





# Multi-Modal Video Dialog State Tracking in the Wild

Adnen Abdessaied , Lei Shi , and Andreas Bulling 

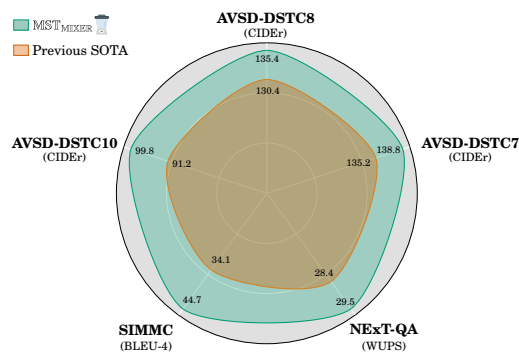
University of Stuttgart, Germany


{adnen.abdessaied, lei.shi, andreas.bulling}@vis.uni-stuttgart.de

 [https://perceptualui.org/publications/abdessaied24\\_eccv](https://perceptualui.org/publications/abdessaied24_eccv)

**Abstract.** We present  $MIST_{MIXER}$   – a novel video dialog model operating over a generic multi-modal state tracking scheme. Current models that claim to perform multi-modal state tracking fall short in two major aspects: (1) They either track only one modality (mostly the visual input) or (2) they target synthetic datasets that do not reflect the complexity of real-world in-the-wild scenarios. Our model addresses these two limitations in an attempt to close this crucial research gap. Specifically,  $MIST_{MIXER}$  first tracks the most important constituents of each input modality. Then, it predicts the *missing* underlying structure of the selected constituents of each modality by learning local latent graphs using a novel multi-modal graph structure learning method. Subsequently, the learned local graphs and features are parsed together to form a global graph operating on the mix of all modalities, further refining its structure and node embeddings. Finally, the fine-grained graph node features are used to enhance the hidden states of the backbone Vision-Language Model (VLM).  $MIST_{MIXER}$  achieves new state-of-the-art results on *five* challenging benchmarks.

**Keywords:** Video Dialog · Vision & Language · Multi-Modal Learning



**Fig. 1:**  $MIST_{MIXER}$   achieves SOTA results on various video-language tasks.

## 1 Introduction

Multi-modal tasks at the intersection of computer vision and natural language processing have been introduced to develop intelligent agents capable of assisting humans in understanding a visual premise through language. Among these tasks, video dialog is considered to be one of the most challenging. In contrast to visual [8] and video [68] question answering, which only require reasoning about a single question, video dialog models have to reason over the entire dialog history in addition to the current question. Furthermore, in contrast to visual dialog [15], video dialog involves reasoning over a video instead of a static image. Thus, a crucial part of a video dialog model is Dialog State Tracking (DST), which was originally introduced to track and update users’ goals in the form of dialog states [42, 64]. Nowadays, it is broadly used when a model keeps track of what it believes to be relevant for answering the question at hand.

Until now, research on DST has been predominately uni-modal in the form of slot-filling tasks [39, 51, 70] where the slots and slot values are constrained by a knowledge domain (e.g. hotel domain) and database schema (e.g. tabular data). However, the current landscape of the field necessitates extending to a multi-modal framework. Current models that claim to perform multi-modal state tracking fall short in two major aspects: (1) Some works track the constituents of only one modality to help the model focus on the most salient ones within a multi-modal context (e.g. video dialog [50], visual dialog [52], image retrieval [20], recommender systems [66]) rendering their state tracking approach uni-modal. More recently, Le *et al.* [34] have proposed VDTN, which extended the slot-filling paradigm to predict the visual attributes of CATER objects [19] from a pool of pre-defined textual values, but their approach suffers from the same aforementioned limitation. (2) Other works [1, 31, 49] have moved closer to performing multi-modal state tracking but have been limited to synthetic datasets that do not reflect the complexity of real-world scenarios.

We present  $\text{MST}_{\text{MIXER}}$  as a step towards addressing the aforementioned limitations. Specifically,  $\text{MST}_{\text{MIXER}}$  uses a backbone VLM and attention-based modality-specific tracking blocks to identify the most relevant constituents of each modality. Then, it uses a multi-modal GNN-based approach to learn the missing underlying structure between the mix of modalities in the form of latent graphs. Finally, it uses the fine-grained GNN features to enhance the hidden states of the backbone VLM to answer the question at hand more efficiently. To summarize, the contributions of our work are three-fold: (1) We propose  $\text{MST}_{\text{MIXER}}$ — a novel video dialog model that, unlike previous works, performs multi-modal state tracking on each input modality separately. Our model is generic by nature and could be easily adapted to deal with a wide range of tasks and datasets. (2) We equip our model with a novel divide-and-conquer GNN-based mechanism that dynamically learns the missing underlying structure of the mix of all modalities. First, it selects the most important constituents of each modality and learns their respective local structures using latent graphs. Then, it parses all individual graphs and features into a global modality-agnostic graph to further refine its structure and node features that we use to enhance

the hidden states of the backbone VLM. (3) As seen in [Figure 1](#), MST<sub>MIXER</sub> sets new state-of-the-art results across a broad range of video-language tasks.

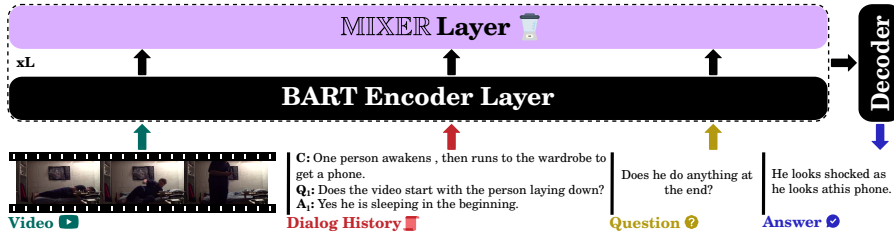
## 2 Related Work

**Video Dialog.** Video dialog has emerged as a natural extension to visual question answering [8], video question answering [69], and visual dialog [15]. Almari *et al.* [4] proposed AVSD – one of the first video dialog datasets based on the Charades videos [59], which has become the default dataset for the task. Later works [35,45] achieved new state-of-the-art results by leveraging pre-trained large language models [43,56] and fine-tuning them on the downstream video dialog task. Others used GNNs to perform reasoning on the dialog history [32] or on the visual scene [27] in an attempt to improve performance. Pham *et al.* [55] proposed an object-centric model to track object-associated dialog states upon receiving new questions. Inspired by the success of neural module networks [6,7], Le *et al.* [33] introduced VGNMN to model the information retrieval process in video-grounded language tasks as a pipeline of neural modules. More recently, Yoon *et al.* [73] introduced a text hallucination mitigation framework based on a hallucination regularization loss.

Despite the high multi-modality of the task in general and the AVSD dataset in particular, all previous works missed out on the idea of performing explicit multi-modal dialog state tracking. Instead, they focused on general vanilla attention methods that particularly tracked *only* one modality (mostly the visual input) at the expense of the others. MST<sub>MIXER</sub> closes this gap by performing multi-modal state tracking on each input modality separately.

**Dialog State Tracking.** Traditional state tracking approaches predicted slot values (e.g. meals offered by a restaurant) from a pre-defined set at each dialog, which is conditioned on some context. As a result, these approaches remained predominately uni-modal even though they were applied within a multi-modal context (e.g., video dialog [50], visual dialog [52], image retrieval [20], recommender systems [66]). However, the current landscape of dialog research necessitates the transition to multi-modal dialog state tracking to cope with the complexity of recent datasets. Some works have already been proposed to address this problem. For example, SIMMC [31,49] was introduced to develop agents capable of helping a human in a shopping scenario and, therefore, need to track the multi-modal state of the dialog to fulfill its task efficiently. More recently, Le *et al.* [34] suggested performing video dialog state tracking by extending the slot-filling task to predict predefined attributes of CATER [19] objects, limiting their approach to only the DVD dataset [38].

As such, all of these works focused only on synthetic and automatically generated datasets. To the best of our knowledge, MST<sub>MIXER</sub> is the first model to perform genuine multi-modal state tracking in the wild for video dialog by being able to deal with complex real-world scenarios.



**Fig. 2:**  $\text{MST}_{\text{MIXER}}$  takes a video, a dialog history, and a question as input and autoregressively generates an answer as output. It uses a BART backbone adapted to deal with multi-modal input features and enhanced via our graph-based mixing approach.

**Graph Structure Learning.** Early works on graph structure learning leveraged bilevel programming [14] to simultaneously learn GNN parameters and topology [17]. Yu *et al.* [75] proposed applying the linear structure equation model in conjunction with a variational autoencoder [57] to learn directed acyclic graphs. Subsequently, Elinas *et al.* [16] suggested using a stochastic variational inference model to jointly estimate the graph posterior and the GNN parameters. Chen *et al.* [11] proposed iteratively refining the graph topology in an end-to-end manner using graph similarity metric learning. Wu *et al.* [65] suggested an all-pair message passing method to propagate signals between arbitrary nodes for classification efficiently.

Our method differs from the aforementioned works in three distinct aspects: (1) We propose a novel multi-modal graph structure learning method that relies on a two-stage divide-and-conquer procedure that first predicts local modality-specific latent graphs before tackling the global graph consisting of the mix of all available modalities. (2) We use our graph learning approach to enhance the hidden states of a backbone VLM. (3) Instead of dealing with uni-modal graph-based tasks (node, edge, or graph classification), we investigate the effect of our method on the multi-modal, non-graph-related downstream task of video dialog.

## 3 Method

### 3.1 Problem Formulation

Given a question  $Q_t$  grounded on a video  $V$  at  $t$ -th dialog turn, a dialog history  $H_t = \{C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$  composed of previous question-answer pairs and a video caption  $C$ , a video dialog model is tasked of autoregressively generating a free-form answer  $A_t$  to the question at hand, i.e. each answer token  $a_t^i$  satisfies

$$a_t^i = \arg \max_{a \in \mathcal{V}} [P(a | V, Q_t, H_t, A_t^{<i})], \quad (1)$$

where  $A_t^{<i}$  and  $\mathcal{V}$  denote the previously predicted answer tokens and the vocabulary, respectively.

### 3.2 Input Representation Learning

As can be seen from [Figure 2](#), MST<sub>MIXER</sub> is based on BART [43] and adapted to handle data from multiple input modalities.

**Visual Representations.** As it is standard for this task, the visual representations are extracted for a given video using I3D-rgb and I3D-flow models [10] pre-trained on YouTube videos and the Kinetics dataset [26]. Formally, a video  $V$  is first split into  $l_v$  segments using a sliding window of  $n$  frames. Then, each segment  $S = \{f_1, f_2, \dots, f_n\}$ , where  $f_i$  represents one video frame, are fed to the pre-trained I3D models to extract the  $d_v$ -dimensional video features  $V_{\text{rgb}}, V_{\text{flow}} \in \mathbb{R}^{l_v \times d_v}$ . Finally, we extracted object features  $V_{\text{sam}} \in \mathbb{R}^{l_v \times d_s}$  from the middle frame of the video using SAM [30]. We mapped these features to match the hidden dimension  $d$  of BART using linear projections with weights matrices  $W_{\text{rgb}}, W_{\text{flow}}, W_{\text{sam}}$ .

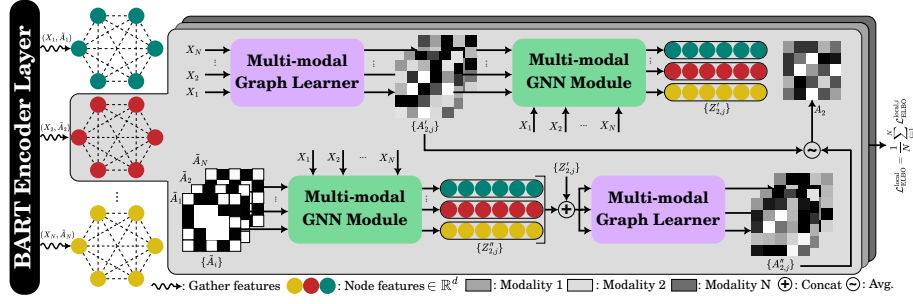
**Audio Representations.** Similar the previous works [32, 45, 73], we used audio features extracted from a pre-trained VGGish model [60]. Since video and audio are synchronous, the same splits were used to generate the  $d_a$ -dimensional audio features  $A_{\text{vggish}} \in \mathbb{R}^{l_v \times d_a}$ . As for the video feature, we mapped the audio features to the BART embedding space using a linear projection with a weight matrix  $W_a \in \mathbb{R}^{d \times d_a}$ . We refer to [22] for further details about feature extraction.

**Textual Representations.** We used the dialog history composed of the video caption, the previous question-answer pairs, and the current question as additional input to the encoder. We separated each segment with the special token  $\langle /s \rangle$ . Subsequently, we embedded their concatenation into a dense representation  $T = [T_H, T_Q] \in \mathbb{R}^{l_{\text{txt}} \times d}$  using a word embedding matrix  $W_{\text{txt}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ , where  $l_{\text{txt}}, \mathcal{V}, T_H$ , and  $T_Q$  are the length of the textual input, the vocabulary, the dense representation of the history and question, respectively. Finally, we input a shifted ground truth into the decoder and embed it using the same word matrix.

**State Tokens.** We inserted special *state* tokens  $\langle s_i \rangle$  at the beginning of each modality ( $V_{\text{rgb}}, V_{\text{flow}}, V_{\text{sam}}, A_{\text{vggish}}, T_H, T_Q$ ) and used them to keep track of the most relevant constituents.

### 3.3 MST<sub>MIXER</sub>: Multi-Modal Feature Mixing

The main idea of MST<sub>MIXER</sub> is to keep track of the most relevant constituents at different semantic levels (e.g. across modalities and encoder layers) and use them to refine the multi-modal state of the model. Specifically, we insert a MIXER layer after every  $\Delta$  encoder layer. Our approach follows a two-stage divide and



**Fig. 3:** In Stage I,  $\text{MST}_{\text{MIXER}}$  first gathers multi-modal features  $\{X_i\}$  from the previous BART layer and computes their respective initial local structures  $\{\hat{A}_i\}$ . Then, it simultaneously learns the local latent multi-modal graphs and refines the features using a two-stream framework, i.e.,  $\{A'_{i,j}, A''_{i,j}\}_j$  and  $\{Z'_{i,j}, Z''_{i,j}\}_j$ , respectively. Finally, it outputs the final multi-modal latent graph  $A_i$  used to compute the local ELBO loss  $\mathcal{L}_{\text{ELBO}}^{\text{local}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{ELBO}}^{\text{local},i}$ .

conquer scheme where we first learn the underlying *local* structures of the individual modalities before learning the *global* inter-modal structure of the mix of all available modalities. We posit that directly learning the latter might be daunting for such a high multi-modal task.

**Multi-Modal Feature Tracking.** We take advantage of the special state tokens  $\langle s_i \rangle$  to keep track of the most relevant modality-specific features at different embedding levels of the encoder. Specifically, for each modality, we select the  $K$  tokens with the highest attention values concerning the respective state token, i.e.

$$X_i = \text{top}_K(\alpha_{\text{avg}}(h_{\langle s_i \rangle}, H_i)) \in \mathbb{R}^{K \times d}, \quad (2)$$

where  $\alpha_{\text{avg}}(h_{\langle s_i \rangle}, H_i)$  is the attention values between the state embedding and the remaining tokens embeddings  $H_i$  of the  $i$ -th modality averaged across heads.

**Mixing Stage I (Divide).** We posit that the selected features  $\{X_i\}$  of each modality encapsulate rich information that could be leveraged to improve the learning capabilities of our model. A viable approach is to take advantage of the power of GNNs to refine these features based on their local structures, as prior works have highlighted the merit of integrating GNNs with transformer-based models [2, 71, 72]. However, the underlying structures that govern  $\{X_i\}$  are missing in our case. To this end, we propose a novel multi-modal graph structure learning approach that simultaneously learns the graph weights and the adjacency matrix in the form of latent graphs. We posit that we can split the adjacency matrix  $A_i$  of the  $i$ -th modality into an initial (observable) part  $\hat{A}_i$  and a missing (sought-after) part  $A'_i$  where  $\hat{A}_i$  is a *binary matrix* constructed using a  $k$ NN ( $k = 4$ ) approach based on  $X_i$ . Thus,

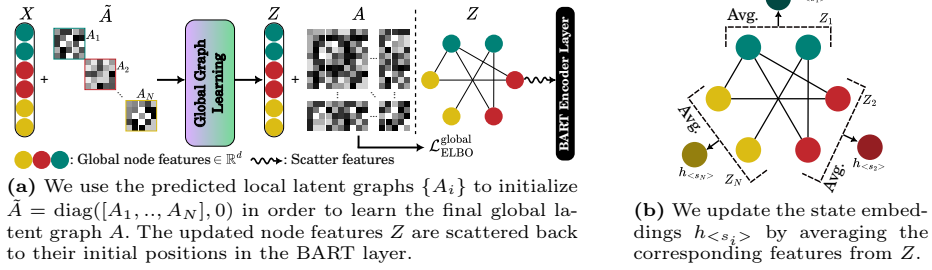


Fig. 4: Overview of mixing stage II.

$$P(X_i, A_i) = P(A_i | X_i) P(X_i) \quad (3)$$

$$= P(A'_i, \tilde{A}_i | X_i) P(X_i). \quad (4)$$

Although the conditional distribution  $P(A'_i, \tilde{A}_i | X)$  can be modeled by a parametric families of distributions  $p_\theta^i(A'_i, \tilde{A}_i | X)$ , the optimal parameter set  $\bar{\theta}$  is not known making the computations of the marginal

$$p_{\bar{\theta}}^i(\tilde{A}_i | X_i) = \int p_\theta^i(A'_i, \tilde{A}_i | X_i) d(A'_i) \quad (5)$$

and therefore, the posterior of each modality

$$p_\theta^i(A'_i | \tilde{A}_i, X_i) = \frac{p_\theta^i(A'_i, \tilde{A}_i | X_i)}{p_\theta^i(\tilde{A}_i | X_i)} \quad (6)$$

intractable. To be able to infer the missing part of the local adjacency matrix, we take advantage of Variational Inference (VI) to learn an approximation  $q_\phi^i(A'_i | \tilde{A}_i, X_i)$  of the posterior. We postulate that the missing adjacency matrix of modality  $i$  depends on its own features  $X_i$  and the features of other modalities  $X_{j \neq i}$ . Therefore, we propose a multi-modal conditioning (MMC) of Equation 6 on all  $X_{j \neq i}$  in addition to  $X_i$ . We also follow the idea of [11] that better graph structures lead to better features, and better features lead to better graph structures. Therefore, as shown in Figure 3, we use a two-stream approach where one stream uses enhanced features to learn the latent multi-modal graphs, and the other uses the predicted graphs to infer fine-grained features to learn both  $q_\phi^i$  and  $p_\theta^i$  for each modality. Specifically, in the purple module of the upper stream, we estimate an edge of latent graph  $A'_{i,j}$  using cosine similarity as

$$a'_{mn} = \frac{1}{K} \sum_{k=1}^K \cos(w_j^k \odot x_m, w_j^k \odot x_n), \quad (7)$$

where  $x_m, x_n \in X_i$ ,  $\{w_j^k\}$  are learnable weights for each modality, and  $\odot$  denotes element-wise multiplication. Then, in the green module, we update the

multi-modal node features using an APPNP [18] module and the predicted latent graphs for modality  $i$  to get  $\{Z'_{i,j}\}_j$ . For the lower stream, we first start by updating the node features similarly to the upper stream by using the initial graphs  $\{\tilde{A}_i\}$  to get  $\{Z''_{i,j}\}_j$ . Then, we use the enhanced node features  $\{[Z'_{i,j}, Z''_{i,j}]\}_j$  to predict the second set of local latent graphs  $\{A''_{i,j}\}_j$ . At the end, we output the final local latent graph of modality  $i$  as

$$A_i = \underbrace{\frac{1}{2}\tilde{A}_i}_{\text{Initialization Bias (IB)}} + \underbrace{\frac{1}{2}\sum_{j=1}^N \frac{1}{N}(A'_{i,j} + A''_{i,j})}_{\text{VI approximation via MMC}} \in \mathbb{R}^{K \times K}. \quad (8)$$

**Mixing Stage II (Conquer).** This stage tries to infer the global latent graph structure governing the mix of all modalities  $\{X_i\}$ . As seen in Figure 4a, it depends on the previously predicted local latent graphs to build the initial global graphs as

$$\tilde{A} = \text{diag}([A_1, \dots, A_N], 0) \in \mathbb{R}^{NK \times NK}. \quad (9)$$

Similar to Stage I, we use a two-stream approach to learn the global  $p_\theta$  and  $q_\phi$  and thus the global latent graph  $A$  and node features

$$Z = \frac{1}{2}(Z' + Z''), \quad (10)$$

where  $Z'$  and  $Z''$  are obtained from the upper and lower streams, respectively. Finally, we update the state tokens embeddings  $h_{\langle s_i \rangle}$  by averaging the corresponding features from  $Z$  (see Figure 4b) and integrate the latter back into the hidden state of the corresponding BART layer following

$$H = (1 - \lambda)(H \oslash (Z, \text{Idx})) + \lambda H, \quad (11)$$

where  $\lambda \in (0, 1)$  is a hyper-parameter and  $\oslash$ ,  $H$ , and  $\text{Idx}$  denote the scatter operation, the hidden state of the BART layer and the indices of the nodes features  $Z$  relative to  $H$ , respectively.

*Loss Function.* Since we rely on VI to infer the local and global latent graphs, we used two ELBO losses to optimize (1) the local multi-modal graph learners  $\{q_\phi^i, p_\theta^i\}$  and (2) the global learners  $q_\phi, p_\theta$ . Please refer to the supplementary material for the derivation of these losses. We trained our model end-to-end using a combination of the generative loss of the video dialog task  $\mathcal{L}_{\text{gen}}$  and both ELBO losses, i.e.


$$\mathcal{L} = \alpha_1 \mathcal{L}_{\text{gen}} - \alpha_2 \mathcal{L}_{\text{ELBO}}^{\text{local}} - \alpha_3 \mathcal{L}_{\text{ELBO}}^{\text{global}}, \quad (12)$$

$$\mathcal{L}_{\text{ELBO}}^{\text{local}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{ELBO}}^{\text{local}, i}, \quad (13)$$

where  $\{\alpha_k\}$  are hyper-parameters and  $\mathcal{L}_{\text{ELBO}}^{\text{local}, i}$  is the local ELBO loss for the  $i$ -th modality.



**Table 1:** Results on AVSD-DSTC7 and AVSD-DSTC8. Best and second best performances are in **bold** and underlined, respectively. ♠ = Two-stage training.

Model	Venue	AVSD-DSTC7						AVSD-DSTC8							
		B-1	B-2	B-3	B-4	M	R	C	B-1	B-2	B-3	B-4	M	R	C
Baseline [22]	<i>ICASSP'19</i>	62.1	48.0	37.9	30.5	21.7	48.1	73.3	61.4	46.7	36.5	28.9	21.0	48.0	65.1
MTN [36]	<i>ACL'19</i>	71.5	58.1	47.6	39.2	26.9	55.9	106.6	—	—	—	—	—	—	—
JMAN [13]	<i>AAAI'20</i>	66.7	52.1	41.3	33.4	23.9	53.3	94.1	64.5	50.4	40.2	32.4	23.2	52.1	87.5
VGD [35]	<i>ACL'20</i>	74.9	62.0	52.0	43.6	28.2	58.2	119.4	—	—	—	—	—	—	—
BiST [37]	<i>EMNLP'20</i>	75.5	61.9	51.0	42.9	28.4	58.1	119.2	68.4	54.8	45.7	37.6	27.3	56.3	101.7
SCGA [27]	<i>AAAI'21</i>	74.5	62.2	51.7	43.0	28.5	57.8	120.1	71.1	59.3	49.7	41.6	27.6	56.6	112.3
RLM [45]	<i>TASLP'21</i>	76.5	64.3	54.3	45.9	29.4	60.6	130.8	74.6	62.6	52.8	44.5	28.6	59.8	124.0
PDC [32]	<i>ICLR'21</i>	77.0	65.3	53.9	44.9	29.2	60.6	129.5	74.9	62.9	52.8	43.9	28.5	59.2	120.1
AV-TRN [58]	<i>ICASSP'22</i>	—	—	—	40.6	26.2	55.4	107.9	—	—	—	39.4	25.0	54.5	99.7
VGNMN [33]	<i>NAACL'22</i>	—	—	—	42.9	27.8	57.8	118.8	—	—	—	—	—	—	—
COST [55]	<i>ECCV'22</i>	72.3	58.9	48.3	40.0	26.6	56.1	108.5	69.5	55.9	46.5	3.82	27.8	57.4	105.1
MRLV [3]	<i>NeurIPS'22</i>	—	59.2	49.3	41.5	26.9	56.9	115.9	—	—	—	—	—	—	—
♠THAM [73]	<i>EMNLP'22</i>	77.8	65.4	54.9	46.8	<u>30.8</u>	<u>61.9</u>	133.5	<u>76.4</u>	<u>64.1</u>	53.8	45.5	<u>30.1</u>	<u>61.0</u>	<u>130.4</u>
DialogMCF [12]	<i>TASLP'23</i>	77.7	65.3	54.7	45.7	30.6	61.3	<u>135.2</u>	75.6	63.3	53.2	44.9	29.3	60.1	125.3
ITR [76]	<i>PAMI'23</i>	<u>78.2</u>	<u>65.5</u>	<u>55.2</u>	<u>46.9</u>	30.5	<u>61.9</u>	133.1	76.2	<u>64.1</u>	<u>54.3</u>	<u>46.0</u>	29.8	60.7	128.5
MST <sub>MIXER</sub> 		<b>78.7</b>	<b>66.5</b>	<b>56.3</b>	<b>47.6</b>	<b>31.3</b>	<b>62.5</b>	<b>138.8</b>	<b>77.5</b>	<b>66.0</b>	<b>56.1</b>	<b>47.7</b>	<b>30.6</b>	<b>62.4</b>	<b>135.4</b>
w/o $V_{sam}$	<i>ECCV'24</i>	78.6	66.3	56.0	47.4	31.2	62.2	137.3	77.4	65.8	56.0	47.3	30.6	62.1	134.8
w/o $A_{vggish}$		78.4	66.0	55.8	47.1	31.0	62.0	136.5	77.1	65.6	55.7	47.1	30.2	61.8	133.6

## 4 Experiments

### 4.1 Datasets

We mainly evaluated our model on the popular and challenging Audio-Visual Scene Aware Dialog (AVSD) dataset [4]. Each of its dialogs comes with 10 question-answer pairs as well as a short description/caption based on a video. Each video is collected from the Charades dataset [59] and the dialogs are generated by human annotators. We considered all three benchmarks of the dataset, i.e. AVSD-DSTC7 [74], AVSD-DSTC8 [28], and AVSD-DSTC10 [58], which were respectively released for the Dialog System Technology Challenge (DSTC). To assess the generalizability of our model, we not only experimented with the generative task of SIMMC 2.0 [31] but also with the recent and challenging open-ended video question answering NExT-QA dataset [67]. We refer to the supplementary material for more details about all *five* benchmarks.

### 4.2 Metrics

We used the established official metrics for each dataset in order to fairly compare MST<sub>MIXER</sub> with the previous models. Specifically, for all *three* AVSD datasets, we used BLEU (B-n) [53], ROUGE-L (R) [46], METEOR (M) [9], and CIDEr (C) [62]. Whereas for SIMMC and NExT-QA, we used B-4 and WUPS [48] scores, respectively.

**Table 2:** Results on AVSD-DSTC10.

Model	Venue	B-1	B-2	B-3	B-4	M	R	C
AV-TRN [58]	<i>ICASSP'22</i>	–	–	–	24.7	19.1	43.7	56.6
+ Ext. [58]	<i>ICASSP'22</i>	–	–	–	37.1	24.5	53.5	86.9
DSTC10 [23]	<i>AAAI'22</i>	67.3	54.5	44.8	<u>37.2</u>	24.3	53.0	<u>91.2</u>
DialogMCF [12]	<i>TASLP'23</i>	<u>69.3</u>	<u>55.6</u>	<u>45.0</u>	36.9	<u>24.9</u>	53.6	<u>91.2</u>
<b>MST<sub>MIXER</sub><sup>C</sup></b>		<b>70.0</b>	<b>57.4</b>	<b>47.6</b>	<b>40.0</b>	<b>25.7</b>	<b>54.5</b>	<b>99.8</b>
w/o $V_{\text{sam}}$	<i>ECCV'24</i>	69.8	57.4	47.5	39.8	25.6	54.3	97.6
w/o $A_{\text{vggish}}$		69.7	57.1	47.2	39.5	25.1	54.0	96.9

**Table 3:** Results on SIMMC.

Model	Venue	B-4
MTN [36]	<i>ACL'19</i>	21.7
GPT-2 [31]	<i>EMNLP'21</i>	19.2
BART [41]	<i>NAACL'22</i>	33.1
PaCE [44]	<i>ACL'23</i>	<u>34.1</u>
<b>MST<sub>MIXER</sub><sup>C</sup></b>	<i>ECCV'24</i>	<b>44.7</b>

**Table 4:** Results on open-ended NExT-QA<sup>◇</sup>.

Model	Venue	WUPS <sub>C</sub>	WUPS <sub>T</sub>	WUPS <sub>D</sub>	WUPS
HCRN [40]	<i>CVPR'20</i>	16.05	17.68	49.78	23.92
HGA [24]	<i>AAAI'20</i>	<u>17.98</u>	<u>17.95</u>	<u>50.84</u>	24.06
Flamingo [5]	<i>NeurIPS'22</i>	–	–	–	<u>28.40</u>
KcGA [25]	<i>AAAI'23</i>	–	–	–	28.20
EMU [61]	<i>arXiv'23</i>	–	–	–	23.40
<b>MST<sub>MIXER</sub><sup>C</sup></b>	<i>ECCV'24</i>	<b>22.12</b>	<b>22.20</b>	<b>55.64</b>	<b>29.50</b>

### 4.3 Main Results

**AVSD-DSTC7.** As can be seen in Table 1, our model managed to achieve new SOTA results across all evaluation metrics, thereby outperforming the latest baselines, including PDC [32], DialogMCF [12], THAM [73], and ITR [76]. Specifically, MST<sub>MIXER</sub> outperformed the latest ITR [76] model by over 1.5% (relative improvement) on B-2, B-3, B-4, and M scores. Since some previous models did not use SAM [30] and audio features, we trained two additional versions of our model where we only removed SAM features before additionally removing the audio features. Both versions are denoted by “w/o  $V_{\text{sam}}$ ” and “w/o  $A_{\text{vggish}}$ ”, respectively. As seen from Table 1, both versions still outperform all previous models across all evaluation metrics.

**AVSD-DSTC8.** As depicted in Table 1, models tend to struggle more on this more recent benchmark. However, MST<sub>MIXER</sub> scored new SOTA results with higher relative improvements compared to DSTC7, thereby lifting the B-2, B-3, B-4, and C scores by over 3% relative to the second best models ITR [76] and THAM [73]. Similarly to AVSD-DSTC7, our ablated versions surpassed these models on all evaluation metrics and marginally underperformed our full model.

**AVSD-DSTC10.** We then evaluated MST<sub>MIXER</sub> on the latest AVSD-DSTC10 benchmark. Contrary to the previous versions, AVSD-DSTC10 does not include human-generated video descriptions during inference since these are unavailable in real-world applications. As depicted in Table 2, models struggle the most on

<sup>◇</sup> C, T, and D denote causal, temporal, and descriptive questions, respectively.

**Table 5:** Influence of the value of  $\lambda$ .

$\lambda$	PPL (val)	AVSD-DSTC7			AVSD-DSTC8		
		B-4	R	C	B-4	R	C
0.0	Training unstable						
0.1	11.03	17.3	29.0	35.1	11.4	24.3	21.2
0.5	5.48	44.6	60.3	126.4	44.7	59.4	123.8
0.9	<b>5.16</b>	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>47.7</b>	<b>62.4</b>	<b>135.4</b>
1.0	5.30	45.1	60.8	131.3	42.3	61.1	126.9

**Table 6:** Influence of the value of  $\Delta$ .

$\Delta$	PPL (val)	AVSD-DSTC7			AVSD-DSTC8		
		B-4	R	C	B-4	R	C
$\leq 2$	Training too long						
3	5.19	45.7	61.5	134.1	46.7	61.5	131.8
4	<b>5.16</b>	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>47.7</b>	<b>62.4</b>	<b>135.4</b>
5	5.21	45.0	61.1	133.6	44.6	60.5	129.1

this challenge version. However, not only our full MST<sub>MIXER</sub> model but also its two ablated versions managed to outperform the latest models on all evaluation metrics.

♣**SIMMC.** To assess the generalizability of our model, we additionally tested it on the generative task of SIMMC 2.0 [49]. As seen from Table 3, MST<sub>MIXER</sub> outperformed the latest published models such as PaCE [44] by achieving a B-4 score of 44.7.

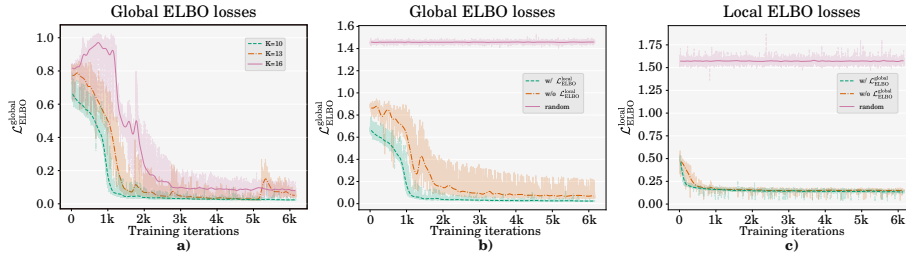
♣**NExT-QA.** Finally, we tested our model on the recent open-ended NExT-QA benchmark [67]. As depicted in Table 4, MST<sub>MIXER</sub> not only outperformed HCRN [40] and HGA [24] on all WUPS scores [48] but also surpassed latest models such as Flamingo [5], KcGA [25], and EMU [61]. Specifically, it lifted the overall WUPS score by 1.1 absolute points compared to the seminal Flamingo-9B model with x18 more parameters.

#### 4.4 Ablation Study

**Effect of  $\lambda$  and  $\Delta$ .** We independently optimized these hyper-parameters based on the validation perplexity (PPL). First, we fixed  $\Delta = 4$  to guarantee a reasonable training time on our hardware setup and varied  $\lambda \in \{0, 0.1, 0.5, 0.9, 1\}$ . As seen in Table 5, the best performance was achieved when using  $\lambda = 0.9$ . Thereafter, we varied  $\Delta \in \{2, 3, 4, 5\}$  while keeping  $\lambda = 0.9$  and achieved the best results for  $\Delta = 4$  as can be seen from Table 6.

**Latent Graph Size  $K$ .** As illustrated in the first section of Table 7, we varied  $K$  from 7 to 16 in three-step intervals. The overall performance of MST<sub>MIXER</sub> peaked when using  $K = 10$  tokens from each modality as the graphs' node features. Using higher values of  $K$  rendered the learning of the global latent graphs with  $K \times N$  nodes more difficult and thus hurt the overall performance of our model. This is underlined by the behavior of the global ELBO loss  $\mathcal{L}_{ELBO}^{\text{global}}$  as illustrated in Figure 5a. Using  $K = 7$  hurt the performance of our model across almost all metrics. We posit that low values of  $K$  are insufficient to capture each modality's most influential constituents. Therefore, we set  $K = 10$  in the rest of the experiments.

♣: Models trained with optimal hyperparameters from AVSD and without  $V_{\text{sam}}$ .



**Fig. 5:** **a)** Larger values of  $K$  make the learning of the global latent graphs more challenging. **b)** The local ELBO loss  $\mathcal{L}_{\text{ELBO}}^{\text{local}}$  facilitates the learning of the global latent graphs. **c)** The global ELBO loss  $\mathcal{L}_{\text{ELBO}}^{\text{global}}$  facilitates the learning of the local latent graphs. All models use SAM and audio features.

**Multi-Modal State Tracking GNNs.** In each row of the middle section of [Table 7](#), we ablated *one* GNN-based tracking module and kept the remaining ones unchanged. Our full model outperformed all these ablated versions *despite them having access to the same input features*. The comparable results of all these ablated versions validate using a uniform graph size  $K$  for all different modalities. Finally, we replaced all GNNs (local and global) with vanilla transformer layers. As can be seen from the [last row](#) of the middle section, this version was outperformed by our full model as well, underlining the efficacy of our proposed multi-modal graph learning approach.

**ELBO Losses.** As can be seen in the third section of [Table 7](#), we conducted extensive experiments with different combinations of the ELBO losses: (1) We first ablated the learning of both global and local latent graphs and, therefore, both ELBO losses resulting in a plain BART model [43]. (2) We then only used the initial graphs  $\hat{A}_i$  as the final latent graph approximations in both training stages I and II leading to improvements compared to plain BART. (3) Thereafter, we ablated the local ELBO loss and directly learned the global latent graphs. This version of our model underperformed BART, which follows our hypothesis that directly learning the global latent graphs is daunting and might lead to performance drops. As illustrated in [Figure 5b](#),  $\mathcal{L}_{\text{ELBO}}^{\text{global}}$  converged faster and reached lower values when optimized jointly with  $\mathcal{L}_{\text{ELBO}}^{\text{local}}$ . (4) We thereafter ablated the global ELBO loss and only learned the local latent graphs, leading to performance increases compared to the previous versions. This underlines that learning the local latent graphs is less sensitive to  $\mathcal{L}_{\text{ELBO}}^{\text{global}}$  than learning the global latent graphs is to  $\mathcal{L}_{\text{ELBO}}^{\text{local}}$  as can be seen in [Figure 5c](#). (5) We finally evaluated a version with a comparable computational complexity as our full model but used random latent graphs instead of learning them. As can be seen in [Figure 5b](#), [Figure 5c](#), and the [last row](#) of [Table 7](#)), both ELBO losses remained constant and the model reached the worst results among all ablated versions empirically showcasing the importance of our latent graph learning approach.

**Table 7:** Comparison between different ablated versions of our model. All ablations use SAM and audio features. **TRN** means that the model replaces the global and local multi-modal GNNs with vanilla transformer layers, and **RAND** denotes that it uses random latent graphs instead of learning them. Our full model is highlighted in **blue**.

K	GNNs	$\mathcal{L}_{\text{ELBO}}^{\text{local}}$	$\mathcal{L}_{\text{ELBO}}^{\text{global}}$	# Params.	AVSD-DSTC7				AVSD-DSTC8			
					B-1	B-4	R	C	B-1	B-4	R	C
7	All	✓	✓	~ 511M	77.8	47.0	61.8	136.2	76.6	47.0	61.5	131.8
10	All	✓	✓	~ 511M	<b>78.7</b>	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>77.5</b>	<b>47.7</b>	<b>62.4</b>	<b>135.4</b>
13	All	✓	✓	~ 511M	77.0	45.4	60.6	131.9	75.7	45.2	60.4	127.0
16	All	✓	✓	~ 511M	76.6	45.4	60.7	132.6	75.8	45.9	60.5	128.4
10	w/o GNN <sub>rgb</sub>	✓	✓	~ 495M	78.4	47.2	62.4	137.2	77.3	47.4	62.0	133.2
10	w/o GNN <sub>flow</sub>	✓	✓	~ 495M	78.5	47.1	62.5	138.5	76.9	47.2	61.9	134.1
10	w/o GNN <sub>sam</sub>	✓	✓	~ 495M	78.1	46.1	62.2	137.2	77.5	46.5	61.7	132.7
10	w/o GNN <sub>vggish</sub>	✓	✓	~ 495M	78.0	45.8	61.4	134.9	76.8	46.5	61.0	131.0
10	w/o GNN <sub>H</sub>	✓	✓	~ 495M	78.1	45.7	61.8	134.1	77.4	46.7	62.2	134.0
10	w/o GNN <sub>Q</sub>	✓	✓	~ 495M	78.2	47.1	62.1	138.5	77.0	47.0	61.8	133.6
10	TRN	✗	✗	~ 500M	77.8	46.9	61.8	136.6	76.8	46.7	61.4	131.8
–	–	✗	✗	~ 411M	76.6	45.1	60.8	131.3	74.2	42.3	61.1	126.9
–	w/ only $\tilde{A}_i$	✗	✗	~ 413M	76.5	45.4	60.9	131.7	75.2	45.5	60.7	130.3
10	All	✗	✓	~ 416M	75.9	44.5	59.8	127.8	74.3	44.2	59.2	122.8
10	All	✓	✗	~ 506M	77.5	46.4	61.4	134.9	76.2	46.6	60.9	130.6
10	All	RAND	RAND	~ 448M	73.0	42.1	57.3	119.2	71.4	41.6	57.1	114.2

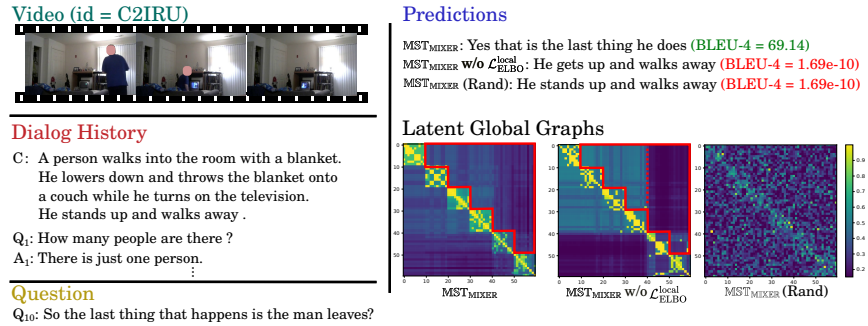
**Table 8:** Comparison between different ablated versions of our model. All ablations were trained with SAM and audio features and with the optimal hyper-parameters as the full model. **IB** = Initialization Bias, **MMC** = Multi-Modal Conditioning.

MST <sub>MIXER</sub>	# Params.	AVSD-DSTC7				AVSD-DSTC8			
		B-1	B-4	R	C	B-1	B-4	R	C
w/o MMC	~ 500M	76.9	46.6	61.4	135.5	75.8	46.1	60.5	130.9
w/o IB	~ 511M	77.6	47.0	61.8	136.2	76.3	46.2	61.2	131.1
Full	~ 511M	<b>78.7</b>	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>77.5</b>	<b>47.7</b>	<b>62.4</b>	<b>135.4</b>

**Latent Graph Learning.** Lastly, we considered two additional ablations of MST<sub>MIXER</sub>. Specifically, we first ablated the multi-modal conditioning (MMC) of Equation 6 and learned the local latent graphs of modality  $i$  based only on its features  $X_i$ . This reduces Equation 8 to

$$A_i = \frac{1}{2}\tilde{A}_i + \frac{1}{2}(A'_i + A''_i). \quad (14)$$

Then, we trained a version without the initialization bias (IB) of Equation 8. As can be seen in Table 8, MMC is essential for high performance. Without it MST<sub>MIXER</sub> achieved the lowest performance across all metrics. The same applies to IB since not incorporating  $\tilde{A}_i$  and only using the posterior approximation impeded the performance across all evaluation metrics.



**Fig. 6:** Qualitative comparison of different model ablations. on response generation and latent global graph inference of  $q_\phi$  obtained from the last encoder layer. The diagonal blocks (from upper left to lower right) correspond to  $V_{rgb}$ ,  $V_{flow}$ ,  $V_{sam}$ ,  $A_{vggish}$ ,  $T_H$ , and  $T_Q$ , respectively.

## 4.5 Qualitative Results

Finally, in [Figure 6](#) we give a qualitative comparison of MST<sub>MIXER</sub> with different ablated versions on response generation and global latent graph inference: Our full model managed to accurately answer the question whereas both ablated version failed to generate reliable responses. Furthermore, we can see how our full model better captured the local interactions within each modality (more structured diagonal blocks) as well as the global ones across modalities: Whereas the off-diagonal region (bordered in red) of the version “w/o  $\mathcal{L}_{ELBO}^{local}$ ” showed a clear divide between the modalities (dotted line), the full model mitigated this by producing more homogeneous values indicating better inter-modal interactions. We provide more examples and failure cases in the supplementary material.

## 5 Conclusion

We proposed MST<sub>MIXER</sub>— a novel multi-modal state tracking model specifically geared towards video dialog. MST<sub>MIXER</sub> first identifies the most influential constituents at different semantic levels (e.g., across modalities and encoder layers). Then, it relies on a two-stage divide and conquer approach to infer the missing underlying structure of the mix of all modalities and leverages it to augment the hidden states of the backbone VLM using GNNs. Through extensive ablations experiments and evaluations on *five* video-and-language benchmarks, we show our approach’s effectiveness and generalization capabilities.

**Acknowledgments.** L. Shi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2075–390740016.

## A ELBO Derivation & Implementation

In this section, we derive the ELBO loss and show how it can be used as an optimization term in our total loss. Without the loss of generality, we only consider the ELBO in the global setting. Given the intractable posterior  $p_\theta(A'|\tilde{A}, X)$  and the its approximation  $q_\phi(A'|\tilde{A}, X)$ , it holds that

$$\mathcal{D}_{\text{KL}}\left(q_\phi(A'|\tilde{A}, X)||p_\theta(A'|\tilde{A}, X)\right) = \mathbb{E}_{q_\phi(A'|\tilde{A}, X)} \left[ \log \frac{q_\phi(A'|\tilde{A}, X)}{p_\theta(A'|\tilde{A}, X)} \right] \quad (15)$$

$$= \mathbb{E}_{q_\phi(A'|\tilde{A}, X)} \left[ \log \frac{q_\phi(A'|\tilde{A}, X)p_\theta(\tilde{A}|X)}{p_\theta(A', \tilde{A}|X)} \right] \quad (16)$$

$$= \mathbb{E}_{q_\phi(A'|\tilde{A}, X)} \left[ \log \frac{q_\phi(A'|\tilde{A}, X)}{p_\theta(A', \tilde{A}|X)} \right] + p_\theta(\tilde{A}|X) \quad (17)$$

$$= \underbrace{p_\theta(\tilde{A}|X)}_{\text{Evidence}} - \underbrace{\mathbb{E}_{q_\phi(A'|\tilde{A}, X)} \left[ \log \frac{p_\theta(A', \tilde{A}|X)}{q_\phi(A'|\tilde{A}, X)} \right]}_{=:\mathcal{L}_{\text{ELBO}}^{\text{global}}} \geq 0 \quad (18)$$

Thus, as its name suggests, ELBO serves as a lower bound of the evidence. As a results, VI tries to maximize the ELBO which is equivalent to minimizing the Kullback-Leibner Divergence between  $q_\phi(A'|\tilde{A}, X)$  and the intractable posterior  $p_\theta(A'|\tilde{A}, X)$  leading to better estimation of the latter. Since we used the ELBOs as terms in the total loss  $\mathcal{L}$  to be minimized, we had to use the opposite value of each one of them. This explains the minus sign in [Equation 12](#) in the main text. Since  $q_\phi$  and  $p_\theta$  only output normalized scores as the prediction for each edge, we appended the zero vectors to both predictions in order to convert the raw scores to a two-value probability before applying the log-softmax function. We provide in [Listing 1.1](#) a code-snippet of our implementation of the ELBO loss.

## B Generative Loss

In addition to the ELBO losses, we used the generative loss  $L_{gen}$  to train our model. It employs teacher forcing and teaches the BART decoder to predict the next response token  $\hat{y}_{j+1}$  conditioned on the previous ground-truth response tokens  $Y_j = [y_1, \dots, y_j]$  and the output of the encoder  $H_{enc}$ . Specifically, the next predicted token satisfies

$$\hat{y}_{j+1} = \arg \max_{y \in V} [\log P(y|Y_j, H_{enc})], \quad (19)$$

where  $V$  and  $P$  denote the vocabulary and the softmax of the logits of the last decoder layer, respectively.

**Table 9:** Summary of the AVSD dataset with all test splits from DSTC7, DSTC8, and DSTC10.

	Train	Val	Test		
			DSTC7	DSTC8	DSTC10
# <b>Dialogs/Videos</b>	7,659	1,787	1,710	1,710	1,804
# <b>Questions/Answers</b>	153,180	35,740	13,490	18,810	28,406
# <b>Words</b>	1,450,754	339,006	110,252	162,226	272,606

**Table 10:** Summary of the open-ended NExT-QA dataset.

	Train	Val	Test
# <b>Videos</b>	3,870	570	1,000
# <b>Questions</b>	37,523	5,343	9,178

## C Datasets

### C.1 AVSD

The AVSD dataset [4] was released in the 7th Dialogue System Technology Challenge (DSTC7) [74]. As can be seen from Table 9, it contains 7,659, 1,787, and 1,710 dialogs for training, validation and testing, respectively. The data for DSTC8 [28] and DSTC10 [58] were only released with 1,710 and 1,804 dialogs for testing, respectively. For all testing splits, six human-generated reference answers were provided for each dialog in order to compute the generation metrics.

### C.2 SIMMC2.0

SIMMC 2.0 [31] is a task-oriented dataset that was proposed for virtual assistance scenarios and contains 11k dialogs with 52,044 unique questions grounded in 5,440 videos from the shopping domain. Its visual and textual data were automatically generated in constrained and pre-defined settings resulting in less complex and challenging scenes compared to AVSD. As can be seen in Figure 7, AVSD features a larger variety of objects that humans interact with daily, more complex dynamics, and more challenging illumination conditions. On the other hand, SIMMC 2.0 only comes with simple items linked to the shopping domain.

### C.3 NExT-QA

NExT-QA [67] was recently introduced as a next generation video question answering benchmark that was introduced to advance video understanding from describing to explaining the temporal actions. Table 10 gives more insight about the statistics of the dataset.



**Fig. 7:** Comparison between the visual complexity of AVSD (a) and SIMMC 2.0 (b). For ethical reasons, we blurred the faces of people appearing in the video frames.



## D Experimental Setup

### D.1 Hardware & Environment

We implemented our model in PyTorch [54] and trained them on a cluster consisting of 8 Nvidia Tesla V100 (32GB) GPUs, 2 Intel(R) Xeon(R) Platinum 8160 CPUs, and 1.5TB of RAM.

### D.2 Training

We trained MST<sub>MIXER</sub> end-to-end using AdamW [47] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 1e - 8$  and a linear learning rate schedule with warm-up for a maximum of 12 epochs. We utilized a learning rate  $lr_{\text{BART}} = 1e - 5$  for the weights of the BART model and a learning rate  $lr_{\text{rest}} = 1e - 4$  for the rest of the parameters of our model. Similarly to  $\lambda$  and  $\Delta$ , we validated the choice of the ELBO loss coefficients  $\alpha_2$  and  $\alpha_3$  based on the validation perplexity. Specifically, we performed a grid search using the value set  $\{1, 10, 100, 1000\}$  while keeping  $\lambda = 0.9$ ,  $\Delta = 4$ , and  $K = 10$ . The training of our full model takes approximately 20 hours to finish. Complete details about the hyperparameter values are listed in Table 12.

### D.3 Inference

Similar to previous works, we utilized beam search with a depth of 5 and a lengths penalty of 0.3 to generate the answers. Each answer is composed of at most 20 tokens. The inference time of our model takes about 2s to answer one question.

## E Additional Ablations

*GNN Types.* We experimented with different types of GNNs within our full model. As depicted in Table 11a, the combination of MST<sub>MIXER</sub> with APPNP [18] led to the best overall performance compared to other GNNs such as GAT [63], GCN [29], and SAGE [21].

**Table 11:** Additional ablations of  $\text{MST}_{\text{MIXER}}$ .

(a) Performance comparison of our best model using different GNN types.

$\text{MST}_{\text{MIXER}}$	AVSD-DSTC7			AVSD-DSTC8		
	B-4	R	C	B-4	R	C
w/ GAT	<u>46.7</u>	61.5	135.4	46.5	60.9	129.4
w/ GCN	46.6	<u>61.9</u>	<u>136.7</u>	<u>46.7</u>	<u>61.6</u>	<u>131.6</u>
w/ SAGE	46.0	61.2	133.4	45.8	60.9	129.3
w/ APPNP	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>47.7</b>	<b>62.3</b>	<b>134.9</b>

(b) Performance comparison between different model sizes. “Base” and “Large” mean that  $\text{MST}_{\text{MIXER}}$  uses a base or a large backbone, respectively.

$\text{MST}_{\text{MIXER}}$	AVSD-DSTC7			AVSD-DSTC8		
	B-4	R	C	B-4	R	C
Base ( $\Delta = 2$ )	39.8	60.0	113.9	40.1	55.4	110.2
Large ( $\Delta = 4$ )	<b>47.6</b>	<b>62.5</b>	<b>138.8</b>	<b>46.7</b>	<b>61.6</b>	<b>131.6</b>

*Mode Size.* Moreover, we experimented with different sizes our model. As depicted in [Table 11b](#), the variant of  $\text{MST}_{\text{MIXER}}$  that is based on BART-base significantly under-performed the large variant across all evaluation metrics of both datasets.

## F Qualitative Results

We provide additional extensive qualitative examples of our best model and some of its ablated versions for comparison in [Figure 8](#). Finally, we give some failure cases in [Figure 9](#).

## References

- Abdessaïed, A., Hochmeister, M., Bulling, A.: OLViT: Multi-modal state tracking via attention-based embeddings for video-grounded dialog. In: LREC-COLING (2024) [2](#)
- Abdessaïed, A., Shi, L., Bulling, A.: VD-GR: Boosting Visual Dialog With Cascaded Spatial-Temporal Multi-Modal Graphs. In: WACV (2024) [6](#)
- Alamri, H., Bilic, A., Hu, M., Beedu, A., Essa, I.: End-to-end multimodal representation learning for video dialog. In: NeurIPS (2022) [9](#)
- Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al.: Audio visual scene-aware dialog. In: CVPR (2019) [3](#), [9](#), [16](#)
- Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) [10](#), [11](#)
- Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Deep compositional question answering with neural module networks. In: CVPR (2016) [3](#)
- Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: NAACL (2016) [3](#)
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015) [2](#), [3](#)

**Table 12:** Detailed hyperparameter setting of the training and inference of our best MST<sub>MIXER</sub> model.

Category	Hyperparameter	
<b>Model Architecture</b>	Dimension of I3D rgb / I3D flow / SAM features $d_v$	2048
	Dimension of SAM features $d_s$	512
	Maximum length of I3D rgb / I3D flow / SAM features $d_l$	36
	Dimension of audio features $d_a$	128
	Maximum length of audio features $l_a = l_v$	36
	Maximum total length of multi-modal input	1024
	Dimension of hidden features $d$	1024/768
	Number of node features in local GNNs $K$	10
	Number of node features in local GNNs $K$	10
	Number for kNNs in $\{\tilde{A}_i\}$	4
	Number of learnable weights of Equation 7	8
	Input dimension of GNNs in Table 11a	1024
	Output dimension of GNNs in Table 11a	1024
	$K$ value of APPNP	2
	$\alpha$ value of APPNP	0.1
	<b>Optimization</b>	Number of attention heads in local GATs
Number of attention heads in global GATs		4
$\lambda$ value		0.9
$\Delta$ value		4
Optimizer		AdamW
Learning rate of parameters in the VLM backbone $lr_{\text{BART}}$		1e-5
Learning rate of other parameters $lr_{\text{rest}}$		1e-4
Values of $\{\alpha_1, \alpha_2, \alpha_3\}$		{1, 100, 100}
Learning rate schedule		linear
Dropout rate		0.1
<b>Hardware</b>	Value of gradient clipping	1.0
	Effective batch size	96
	Number of epochs	12
	GPU model	Tesla V100-32GB
	Number of GPUs	8
	Distributed training	PyTorch DDP
	<b>Inference</b>	Maximum number of response tokens
Depth of beam search		5
Length penalty in beam search		0.3
Batch size		1

9. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (2005) 9
10. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 5
11. Chen, Y., Wu, L., Zaki, M.: Iterative deep graph learning for graph neural networks: Better and robust node embeddings. NeurIPS (2020) 4, 7
12. Chen, Z., Liu, H., Wang, Y.: DialogMCF: Multimodal Context Flow for Audio Visual Scene-Aware Dialog. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023) 9, 10
13. Chu, Y.W., Lin, K.Y., Hsu, C.C., Ku, L.W.: Multi-step joint-modality attention network for scene-aware dialogue system. In: DSTC Workshop @ AAI (2020) 9

14. Colson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. *Annals of operations research* (2007) [4](#)
15. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual Dialog. In: *CVPR* (2017) [2, 3](#)
16. Elinas, P., Bonilla, E.V., Tiao, L.: Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. *NeurIPS* (2020) [4](#)
17. Franceschi, L., Niepert, M., Pontil, M., He, X.: Learning discrete structures for graph neural networks. In: *ICML* (2019) [4](#)
18. Gasteiger, J., Bojchevski, A., Günnemann, S.: Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In: *ICLR* (2019) [8, 17](#)
19. Girdhar, R., Ramanan, D.: CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In: *ICLR* (2020) [2, 3](#)
20. Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., Feris, R.: Dialog-based interactive image retrieval. *NeurIPS* **31** (2018) [2, 3](#)
21. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs (2017) [17](#)
22. Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A., Essa, I., Batra, D., Parikh, D.: End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In: *ICASSP* (2019) [5, 9](#)
23. Huang, X., Tan, H.L., Leong, M.C., Sun, Y., Li, L., Jiang, R., Kim, J.: Investigation on transformer-based multi-modal fusion for audio-visual scene-aware dialog. In: *DSTC10 Workshop @ AAAI* (2022) [10](#)
24. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020) [10, 11](#)
25. Jin, Y., Niu, G., Xiao, X., Zhang, J., Peng, X., Yu, J.: Knowledge-Constrained Answer Generation for Open-Ended Video Question Answering. In: *AAAI* (2023) [10, 11](#)
26. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017) [5](#)
27. Kim, J., Yoon, S., Kim, D., Yoo, C.D.: Structured co-reference graph attention for video-grounded dialogue. In: *AAAI* (2021) [3, 9](#)
28. Kim, S., Galley, M., Gunasekara, C., Lee, S., Atkinson, A., Peng, B., Schulz, H., Gao, J., Li, J., Adada, M., et al.: The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394* (2019) [9, 16](#)
29. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. In: *ICLR* (2017) [17](#)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. *arXiv preprint arXiv:2304.02643* (2023) [5, 10](#)
31. Kottur, S., Moon, S., Geramifard, A., Damavandi, B.: SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In: *EMNLP* (2021) [2, 3, 9, 10, 16](#)
32. Le, H., Chen, N.F., Hoi, S.: Learning reasoning paths over semantic graphs for video-grounded dialogues. In: *ICLR* (2021) [3, 5, 9, 10](#)
33. Le, H., Chen, N.F., Hoi, S.C.H.: VGNMN: video-grounded neural module network to video-grounded language tasks. In: *NAACL* (2022) [3, 9](#)
34. Le, H., Chen, N.F., Hoi, S.C.: Multimodal Dialogue State Tracking. In: *NAACL* (2022) [2, 3](#)

35. Le, H., Hoi, S.C.: Video-Grounded Dialogues with Pretrained Generation Language Models. In: ACL (2020) [3](#), [9](#)
36. Le, H., Sahoo, D., Chen, N., Hoi, S.: Multimodal transformer networks for end-to-end video-grounded dialogue systems. In: ACL (2019) [9](#), [10](#)
37. Le, H., Sahoo, D., Chen, N., Hoi, S.C.: BiST: Bi-directional Spatio-Temporal Reasoning for Video-Grounded Dialogues. In: EMNLP (2020) [9](#)
38. Le, H., Sankar, C., Moon, S., Beirami, A., Geramifard, A., Kottur, S.: DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. In: ACL (2021) [3](#)
39. Le, H., Socher, R., Hoi, S.C.: Non-autoregressive dialog state tracking. In: ICLR (2020) [2](#)
40. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: CVPR (2020) [10](#), [11](#)
41. Lee, H., Kwon, O.J., Choi, Y., Park, M., Han, R., Kim, Y., Kim, J., Lee, Y., Shin, H., Lee, K., Kim, K.E.: Learning to embed multi-modal contexts for situated conversational agents. In: NAACL-Findings (Jul 2022) [10](#)
42. Lee, H., Lee, J., Kim, T.Y.: SUMBT: Slot-utterance matching for universal and scalable belief tracking. In: ACL (2019) [2](#)
43. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020) [3](#), [5](#), [12](#)
44. Li, Y., Hui, B., Yin, Z., Yang, M., Huang, F., Li, Y.: PaCE: Unified Multi-modal Dialogue Pre-training with Progressive and Compositional Experts. In: ACL (2023) [10](#), [11](#)
45. Li, Z., Li, Z., Zhang, J., Feng, Y., Zhou, J.: Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. Transactions on Audio, Speech, and Language Processing (2021) [3](#), [5](#), [9](#)
46. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004) [9](#)
47. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. In: ICLR (2019) [17](#)
48. Malinowski, M., Fritz, M.: A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In: NeurIPS (2014) [9](#), [11](#)
49. Moon, S., Kottur, S., Crook, P.A., De, A., Poddar, S., Levin, T., Whitney, D., Difrancio, D., Beirami, A., Cho, E., Subba, R., Geramifard, A.: Situated and interactive multimodal conversations. In: COLING (2020) [2](#), [3](#), [11](#)
50. Mou, X., Sigouin, B., Steenstra, I., Su, H.: Multimodal dialogue state tracking by QA approach with data augmentation. In: DSTC8 Workshop @ AAI (2020) [2](#), [3](#)
51. Mrkšić, N., Ó Séaghdha, D., Wen, T.H., Thomson, B., Young, S.: Neural belief tracker: Data-driven dialogue state tracking. In: ACL (2017) [2](#)
52. Pang, W., Wang, X.: Visual dialogue state tracking for question generation. In: AAI (2020) [2](#), [3](#)
53. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002) [9](#)
54. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: NeurIPS (2019) [17](#)

55. Pham, H.A., Le, T.M., Le, V., Phuong, T.M., Tran, T.: Video Dialog as Conversation about Objects Living in Space-Time. In: ECCV (2022) [3](#), [9](#)
56. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019) [3](#)
57. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014) [4](#)
58. Shah, A., Geng, S., Gao, P., Cherian, A., Hori, T., Marks, T.K., Le Roux, J., Hori, C.: Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In: ICASSP (2022) [9](#), [10](#), [16](#)
59. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016) [3](#), [9](#)
60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) [5](#)
61. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative Pretraining in Multimodality. arXiv preprint arXiv:2307.05222 (2023) [10](#), [11](#)
62. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. In: CVPR (2015) [9](#)
63. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. In: ICLR (2018) [17](#)
64. Wu, C.S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R., Fung, P.: Transferable multi-domain state generator for task-oriented dialogue systems. In: ACL (2019) [2](#)
65. Wu, Q., Zhao, W., Li, Z., Wipf, D., Yan, J.: Nodeformer: A scalable graph structure learning transformer for node classification. In: NeurIPS (2022) [4](#)
66. Wu, Y., Macdonald, C., Ounis, I.: Multi-modal dialog state tracking for interactive fashion recommendation. In: ACM RecSys (2022) [2](#), [3](#)
67. Xiao, J., Shang, X., Yao, A., Chua, T.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR (2021) [9](#), [11](#), [16](#)
68. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM MM (2017) [2](#)
69. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016) [3](#)
70. Xu, P., Hu, Q.: An end-to-end approach for handling unknown slot values in dialogue state tracking. In: ACL (2018) [2](#)
71. Yang, J., Liu, Z., Xiao, S., Li, C., Lian, D., Agrawal, S., Singh, A., Sun, G., Xie, X.: GraphFormers: GNN-nested Transformers for Representation Learning on Textual Graph. In: NeurIPS (2021) [6](#)
72. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? In: NeurIPS (2021) [6](#)
73. Yoon, S., Yoon, E., Yoon, H.S., Kim, J., Yoo, C.: Information-theoretic text hallucination reduction for video-grounded dialogue. In: EMNLP (2022) [3](#), [5](#), [9](#), [10](#)
74. Yoshino, K., Hori, C., Perez, J., D’Haro, L.F., Polymenakos, L., Gunasekara, C., Lasecki, W.S., Kummerfeld, J.K., Galley, M., Brockett, C., et al.: Dialog system technology challenge 7. arXiv preprint arXiv:1901.03461 (2019) [9](#), [16](#)
75. Yu, Y., Chen, J., Gao, T., Yu, M.: DAG-GNN: DAG structure learning with graph neural networks. In: ICML (2019) [4](#)

76. Zhang, H., Liu, M., Wang, Y., Cao, D., Guan, W., Nie, L.: Uncovering hidden connections: Iterative tracking and reasoning for video-grounded dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) [9](#), [10](#)

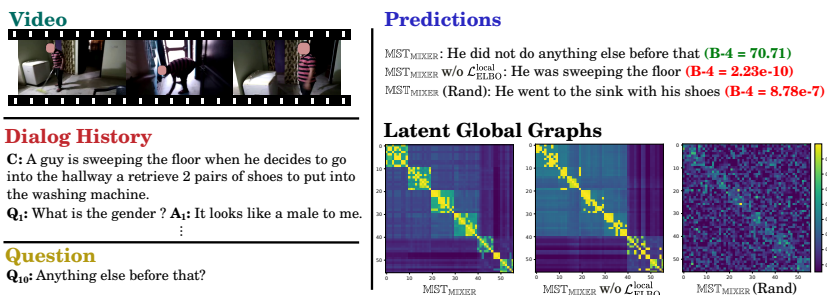
```

1 # -----
2 # Implementation of the ELBO loss
3 # -----
4 import torch
5 import torch.nn as nn
6 import torch.nn.functional as F
7
8 class ELBO(nn.Module):
9     def __init__(self):
10         super(ELBO, self).__init__()
11
12     def forward(self, Aq, Ap):
13         """
14         Args:
15             Aq: The predicted latent graph of q_phi
16                 shape = (batch_size, K, K) -- local graphs
17                 shape = (batch_size, NK, NK) -- global graphs
18
19             Ap: The predicted latent graph of p_theta
20                 shape = (batch_size, K, K) -- local graphs
21                 shape = (batch_size, NK, NK) -- global graphs
22
23         Returns:
24             The ELBO loss
25         """
26         Aq_flat = Aq.view(-1).unsqueeze(-1)
27         Ap_flat = Ap.view(-1).unsqueeze(-1)
28
29         Aq_flat = torch.cat(
30             [torch.zeros_like(Aq_flat), Aq_flat], dim=-1)
31         Ap_flat = torch.cat(
32             [torch.zeros_like(Ap_flat), Ap_flat], dim=-1)
33
34         log_Aq = F.log_softmax(QA_flattened, dim=1)
35         log_Ap = F.log_softmax(PA_flattened, dim=1)
36
37         Aq_dist = torch.exp(log_Aq)
38
39         loss_Aq = torch.mean(log_Aq * Aq_dist)
40         loss_Ap = torch.mean(log_Ap * Ap_dist)
41
42         elbo_loss = loss_Aq - loss_Ap
43
44         return elbo_loss
45

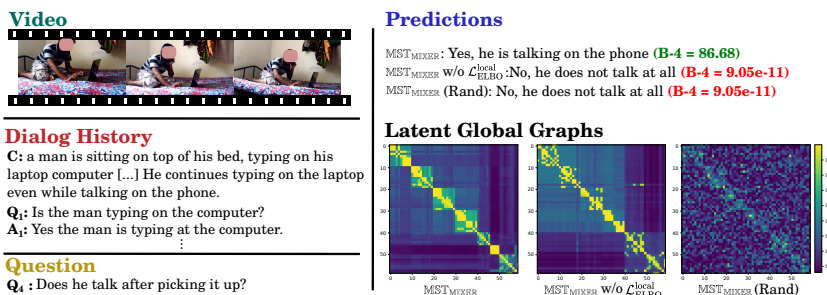
```

**Listing 1.1:** PyTorch implementation of the ELBO loss. Since  $q_\phi$  and  $p_\theta$  only output normalized scores as the prediction for each edge, we append the zero vectors to both predictions in lines 29-30 to convert the raw scores to a two-value probability before applying softmax.

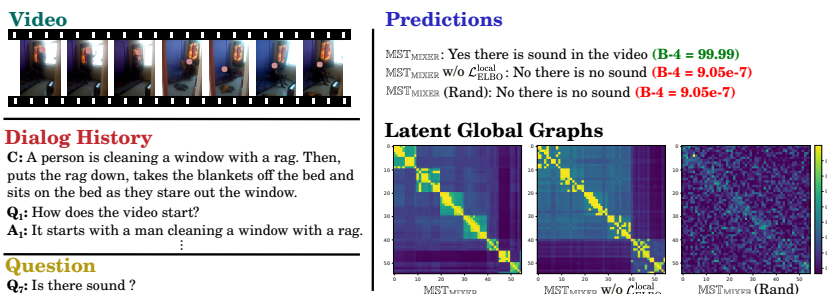




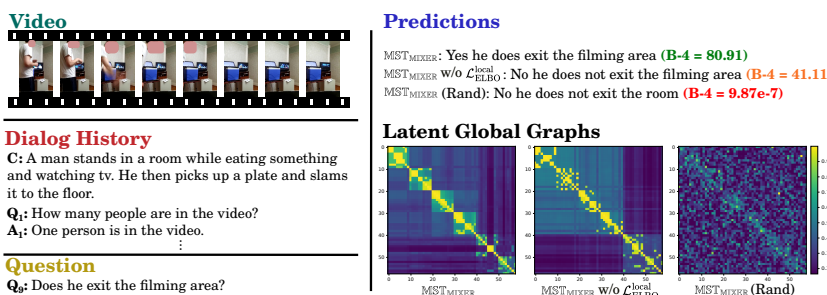
(a) Dialog with video-id = 3A91C.



(b) Dialog with video-id = 2EW71.

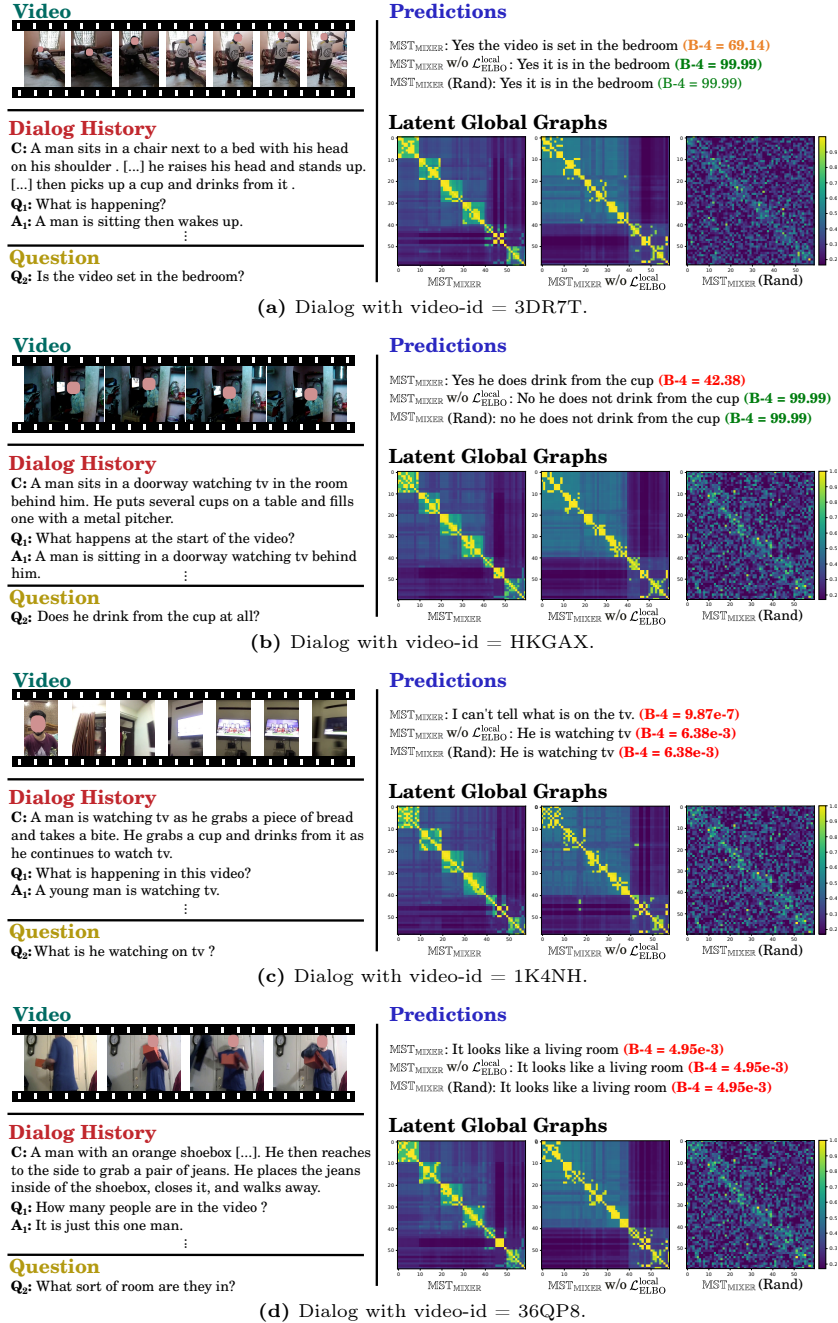


(c) Dialog with video-id = K2XKT.



(d) Dialog with video-id = UO7PC. Although the ablated version ( $MSTMIXER w/o \mathcal{L}_{ELBO}^{local}$ ) reached a BLEU-4 score of 41.11, it incorrectly answered the question since the person did leave the filming area as can be seen from the last frames of the video.

**Fig. 8:** Qualitative results on data samples from the test split of AVSD-DSTC7. For ethical reasons, we blurred the faces of people appearing in the video frames.



**Fig. 9:** Negative qualitative results on data samples form the test split of AVSD-DSTC7. For ethical reasons, we blurred the faces of people appearing in the video frames.