in z , we introduce a mixture-of-decoders mechanism to output the probability of a response. This mechanism draws inspiration from

the Mixture-of-Softmaxes (MoS) trick proposed by [49]. The proposed method incorporates three decoding modules corresponding to various settings of model outputs by

$$p_{\theta}(y_t = x | y_{<t}, c, k, z) = \sum_{m \in \{k, c, ck\}} \pi_t^m \frac{\exp o_t^m \top w_x}{\sum_{x'} \exp o_t^m \top w_{x'}} \quad (6)$$

where w_x is the embedding vector of word w_x , and $o_t^m, m = \{k, c, ck\}$ represent three output vectors corresponding to knowledge, context and a mixture of knowledge and context information. The dimension of o_t^m aligns with the dimension of w_x . The term π_t^m is the mixture weight of the m -th component, subject to the constraint $\sum_m \pi_t^m = 1$. In our model, we define the output vector o_t^m as follows:

$$\begin{aligned} o_t^k &= W_k[h_t^{dec}, z, k], o_t^c = W_c[h_t^{dec}, z, c], \\ o_t^{ck} &= W_{ck}[h_t^{dec}, z, k, c], \end{aligned} \quad (7)$$

where $W_{c,k,ck}$ are weights that transform inputs into vectors with the same dimension as the word embedding vector. The mixture weight π_t^m is computed by $\pi_t^m = \frac{\exp(w_{\pi,m} \top o_t^m)}{\sum_{k'} \exp(w_{\pi,m} \top o_t^{k'})}$, where $w_{\pi,m}$ represents a trainable weight.

5. Interpolation of Latent Variables

We propose two methods to compute the interpolation weight λ . Initially, we treat the interpolation weight λ as a deterministic variable and pre-compute it based on the relatedness between the response and the knowledge during the training phase. During the testing, we substitute λ with a value predicted by a neural network, as depicted in Figure 1b. However, the same or similar input context and knowledge can yield suitable responses at different levels of knowledge specificity for multiple λ values. To account for the uncertainty and variability of interpolation weights, we propose a second approach where we model the interpolation weight as following a distribution, which is jointly trained and inferred with z_k, z_c , as illustrated in Figure 1c.

5.1. Deterministic Interpolation Weight

One direct method to acquire the interpolation weight λ , is to pre-compute it as a relatedness measure between the response and knowledge. Specifically, λ indicates the relative usage ratio of context and knowledge. We propose calculating λ by determining the relative tf-idf similarity between the response and context/knowledge:

$$\lambda = \frac{\text{sim}(r, k)}{\text{sim}(r, k) + \text{sim}(r, c)}, \quad (8)$$

where $\text{sim}(\cdot, \cdot)$ represents the tf-idf similarity between two text portions. During testing, the ground truth response is unavailable; hence we employ an MLP network with the sigmoid function to predict λ . To train this MLP, we construct training data with inputs as the concatenation of context and knowledge representations, represented as: $\hat{\lambda} = \sigma(\text{MLP}([c, k]))$. During the training phase, to bridge the discrepancy between the predicted weight and the ground truth, we aim to minimize the Mean Squared Error (MSE) loss, denoted as $\mathcal{L}_{\lambda}^{\text{MSE}}$, between $\hat{\lambda}$ and λ computed using Equation (8). We refer to the KIV model that uses deterministic interpolation weight as KIV_d, and the objective for this model is:

$$\mathcal{L}_{\text{KIV}}^d = \mathcal{L}_{\text{KIV}} - \mathcal{L}_{\lambda}^{\text{MSE}}. \quad (9)$$

5.2. Probabilistic Interpolation Weight

In order to adaptively utilize background knowledge for response generation, emulating human-like knowledge-grounded conversation behaviors, we assume that the interpolation weight λ follows a Logistic Normal distribution. This distribution is known

for its flexibility in approximating the Dirichlet distribution, and it can effectively capture correlations between components of probability vectors [50]. Each response is generated by sampling an interpolation weight from this Logistic-Normal distribution:

$$\begin{aligned} p(\lambda|\mu^\lambda, \sigma^\lambda) &\sim \mathcal{LN}(\mu^\lambda, \sigma^\lambda) \\ &= \frac{1}{\sigma^\lambda \sqrt{2\pi}} \frac{1}{\lambda(1-\lambda)} \exp\left(-\frac{(\text{logit}(\lambda) - \mu^\lambda)^2}{2(\sigma^\lambda)^2}\right), \end{aligned} \quad (10)$$

where μ^λ and σ^λ are parameters of the Logistic Normal distribution. To parameterize $p(\lambda|\mu^\lambda, \sigma^\lambda)$, we obtain the posterior and priori latent variables by reparameterizing $\mu^\lambda, \sigma^\lambda$ as follows:

$$\begin{aligned} [\mu_{pst}^\lambda, \sigma_{pst}^\lambda] &= f_{pst}^\lambda([c, k, y]), \\ [\mu_{pri}^\lambda, \sigma_{pri}^\lambda] &= f_{pri}^\lambda([c, k]), \\ p(\lambda^{pst} | z_k, z_c) &\sim \mathcal{LN}(\mu_{pst}^\lambda, \sigma_{pst}^\lambda), \\ p(\lambda^{pri} | z_k, z_c) &\sim \mathcal{LN}(\mu_{pri}^\lambda, \sigma_{pri}^\lambda). \end{aligned} \quad (11)$$

The posterior and prior network $f_{pst}^\lambda, f_{pri}^\lambda$ are MLP networks with tanh activation functions. To obtain a sample from the reparameterized Logistic-Normal distribution, we first draw a sample from the Normal distribution $\mathcal{N}(\mu^\lambda, \sigma^\lambda)$ and then apply the logistic function to transform the sample into the Logistic-Normal distribution space. The KL-divergence between the posterior and prior weight can be computed using the closed-form formula for the Gaussian distribution [50]. By injecting the probabilistic interpolation weight λ , the posterior distribution during training can be factorized as: $q_\phi(z, \lambda | y, c, k) = q_\phi(z | y, c, k) q_\phi(\lambda | y, c, k)$. The prior distribution $p_\theta(z | c, k)$ can be decomposed in the same way. Injecting the factorized distribution in Equation (3), the final objective can be rewritten as:

$$\mathcal{L}_{KIV}^p = -\text{KL}(q_\phi(\lambda | c, k, y) || p_\theta(\lambda | c, k)) + \mathcal{L}_{KIV}. \quad (12)$$

The KIV model with the above objective is denoted as KIV_p, as shown in Figure 1c.

5.3. Model Training

In the scenario of deterministic interpolation weights, we use the loss function $\mathcal{L}_{KIV}^d$ as defined in Equation (9) for training. On the other hand, when employing probabilistic interpolation weights, the training objective is defined as in Equation (12). To mitigate the issue of posterior collapse in response generation, we implement techniques such as the KL annealing trick and the bag-of-words loss, as proposed in previous work [28,46].

6. Experiments and Analysis

This section is organized as follows: First, in Section 6.1, we discuss experimental settings, covering aspects such as the dataset, baseline models, and metrics. Subsequently, in Section 6.2, we delve into the implementation specifics of the proposed model. Finally, we present and analyze our experimental results in sections ranging from Sections 6.3–6.5.

6.1. Experimental Settings

The programming environment is set up with Python version 3.7.16 and Cudatoolkit version 11.7. It utilizes a Tesla V100 32 GB GPU for accelerated computations. The required packages are installed using pip, including torch version 1.13.1 for deep learning tasks, numpy for numerical computations, spacy with the en-core-web-trf model for natural language processing, and pandas version 1.3.5 for data manipulation and analysis.

6.1.1. Dataset

We evaluated our model on two commonly used public benchmark datasets for the knowledge grounded dialog system, Wizard-of-Wikipedia (WoW) [7] and Holl-E [10]. WoW is an open-domain knowledge graph dataset created using Wikipedia passages as a source of background knowledge. It has fine-grained annotations of selected knowledge. The test set of the Wizard dataset is divided into two subsets: *Test Seen* and *Test Unseen*. *Test Seen* contains 3619 conversation turns on topics overlapping with those in the training set. In contrast, *Test Unseen* includes 3689 turns on topics never encountered in the training or validation sets. In total, there are 68,931/3686/7308 conversations used for training/validation/testing. In each test set, we assess the proposed model and baselines under two scenarios: (1) the ground-truth knowledge selected by the model is known; (2) a separately trained knowledge selection model predicts the knowledge. To ensure a fair comparison in the predicted knowledge setting, our proposed model and all baselines utilize the knowledge determined by a pre-trained transformer memory network. This is the same knowledge selection module employed in the two-state generative model [7]. Holl-E is a specialized language model that focuses on the movie domain. It has been trained on a diverse range of data, including plots, comments, and movie reviews from various websites. The model has two versions of the test set: *Single reference test* and *Multi-reference test*. The *Single reference test* contains one annotated response per conversation, while the *Multi-reference test* includes multiple human-annotated ground-truth knowledge and corresponding responses for each instance. In total, there are 7228/930/913 dialogues used for training/validation/testing.

6.1.2. Baselines

We compare our proposed models with the following four baselines:

- **HRED** [3]: A general knowledge-free model encodes the context at two hierarchical levels.
- **CVAE** [32]: This model can be considered a modified kg-CVAE model [28] as depicted in Figure 1a.
- **GTPP** [51]: This model, based on HRED, incorporates grounded knowledge through a copying mechanism, enabling it to copy phrases from the knowledge at the appropriate decoding step.
- **TMem** [7]: Transformer Memory Network first concatenates the representations of context and knowledge and employs a transformer-based framework to generate knowledge grounded responses.
- **SKT** [25]: The sequential latent variable model is utilized to capture the knowledge selection process in multi-turn dialogue generation.
- **KIV_c**: This is a variant of our proposed model which directly concatenates the context latent variable z_c and the knowledge latent variable z_k . This conditions the response generation without any variable interpolation, i.e., Equation (2) is replaced with $z = [z_k, z_c]$ with the encoders and decoder remaining the same.

6.1.3. Metrics

We employ four types of automatic metrics to evaluate our proposed model and the baseline models: Per Perplexity (PPL), three embedding metrics (*Embedding Average* (AVE), *Embedding Extreme* (EXT), and *Greedy Matching* (GRY)), *Distinct₁* (*Dist₁*) and *Distinct₂* (*Dist₂*), and our calculated Bilingual Evaluation Understudy metrics ΔBLEU_k .

PPL: It is the exponentiation of the word entropy, and describes how well the generative model predicts the expected responses. It does not directly capture coherence, as a low perplexity value indicates an accurate prediction, but can not guarantee a coherent text. Coherence depends on factors like transitions, readability, consistent topics, and logical structures. Diversity in text generation refers to variations and novelty, which perplexity does not capture.

AVE, EXT, and GRY: Rather than using n-gram overlapping-based metrics like Bilingual Evaluation Understudy (BLEU) or Recall-Oriented Understudy for Gisting Evaluation

(ROUGE), we report word embedding-based similarity metrics [52] to capture the semantic alignment between generated responses and ground truth. Specifically, we adopt three embedding metrics: *Embedding Average* (AVE), *Embedding Extreme* (EXT), and *Greedy Matching* (GRY).

AVE calculates the average similarity between consecutive sentence embeddings in generated text, indirectly enhancing coherence/diversity. It incorporates contextual information from previous words, aligning the generated text with the overall topic/theme and improving coherence/diversity.

EXT involves selecting the most unique or extreme word embeddings to generate text that deviates from typical language patterns. It aims to enhance diversity but may result in inconsistencies and reduced coherence. A low embedding extreme score suggests repetition or redundant sentences, while a high score indicates incoherence.

GRY is a method employed to identify the most pertinent sentences from a larger body of text. Its purpose is to ensure consistency by selecting text segments that are highly relevant to a given query. While emphasizing coherence, greedy matching tends to prioritize closely associated sentences, potentially leading to a dearth of diversity in the generated text.

(Dist₁) and (Dist₂) [53]: We use (Dist₁) and (Dist₂) to evaluate the diversity of responses, which calculates the ratio of unique unigrams (or bigrams) to the total number of generated words. It determines the ratio of unique n-grams to the total number of n-grams produced. A higher distinct value suggests a greater range of content. However, in certain instances, a higher distinctness can result in reduced coherence as it may introduce unrelated ideas that disrupt the overall flow and cohesiveness of the text.

ΔBLEU_k: We use ΔBLEU_k to measure the engagement of knowledge. To assess whether the knowledge is articulated diversely and engagingly, we define a unique automatic metric that measures the ratio of knowledge utilization in responses based on the word overlap between responses and provided knowledge. We initially compute BLEU_k($\hat{y}$) as the average BLEU score, considering the generated response $\hat{y}$ as the hypothesis and given knowledge as the reference. An exceptionally high BLEU_k($\hat{y}$) signifies an excessive copy of external knowledge and makes the response less engaging. We further consider BLEU_k(y) of the gold standard response created by human y as the ground truth and compute the average absolute difference between BLEU_k($\hat{y}$) and BLEU_k(y), namely ΔBLEU_k, to measure the engagement of knowledge:

$$\Delta\text{BLEU}_k = \frac{1}{N} \sum_{y, \hat{y}} |\text{BLEU}_k(y) - \text{BLEU}_k(\hat{y})|, \quad (13)$$

where N is the number of samples in the test set, y and $\hat{y}$ are gold and generated responses from the same context and knowledge. A lower ΔBLEU_k indicates a closer knowledge copy ratio of generated responses to ground truths, showing better knowledge engagement.

6.2. Implementation Details

The vocabulary size in our model is limited to 20,000, encompassing 95.75% of words in the dataset. The embedding size that has been shared with both the encoder and decoder is 22.89 MB. We use the pre-trained GloVe 300-dimensional word embeddings for both the encoder and the decoder. The encoder size is 42.69 MB, and the decoder size is 58.24 MB. Single-layer bi-directional RNNs with GRU [54] are used for knowledge/response encoders and word-level networks in context encoders. A single-layer uni-directional GRU is employed for the utterance-level network in the context encoder. Another single-layer GRU is used for the decoder. The dimension of all hidden states in the GRU network is set to 512. The size of the latent variables is set to 128. The inference and prior networks consist of a single-layer feed-forward network with a tanh activation function.

We apply Layer Normalization when training the decoder. All weights are initialized by the Xavier method [55]. The model is trained end-to-end by Adam optimizer [56], with the learning rate set to 10^{-4} and gradient clipped applied at 1. During text generation, we

use a greedy strategy along with the KL-annealing strategy, in which the temperature is increased by 10^{-5} after each batch update iteration, varying from 0 to 1.

6.3. Metric-Based Evaluation

As shown in Table 1, test seen refers to evaluating a model's performance on familiar data it has encountered or been trained on, while test unseen assesses the model's ability to generalize to new, unseen data. Test seen measures the model's recall and application of learned patterns, while test unseen evaluates its adaptability and accuracy in novel scenarios. Gold knowledge refers to the reference or target information used for evaluation or training. It represents correct answers and serves as a benchmark for measuring model performance. Predicted knowledge refers to the knowledge generated by the model during response generation. Our experiment aims to evaluate how well the model utilizes the gold knowledge in its responses. The averaged BLEU_k score is used as a metric to measure the quality of incorporating the gold knowledge into the generated responses.

Table 1. Automatic evaluation results on WoW Dataset for Test Seen and Test Unseen setting. ΔB_k in table denotes ΔBLEU_k .

Test Seen		Gold Knowledge							Predicted Knowledge						
Method	PPL	AVE	EXT	GRY	Dist ₁	Dist ₂	ΔB_k	PPL	AVE	EXT	GRY	Dist ₁	Dist ₂	ΔB_k	
HRED	25.1	88.6	53.7	83.9	3.7	22.3	9.3	36.5	88.1	53.6	84.9	3.5	21.6	11.1	
CVAE	27.6	87.0	51.2	82.1	5.2	30.2	9.2	42.1	87.4	52.1	83.6	5.4	31.7	8.0	
GTTP	20.5	89.9	54.5	85.2	4.8	26.3	11.4	38.8	88.9	53.2	84.2	4.2	25.1	10.7	
TMem	23.1	87.4	52.9	83.4	4.1	23.9	10.9	40.5	88.2	53.7	84.6	3.5	26.6	11.7	
SKT	53.3	85.5	90.1	64.8	8.4	30.0	2.17	59.2	86.3	88.2	66.3	7.8	29.1	2.21	
KIV _c	20.8	91.5	54.7	84.3	6.2	32.8	10.4	29.7	90.2	53.9	83.3	5.3	31.7	12.0	
KIV _d	15.6	92.8	56.4	85.8	7.1	33.7	3.5	25.3	90.4	55.3	86.6	6.5	34.1	4.5	
KIV _p	19.3	91.6	55.8	86.9	8.4	37.2	2.8	31.5	89.9	54.0	86.4	7.6	35.8	2.7	
Test Unseen		Gold Knowledge							Predicted Knowledge						
Method	PPL	AVE	EXT	GRY	Dist ₁	Dist ₂	ΔB_k	PPL	AVE	EXT	GRY	Dist ₁	Dist ₂	ΔB_k	
HRED	56.2	88.3	52.5	84.0	2.4	19.4	8.9	88.5	87.2	51.4	82.0	2.2	18.9	11.8	
CVAE	68.7	85.2	50.5	81.9	3.0	24.9	10.3	75.3	84.6	50.3	81.8	2.8	22.6	9.2	
GTTP	49.1	86.9	51.8	83.8	2.8	26.7	13.0	81.9	85.9	52.3	83.5	2.5	22.4	11.0	
TMem	47.3	86.3	52.2	81.5	2.6	21.2	9.8	70.6	87.1	52.9	82.4	2.5	20.4	10.9	
SKT	83.0	84.5	89.3	62.8	5.76	19.3	2.74	92.1	82.5	87.6	63.5	5.21	17.3	3.16	
KIV _c	49.6	89.1	52.7	84.3	2.8	22.6	9.6	65.4	88.7	52.5	83.9	2.7	21.8	9.2	
KIV _d	44.2	90.2	53.6	84.5	3.5	23.8	4.2	57.2	89.3	52.8	83.7	3.3	24.6	4.7	
KIV _p	38.7	89.7	55.9	82.8	4.8	30.2	3.1	50.9	90.2	53.6	82.4	4.0	29.5	3.5	

The numbers in **Bold** represent the best results for the corresponding measure.

Automatic evaluations in the *Test Seen* setting reveal that the responses generated by KIV_d and KIV_p are considerably more coherent and relevant than those produced by all baseline models, as indicated by the PPL and word embedding-based similarity metrics. Regarding diversity metrics, KIV_p outperforms all others, except one: Dist_{1,2} has the same or slightly lower result compared with SKT on Gold and Predicted Knowledge. The EXT score on SKT is higher than our model, which confirms the Dist₁ measure that the diversity generated by SKT seems slightly better than ours. However, an emphasis on higher EXT score may bring incoherence since the model may produce more unique word embeddings. Having such a result from SKT, we think their model put a great emphasis on dealing with the diversity in the knowledge selection of conversations. Regarding the ΔBLEU_k indicator, SKT adopted a copy mechanism to maximize the effect of knowledge for response generation, showing a slight improvement in knowledge engagement.

In the scenario of *Test Unseen*, KIV_d and KIV_p demonstrate a similar pattern as *Test Seen* in terms of relevance and diversity metrics. These results substantiate our assertion

that interpolating two latent variables, conditioned on knowledge and context, aids in generating more appropriate and informative responses. A comparison between interpolation fusion methods ($KIV_{d,p}$) and non-interpolation fusion (KIV_c) reveals that interpolation significantly enhances the quality of responses. This indicates that interpolation is more appropriate for fusing knowledge and context in knowledge grounded conversations.

Our analysis of knowledge engagement metric $\Delta BLEU_k$ shows that KIV_p achieves the lowest score, except that it is slightly higher than SKT. This indicates that interpolating latent variables using probabilistic weights can adaptively learn the joint representation of context and knowledge, leading to the generation of more engaging responses with respect to the incorporated knowledge.

Similar results were also observed on Holl-E in Table 2. The $Dist_{1,2}$ in KIV_p is lower than $KIV_{c,d}$, suggesting that KIV_p use more copied knowledge, while $KIV_{c,d}$ generate a greater range of content that may not be from the knowledge base. The $KIV_{c,d}$ AVG score is higher than KIV_p suggesting that the former generates consecutive sentences that are more similar to each other. Higher score in EXT by $KIV_{c,d}$ indicate generated works may be incoherent since they may produce more unique or extreme word embeddings. SKT model has highest EXT score in our experiments. A higher GRY in $KIV_{c,d}$ identifies the most pertinent sentences chosen from the knowledge base, therefore showing more emphasized coherence. However, a lower score by KIV_p may produce more diversity in the generated sentences. The PPL scores in KIV_p is much higher than $KIV_{c,d}$, indicating more accurate in the text prediction, and $BLEU_k$ scores are lower in KIV_p than $KIV_{c,d}$ suggesting the model is better at engagement of the knowledge, i.e., a closer knowledge copy. Our experiments show a similar result in Multiple Reference. We have not calculated the SKT's $\Delta BLEU_k$ score, since our KIV_p $\Delta BLEU_k$ score has almost achieved a perfect score. In addition, following previous work [10], we calculate the scores for the multi-reference dataset by taking the maximum score over multiple reference responses. Since the $Dist_{1,2}$ score is calculated according to the repeated n-gram of generated responses instead of the multiple reference responses, we exclude $Dist_{1,2}$.

Table 2. Automatic evaluation results on HOLL-E Dataset. ΔB_k in table denotes $\Delta BLEU_k$.

Test Seen	Single Reference							Multiple References				
	PPL	AVE	EXT	GRY	$Dist_1$	$Dist_2$	ΔB_k	PPL	AVE	EXT	GRY	ΔB_k
SKT	46.3	88.1	90.7	70.6	6.76	28.5	-	52.1	94.5	96.1	77.2	-
KIV_c	9.251	84.9	42.0	77.3	0.4	2.4	3.6	7.204	87.5	41.4	68.1	10.3
KIV_d	3.141	97.0	88.1	96.1	6.5	34.8	3.5	0.502	98.3	94.0	97.3	8.0
KIV_p	2.665	97.0	88.7	96.2	6.6	34.8	0.04	0.415	98.9	94.2	97.5	3.2

The numbers in **Bold** represent the best results for the corresponding measure.

6.4. Human Evaluation

In addition to the automated evaluation, we also carried out a human evaluation on the WoW dataset to gauge the quality of responses produced by our model and the baseline models. We hired three professional annotators to assess the generated responses based on four criteria, which fall into two categories: (1) *Coherence (C)/Fluency (F)*: This assesses whether the response is coherent with the dialogue context and fluent to read. (2) *Knowledge Correctness (KC)/Knowledge Diversity (KD)*: This assesses whether the response is consistent with the provided knowledge and presents relevant knowledge diversely and engagingly. For both the *Test Seen* and *Test Unseen* settings, each annotator rated 100 randomly sampled responses generated by each model using a five-point scale (1–5).

Human evaluation results are listed in Table 3. We observe that GTTP is a robust baseline method, except for the metric *Knowledge Diversity (KD)*, since GTTP tends to copy the whole sentence of the given knowledge. The proposed $KIV_{d,p}$ achieve the best performance in terms of most metrics, in which KIV_p performs slightly lower on *Knowledge Diversity* than SKT, but outperforms by a substantial margin compared with the rest of the

baselines. This result verifies that the proposed interpolation-based models, especially KIV_p, can present relevant knowledge excitingly and engagingly while keeping the response coherent with the context.

Table 3. Human Evaluation results on WoW Dataset with respect to *Coherence (C)*, *Fluency (F)*, *Knowledge Correctness (KC)*/*Diversity (KD)*. The numbers in **Bold** represent the best results for the corresponding measure.

Models	Test Seen								Test Unseen							
	Gold Knowledge				Predicted Knowledge				Gold Knowledge				Predicted Knowledge			
	C	F	KC	KD	C	F	KC	KD	C	F	KC	KD	C	F	KC	KD
Gold	4.08	4.12	3.60	3.05	4.08	4.12	3.60	3.05	4.19	4.26	3.73	3.02	4.19	4.26	3.73	3.02
HRED	3.39	3.52	2.14	2.45	3.15	3.17	1.71	2.26	3.17	3.29	1.92	2.38	3.09	2.95	1.69	1.84
CVAE	3.14	3.08	2.62	2.58	3.00	2.87	2.06	2.47	3.08	2.99	2.51	2.31	3.07	2.82	2.63	2.19
GTTP	3.31	3.47	3.26	1.94	3.28	3.28	3.19	1.88	3.09	3.30	3.15	2.01	3.05	3.29	3.03	1.97
TMem	3.32	3.37	3.01	2.46	3.18	3.22	2.86	2.29	2.91	3.04	2.76	2.24	2.75	2.83	2.65	2.10
SKT	3.12	3.09	3.21	2.89	3.10	3.02	2.61	2.13	2.12	3.13	3.24	3.15	2.23	2.31	3.10	2.45
KIV _c	3.39	3.14	3.19	2.52	3.15	3.22	2.96	2.46	2.94	2.82	2.79	2.35	2.87	2.72	2.58	2.22
KIV _d	3.53	3.67	3.21	2.58	3.36	3.18	3.16	2.37	3.19	3.43	3.16	2.36	3.18	3.14	3.08	2.27
KIV _p	3.49	3.62	3.31	2.87	3.35	3.27	3.14	2.59	3.13	3.31	3.21	2.48	3.19	3.09	3.16	2.39

The numbers in **Bold** represent the best results for the corresponding measure.

6.5. Qualitative Analysis

6.5.1. Impact of Interpolation Weight

We first investigate the relationship between the interpolation weight λ employed in our model and the ratio of knowledge utilization in the generated responses. We group the test samples into separate bins according to their λ value and calculate the averaged knowledge BLEU score, $BLEU_K$, for responses within each bin. The relationship between λ and the averaged $BLEU_K$ is illustrated in the line plot in Figure 3. The x-axis represents the values of λ , ranging from 0 to 1, while the y-axis represents the averaged BLEU score ($BLEU_K$). Each data point on the line plot represents the average $BLEU_K$ for responses falling within a specific range of λ values. When $\lambda = 0$, the output is determined solely by the context variable z_c , with no knowledge-specific influence; this can be interpreted as a response that relies solely on the immediate context and does not make use of any external knowledge. When $\lambda = 1$, the output is determined solely by the knowledge specific variable z_k , without any contribution from the context variable z_c . This means that the response will be generated solely based on the acquired knowledge, disregarding the context provided. Intermediate values of λ (between 0 and 1) blend both variables, enabling smooth transitions and combinations of knowledge and context.

The line plot in Figure 3 shows the relationship between λ and the averaged $BLEU_K$ (knowledge BLEU score) for different experimental settings. It indicates that as the interpolation weight λ increases, the $BLEU_K$ also tends to increase, implying a positive correlation between knowledge utilization (represented by λ) and the quality of responses (measured by $BLEU_K$). This suggests that the learned interpolation weight can effectively represent the knowledge utilization ratio, as higher values of λ indicate a stronger influence of knowledge in generating responses.

The line plot helps visualize the trend that increasing the emphasis on knowledge (higher λ) leads to improved knowledge utilization and better quality responses. It provides evidence that the interpolation weight λ plays a crucial role in controlling the balance between knowledge and context in the model's responses. It demonstrates that, as the model assigns higher weights to the knowledge variable, it generates responses with better utilization of knowledge, leading to improved $BLEU_K$ scores.

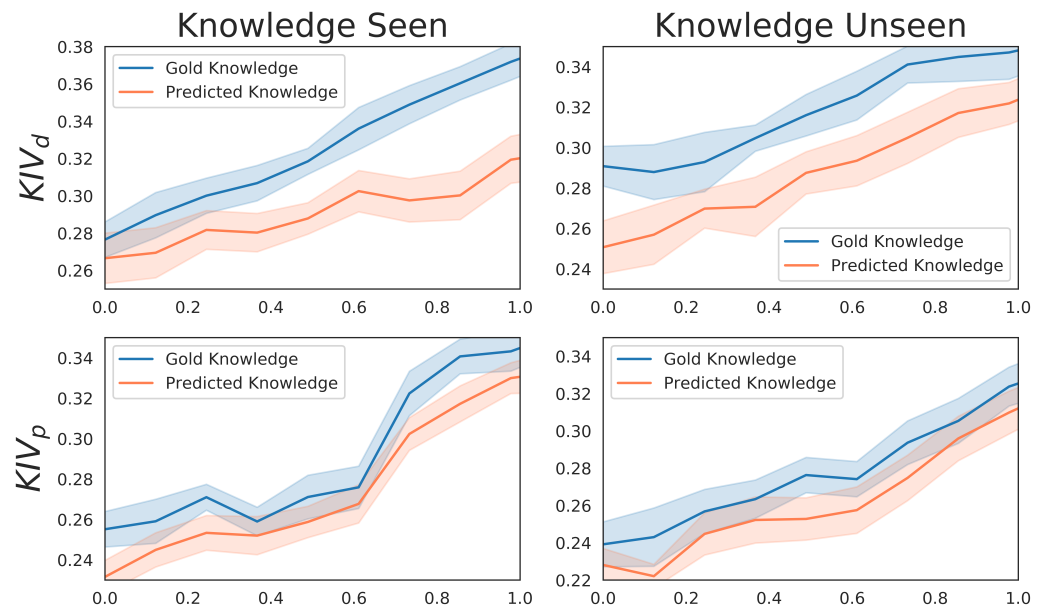


Figure 3. BLEU_k against interpolation weight λ on WoW dataset.

6.5.2. Case Study

Figure 4a presents an example of responses generated by different models given the same context and knowledge. As can be observed, our proposed model tends to generate more reasonable and engaging responses by considering both the context and the provided knowledge. We also notice that KIV_c leans towards addressing the query by delivering relevant knowledge, whereas KIV_p is more inclined to create more diverse and indirect responses, utilizing the knowledge as a conversation guide. Figure 4b,c displays responses sampled with different interpolation weights λ 's in KIV_d and KIV_p. The learned latent variable in our model facilitates a smooth transition from solely relying on the query to parroting the external knowledge verbatim, demonstrating that the model can integrate knowledge and context at a semantic level within the latent space.

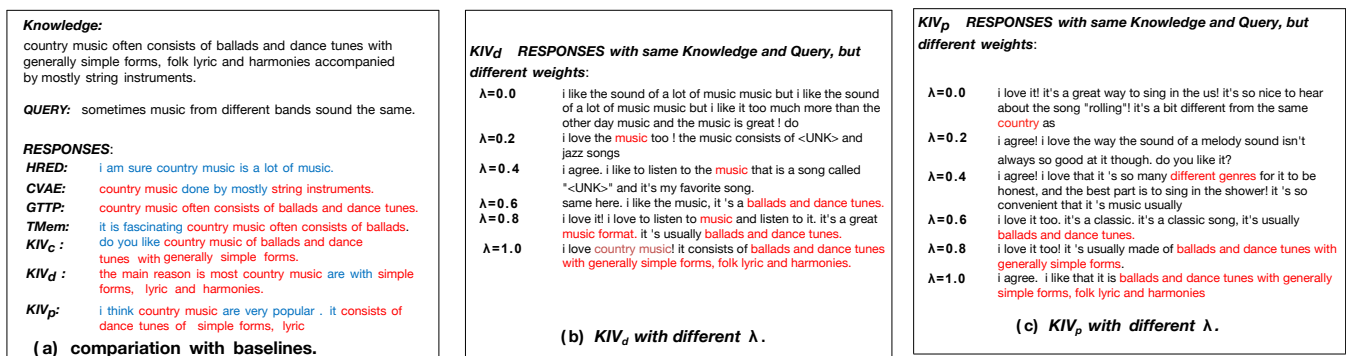


Figure 4. Generated examples of our model and baselines on WoW dataset. The red color shows the copy parts of knowledge indicating a higher propensity to provide knowledge, the blue color indicates the generated pieces corresponding to the replies to queries. (a) Responses generated by KIV and baseline models. (b,c) Responses generated by KIV_d and KIV_p with different interpolation weights.

6.5.3. More Generated Examples

In this section, we present additional examples of responses generated by our model and baseline models. In Figure 5, we have included three new topics: heavy metal music, Chevrolet corvette, and coffee. As can be seen, both KIV_d and KIV_p exhibit the ability to generate responses that are not only more relevant but also more engaging.

In Figure 6, we have included a multi-turn dialogue generated by our model KIV_p. During the multi-turn dialogue initiated by the wizard, the topic of “Science Fiction” was introduced, leading to an exploration of various aspects of the genre. As the conversation progressed into its second round, the model smoothly transitioned the focus towards political and philosophical issues within science fiction. At this point, the human participant engaged by discussing time travel and mentioning the Harry Potter movies. Recognizing the significance of Harry Potter as a book that has had a transformative impact, the model suggested it as further reading on the topic, which the human enthusiastically accepted. This seamless exchange of multiple turns created a highly engaging and enjoyable conversational experience.

Knowledge: country music often consists of ballads and dance tunes with generally simple forms, folk lyric and harmonies accompanied by mostly string instruments. QUERY: sometimes music from different bands sound the same. RESPONSES: HRED: i am sure country music is a lot of music. CVAE: country music done by mostly string instruments. GTTP: country music often consists of ballads and dance tunes TMem: it is fascinating country music often consists of ballads. KIV_c: do you like country music of ballads and dance tunes with generally simple forms. KIV_d: the main reason is most country music are with simple forms, lyric and harmonies. KIV_p: i think country music are very popular. it consists of dance tunes of simple forms, lyric	Knowledge: the chevrolet corvette, known colloquially as the vette, is a sports car manufactured by chevrolet. QUERY: what do you know about the chevrolet corvette? RESPONSES: HRED: i not sure chevrolet corvette is a car car. CVAE: i've been a chevrolet corvette , so much sports GTTP: the vette, is a sports car manufactured by chevrolet. TMem: the vette or chevy corvette , is a sports car. KIV_c: chevrolet corvette is known sports car. KIV_d: i love the sports car manufactured by chevrolet and like car which is made by chevrolet. KIV_p: when i was a kid, i know it 's cool sports car and enjoy the classic corvette . do you like corvette?
(a) Topic: country music.	(b) Topic : chevrolet corvette.
Knowledge: dave grohl was the longest serving drummer for the grunge band nirvana and the frontman and founder of the rock band foo fighters, of which he is the lead vocalist, rhythm and lead guitarist, and primary songwriter. QUERY: that s awesome. who is your favorite metal musician? RESPONSES: HRED: do you know i like you . CVAE: lead vocalist , rhythm and lead vocalist is my favor GTTP: it 's dave grohl was the longest serving foo fighters TMem: i really like dave groh bob ross. KIV_c: do you like grohl of rock band foo foo fighter KIV_d: dave grohl who played in grunge rock bands nirvana and foo fighter KIV_p: well, dave grohl is often a drummer of grunge band. It was pioneers with the heavy metal music.	Knowledge: coffee plants are now cultivated in over 70 countries, primarily in the equatorial regions of the americas, southeast asia, india, and africa. QUERY: i really like coffee from kenya. RESPONSES: HRED: i like kenya coffee too. CVAE: coffee is bit cultivated in over cuntries in americas, GTTP: i love coffee plants! coffee plants are cultivated in over 70 countries! TMem: over 70 countries , primarily in the equatorial regions of the americas, southeast asia, india, and africa. KIV_c: coffee plants are now cultivated in over 70 countries with the equatorial regions like the us. KIV_d: i love coffee in over 70 countries including kenya, do you prefer coffer of the equatorial KIV_p: regions of the americas, southeast asia, india, and africa.
(c) Topic: heavy metal music.	(d) Topic: coffee.

Figure 5. More generated examples of our model and baseline models on WoW dataset.

Model: KIV _p
Topic: Science Fiction Human: science fiction Model: i love science fiction, i love the idea of futuristic science and technology. Human: i'm a huge fan of science fiction myself! Model: you're right! i love how they focus on political issues and philosophical issues Human: i agree. one of my favorite forms of science fiction is anything related to time travel! i find it fascinating. Model: me too! i love how many different genres of changing things are changing. i think the story is really interesting. Human: and that's difficult to do when dealing with time travel. i actually haven't seen the latest harry potter movies. guess it 's time to check them out! Model: you should definitely check out the science fiction fiction, it's fascinating how the book has changed many things . Human: thank you for the suggestion, i will definitely check it out! Model: it's been a very successful film, but i don't think i'd have to be a writer .

Figure 6. More generated examples of our model on multi-turn dialogue on WoW dataset.

7. Conclusions

This paper introduces KIV, a novel interpolation-based CVAE designed to generate knowledge-grounded responses. This approach incorporates two distinct latent variables for modeling context and knowledge. These latent variables are seamlessly integrated via linear interpolation, where the interpolation weight is tied to the degree of knowledge specificity. Specifically, we propose two interpolation strategies: deterministic interpolation, which uses semantic similarity as interpolation weight, and probabilistic interpolation, which treats the interpolation weight as a probabilistic variable sampled from a Logistic Normal distribution. Based on both automatic and human evaluations, our experimental results demonstrate that both interpolation strategies outperform in relevance and fluency. Moreover, probabilistic interpolation significantly enhances knowledge engagement and diversity. Future research will concentrate on exploring more sophisticated methods for interpolation weight optimization and handling multimodal data such as images or sound.

Author Contributions: Conceptualization, X.L. and J.D.; Software, T.N.; Formal Analysis, J.D.; Writing—Review and Editing, X.L.; Supervision, L.Z. and R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (62006062, 62176076), the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies(2022B1212010005), Natural Science Foundation of Guangdong (2023A1515012922), Key Technologies Research and Development Program of Shenzhen JSGG20210802154400001, and HIT-KONKA Joint Lab.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A. Derivation of Lower Bound

In this section we will show $\mathcal{L}_{\text{KIV}}$ is a lower bound estimation of $\log p(y|c, k)$. The loss function of CVAE [27] can be formulated as maximizing by the Evidence Lower Bound (ELBO) of logarithm likelihood $\log p(y|c, k)$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} &= -\text{KL}(q_\phi(z|y, c, k)||p_\theta(z|y, c, k)) \\ &\quad + \mathbb{E}_{z \sim q_\phi(z|c, k)} \log p_\theta(y|c, k, z) \\ &\leq \log p_\theta(y|c, k).\end{aligned}\tag{A1}$$

By injecting the interpolated latent variable, the loss function of our model $\mathcal{L}_{\text{KIV}}$ now is:

$$\begin{aligned}\mathcal{L}_{\text{KIV}} &= -\lambda \text{KL}(q_\phi(z_k|y, k)||p_\theta(z_k|k)) \\ &\quad - (1 - \lambda) \text{KL}(q_\phi(z_c|y, c)||p_\theta(z_c|c)) \\ &\quad + \mathbb{E}_{z_c \sim q_\phi(z_c|c), z_k \sim q_\phi(z_k|k)} \log p_\theta(y|c, k, z).\end{aligned}\tag{A2}$$

To verify the loss function $\mathcal{L}_{\text{KIV}}$ still remains the lower bound of $\log p(y|c, k)$, we only need to show that

$$\mathcal{L}_{\text{KIV}} \leq \mathcal{L}_{\text{CVAE}}.\tag{A3}$$

To verify Inequality (A3) holds, we notice that the reconstruction loss function in Equations (A1) and (A2) is the same term. We only need to verify the sum of first two negative KL-divergence terms in Equation (A2) is a lower bound of negative KL-divergence

term in Equation (A1). We leverage the convex property of KL-divergence [57] to derive it, i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$\begin{aligned} & \text{KL}(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \\ & \leq \lambda \text{KL}(p_1 \parallel q_1) + (1 - \lambda) \text{KL}(p_2 \parallel q_2), \end{aligned} \quad (\text{A4})$$

for any $\lambda \in [0, 1]$. By replacing $z = \lambda z_k + (1 - \lambda)z_c$ in the negative KL-divergence term in $\mathcal{L}_{\text{CVAE}}$. We can easily verify that:

$$\begin{aligned} & -\text{KL}(q_\phi(z|y, c, k) \parallel p_\theta(z|y, c, k)) \leq \\ & -\lambda \text{KL}(q_\phi(z_k|y, k) \parallel p_\theta(z_k|k)) + \\ & -(1 - \lambda) \text{KL}(q_\phi(z_c|y, c) \parallel p_\theta(z_c|c)). \end{aligned} \quad (\text{A5})$$

Therefore $\mathcal{L}_{\text{KIV}}$ is the lower bound estimation of logarithm likelihood $\log p(y|c, k)$.

References

- Shang, L.; Lu, Z.; Li, H. Neural Responding Machine for Short-Text Conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, Beijing, China, 26–31 July 2015; pp. 1577–1586.
- Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.Y.; Gao, J.; Dolan, B. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Denver, CO, USA, 31 May–5 June 2015; pp. 196–205.
- Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI 2016, Phoenix, AZ, USA, 12–17 February 2016; pp. 3776–3783.
- Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; Jurafsky, D. Adversarial Learning for Neural Dialogue Generation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 7–11 September 2017; pp. 2157–2169.
- Shen, L.; Feng, Y. CDL: Curriculum dual learning for emotion-controllable response generation. *arXiv* **2020**, arXiv:2005.00329.
- Xu, J.; Wang, H.; Niu, Z.; Wu, H.; Che, W. Knowledge graph grounded goal planning for open-domain conversation generation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 9338–9345.
- Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; Weston, J. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
- Qin, L.; Galley, M.; Brockett, C.; Liu, X.; Gao, X.; Dolan, B.; Choi, Y.; Gao, J. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019; pp. 5427–5436.
- Ghazvininejad, M.; Brockett, C.; Chang, M.; Dolan, B.; Gao, J.; Yih, W.; Galley, M. A Knowledge-Grounded Neural Conversation Model. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LA, USA, 2–7 February 2018; pp. 5110–5117.
- Moghe, N.; Arora, S.; Banerjee, S.; Khapra, M.M. Towards Exploiting Background Knowledge for Building Conversation Systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, 31 October–4 November 2018; pp. 2322–2332.
- Lian, R.; Xie, M.; Wang, F.; Peng, J.; Wu, H. Learning to select knowledge for response generation in dialog systems. *arXiv* **2019**, arXiv:1902.04911.
- Xu, Y.; Ishii, E.; Winata, G.I.; Lin, Z.; Madotto, A.; Liu, Z.; Xu, P.; Fung, P. CAiRE in DialDoc21: Data augmentation for information seeking dialogue system. In Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021), Online, 5 August 2021; pp. 46–51.
- Yang, C.; Lin, Z.; Li, J.; Meng, F.; Wang, W.; Wang, L.; Zhou, J. TAKE: Topic-shift aware knowledge selection for dialogue generation. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 253–265.
- Komeili, M.; Shuster, K.; Weston, J. Internet-augmented dialogue generation. *arXiv* **2021**, arXiv:2107.07566.
- Varshney, D.; Prabhakar, A.; Ekbal, A. Commonsense and named entity aware knowledge grounded dialogue generation. *arXiv* **2022**, arXiv:2205.13928.
- Yang, Y.; Gao, Y.; Li, J.; Huang, H. G²: Enhance Knowledge Grounded Dialogue via Ground Graph. *arXiv* **2022**, arXiv:2204.12681.
- Sun, Q.; Xu, C.; Hu, H.; Wang, Y.; Miao, J.; Geng, X.; Chen, Y.; Xu, F.; Jiang, D. Stylized knowledge-grounded dialogue generation via disentangled template rewriting. *arXiv* **2022**, arXiv:2204.05610.

18. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, TX, USA, 1–5 November 2016; pp. 2383–2392.
19. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1870–1879.
20. Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, PR, USA, 2–4 May 2016.
21. Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; Yin, D. Knowledge diffusion for neural dialogue generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018; pp. 1489–1498.
22. Ren, P.; Chen, Z.; Monz, C.; Ma, J.; de Rijke, M. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019.
23. Meng, C.; Ren, P.; Chen, Z.; Monz, C.; Ma, J.; de Rijke, M. RefNet: A reference-aware network for background based conversation. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019.
24. Li, Y.; Zhao, J.; Lyu, M.R.; Wang, L. Eliciting Knowledge from Large Pre-Trained Models for Unsupervised Knowledge-Grounded Conversation. *arXiv* **2022**, arXiv:2211.01587.
25. Kim, B.; Ahn, J.; Kim, G. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv* **2020**, arXiv:2002.07510.
26. Lin, X.; Jian, W.; He, J.; Wang, T.; Chu, W. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–8 July 2020; pp. 41–52.
27. Kingma, D.P.; Mohamed, S.; Rezende, D.J.; Welling, M. Semi-supervised learning with deep generative models. In Proceedings of the 28th Conference on Neural Information Processing Systems, NIPS 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 3581–3589.
28. Zhao, T.; Zhao, R.; Eskenazi, M. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 654–664.
29. Song, H.; Zhang, W.; Cui, Y.; Wang, D.; Liu, T. Exploiting Persona Information for Diverse Generation of Conversational Responses. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 5190–5196.
30. Massiceti, D.; Siddharth, N.; Dokania, P.K.; Torr, P.H. FlipDial: A generative model for two-way visual dialogue. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6097–6105.
31. Locatello, F.; Bauer, S.; Lucic, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 10–15 June 2019; pp. 4114–4124.
32. Ye, H.; Lo, K.; Su, S.; Chen, Y. Knowledge-Grounded Response Generation with Deep Attentional Latent-Variable Model. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019.
33. Zheng, W.; Zhou, K. Enhancing Conversational Dialogue Models with Grounded Knowledge. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, 3–7 November 2019; pp. 709–718.
34. Sun, W.; Ren, P.; Ren, Z. Generative Knowledge Selection for Knowledge-Grounded Dialogues. *arXiv* **2023**, arXiv:2304.04836.
35. Prabhumoye, S.; Hashimoto, K.; Zhou, Y.; Black, A.W.; Salakhutdinov, R. Focused attention improves document-grounded generation. *arXiv* **2021**, arXiv:2104.12714.
36. Chen, Q.; Lin, J.; Zhang, Y.; Ding, M.; Cen, Y.; Yang, H.; Tang, J. Towards knowledge-based recommender dialog system. *arXiv* **2019**, arXiv:1908.05391.
37. Wang, L.; Hu, H.; Sha, L.; Xu, C.; Wong, K.F.; Jiang, D. Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph. *arXiv* **2021**, arXiv:2110.07477.
38. Zhou, K.; Zhao, W.X.; Bian, S.; Zhou, Y.; Wen, J.R.; Yu, J. Improving conversational recommender systems via knowledge graph based semantic fusion. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, USA, 6–10 July 2020; pp. 1006–1014.
39. Li, S.; Namazifar, M.; Jin, D.; Bansal, M.; Ji, H.; Liu, Y.; Hakkani-Tur, D. Enhanced knowledge selection for grounded dialogues via document semantic graphs. *arXiv* **2022**, arXiv:2206.07296.
40. Li, Z.; Niu, C.; Meng, F.; Feng, Y.; Li, Q.; Zhou, J. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019; pp. 12–21.

41. Wu, Z.; Lu, B.R.; Hajishirzi, H.; Ostendorf, M. DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization. *arXiv* **2021**, arXiv:2109.04673.
42. Mi, L.; He, T.; Park, C.F.; Wang, H.; Wang, Y.; Shavit, N. Revisiting Latent-Space Interpolation via a Quantitative Evaluation Framework. *arXiv* **2021**, arXiv:2110.06421.
43. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [[CrossRef](#)] [[PubMed](#)]
44. Qian, S.; Li, G.; Cao, W.M.; Liu, C.; Wu, S.; Wong, H.S. Improving representation learning in autoencoders via multidimensional interpolation and dual regularizations. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 3268–3274.
45. Verma, V.; Lamb, A.; Kannala, J.; Bengio, Y.; Lopez-Paz, D. Interpolation Consistency Training for Semi-supervised Learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 3635–3641.
46. Bowman, S.R.; Vilnis, L.; Vinyals, O.; Dai, A.M.; Jozefowicz, R.; Bengio, S. Generating Sentences from a Continuous Space. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, 11–12 August 2016; pp. 10–21.
47. Zhao, J.J.; Kim, Y.; Zhang, K.; Rush, A.M.; LeCun, Y. Adversarially Regularized Autoencoders. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, 10–15 July 2018; pp. 5902–5911.
48. Gao, X.; Lee, S.; Zhang, Y.; Brockett, C.; Galley, M.; Gao, J.; Dolan, B. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 1229–1238.
49. Ganea, O.; Gelly, S.; Bécigneul, G.; Severyn, A. Breaking the Softmax Bottleneck via Learnable Monotonic Pointwise Non-linearities. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 10–15 June 2019; pp. 2073–2082.
50. Srivastava, A.; Sutton, C.A. Autoencoding Variational Inference For Topic Models. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
51. See, A.; Liu, P.J.; Manning, C.D. Get to the point: Summarization with pointer-generator networks. *arXiv* **2017**, arXiv:1704.04368.
52. Liu, C.W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; Pineau, J. 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, Austin, TX, USA, 1–5 November 2016; pp. 2122–2132.
53. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. 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, NAACL 2016, San Diego, CA, USA, 12–17 June 2016; pp. 110–119.
54. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
55. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
57. Kullback, S. *Information Theory and Statistics*; Courier Corporation: North Chelmsford, MA, USA, 1997.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.