Information-Theoretic Text Hallucination Reduction for Video-grounded Dialogue

Sunjae Yoon†, Eunseop Yoon†, Hee Suk Yoon†, Junyeong Kim‡, Chang D. Yoo†*
†Korea Advanced Institute of Science and Technology (KAIST)
‡Chung-Ang University
{sunjae.yoon,esyoon97,hskyoon,cd_yoo}@kaist.ac.kr
‡junyeongkim@cau.ac.kr

Abstract

Video-grounded Dialogue (VGD) aims to decode an answer sentence to a question regarding a given video and dialogue context. Despite the recent success of multi-modal reasoning to generate answer sentences, existing dialogue systems still suffer from a text hallucination problem, which denotes indiscriminate text-copying from input texts without an understanding of the question. This is due to learning spurious correlations from the fact that answer sentences in the dataset usually include the words of input texts, thus the VGD system excessively relies on copying words from input texts by hoping those words to overlap with ground-truth texts. Hence, we design Text Hallucination Mitigating (THAM) framework, which incorporates Text Hallucination Regularization (THR) loss derived from the proposed information-theoretic text hallucination measurement approach. Applying THAM with current dialogue systems validates the effectiveness on VGD benchmarks (i.e., AVSD@DSTC7 and AVSD@DSTC8) and shows enhanced interpretability.

1 Introduction

Achieving a natural conversational agent that can do ‘look’ (i.e., understand what they are seeing) and ‘tell’ (i.e., converse what they are thinking) is desiderata in our vision-language community. By the broad application of conversational agent, it can potentially assist various subsections of our environment including education, entertainment, security, and visual or other impairments. For the natural conversation between humans and computers, a video-grounded dialogue (VGD) task (Alamri et al., 2019; Hori et al., 2020) has been introduced to generate adequate conversational responses to the queries of humans, while following up on video and dialogue context, which gives more challenging than traditional image-grounded or text-grounded dialogue tasks. To be specific, given video $V$, video caption $C$, dialogue history of past Q&A pairs: $H = \{(Q^1, A^1), \ldots, (Q^{r-1}, A^{r-1})\}$, and current $r$-th round question $Q^r$, VGD system is expected to make free-form answer sentence $A^r$ to given question. Despite recent advancements in multi-modal interactions including transformer (Vaswani et al., 2017), current VGD systems still suffer text hallucination problem, which denotes indiscriminate text-copying from input texts (i.e., question, caption, and dialogue history) to decode answer tokens, but the generated answer sentences are rather inadequate and not related to the question. This is because current VGD systems learn spurious correlations from the fact that many ground-truth answers in the dataset include partial input texts, thus they perform incorrect text-copy from input texts, namely text hallucination, even in answers where input texts are unnecessary.

Figure 1 gives two indiscriminate text hallucinating cases confounded by spurious correlations in VGD. As shown in Figure 1(a), for the given question ‘does he place the towel and clothes anywhere?’, we human identify where the man placed the towel and clothes, and if it cannot be confirmed, we give a sentence meaning ‘unknown’. However, in many cases, VGD systems are optimized in situations where they could find clues in video and dialogue, so for a case that they can not find clues, they simply pretend to know the answer by copying texts from input sentences without reasoning why the question is not answerable. Thus, the VGD systems depend on indiscriminate text hallucination, copying input sentences (i.e., questions, caption, dialogue), hoping the copied answer words to overlap the ground-truth words. Figure 1(b) presents another dependence on this text hallucination even in the answerable question. Given the question of ‘does the man wear glasses?’, the current VGD system provides incorrect answer without referring to the video and focuses on pretending to know the
a smiling person grabs a towel, and fresh clothes. then the person begins undressing.

where does the video take place?
looks like a bedroom maybe.

what happens after he enters the room?
he grabs a towel and clothes and then goes off screen.

does he place the towel and clothes anywhere?

yes, he grabs a towel and clothes.
not that i can see because he goes off screen with them.

Video (V)
Dialogue History (H), Caption (C)

C: a man is using a sink and is opening a bottom drawer to take out a red bag.

Q^1: what is happening in the video?
A^1: a man is using the sink and getting a red bag from the bottom drawer.

Q^r: can you hear any sounds?
A^r: you can hear the faucet when he turns it on.

Question (Q^r): does he wear glasses?
Answer (A^r)

Pred: he does not wear glasses.
GT: he has a pair of glasses on during the video.

(a) Text hallucination by non-answerable question

(b) Text hallucination by non-referring video

Figure 1: Illustration of video-grounded dialogue system including incorrect answer generation by (a) non-answerable question and (b) non-referring video.

Figure 2: (a) sentence similarity scores (BLEU score) between input sentences (caption, dialogue, question) and answer sentence (prediction, ground-truth). (b) sentence similarity between input sentences and answer sentences (incorrect predictions, ground-truth), which tells that incorrect answers have made mistakes by hallucinating input sentences.

answer via copying input texts. This is because the system is holding overconfidence in the text hallucination, such that it ignores the meaning of the question and video. Therefore, current VGD systems are prone to rely on language model tainted with incorrect text hallucination, which hinders them from accurately learning question-answer association.

Our manual studies in Figure 2 give experimental evidence that the answer sentences predicted by current VGD systems (Le et al., 2019b; Li et al., 2021) are dependent on indiscriminate text hallucination. Figure 2(a) presents sentence similarity score, BLEU (Papineni et al., 2002), which computes word overlapping between (1) predicted answers and input texts (i.e., caption, dialogue and question), and (2) ground-truth answers and input texts from AVSD validation dataset. The higher scores between predicted answers and input texts explain the reliance on input texts for decoding answer tokens. We may take this for granted, but as shown in Figure 2(b), the problem gets distinguishable when collecting all the ‘incorrect’ predictions. Many failure cases (i.e., incorrect predictions) include that the predicted answers are more similar to input texts, which proves indiscriminate text hallucination without the understanding of given questions and videos.

One straightforward solution to mitigate this indiscriminate text hallucination is to extend the dataset using augmentations or modulating answer descriptions to be more stereoscopic. However, the augmentation has limitations in terms of diversity and the modulated descriptions can be sometimes ad-hoc and unnecessarily extravagant. Intrigued by the current overconfidence in text hallucination of VGD systems, we contrive to build Text Hallucination Mitigating (THAM) framework that mitigates feature-level hallucination effects via introducing information-theoretic regularization. THAM framework incorporates Text Hallucination Regularization (THR) loss derived from the mutual information between the response language model and the proposed hallucination lan-

\footnote{Audio-Visual Scene Aware Dialog (Alamri et al., 2019)}
guage model. Minimizing THR loss contributes to reducing indiscriminate text copying and boosting dialogue performances. THAM validates effectiveness with steady performance gain on top of the current several runner models (Hori et al., 2019a; Le et al., 2019b; Kim et al., 2021; Li et al., 2021) via a model-agnostic approach. Experimental results show state-of-the-art performances on two VGD benchmarks (i.e., AVSD@DSTC7 and AVSD@DSTC8) and enhanced interpretability.

2 Related Work

2.1 Video-grounded Dialogues

Visual Question Answering (VQA) (Antol et al., 2015; Li et al., 2022; Xiao et al., 2022) is one of the proxy tasks for evaluating multi-modal understanding of vision-language systems. The recent success of natural language processing (Devlin et al., 2018; Radford et al., 2019) gives a bridge to advance VQA for video-grounded dialogue (VGD) system (Alamri et al., 2019; Hori et al., 2020), which aims to generate open-ended answer sentences founded on video and dialogue of human. For this VGD, many recurrent neural networks (Nguyen et al., 2015; Li et al., 2022; Xiao et al., 2022) is one of the effective dialogues, and a transformer-based VGD system (Li et al., 2021) has also been introduced to enhance multi-modal interaction between video and text, including word-embedding attention (Lee et al., 2020), hierarchical attention (Le et al., 2019a) and pointer-augmented decoding (Le and Chen, 2020). Furthermore, graph representation is considered to connect common semantics among intra-frames and inter-frames (Geng et al., 2021) and to uncover co-referencing between frames and texts (Kim et al., 2021). However, these systems still suffer from the hallucination problem in generating answer sentences and for this problem, we proposed an information-theoretic text hallucination mitigating framework.

3 Preliminaries

3.1 Estimating Mutual Information

To identify the feature-level text hallucination, we first introduce the mutual information $I(\cdot;\cdot)$, which measures co-dependence between two random variables $X$ and $Y$ over the space $\mathcal{X} \times \mathcal{Y}$ like below:

$$I(X; Y) := H(X) - H(X|Y),$$

where $H(\cdot)$ is the Shannon entropy and $H(X|Y)$ is the conditional entropy of $X$ given $Y$. This mutual information is also equal to the Kullback-Leibler (KL-) divergence $D_{KL}(\cdot||\cdot)$ between joint probability distribution $P_{XY}$ and the product of marginals $P_{X} \otimes P_{Y}$ like below:

$$I(X; Y) = D_{KL}(P_{XY}||P_{X} \otimes P_{Y}),$$

where, given two probability distributions $p(x)$ and $q(x)$ on variable $x$, KL divergence is defined as:

$$D_{KL}(p||q) := \mathbb{E}_{x \sim p}[\log\left(\frac{p(x)}{q(x)}\right)].$$

As the KL divergence increases, the co-dependence between $X$ and $Y$ becomes stronger. However, calculating KL divergence is tractable for only a few cases (i.e., discrete variables), as it is unavailable to hold exact distributions of the training dataset. Recent approach (Belghazi et al., 2018) is performed on estimating mutual information for continuous high-dimensional variables using neural network founded on the Donsker-Varadhan representation (Donsker and Varadhan, 1975) defined below:

$$I(X; Y) \leq I_{\phi}(X; Y) = \sup_{\phi \in \Phi} \mathbb{E}_{P_{XY}}[T_{\phi}] - \log(\mathbb{E}_{P_{X} \otimes P_{Y}}[e^{T_{\phi}}]),$$

where $T_{\phi} : \mathbb{R}^{D} \rightarrow \mathbb{R}$ is a neural network parameterized by $\phi \in \Phi$, and the expectations of $\mathbb{E}_{P_{XY}}$ and $\mathbb{E}_{P_{X} \otimes P_{Y}}$ are approximated by empirical sampling. Thus, maximizing $I_{\phi}(X; Y)$ provides a tight lower bound of original mutual information$^4$I$(X; Y)$.

3.2 Video-grounded Dialogue Task

Video-grounded Dialogue (VGD) aims to produce free-form natural language answer for a given question. In the formal definition of the VGD task (Alamri et al., 2019), VGD system takes answer pairs of previous rounds and caption as inputs and produces answer sentence $a_{r}$, where $v$ is video, $h$ is dialogue history and $q_{r}$ question asked at current round $r \in \{1, \cdots, R\}$. Here, the dialogue history $h = \{e, (q^{1}, a^{1}), \cdots, (q^{r-1}, a^{r-1})\}$ is a set of question-answer pairs of previous rounds and caption $c$ describing the summary of the video. For training of the VGD system, we perform next-word prediction, where it is trained to predict next-word answer token $a_{r}^{t}$ for given inputs of tuples $(v, h, q^{r})$ and partial answer word tokens $a_{r}^{<t}$ before $t$-th.

$^3$It provides a supremum of the KL divergence over all functions $T : D_{KL}(P||Q) = \sup_{T : \mathbb{R}^{D} \rightarrow \mathbb{R}} \mathbb{E}_{P}[T] - \log(\mathbb{E}_{Q}[\exp(T)])$.

$^4$Refer proofs in the appendices.
4 Text Hallucination Mitigating Framework

In Figure 3, to build Text Hallucination Mitigating (THAM) framework, we prepare three different language models composed of encoder-decoder pairs: (1) Response Language Model (RLM), (2) Hallucination Language Model (HLM), and (3) Language Model (LM). RLM is a naive VGD model, such that it is given complete samples of $v, h, q$, and partial answer $a_{r<t}$ to predict the next answer token $a_r$. HLM is designed to generate answer tokens relying on the text hallucination, where HLM is given deficient input texts (i.e., $h, a_{r<t}$) without question, which is unavailable to reason the correct answer and inevitably relies on hallucinating sentence to overlap with ground-truth words via copying input texts without knowledge of the question. Using this HLM, our proposed Text Hallucination Regularization (THR) mitigates feature-level hallucination effects in the RLM via minimizing the mutual information between the features of RLM encoder and hallucinating features of HLM encoder. However, not all the features of HLM are bad, because HLM, as a language model, is also trained to make a grammatically complete sentence, where those grammatical knowledge should be removed before performing THR. Therefore we train another language model (LM), which predicts the next answer token $a_r'$ from only given partial answer $a_{r<t}$. We remove encoder features of LM from those of HLM in advance and apply the THR loss.

4.1 Input representations

We give formal feature definitions of $v$, $h$, $q^r$ and $a^r$ embedded into $d$-dimensional space. Following (Hori et al., 2019b; Li et al., 2021), for the video features, we utilize the I3D model (Carreira and Zisserman, 2017) pre-trained on YouTube videos and the Kinetics dataset (Kay et al., 2017) to get 2048-dimensional rgb features $v_{rgb} \in \mathbb{R}^{L \times 2048}$ and optical flow features $v_{opt} \in \mathbb{R}^{L \times 2048}$ in the images, where $L$ is the number of video frames. Audio features are also available in the video of the AVSD dataset, we get 128-dimensional features $v_{aud} \in \mathbb{R}^{L \times 128}$ using pre-trained VGGish (Hershey et al., 2017). The aforementioned three features are concatenated along feature dimension axis and embedded into $d$-dimensional space as:

$$v = [v_{rgb}||v_{opt}||v_{aud}] W_v \in \mathbb{R}^{L \times d}, \quad (5)$$

where $W_v \in \mathbb{R}^{(2048+2048+128) \times d}$ is $d$-dimensional embedder and $\cdot||\cdot$ denotes concatenation.

For the text features, we follow the T5-base Transformer (Raffel et al., 2020) and tokenize all the sentences (i.e., $q^r, h, a^r$) into a series of Word-Pieces (Wu et al., 2016). The final representations for each sub-word token are obtained by summing up their token embeddings and relative positional embeddings, followed by a layer normalization (Ba et al., 2016). We give formal definitions of them as: history $h \in \mathbb{R}^{L_h \times d}$, question $q \in \mathbb{R}^{L_q \times d}$ and answer $a \in \mathbb{R}^{L_a \times d}$, where $L_h, L_q$ and $L_a$ are the number of tokens of each text.

4.2 Text Hallucination Regularization

Text Hallucination Regularization (THR) is designed for the VGD model (i.e., RLM) to mitigate...
indiscriminate text hallucination (i.e., text or word copying) from input texts without understanding of the question. As we describe the methodology of THAM in Section 4, here, we focused on mathematical formulations for the reproducibility of THAM with proposed THR.

Training language models. To prepare own purpose of three language models (i.e., RLM, HLM, LM), as the first stage, we train them with their defined inputs in the followings. Response Language Model (RLM) is designed for original purpose of VGD, where it is given complete input sample as \( X_{<t} = [v|h|q|a_{<t}] \) and trained to generate next word tokens for answer sentence \( a^r = \{a_1', \ldots, a_m'\} \) with sentence length \( m \) using cross-entropy loss like below:

\[
\mathcal{L}_{RLM}(\theta) = \log \prod_{t=1}^{m} P(a_t'|X_{<t};\theta). \tag{6}
\]

Hallucination Language Model (HLM) is intended to learn reliance on text hallucination effects for generating an answer. To train HLM, we utilize the fact that ground-truth answer sentences of VGD are usually similar to the partial texts of inputs. Therefore, we give the HLM with deficient input texts \( X_{<t} = [v|h|a_{<t}] \) without question like:

\[
\mathcal{L}_{HLM}(\theta^*) = \log \prod_{t=1}^{m} P(a_t'|X_{<t}^*;\theta^*), \tag{7}
\]

where the deficient input texts make it difficult for HLM to perform correct answer reasoning. (See more results in the ablation studies of Table 3.) In the optimization, although the HLM can identify the similarities between partial texts of inputs and ground-truth answers, but it is unavailable to learn why the answers are similar to input texts, which results in training of the text hallucination. Using this overconfidence in text hallucination of HLM, we build Text Hallucination Regularization (THR) loss to mitigate the text hallucinating effect in naive RLM in the following.

Text Hallucination Regularization. Text Hallucination Regularization (THR) is introduced to mitigate indiscriminate text hallucination of VGD models to answer the question. THR loss is defined by feature-level mutual information between RLM and HLM. To this, we first define encoder features of each trained model: (1) RLM’s encoder features as \( F_{<t} = f_{RLM}(X_{<t}, \theta) \in \mathbb{R}^d \) and (2) HLM’s encoder features as \( F_{<t}^* = f_{HLM}(X_{<t}^*, \theta^*) \in \mathbb{R}^d \), where \( f \) denotes the transformer encoders of each model. These two features (i.e., \( F_{<t}, F_{<t}^* \)) are outputs from the position of \( a_{t-1} \) in the transformer. Here, we refer to \( F_{<t} \) as ‘factual’ features and \( F_{<t}^* \) as ‘hallucinating’ features. Our proposed THR aims to hold feature-level independence between factual features and hallucinating features via minimizing mutual information among them. However the grammatical knowledge in \( F_{<t}^* \) to build language sentence still should be correlated with \( F_{<t} \), as both language models are trained from grammatically complete ground-truth language sentences. Thus, we prepare pure language model (LM), which predicts answer token with only given partial answer tokens \( X_{<t}^1 = [a_{<t}] \):

\[
\mathcal{L}_{LM}(\theta^1) = \log \prod_{t=1}^{m} P(a_t^1|X_{<t}^1;\theta^1), \tag{8}
\]

where we get pure language features \( F_{<t}^1 = f_{LM}(X_{<t}^1, \theta^1) \in \mathbb{R}^d \) from the LM’s encoder, which has the only grammatical knowledge to make complete language. We remain pure hallucinating effects via subtracting the language features \( F_{<t}^1 \) from the hallucinating features \( F_{<t}^* \):

\[
G_{<t} = F_{<t}^* - F_{<t}^1 \in \mathbb{R}^d, \tag{9}
\]

where the \( G_{<t} \) is the pure hallucinating (pure-h) features, which hold hallucinating effects without grammatical knowledge. Founded on factual features \( F_{<t} \) and pure-h features \( G_{<t} \), we finally define THR loss. THR loss calculates feature-level mutual information between \( F_{<t} \) and \( G_{<t} \). Thanks to the mutual information neural estimator (MINE) (Belghazi et al., 2018), we get high-dimensional mutual information between the \( F_{<t} \) and the \( G_{<t} \), where we utilize it as THR loss for a regularization:

\[
\mathcal{L}_{THR}(\theta, \phi) = I_{\phi}(f_{RLM}(X_{<t}, \theta); G_{<t}) \tag{10}
\]

By minimizing \( \mathcal{L}_{THR}(\theta, \phi) \) with respect to the parameter \( \theta \), we train the RLM to be independent of HLM’s indiscriminate text hallucination\(^7\). Following the maximizing lower bound of estimated mutual information in Equation 4, the final objective function is formulated as:

\[
\min_{\theta} \max_{\phi} \mathcal{L}_{RLM}(\theta) + \alpha \mathcal{L}_{THR}(\theta, \phi) \tag{11}
\]

where \( \alpha \) is a hyperparameter and the objective function is a minimax problem, we alternate to train and update the parameters \( \theta \) and \( \phi \) in every epoch.

\(^7\)The \( \theta^*, \theta^1 \) in \( G_{<t} \) are different parameters with \( \theta \) in \( F_{<t} \).
### Table 1: Experimental results on the test split of AVSD benchmark at DSTC7 and DSTC8 challenges (B: BLEU, M: METEOR, R: ROUGE-L, C: CIDEr, cp: caption, ⋆: reported in (Kim et al., 2021)).

<table>
<thead>
<tr>
<th>Methods</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>M</th>
<th>R</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Hori et al., 2019a)</td>
<td>0.621</td>
<td>0.480</td>
<td>0.379</td>
<td>0.305</td>
<td>0.217</td>
<td>0.481</td>
<td>0.733</td>
</tr>
<tr>
<td>HMA (Le et al., 2019a)</td>
<td>0.633</td>
<td>0.490</td>
<td>0.386</td>
<td>0.310</td>
<td>0.242</td>
<td>0.515</td>
<td>0.856</td>
</tr>
<tr>
<td>RMFF (Yeh et al., 2019)</td>
<td>0.636</td>
<td>0.510</td>
<td>0.417</td>
<td>0.345</td>
<td>0.224</td>
<td>0.505</td>
<td>0.877</td>
</tr>
<tr>
<td>EE-DMN (Lin et al., 2019)</td>
<td>0.641</td>
<td>0.493</td>
<td>0.388</td>
<td>0.310</td>
<td>0.241</td>
<td>0.527</td>
<td>0.912</td>
</tr>
<tr>
<td>JMAN (Chu et al., 2020)</td>
<td>0.667</td>
<td>0.521</td>
<td>0.413</td>
<td>0.334</td>
<td>0.239</td>
<td>0.533</td>
<td>0.941</td>
</tr>
<tr>
<td>FA-HRED (Nguyen et al., 2019)</td>
<td>0.695</td>
<td>0.553</td>
<td>0.444</td>
<td>0.360</td>
<td>0.249</td>
<td>0.544</td>
<td>0.997</td>
</tr>
<tr>
<td>CMU (Sanabria et al., 2019)</td>
<td>0.718</td>
<td>0.584</td>
<td>0.478</td>
<td>0.394</td>
<td>0.267</td>
<td>0.563</td>
<td>1.094</td>
</tr>
<tr>
<td>MSTN (Lee et al., 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.377</td>
<td>0.275</td>
<td>0.566</td>
<td>1.115</td>
</tr>
<tr>
<td>JSTL (Hori et al., 2019c) w/o cp</td>
<td>0.675</td>
<td>0.543</td>
<td>0.446</td>
<td>0.371</td>
<td>0.248</td>
<td>0.527</td>
<td>0.966</td>
</tr>
<tr>
<td>JSTL (Hori et al., 2019c)</td>
<td>0.727</td>
<td>0.593</td>
<td>0.488</td>
<td>0.405</td>
<td>0.273</td>
<td>0.566</td>
<td>1.118</td>
</tr>
<tr>
<td>MTN* (Le et al., 2019b)</td>
<td>0.731</td>
<td>0.597</td>
<td>0.490</td>
<td>0.406</td>
<td>0.271</td>
<td>0.564</td>
<td>1.127</td>
</tr>
<tr>
<td>MTN-P (Le and Chen, 2020)</td>
<td>0.750</td>
<td>0.619</td>
<td>0.514</td>
<td>0.427</td>
<td>0.280</td>
<td>0.580</td>
<td>1.189</td>
</tr>
<tr>
<td>VGNMN (Le et al., 2022)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.429</td>
<td>0.278</td>
<td>0.578</td>
<td>1.188</td>
</tr>
<tr>
<td>SCGA (Kim et al., 2021)</td>
<td>0.745</td>
<td>0.622</td>
<td>0.517</td>
<td>0.430</td>
<td>0.285</td>
<td>0.578</td>
<td>1.201</td>
</tr>
<tr>
<td>RLM (Li et al., 2021)</td>
<td>0.765</td>
<td>0.643</td>
<td>0.543</td>
<td>0.459</td>
<td>0.294</td>
<td>0.606</td>
<td>1.308</td>
</tr>
<tr>
<td>T5RLM (Ours)</td>
<td>0.767</td>
<td>0.644</td>
<td>0.542</td>
<td>0.461</td>
<td>0.296</td>
<td>0.608</td>
<td>1.311</td>
</tr>
<tr>
<td>THAM (T5RLM + THR loss)</td>
<td>0.778</td>
<td>0.654</td>
<td>0.549</td>
<td>0.468</td>
<td>0.308</td>
<td>0.619</td>
<td>1.335</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>M</th>
<th>R</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDMN (Xie and Iacobacci, 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.296</td>
<td>0.214</td>
<td>0.496</td>
<td>0.761</td>
</tr>
<tr>
<td>JMAN (Chu et al., 2020)</td>
<td>0.645</td>
<td>0.504</td>
<td>0.402</td>
<td>0.324</td>
<td>0.232</td>
<td>0.521</td>
<td>0.875</td>
</tr>
<tr>
<td>STSGR (Geng et al., 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.357</td>
<td>0.267</td>
<td>0.553</td>
<td>1.004</td>
</tr>
<tr>
<td>MSTN (Lee et al., 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.385</td>
<td>0.270</td>
<td>0.564</td>
<td>1.073</td>
</tr>
<tr>
<td>MTN-P (Le and Chen, 2020)</td>
<td>0.701</td>
<td>0.587</td>
<td>0.494</td>
<td>0.419</td>
<td>0.263</td>
<td>0.564</td>
<td>1.097</td>
</tr>
<tr>
<td>SCGA (Kim et al., 2021) w/o cp</td>
<td>0.675</td>
<td>0.559</td>
<td>0.459</td>
<td>0.377</td>
<td>0.269</td>
<td>0.555</td>
<td>1.024</td>
</tr>
<tr>
<td>SCGA (Kim et al., 2021)</td>
<td>0.711</td>
<td>0.593</td>
<td>0.497</td>
<td>0.416</td>
<td>0.276</td>
<td>0.566</td>
<td>1.123</td>
</tr>
<tr>
<td>RLM (Li et al., 2021)</td>
<td>0.746</td>
<td>0.626</td>
<td>0.528</td>
<td>0.445</td>
<td>0.286</td>
<td>0.598</td>
<td>1.240</td>
</tr>
<tr>
<td>T5RLM (Ours)</td>
<td>0.749</td>
<td>0.631</td>
<td>0.529</td>
<td>0.445</td>
<td>0.290</td>
<td>0.600</td>
<td>1.263</td>
</tr>
<tr>
<td>THAM (T5RLM + THR loss)</td>
<td>0.764</td>
<td>0.641</td>
<td>0.538</td>
<td>0.455</td>
<td>0.301</td>
<td>0.610</td>
<td>1.304</td>
</tr>
</tbody>
</table>

## 5 Experiments

### 5.1 Datasets

**AVSD@DSTC7 and AVSD@DSTC8.** (Audio-Visual Scene Aware Dialog) (Alamri et al., 2019; Hori et al., 2020) is a popular benchmark dataset for VGD, where each dialogue includes 10 pairs of question and answer for one video. The video is collected from Charades (Sigurdsson et al., 2016) human-activity dataset and has a short description summarizing overall scenes in the video. AVSD@DSTC 7 and 8 are released for Dialogue System Technology Challenge (DSTC), where AVSD@DSTC7 contains 7,659, 1,787, and 1,710 dialogues for training, validation and test, but AVSD@DSTC8 is only provided with 1,710 dialogues for test in the second challenge. For test-set evaluation, 6 reference answers are provided.

### 5.2 Metrics

We follow official natural language generation metrics for AVSD benchmark (i.e., BLEU, METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015)). The metrics are provided by challenge organizers and formulated to compute the word overlapping between each generated answer and reference answer.

### 5.3 Results on AVSD benchmark

Table 1 summarizes the experimental results on the AVSD dataset. THAM is compared to several previous results of VGD systems (Please refer the descriptions about these VGD systems in the...)

---

8github.com/dialogtekgeek/DSTC8-A VSD_official
Table 2: Experimental results on the test split of AVSD benchmark at DSTC7 and DSTC8 challenges for applying THR loss on VGD runner models (B1 = BELU1, *: reconstruction-based results, †: single reference results).

<table>
<thead>
<tr>
<th>Methods</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Hori et al., 2019a)</td>
<td>0.621</td>
<td>0.480</td>
<td>0.379</td>
<td>0.305</td>
<td>0.217</td>
<td>0.481</td>
<td>0.733</td>
</tr>
<tr>
<td>Baseline + THR loss</td>
<td>0.635</td>
<td>0.495</td>
<td>0.388</td>
<td>0.313</td>
<td>0.230</td>
<td>0.492</td>
<td>0.762</td>
</tr>
<tr>
<td>MTN† (Le et al., 2019b)</td>
<td>0.357</td>
<td>0.241</td>
<td>0.173</td>
<td>0.128</td>
<td>0.162</td>
<td>0.355</td>
<td>1.249</td>
</tr>
<tr>
<td>MTN† + THR loss</td>
<td>0.371</td>
<td>0.252</td>
<td>0.181</td>
<td>0.136</td>
<td>0.175</td>
<td>0.374</td>
<td>1.265</td>
</tr>
<tr>
<td>SCGA* (Kim et al., 2021)</td>
<td>0.746</td>
<td>0.618</td>
<td>0.514</td>
<td>0.428</td>
<td>0.283</td>
<td>0.575</td>
<td>1.193</td>
</tr>
<tr>
<td>SCGA* + THR loss</td>
<td>0.758</td>
<td>0.629</td>
<td>0.522</td>
<td>0.430</td>
<td>0.295</td>
<td>0.587</td>
<td>1.214</td>
</tr>
<tr>
<td>RLM (Li et al., 2021)</td>
<td>0.765</td>
<td>0.643</td>
<td>0.543</td>
<td>0.459</td>
<td>0.294</td>
<td>0.606</td>
<td>1.308</td>
</tr>
<tr>
<td>RLM + THR loss</td>
<td>0.775</td>
<td>0.651</td>
<td>0.551</td>
<td>0.465</td>
<td>0.305</td>
<td>0.616</td>
<td>1.331</td>
</tr>
<tr>
<td>T5RLM (Ours)</td>
<td>0.767</td>
<td>0.644</td>
<td>0.542</td>
<td>0.461</td>
<td>0.296</td>
<td>0.608</td>
<td>1.311</td>
</tr>
<tr>
<td>T5RLM + THR loss</td>
<td>0.778</td>
<td>0.654</td>
<td>0.549</td>
<td>0.468</td>
<td>0.308</td>
<td>0.619</td>
<td>1.335</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on variants of HLM to learn indiscriminate text hallucination from different text inputs on the valid split of AVSD@DSTC7. (single reference)

<table>
<thead>
<tr>
<th>Input variants on HLM</th>
<th>BELU1</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X^*<em>t = [h][a</em>{&lt;t}])</td>
<td>0.324</td>
<td>1.513</td>
</tr>
<tr>
<td>(X^*<em>t = [q][a</em>{&lt;t}])</td>
<td>0.289</td>
<td>1.329</td>
</tr>
<tr>
<td>(X^*<em>t = [v][a</em>{&lt;t}])</td>
<td>0.275</td>
<td>1.215</td>
</tr>
<tr>
<td>(X^*_t = [v</td>
<td>h][a_{&lt;t}])</td>
<td>0.309</td>
</tr>
<tr>
<td>(X^*<em>t = [h][q][a</em>{&lt;t}])</td>
<td>0.279</td>
<td>1.306</td>
</tr>
</tbody>
</table>

Table 3 summarizes the THAM results on input variants of HLM. HLM is designed to build excessive text conjugating language models via giving inputs that can not infer the correct answer. In the optimization, it is just optimized to learn spurious

5.4 Ablation Study

Figure 4: Illustration of THR loss (left) and cross-entropy loss (right) along the epoch on valid split of AVSD@DSTC7 with and without subtracting the encoder features of LM from the encoder features of HLM.
and an older guy walks past a boy and then washes something. he -0.15 0.35 -0.10 -0.05 0.0 0.05 0.10 0.15 0.20 0.25 0.30

(a) Distributions between $F_{<1}$ and $G_{<1}$. (b) Distributions between $F_{<1}$ and $G_{<1}$. (c) Distributions between $F_{<1}$ and $G_{<1}$. (d) Distributions between $F_{<1}$ and $G_{<1}$. (e) Distributions between $F_{<1}$ and $G_{<1}$. (f) Distributions between $F_{<1}$ and $G_{<1}$. (g) Distributions between $F_{<1}$ and $G_{<1}$. (h) Distributions between $F_{<1}$ and $G_{<1}$. (i) Distributions between $F_{<1}$ and $G_{<1}$.

Figure 5: Joint distributions of encoder features between (a) RLM ($F_{<1}$) and HLM ($F^*_{<1}$), (b) RLM ($F_{<1}$) and HLM with subtracting LM ($G_{<1}$). (a) shows correlations with $F^*_{<1}$ by grammatical knowledge in HLM, and (b) shows relatively independent distributions by $G_{<1}$.

correlations between inputs $X_{<t}$ and outputs $a^r$. Introducing history only for the inputs of HLM shows the most effectiveness. We consider this is because the history (i.e., dialogue history) contains a relatively large amount of texts, but without question, it is just captions that can not infer the answer. Here, the HLM inevitably learns indiscriminate text hallucination as HLM does not know the question: text hallucination as a result of copying a sentence from input sentences can lead to greater overlap with the ground-truth answer than simply generating an answer without knowing the question. Conversely, we also devise the HLM with an input of question without history, which was not effective in THAM performance. We consider that this is because the AVSD dataset includes some samples, where the correct answer can be easily inferred from a question alone without any other modalities, thus text conjugating on the question should be beneficial.

Figure 4 shows THR loss $L_{THR}(\theta, \phi)$ and cross-entropy loss $L_{RLM}(\theta)$ from ablation studies with and without subtracting the encoder features of LM from the encoder features of HLM. THR loss explains the mutual information $I_\phi(F_{<t}, G_{<t})$ between RLM and HLM, and the minimization of it regularizes indiscriminate text hallucination existing in RLM. For the case ‘with subtracting LM’, it shows that both $L_{THR}(\theta, \phi)$ and $L_{RLM}(\theta)$ decrease and converge according to the epoch. However, for the case ‘without subtracting LM’, minimizing the $L_{THR}(\theta, \phi)$ hinders the convergence of $L_{RLM}(\theta)$. This is because the encoder features that contribute a sentence are in both RLM and HLM, minimizing $L_{THR}(\theta, \phi)$ without removing them from HLM becomes adversarial with learning from cross-entropy loss, which degrades the performance of the VGD system.

5.5 Qualitative Results

Figure 5 gives joint distributions among the language models’ encoder features. Here, the RLM is fully trained from THAM framework. From 512 samples of AVSD validation set, we select a single value among the $d$-dimensional space at the same position of each encoder feature (i.e., $F_{<t}, F^*_{<t}, G_{<t}$). Figure 5(a) summarizes joint plots between $F_{<t}$ and $F^*_{<t}$, where the correlations are confirmed due to the common grammatical knowledge from language models. However Figure 5(b) shows uncorrelated distributions between $F_{<t}$ and $G_{<t}$, which means the grammatical knowledge is properly removed from $G_{<t}$. Figure 6 gives responses of naive RLM and THAM (naive RLM + THR loss). For the question of “what are they both wearing”, naive RLM shows the reliance on texts from history without understanding of the question. However, the THAM is generating correct answer sentence pertinent to the given question.

Conclusion

Text Hallucination Mitigating framework is proposed for Video-grounded Dialogue. THAM considers the text hallucination problem, which copies input texts for answer generation without understanding of the question. THAM framework incorporates Text Hallucination Regularization loss derived from proposed information-theoretic text hallucination measurement approach. Empirical
results on VGD benchmarks show that THAM achieves state-of-the-art performances and effectiveness.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics). We would also like to thank the anonymous reviewers for their feedback.

Limitations

The limitations of the Text Hallucination Mitigating Framework are as follows. First, our empirical analysis provides that THAM is facing a failure case about the question of sounds. In Figure 7 in supplemental materials, for the question of “what kind of noise”, THAM is hallucinating response without understanding the question. Although the answer “I can hear some noise” can be plausible, but it also seems just hallucinating by copying from history texts. We speculate this is because the sound features contain less information (128 dimensions) comparing to video (2048 dimensions), which requires more specialized attention (e.g., fine-grained audio processing). For the second limitation, THAM is based on two-stage training mechanism. To perform mitigation of text hallucination, pre-training of each language model is required as a first-stage training. To overcome the aforementioned limitations, we will perform further studies and make an effort on video interpretability improvements.

Ethics Statement

As one of the interactive AI, the Video-grounded Dialogue system is designed for providing assistance to various subsections of our environments including education, entertainment, and visual impairments. Our proposed Text Hallucination Mitigation Framework have contributed to improving response qualities and alleviating abnormalities in the system. We also consider the potential negative societal impact that those who are aware of the VGD system can deliberately manipulate it to get prohibited information. Furthermore, to apply the VGD system in the real environment, fairness and bias issues of dialogue systems should also be addressed.

References


Hung Le, Steven C.H. Hoi, Doyen Sahoo, and Nancy F. Chen. 2019a. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In DSTC7 at AAAI2019 workshop.


Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. AAAI.


A Training Details.

Training. THAM is trained on NVIDIA TITAN V (12GB of memory) GPU with Adam optimizer (Kingma and Ba, 2014) with \( \beta_1 = 0.9, \beta_2 = 0.99, \) and \( \epsilon = 10^{-8} \). We utilized the piece-wise linearly decreased learning rate from 6.25e-4 to 0 and set the learning rate warm-up strategy to 10,000 training steps and trained the model up to 20 epochs. In Section 4.1, the interpolation is conducted via the window overlapping method. The first-stage training is performed on three language models (i.e., RLM, HLM, LM) respectively with a batch size of 8 and a dropout rate of 0.3. For the \( d \)-dimensional space, all language models use \( d=768 \). The second-stage training is performed on RLM with THR loss with the same batch size and dropout rate with the first training. The best model is decided by the lowest validation loss on the validation set with \( \alpha = 0.01 \) in equation (11) of the main paper on the setting \( X_{\leq t} = [h|a_{<t}] \). The training takes about 5 hours to be fully optimized at the losses of about 0.184 on training and 0.284 on validation. Inference time for generating the answer for a single question takes about 2 seconds. Our model is not performed on hyperparameter searching for model fine-tuning.

Inference. In the inference, answer generation adopts a beam search with a beam size of 5 and a length penalty of 1.0, where the maximum length of sentence is set to 30. Every performance of THAM in table 1 and 2 of the main paper is averaging from 5 times random seed validation.

B Donsker-Varadhan representation.

For the probability distribution of \( P \) and \( Q \), the KL divergence admits the following dual representation as:

\[
D_{KL}(P||Q) = \sup_{T: \Omega \rightarrow R} E_P[T] - \log(E_Q[e^T]), \tag{12}
\]

where \( \Omega \) is high-dimensional variables and the supremum is taken over all functions \( T \) such that the two expectations are finite. The proof for this representation is given as follows. For a given function \( T \), consider the Gibbs distribution \( G \) define by \( dG = \frac{1}{Z}e^T dQ \), where \( Z = E_Q[e^T] \). For the construction, we are available to derive 10 as:

\[
E_P[T] - \log Z = E_P[\log \frac{dG}{dQ}] \tag{13}
\]

Let \( \Delta \) be the gap as:

\[
\Delta := D_{KL}(P||Q) - (E_P[T] - \log(E_Q[e^T])) \tag{14}
\]

Using the Equation (2), we can write \( \Delta \) as KL-divergence:

\[
\Delta = E_P[\log \frac{dP}{dQ} - \log \frac{dG}{dQ}] = E_P \log \frac{dP}{dQ} = D_{KL}(P||G) \tag{15}
\]

The positivity of the KL-divergence gives \( \Delta \geq 0 \). Therefore, we are able to show that for any \( T \),

\[
D_{KL}(P||Q) \geq E_P[T] - \log(E_Q[e^T]), \tag{16}
\]

and the inequality is preserved via taking the supremum over the right-hand side, where the identity of Equation (4) also shows that the bound is tight whenever \( G = P \), for optimal functions \( T^* \) taking the form \( T^* \log \frac{dP}{dQ} + C \) for some constant \( C \in \mathbb{R} \).

C Failure case

We also confirmed that the proposed THAM is fragile to the questions of asking sounds in the video, where it copies the input texts of “I can hear some noise” from history texts in Figure 7. While we admit that the above case can produce semantically correct answers, we feel that the VGD systems should be able to generate more rich answers using their own languages. Furthermore, the sound features contain less information (128 dimensions) compared to video (2048 dimensions), which requires more specialized attention.

Figure 7: Failure case on question about sounds

Video (V)

Dialogue History (H) Video Caption (C)

\( C \): a person is opening a jar of food after taking it off the shelf in the kitchen, they go to the dining room while eating, sitting down

\( Q^1 \): how many people in the video?

\( A^1 \): only one person in the whole video

\( Q^2 \): is there sound to the video?

\( A^2 \): I can hear some noise

\( GT \): the woman eating sounds so

Failure case: Questions about sounds