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: $H = \{(Q^1, A^1), ..., (Q^{r-1}, A^{r-1})\}$, 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

*Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4182–4193
December 7-11, 2022

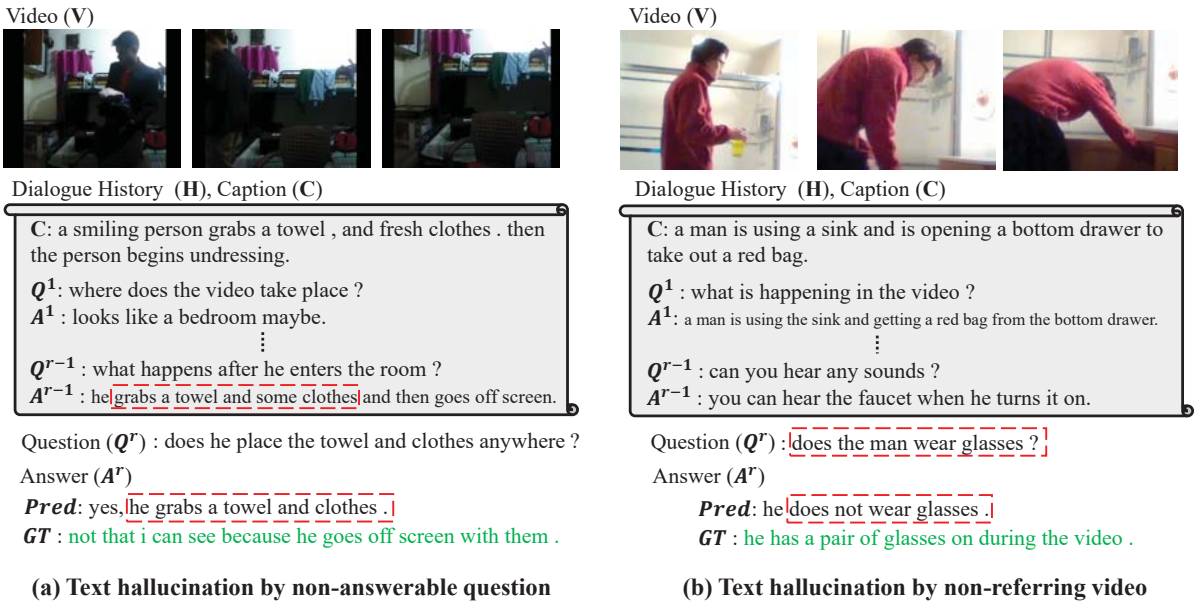


Figure 1: Illustration of video-grounded dialogue system including incorrect answer generation by (a) non-answerable question and (b) non-referring video.

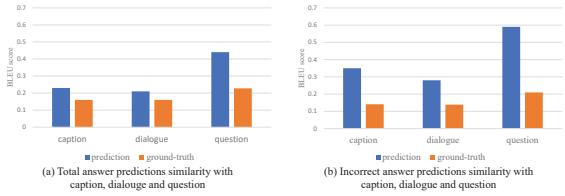


Figure 2: (a) sentence similarity scores (BLEU score) between input sentences (caption, dialogue, question) and answer sentence (prediction, ground-truth). (b) sentence similarity between input sentences and answer sentences (incorrect predictions, ground-truth), which tells that incorrect answers have 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[1] 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'[2] 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-

[1]Audio-Visual Scene Aware Dialog (Alamri et al., 2019)

[2]Here, we regard predictions with a BLEU score of less than 0.1 as 'incorrect'.

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 sentence founded on video and dialogue of human. For this VGD, many recurrent neural networks (Nguyen et al., 2019; Sanabria et al., 2019) have been proposed to hold meaningful semantics along the consecutive 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:

$$I(X;Y) := H(X) - H(X|Y), \tag{1}$$

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:

$$I(X;Y) = D_{KL}(P_{XY}||P_X \otimes P_Y), \tag{2}$$

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

$$D_{KL}(p||q) := \mathbb{E}_{x\sim p}[\log(\frac{p(x)}{q(x)})]. \tag{3}$$

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[3] (Donsker and Varadhan, 1975) defined below:

$$\begin{aligned} I(X;Y) &\leq I_\phi(X;Y) \\ &= \sup_{\phi\in\Phi}\mathbb{E}_{P_{XY}}[T_\phi] - \log(\mathbb{E}_{P_X\otimes P_Y}[e^{T_\phi}]), \end{aligned} \tag{4}$$

where $T_\phi : \mathbb{R}^D \to \mathbb{R}$ 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 tuples (v, h, q^r) 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 $h = \{c, (q^1, a^1), \cdots, (q^{r-1}, a^{r-1})\}$ 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 t-th answer word 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 : $D_{KL}(P||Q) = \sup_{T:\mathbb{R}^D\to\mathbb{R}}\mathbb{E}_P[T] - log(\mathbb{E}_Q[\exp(T)])$.

[4] Refer proofs in the appendices.

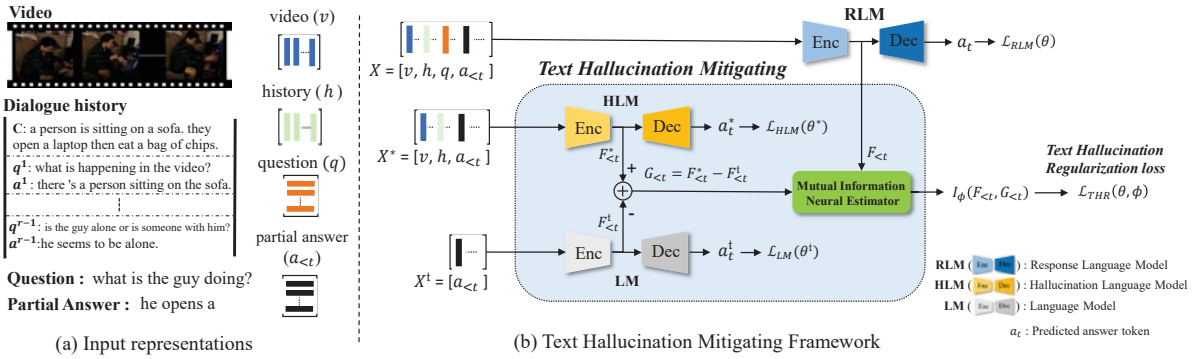


Figure 3: Illustration of Text Hallucination Mitigating Framework (THAM) for video-grounded dialogue. THAM mitigates feature-level hallucination effects in Response Language Model via introducing Text Hallucination Regularization (THR) loss, where THR aims to minimize mutual information between encoder features of RLM and features from Hallucination Language Model.

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_t. 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_t 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 $\mathbf{v}_{rgb} \in \mathbb{R}^{L\times 2048}$ and optical flow features $\mathbf{v}_{opt} \in \mathbb{R}^{L\times 2048}$ in the images, where the L is the number of video frames. Audio features are also available in the video of the AVSD dataset, we get 128-dimensional features[5] $\mathbf{v}_{aud} \in \mathbb{R}^{L\times 128}$ 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:

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

where $W_{\mathbf{v}} \in \mathbb{R}^{(2048+2048+128)\times d}$ is d-dimensional embbeder 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 $\mathbf{h} \in \mathbb{R}^{L_\mathbf{h}\times d}$, question $\mathbf{q} \in \mathbb{R}^{L_\mathbf{q}\times d}$ and answer $\mathbf{a} \in \mathbb{R}^{L_\mathbf{a}\times d}$, where $L_\mathbf{h}, L_\mathbf{q}$ and $L_\mathbf{a}$ are the number of tokens of each text[6].

4.2 Text Hallucination Regularization

Text Hallucination Regularization (THR) is designed for the VGD model (*i.e.*, RLM) to mitigate

[5] Interpolation is considered for audio features to be synchronized with video features.

[6] We delete superscript r in the notations of features for simplicity.

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 $X_{<t} = [\mathbf{v}||\mathbf{h}||\mathbf{q}||\mathbf{a}_{<t}]$ and trained to generate next word tokens for answer sentence $a^r = \{a_1^r, \cdots, a_m^r\}$ with sentence length m using cross-entropy loss like below:

$$\mathcal{L}_{RLM}(\theta) = \log \prod_{t=1}^{m} P(a_t^r | X_{<t}; \theta). \tag{6}$$

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 $X_{<t}^{\star} = [\mathbf{v}||\mathbf{h}||\mathbf{a}_{<t}]$ without question like:

$$\mathcal{L}_{HLM}(\theta^{\star}) = \log \prod_{t=1}^{m} P(a_t^r | X_{<t}^{\star}; \theta^{\star}), \tag{7}$$

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 $F_{<t}^{\star} = f_{HLM}(X_{<t}^{\star}, \theta^{\star}) \in \mathbb{R}^d$, where f denotes the transformer encoders of each model. These two features (*i.e.*, $F_{<t}, F_{<t}^{\star}$) are outputs from the position of $\mathbf{a}_{t-1}$ in the transformer. Here, we refer to $F_{<t}$ as 'factual' features and $F_{<t}^{\star}$ 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 $F_{<t}^{\star}$ 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 $X_{<t}^{\dagger} = [\mathbf{a}_{<t}]$:

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

where we get pure language features $F_{<t}^{\dagger} = f_{LM}(X_{<t}^{\dagger}, \theta^{\dagger}) \in \mathbb{R}^d$ 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 $F_{<t}^{\dagger}$ from the hallucinating features $F_{<t}^{\star}$:

$$G_{<t} = F_{<t}^{\star} - F_{<t}^{\dagger} \in \mathbb{R}^d, \tag{9}$$

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 θ, 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 α is a hyperparameter and the objective function is a minimax problem, we alternate to train and update the parameters θ and ϕ in every epoch.

[7]The $\theta^{\star}$, $\theta^{\dagger}$ in $G_{<t}$ are different parameters with θ in $F_{<t}$.

Table 1: Experimental results on the test split of AVSD benchmark at DSTC7 and DSTC8 challenges (B: BELU, M: METEOR, R: ROUGE-L, C: CIDEr, cp: caption, ⋆: reported in (Kim et al., 2021)).

Methods	B1	B2	B3	B4	M	R	C
AVSD@DSTC7							
Baseline (Hori et al., 2019a)	0.621	0.480	0.379	0.305	0.217	0.481	0.733
HMA (Le et al., 2019a)	0.633	0.490	0.386	0.310	0.242	0.515	0.856
RMFF (Yeh et al., 2019)	0.636	0.510	0.417	0.345	0.224	0.505	0.877
EE-DMN (Lin et al., 2019)	0.641	0.493	0.388	0.310	0.241	0.527	0.912
JMAN (Chu et al., 2020)	0.667	0.521	0.413	0.334	0.239	0.533	0.941
FA-HRED (Nguyen et al., 2019)	0.695	0.553	0.444	0.360	0.249	0.544	0.997
CMU (Sanabria et al., 2019)	0.718	0.584	0.478	0.394	0.267	0.563	1.094
MSTN (Lee et al., 2020)	-	-	-	0.377	0.275	0.566	1.115
JSTL (Hori et al., 2019c) w/o cp	0.675	0.543	0.446	0.371	0.248	0.527	0.966
JSTL (Hori et al., 2019c)	0.727	0.593	0.488	0.405	0.273	0.566	1.118
MTN⋆ (Le et al., 2019b)	0.731	0.597	0.490	0.406	0.271	0.564	1.127
MTN-P (Le and Chen, 2020)	0.750	0.619	0.514	0.427	0.280	0.580	1.189
VGNMN (Le et al., 2022)	-	-	-	0.429	0.278	0.578	1.188
SCGA (Kim et al., 2021)	0.745	0.622	0.517	0.430	0.285	0.578	1.201
RLM (Li et al., 2021)	0.765	0.643	0.543	0.459	0.294	0.606	1.308
T5RLM (Ours)	0.767	0.644	0.542	0.461	0.296	0.608	1.311
THAM (T5RLM + THR loss)	**0.778**	**0.654**	**0.549**	**0.468**	**0.308**	**0.619**	**1.335**
AVSD@DSTC8							
MDMN (Xie and Iacobacci, 2020)	-	-	-	0.296	0.214	0.496	0.761
JMAN (Chu et al., 2020)	0.645	0.504	0.402	0.324	0.232	0.521	0.875
STSGR (Geng et al., 2020)	-	-	-	0.357	0.267	0.553	1.004
MSTN (Lee et al., 2020)	-	-	-	0.385	0.270	0.564	1.073
MTN-P (Le and Chen, 2020)	0.701	0.587	0.494	0.419	0.263	0.564	1.097
SCGA (Kim et al., 2021) w/o cp	0.675	0.559	0.459	0.377	0.269	0.555	1.024
SCGA (Kim et al., 2021)	0.711	0.593	0.497	0.416	0.276	0.566	1.123
RLM (Li et al., 2021)	0.746	0.626	0.528	0.445	0.286	0.598	1.240
T5RLM (Ours)	0.749	0.631	0.529	0.445	0.290	0.600	1.263
THAM (T5RLM + THR loss)	**0.764**	**0.641**	**0.538**	**0.455**	**0.301**	**0.610**	**1.304**

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[8] 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

[8] github.com/dialogtekgeek/DSTC8-AVSD_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).

Methods	B1	B2	B3	B4	METEOR	ROUGE-L	CIDEr
AVSD@DSTC7							
Baseline (Hori et al., 2019a)	0.621	0.480	0.379	0.305	0.217	0.481	0.733
Baseline + THR loss	0.635	0.495	0.388	0.313	0.230	0.492	0.762
MTN† (Le et al., 2019b)	0.357	0.241	0.173	0.128	0.162	0.355	1.249
MTN† + THR loss	0.371	0.252	0.181	0.136	0.175	0.374	1.265
SCGA⋆ (Kim et al., 2021)	0.746	0.618	0.514	0.428	0.283	0.575	1.193
SCGA⋆ + THR loss	0.758	0.629	0.522	0.430	0.295	0.587	1.214
RLM (Li et al., 2021)	0.765	0.643	0.543	0.459	0.294	0.606	1.308
RLM + THR loss	0.775	0.651	**0.551**	0.465	0.305	0.616	1.331
T5RLM (Ours)	0.767	0.644	0.542	0.461	0.296	0.608	1.311
T5RLM + THR loss	**0.778**	**0.654**	0.549	**0.468**	**0.308**	**0.619**	**1.335**
AVSD@DSTC8							
MTN⋆ (Le et al., 2019b)	0.691	0.570	0.471	0.402	0.252	0.549	1.043
MTN⋆ + THR loss	0.707	0.582	0.481	0.409	0.265	0.563	1.079
SCGA⋆ (Kim et al., 2021)	0.706	0.587	0.498	0.412	0.277	0.563	1.113
SCGA⋆ + THR loss	0.727	0.603	0.507	0.425	0.289	0.581	1.169
RLM (Li et al., 2021)	0.746	0.626	0.528	0.445	0.286	0.598	1.240
RLM + THR loss	0.762	0.639	0.537	0.452	0.299	0.607	1.287
T5RLM (Ours)	0.749	0.631	0.529	0.445	0.290	0.600	1.263
T5RLM + THR loss	**0.764**	**0.641**	**0.538**	**0.455**	**0.301**	**0.610**	**1.304**

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)

Input variants on HLM	BELU1	CIDEr
$X^{\star}_{<t} = [\mathbf{h}\|\|\mathbf{a}_{<t}]$	0.324	1.513
$X^{\star}_{<t} = [\mathbf{q}\|\|\mathbf{a}_{<t}]$	0.289	1.329
$X^{\star}_{<t} = [\mathbf{v}\|\|\mathbf{a}_{<t}]$	0.275	1.215
$X^{\star}_{<t} = [\mathbf{v}\|\|\mathbf{h}\|\|\mathbf{a}_{<t}]$	0.309	1.482
$X^{\star}_{<t} = [\mathbf{h}\|\|\mathbf{q}\|\|\mathbf{a}_{<t}]$	0.279	1.306

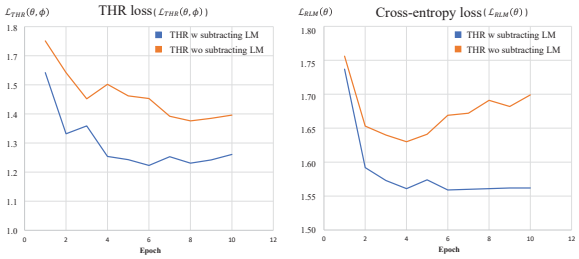
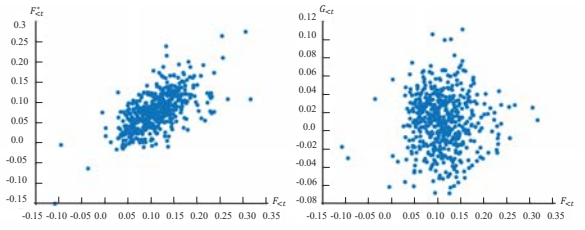


Figure 4: Illustration of THR loss (left) and cross-entropy loss (right) along the epoch on valid split of AVSD@DSTC7 with and without subtracting the encoder features of LM from the encoder features of HLM

sec 2.1 of the Related Work.), where the performances of the official six references are evaluated on AVSD@DSTC7 and AVSD@DSTC8. To validate the effectiveness of proposed our THR loss, we report performances of our naive VGD model (*i.e.*, RLM) based on the T5 Transformer (Raffel et al., 2020). Here, we use 'T5RLM' for the terminology of our RLM to avoid confusion with RLM in (Li et al., 2021) based on GPT2 Transformer (Radford et al., 2019). In the method, we select a Transformer-base encoder for THAM for its simplicity. However, as our framework can be applied to any other VGD systems in a model-agnostic manner, we also validate its effectiveness on recent runner VGD models in Table 2. In detail, we reproduce the MTN, SCGA and RLM from their public papers and codes. For the MTN, we measure predicted answers with a single reference following the original work of it. On top of VGD models, THR loss show steady performance gain on both AVSD datasets.

5.4 Ablation Study

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



(a) Distributions between $F_{<t}$ and $F^{\star}_{<t}$ (b) Distributions between $F_{<t}$ and $G_{<t}$

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

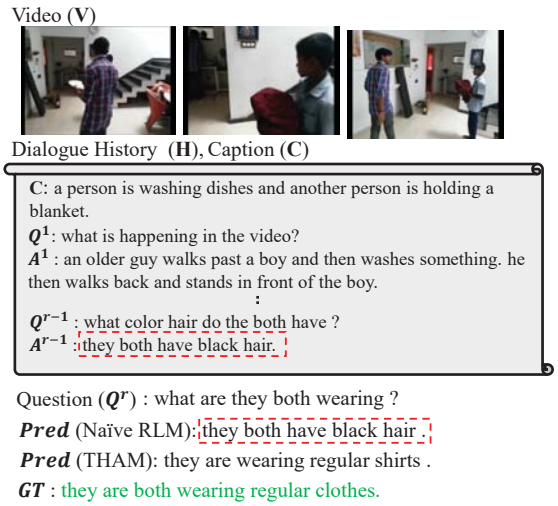


Figure 6: Response comparison between naive RLM and THAM on validation set of AVSD@DSTC7.

correlations between inputs $X^{\star}_{<t}$ 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 $\mathcal{L}_{THR}(\theta, \phi)$ and cross-entropy loss $\mathcal{L}_{RLM}(\theta)$ 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 $\mathcal{L}_{THR}(\theta, \phi)$ and $\mathcal{L}_{RLM}(\theta)$ decrease and converge according to the epoch. However, for the case 'without subtracting LM'[9], minimizing the $\mathcal{L}_{THR}(\theta, \phi)$ hinders the convergence of $\mathcal{L}_{RLM}(\theta)$. This is because the encoder features that contribute a sentence are in both RLM and HLM, minimizing $\mathcal{L}_{THR}(\theta, \phi)$ without removing them from HLM becomes adversarial with learning from cross-entropy loss, which degrades the performance of the VGD system.

[9] $\mathcal{L}_{THR}(\theta, \phi) = I_{\phi}(F_{<t}, F^{\star}_{<t})$ for THR loss without LM

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.*, $F_{<t}, F^{\star}_{<t}, G_{<t}$). Figure 5(a) summarizes joint plots between $F_{<t}$ and $F^{\star}_{<t}$, 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

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. 2020. Multi-step joint-modality attention network for scene-aware dialogue system. In *DSTC8 at AAAI2020 workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Monroe D Donsker and SR Srinivasa Varadhan. 1975. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):183–212.

Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. 2021. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In *Proc. AAAI Conference on Artificial Intelligence*.

Shijie Geng, Peng Gao, Tim Marks, Chiori Hori, and Anoop Cherian. 2020. Spatio-temporal scene grpah reasoning for audio visual scene-aware dialog at dstc8. In *DSTC8 at AAAI2020 workshop*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. 2019a. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019b. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE.

Chiori Hori, Anoop Cherian, Takaaki Hori, and Tim K. Marks. 2020. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc8. In *DSTC8 at AAAI2020 workshop*.

Chiori Hori, Anoop Cherian, Tim K. Marks, and Takaaki Hori. 2019c. Joint student-teacher learning for audio-visual scene-aware dialog. In *Proceedings of the Interspeech*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang D Yoo. 2021. Structured co-reference graph attention for video-grounded dialogue. *arXiv preprint arXiv:2103.13361*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hung Le, Nancy Chen, and Steven Hoi. 2022. Vgnmn: Video-grounded neural module networks for video-grounded dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3377–3393.

Hung Le and Nancy F. Chen. 2020. Multimodal transformer with pointer network for the dstc8. In *DSTC8 at AAAI2020 workshop*.

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*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2019b. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. In *DSTC8 at AAAI2020 workshop*.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483.

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

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI2019 workshop*.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2019. From film to video: Multi-turn question answering with multi-modal context. In *DSTC7 at AAAI2019 workshop*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbad's submissino for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*.

Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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.

Huiyuan Xie and Ignacio Iacobacci. 2020. Audio visual scene-aware dialog system using dynamic memory networks. In *DSTC8 at AAAI2020 workshop*.

Yi-Ting Yeh, Tzu-Chuan Lin, Hsiao-Hua Cheng, Yi-Hsuan Deng, Shang-Yu Su, and Yun-Nung Chen. 2019. Reactive multi-stage feature fusion for multimodal dialogue modeling. In *DSTC7 at AAAI2019 workshop*.

A Training Details.

Training. THAM is trained on NVIDIA TITAN V (12GB of memory) GPU with Adam optimizer (Kingma and Ba, 2014) with β_1 = 0.9, β_2 = 0.99, and ϵ = 10e-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 $X^{\star}_{<t} = [\mathbf{h}||\mathbf{a}_{<t}]$. 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:

$$D_{KL}(P||Q) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_P[T] - \log(\mathbb{E}_Q[e^T]), \quad (12)$$

where Ω 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 $Z = \mathbb{E}_Q[e^T]$. For the construction, we are available to derive[10] as:

$$\mathbb{E}_P[T] - \log Z = \mathbb{E}_P[\log\frac{dG}{dQ}] \quad (13)$$

[10] $\log\frac{dG}{dQ} = \log\frac{1}{Z}e^T = \log\frac{1}{Z} + T = T - \log Z$

Let Δ be the gap as:

$$\Delta := D_{KL}(P||Q) - (\mathbb{E}_P[T] - \log(\mathbb{E}_Q[e^T])) \quad (14)$$

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

$$\begin{aligned}\Delta =& \mathbb{E}_P[\log\frac{dP}{dQ} - \log\frac{dG}{dQ}] \\ =& \mathbb{E}_P\log\frac{dP}{dG} = D_{KL}(P||G)\end{aligned} \quad (15)$$

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

$$D_{KL}(P||Q) \geq \mathbb{E}_P[T] - \log(\mathbb{E}_Q[e^T]), \quad (16)$$

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 $T^{\star}$ taking the form $T^{\star} \log\frac{dP}{dQ} + C$ for some constant $C \in \mathbb{R}$.

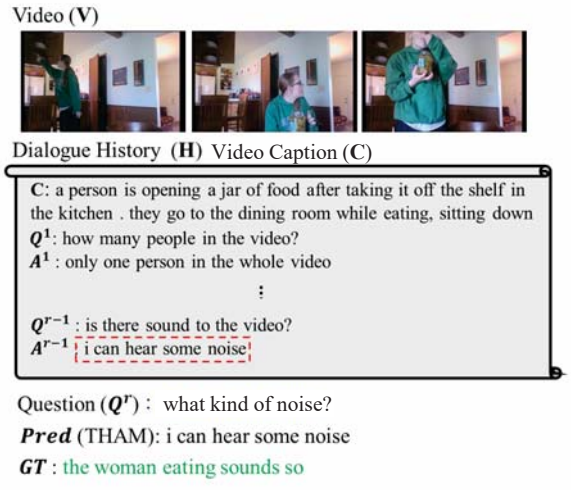


Figure 7: Failure case on question about sounds

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.