University of Stuttgart
Germany

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

MST$_{MIXER}$: Multi-Modal Video Dialog State Tracking in the Wild

Adnen Abdessaied, Lei Shi, Andreas Bulling
University of Stuttgart, Germany
adnen.abdessaied@vis.uni-stuttgart.de

European Conference on Computer Vision (ECCV)

Milano, Italy

02.10.2024

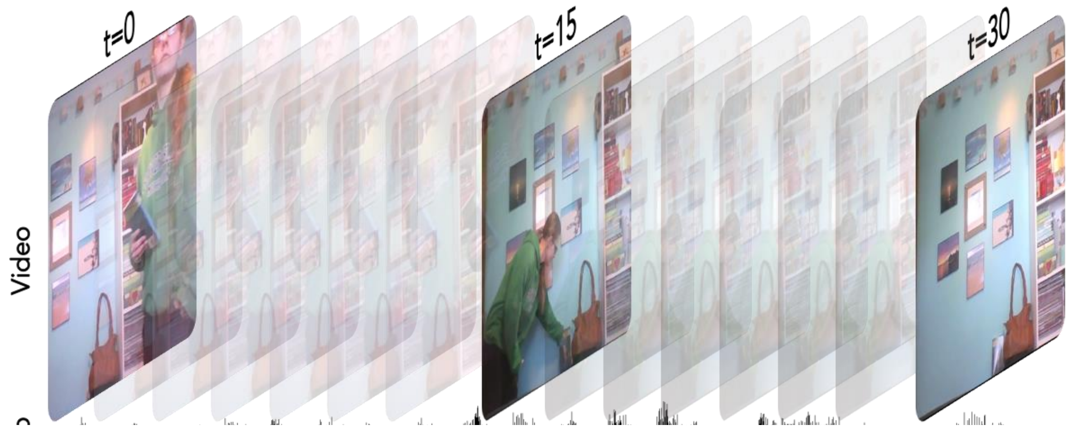
Introduction

Introduction

Video Dialog - Task formulation

-

Introduction



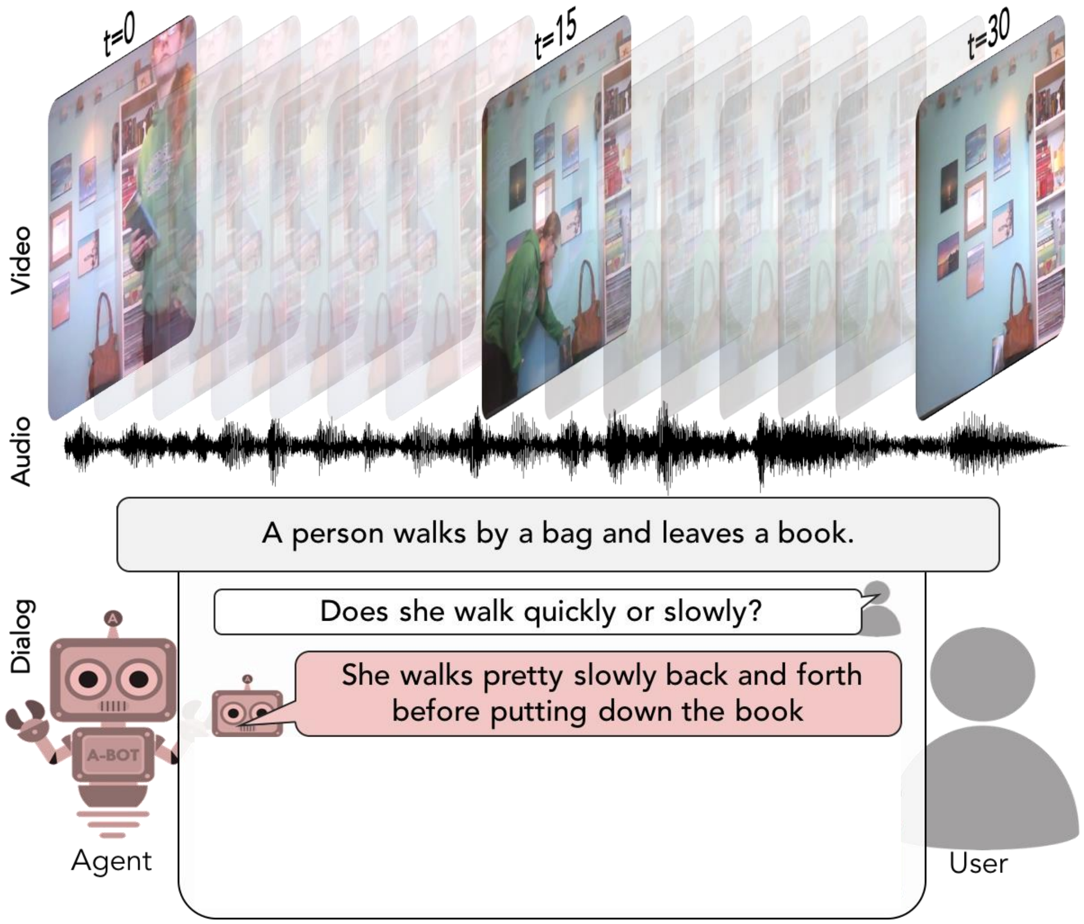
Video Dialog - Task formulation
Given a video,

Introduction

Video Dialog - Task formulation

Given a video, audio data,

Introduction

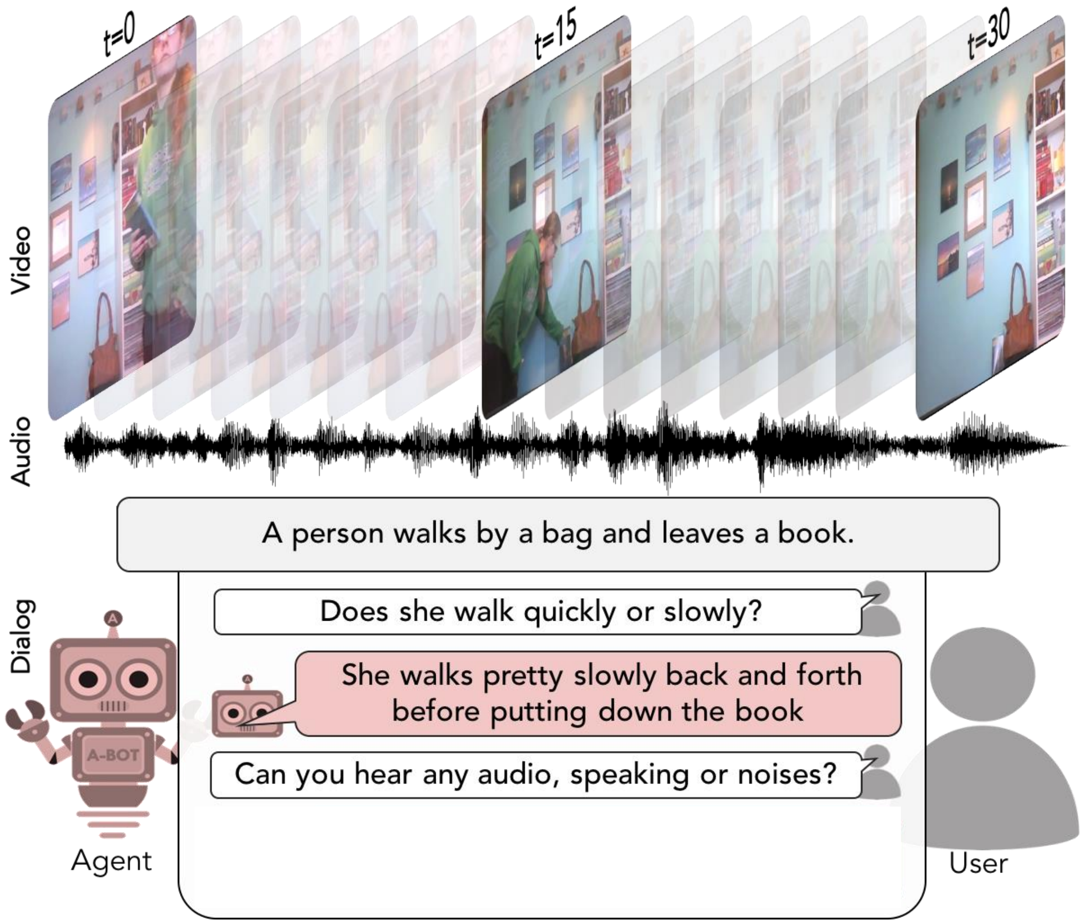


From https://video-dialog.com/

Video Dialog - Task formulation

Given a video, audio data, a dialog history,

Introduction

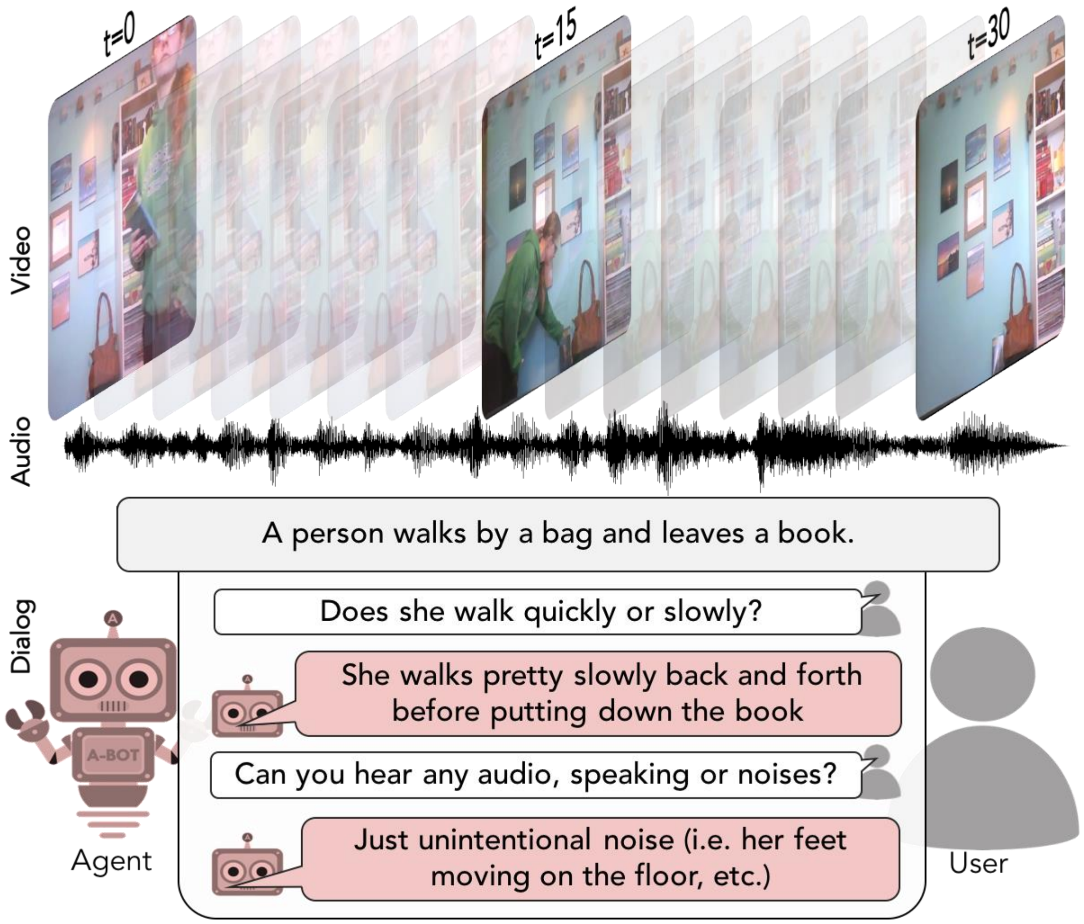


From https://video-dialog.com/

Video Dialog - Task formulation

Given a video, audio data, a dialog history, and a question at time step t,

Introduction

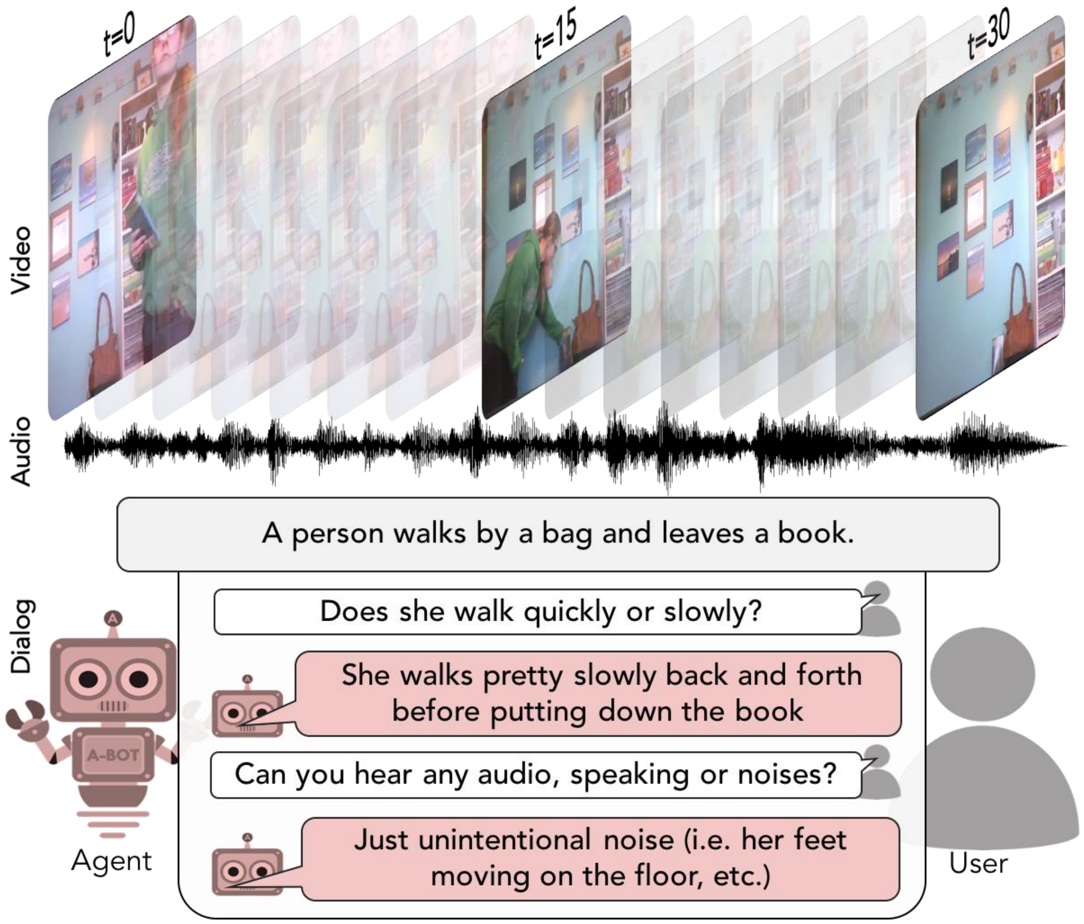


From https://video-dialog.com/

Video Dialog - Task formulation

Given a video, audio data, a dialog history, and a question at time step *t*, generate an appropriate answer auto-regressively

Introduction



From https://video-dialog.com/

Video Dialog - Task formulation
Given a video, audio data, a dialog history, and a question at time step *t*, generate an appropriate answer auto-regressively

Video Dialog is a natural extension to visual dialog

- Video vs image
- Complements videos with audio data

More input modalities

Motivation

Motivation

Video Dialog [1] is a highly multi-modal task → More challenging than similar tasks

- **VQA [2] & VideoQA [3]:** Reasoning about the dialog history in addition to the question
- **Visual Dialog [4]:** Reasoning about a dynamic scene instead of a static image

↓

Dialog State Tracking (DST) is crucial in building capable models

- **DST** was originally introduced to track and update users' goals in form of dialog states [5, 6]

- Now, it is broadly used to describe a model that keeps track of what it believes to be relevant for answering the question at hand

Motivation

Research on DST has been predominately uni-modal in the form of slot-filling tasks [7, 8]

↓

The current landscape necessitates extending DST to the multi-modal domain

Current models with “multi-modal” DST fall short in two major aspects:

- track the constituent of only one modality within a multi-modal task [9, 10]
  → uni-modal DST
- limited to synthetic and automatically-generated datasets [11, 12, 13]
  → do not reflect the complexity of real world scenarios

Motivation

MST_{MIXER} addresses the aforementioned limitations with

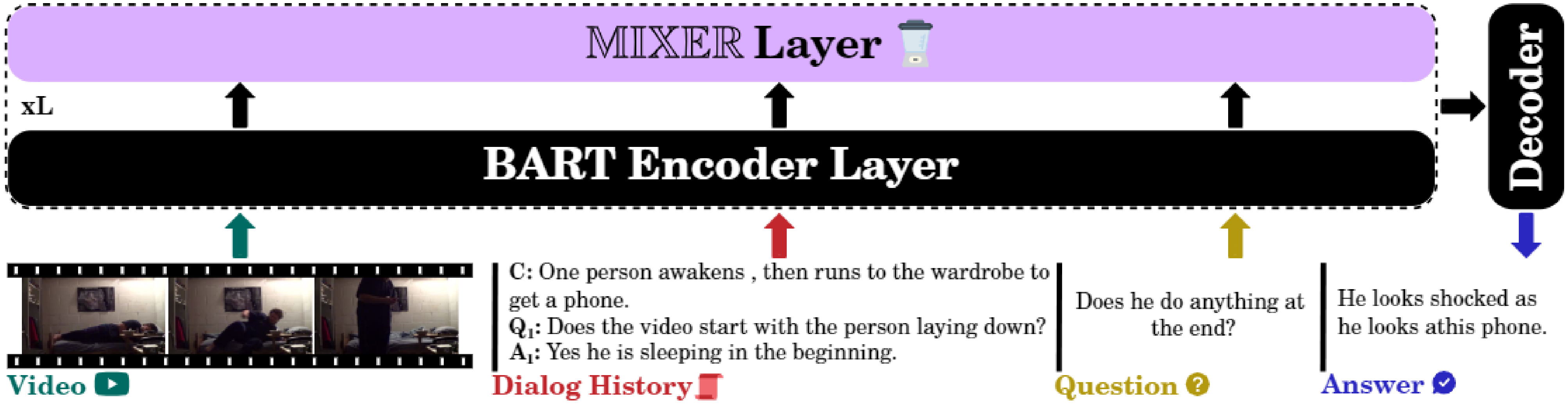
- ✓ modality-specific tracking blocks to identify the most relevant constituents of each modality
- ✓ a multi-modal GNN approach to learn the underlying structure between the mix of modalities

→
- ✓ Performs multi-modal state tracking in the real sense of the word
- ✓ Can tackle a wide-range of real-world datasets and benchmarks

Method

Main Idea

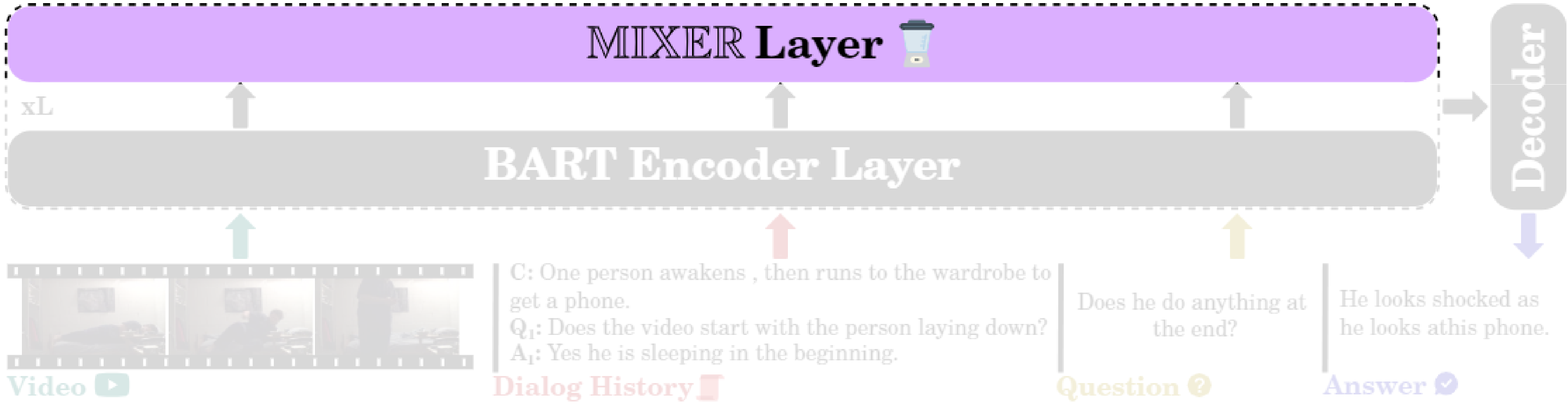
- Perform multi-modal state tracking using MIXER layers
- Interleave BART encoder layers with MIXER layers → Enhance their hidden states



Method

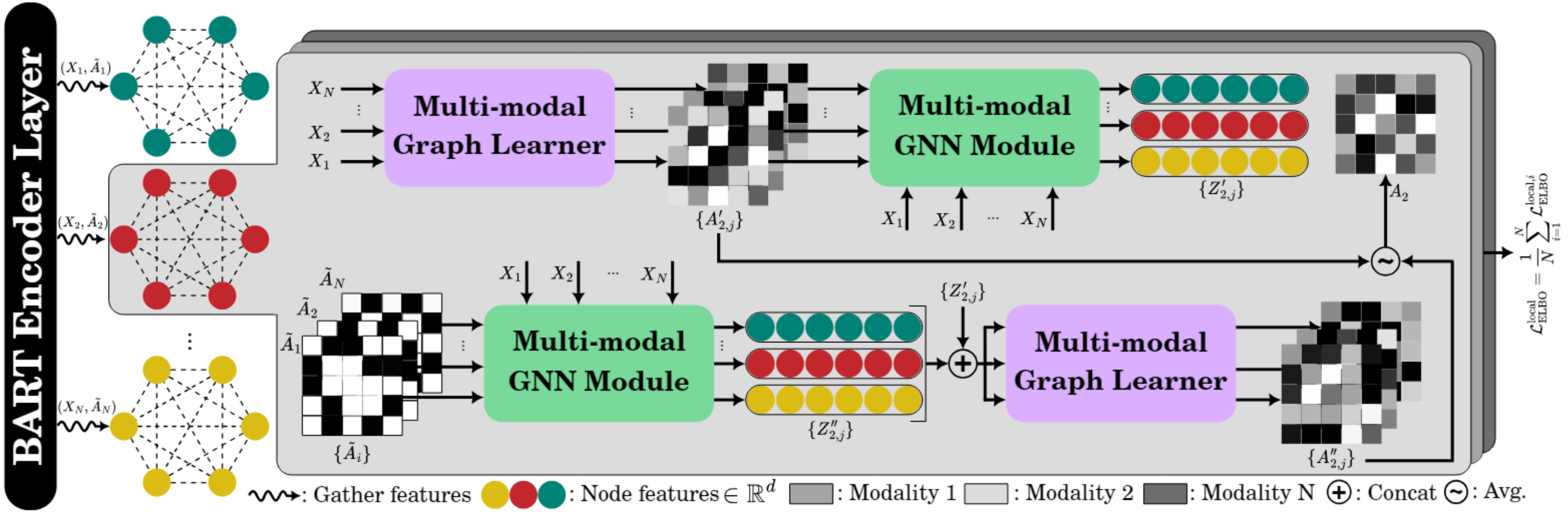
MIXER layer

- Keeps track of the most relevant constituents of each modality at different semantic levels
- Employs a ***divide-and-conquer*** approach:
  local structures of individual modalities → global structure of the mix of all modalities



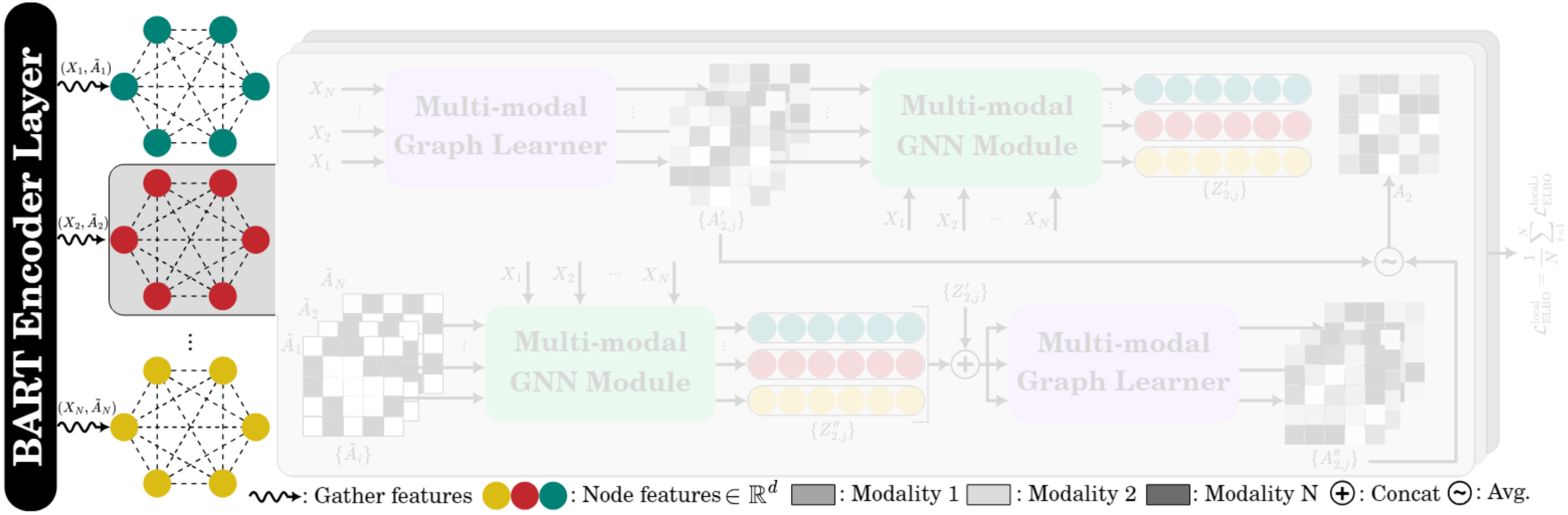
Method

Divide Stage



Method

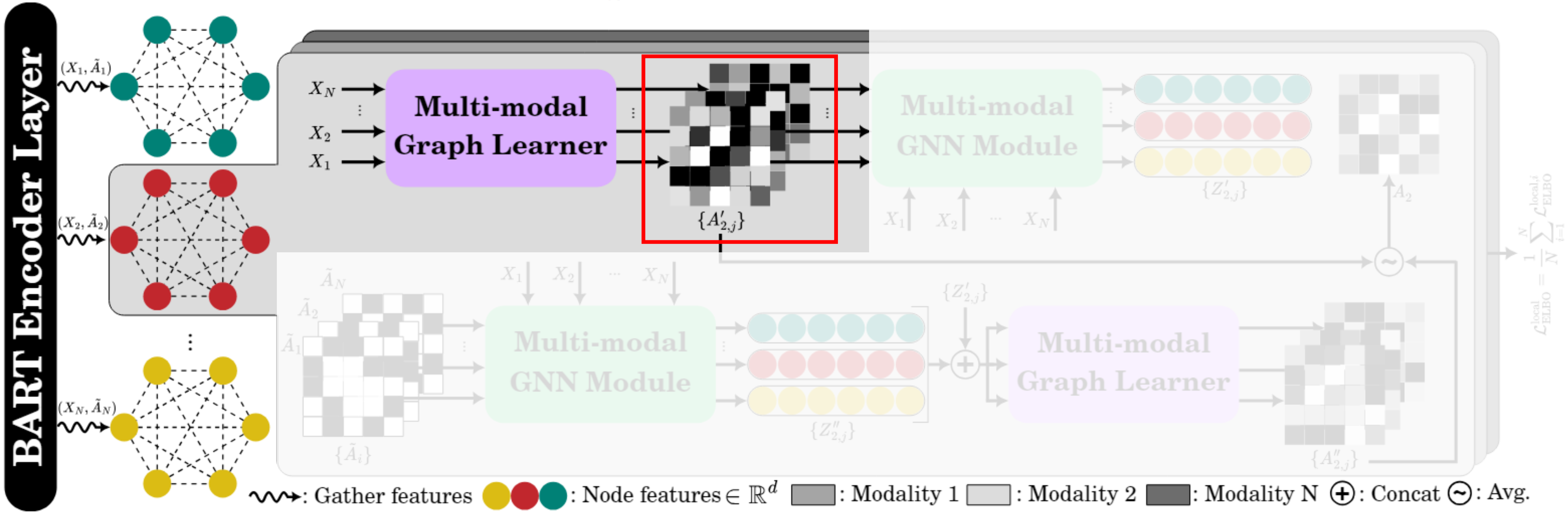
Divide Stage



Method

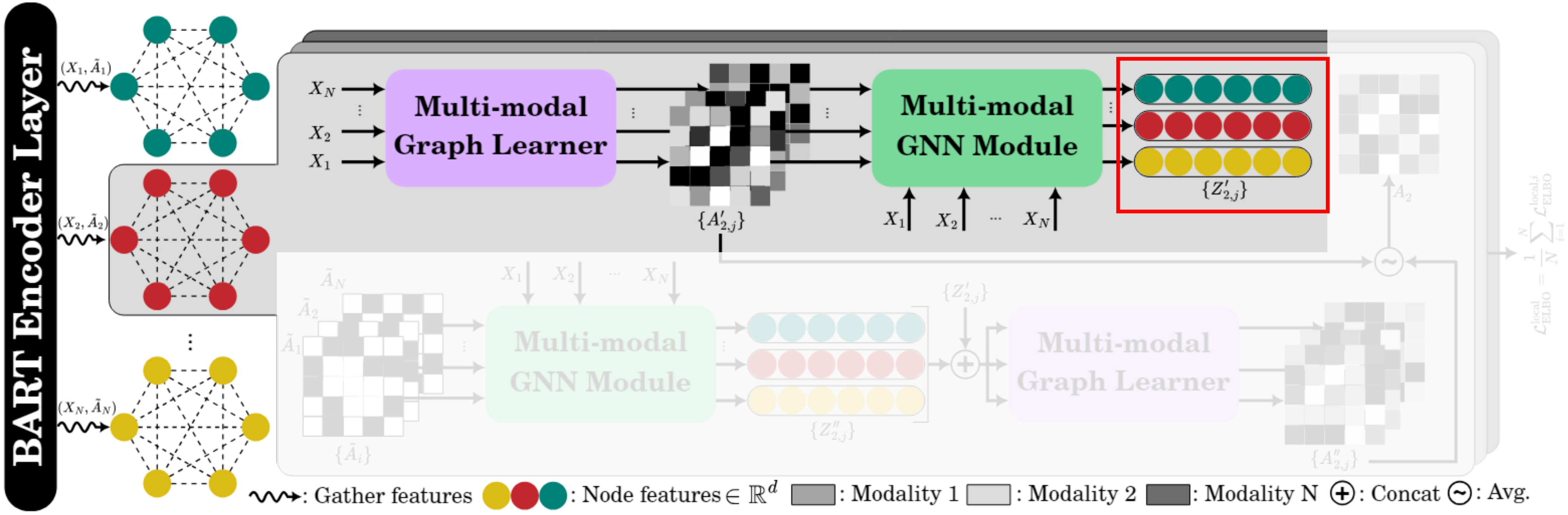
Divide Stage

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



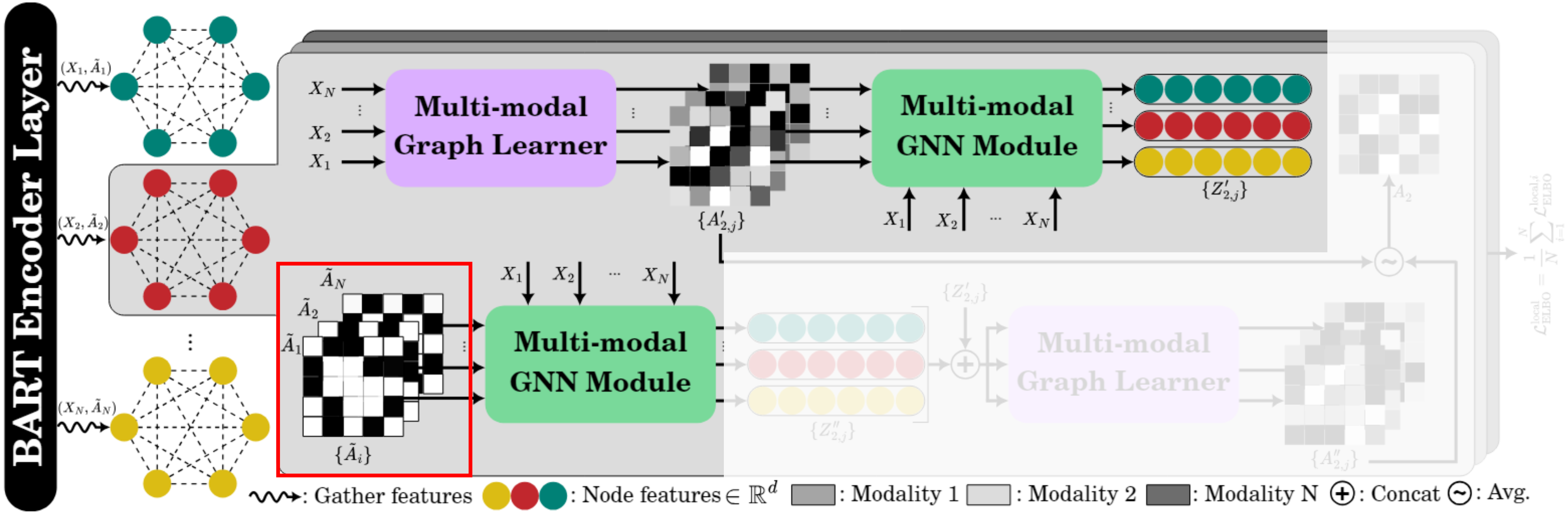
Method

Divide Stage



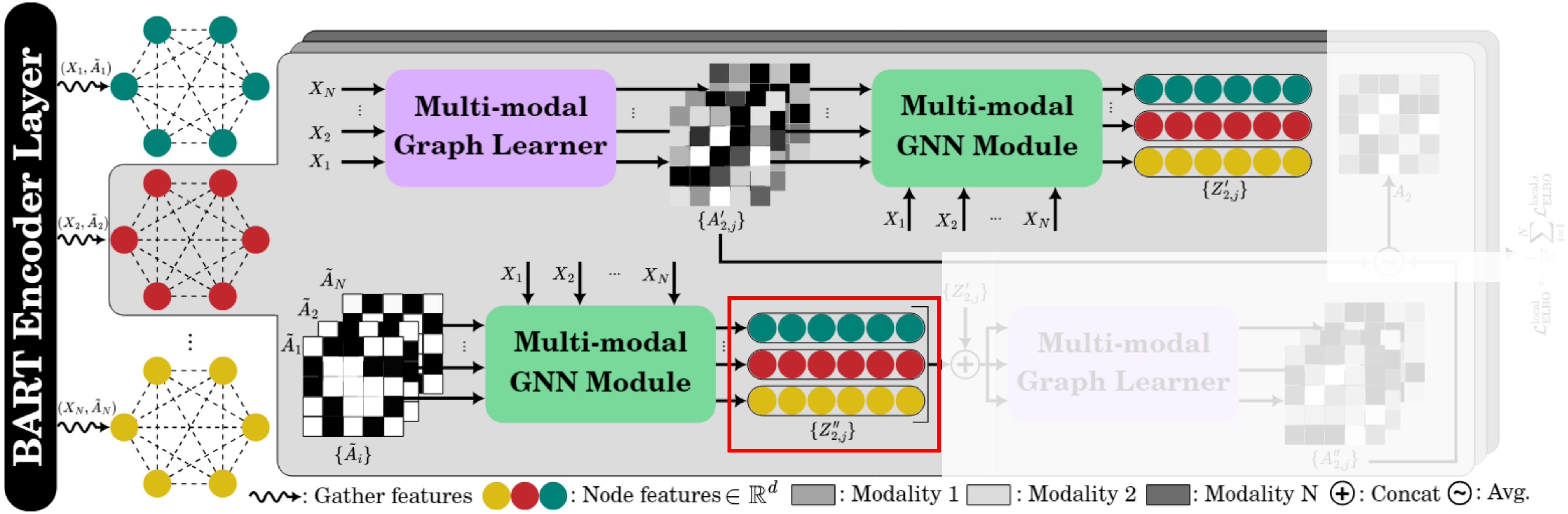
Method

Divide Stage



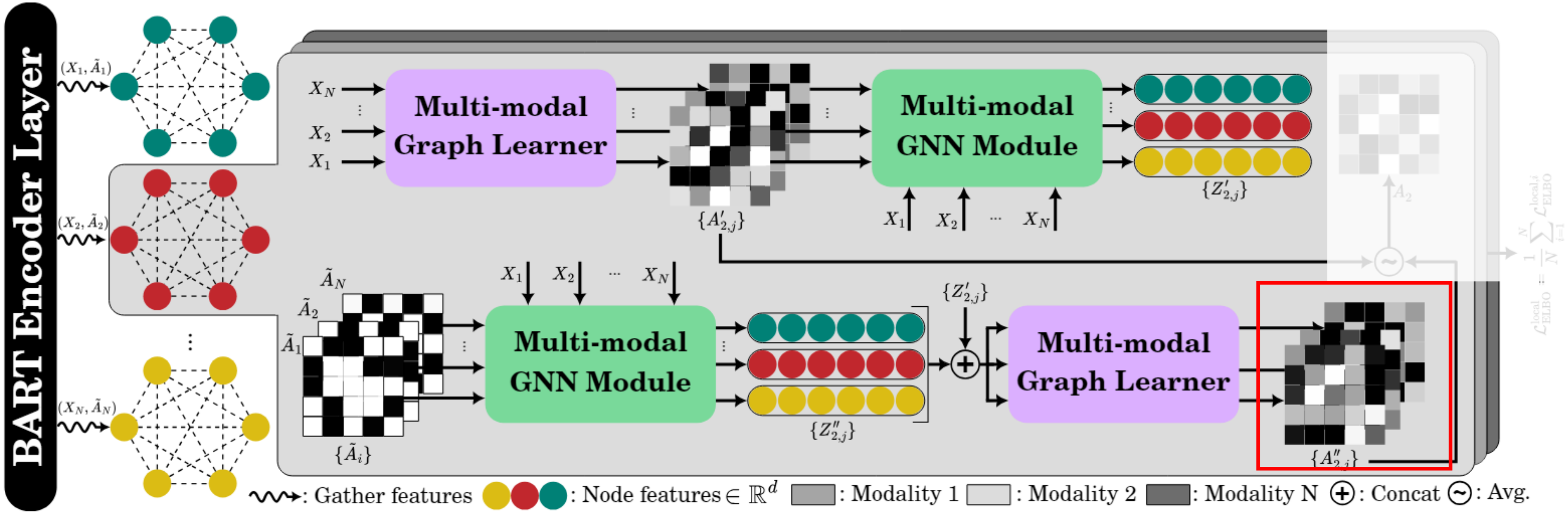
Method

Divide Stage



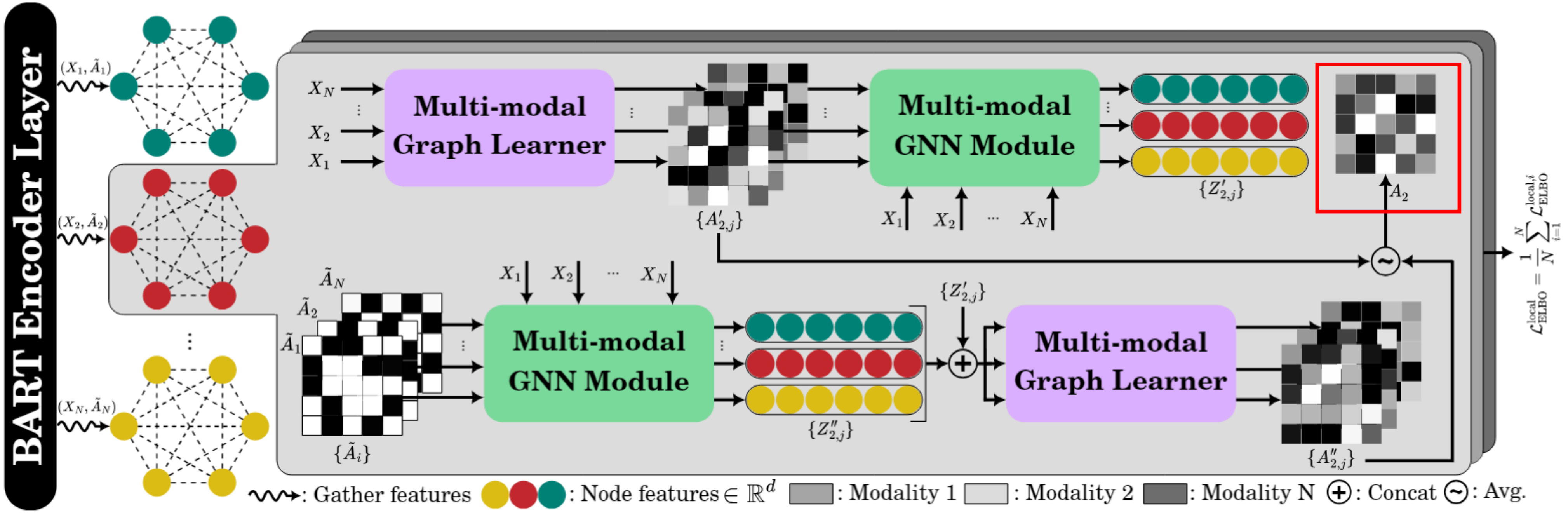
Method

Divide Stage



Method

Divide Stage

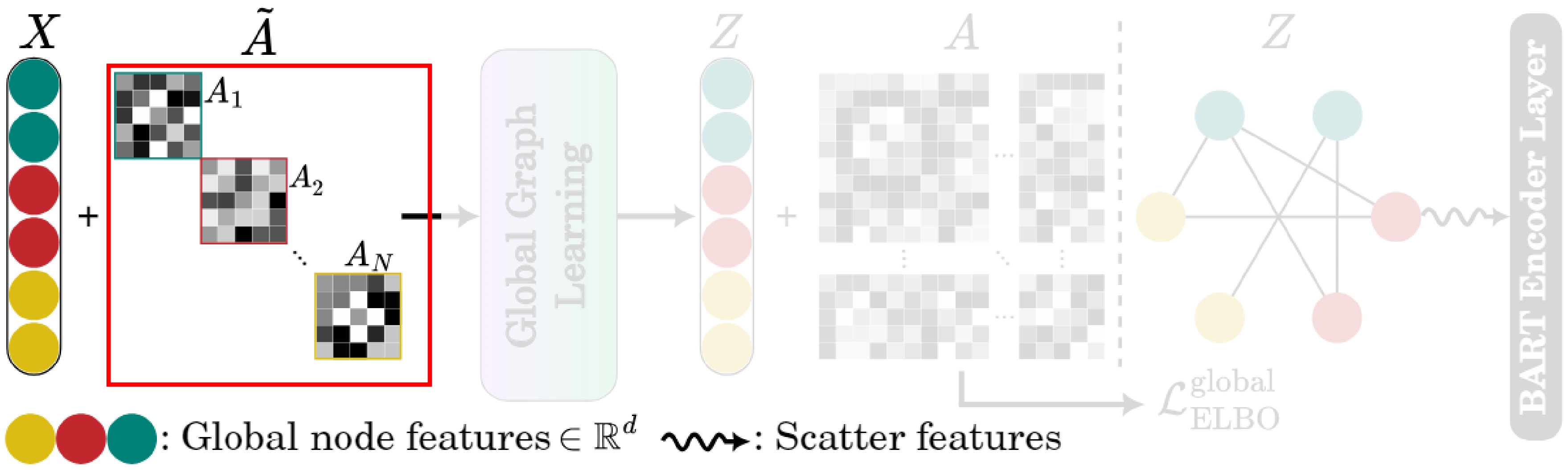


Method

Conquer Stage

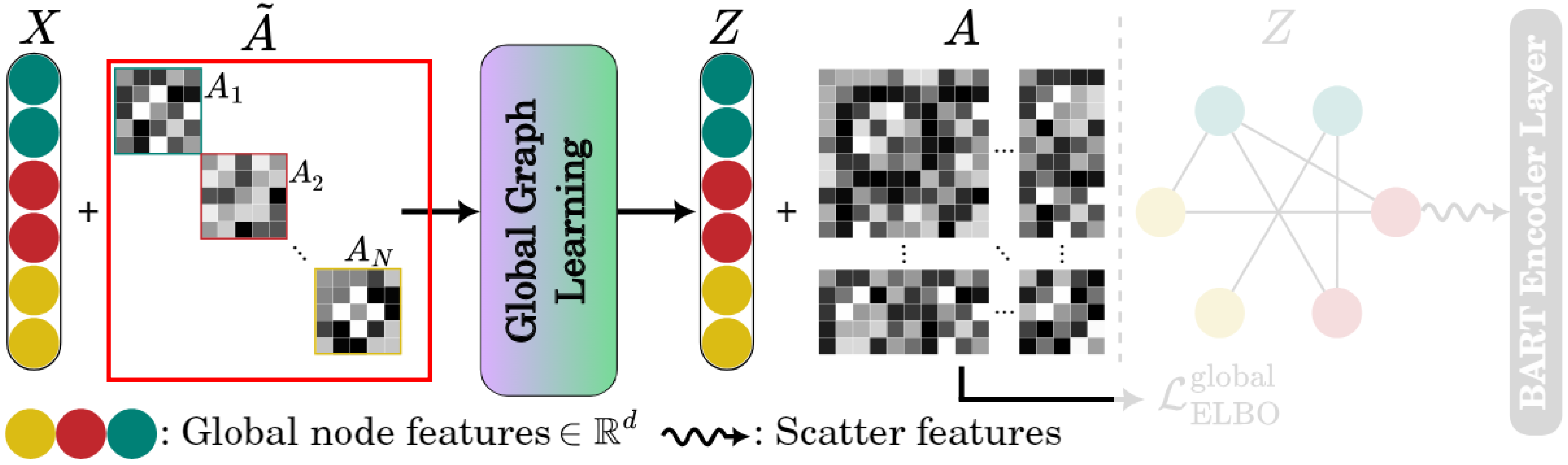
Method

Conquer Stage



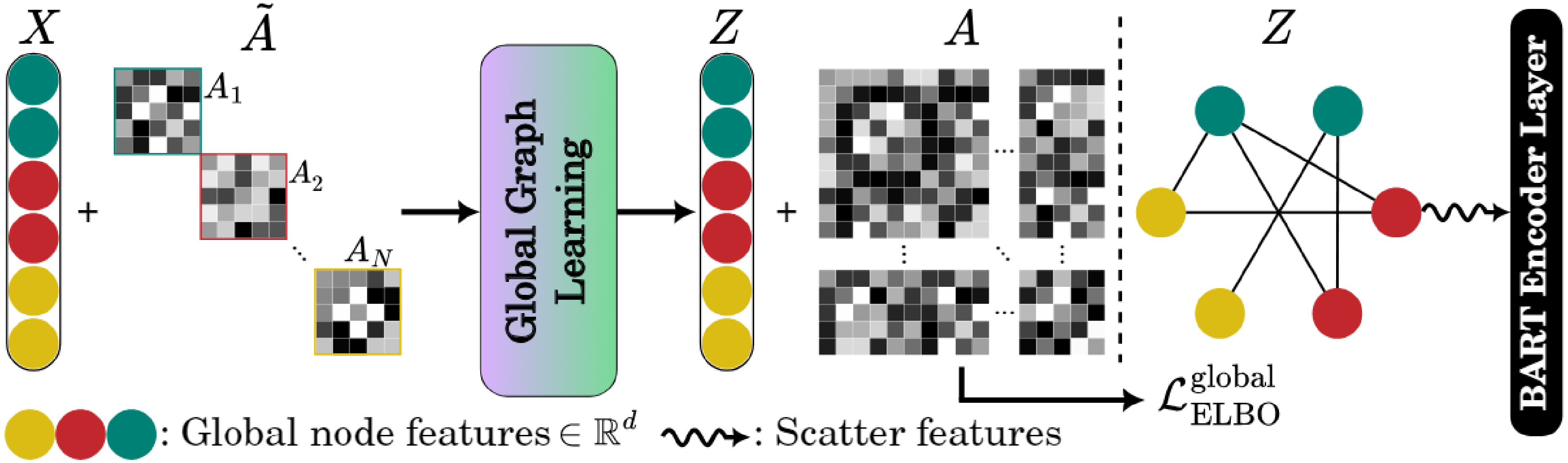
Method

Conquer Stage



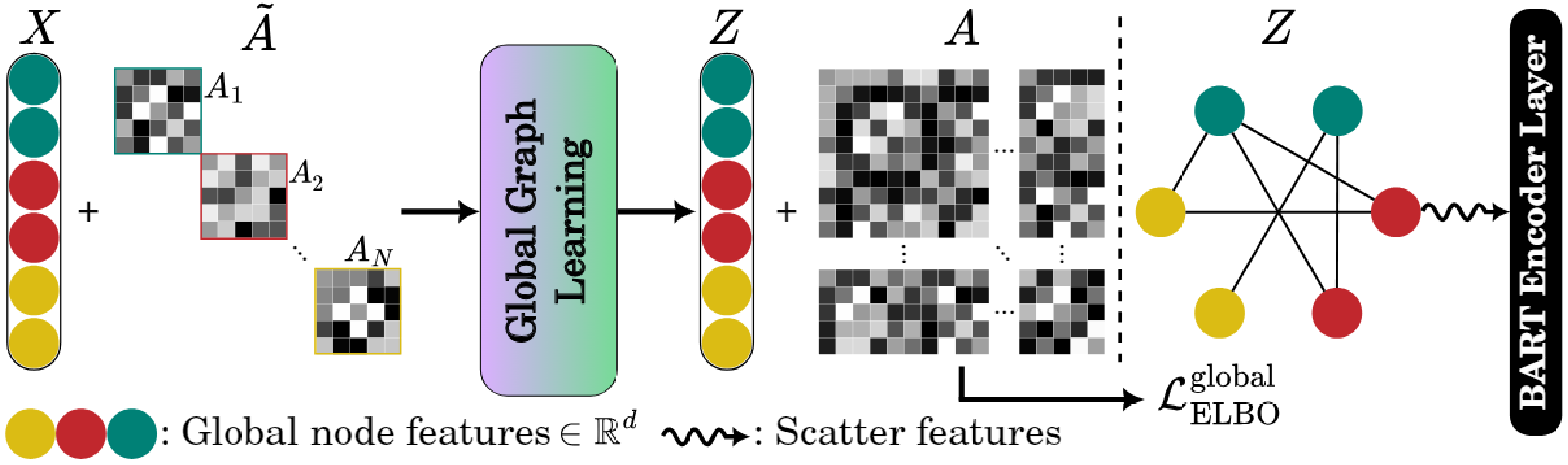
Method

Conquer Stage



Method

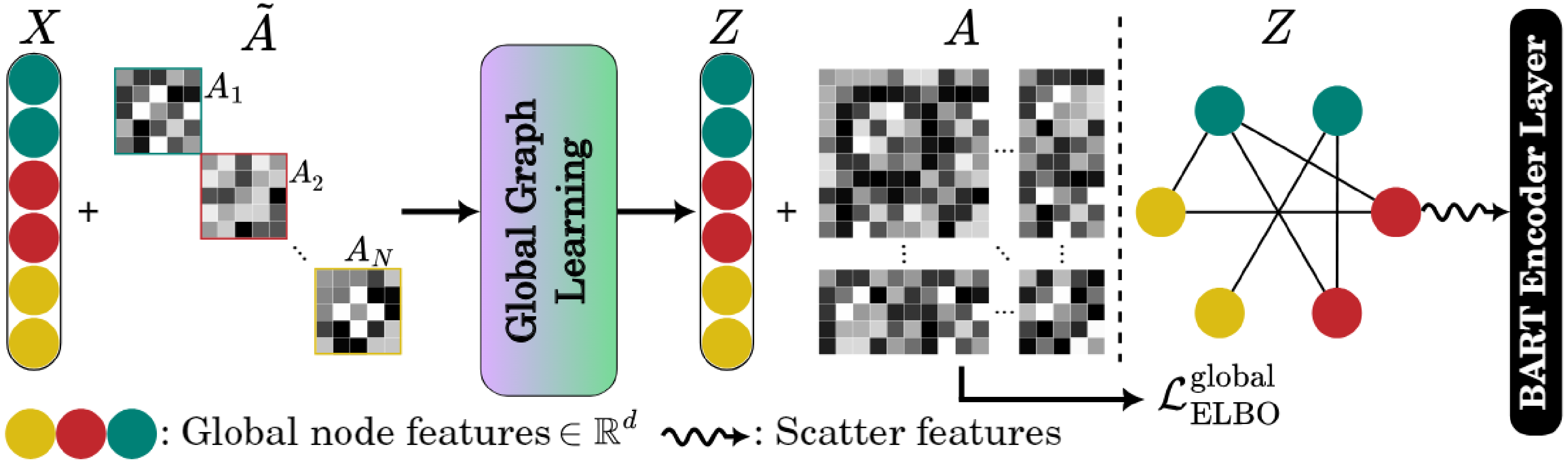
Conquer Stage



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

Method

Conquer Stage



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

- H: hidden states
- $\lambda \in [0, 1]$: hyper-parameters
- $\oslash$: PyTorch scatter operation
- Z: global graph features
- Idx: indices of Z w.r.t. H

Method

Method

Training

- We trained our model end-to-end using the following combination of losses:

Method

Training

- We trained our model end-to-end using the following combination of losses:

$$\mathcal{L} = \alpha_1 \boxed{\mathcal{L}_{\text{gen}}}$$

Next token prediction loss

Method

Training

- We trained our model end-to-end using the following combination of losses:

$$\mathcal{L} = \alpha_1 \boxed{\mathcal{L}_{\text{gen}}} - \alpha_2 \boxed{\mathcal{L}_{\text{ELBO}}^{\text{local}}}$$

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

Next token prediction loss

Average ELBO loss of the local latent graphs

Method

Training

- We trained our model end-to-end using the following combination of losses:

$$\mathcal{L} = \alpha_1 \boxed{\mathcal{L}_{\text{gen}}} - \alpha_2 \boxed{\mathcal{L}_{\text{ELBO}}^{\text{local}}} - \alpha_3 \boxed{\mathcal{L}_{\text{ELBO}}^{\text{global}}},$$

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

Next token prediction loss

Average ELBO loss of the local latent graphs

ELBO loss of the global latent graph

Results

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	AVSD-DSTC7						
		B-1	B-2	B-3	B-4	M	R	C
Baseline	*ICASSP'19*	62.1	48.0	37.9	30.5	21.7	48.1	73.3
MTN	*ACL'19*	71.5	58.1	47.6	39.2	26.9	55.9	106.6
JMAN	*AAAI'20*	66.7	52.1	41.3	33.4	23.9	53.3	94.1
VGD	*ACL'20*	74.9	62.0	52.0	43.6	28.2	58.2	119.4
BiST	*EMNLP'20*	75.5	61.9	51.0	42.9	28.4	58.1	119.2
SCGA	*AAAI'21*	74.5	62.2	51.7	43.0	28.5	57.8	120.1
RLM	*TASLP'21*	76.5	64.3	54.3	45.9	29.4	60.6	130.8
PDC	*ICLR'21*	77.0	65.3	53.9	44.9	29.2	60.6	129.5
AV-TRN	*ICASSP'22*	–	–	–	40.6	26.2	55.4	107.9
VGNMN	*NAACL'22*	–	–	–	42.9	27.8	57.8	118.8
COST	*ECCV'22*	72.3	58.9	48.3	40.0	26.6	56.1	108.5
MRLV	*NeurIPS'22*	–	59.2	49.3	41.5	26.9	56.9	115.9
THAM	*EMNLP'22*	77.8	65.4	54.9	46.8	30.8	61.9	133.5
DialogMCF	*TASLP'23*	77.7	65.3	54.7	45.7	30.6	61.3	135.2
ITR	*PAMI'23*	78.2	65.5	55.2	46.9	30.5	61.9	133.1
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**78.7**	**66.5**	**56.3**	**47.6**	**31.3**	**62.5**	**138.8**

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	AVSD-DSTC7						
		B-1	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
Baseline	*ICASSP'19*	62.1	48.0	37.9	30.5	21.7	48.1	73.3
MTN	*ACL'19*	71.5	58.1	47.6	39.2	26.9	55.9	106.6
JMAN	*AAAI'20*	66.7	52.1	41.3	33.4	23.9	53.3	94.1
VGD	*ACL'20*	74.9	62.0	52.0	43.6	28.2	58.2	119.4
BiST	*EMNLP'20*	75.5	61.9	51.0	42.9	28.4	58.1	119.2
SCGA	*AAAI'21*	74.5	62.2	51.7	43.0	28.5	57.8	120.1
RLM	*TASLP'21*	76.5	64.3	54.3	45.9	29.4	60.6	130.8
PDC	*ICLR'21*	77.0	65.3	53.9	44.9	29.2	60.6	129.5
AV-TRN	*ICASSP'22*	–	–	–	40.6	26.2	55.4	107.9
VGNMN	*NAACL'22*	–	–	–	42.9	27.8	57.8	118.8
COST	*ECCV'22*	72.3	58.9	48.3	40.0	26.6	56.1	108.5
MRLV	*NeurIPS'22*	–	59.2	49.3	41.5	26.9	56.9	115.9
THAM	*EMNLP'22*	77.8	65.4	54.9	46.8	30.8	61.9	133.5
DialogMCF	*TASLP'23*	77.7	65.3	54.7	45.7	30.6	61.3	135.2
ITR	*PAMI'23*	78.2	65.5	55.2	46.9	30.5	61.9	133.1
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**78.7**	**66.5**	**56.3**	**47.6**	**31.3**	**62.5**	**138.8**

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	AVSD-DSTC8						
		B-1	B-2	B-3	B-4	M	R	C
Baseline	*ICASSP'19*	61.4	46.7	36.5	28.9	21.0	48.0	65.1
MTN	*ACL'19*	—	—	—	—	—	—	—
JMAN	*AAAI'20*	64.5	50.4	40.2	32.4	23.2	52.1	87.5
VGD	*ACL'20*	—	—	—	—	—	—	—
BiST	*EMNLP'20*	68.4	54.8	45.7	37.6	27.3	56.3	101.7
SCGA	*AAAI'21*	71.1	59.3	49.7	41.6	27.6	56.6	112.3
RLM	*TASLP'21*	74.6	62.6	52.8	44.5	28.6	59.8	124.0
PDC	*ICLR'21*	74.9	62.9	52.8	43.9	28.5	59.2	120.1
AV-TRN	*ICASSP'22*	—	—	—	39.4	25.0	54.5	99.7
VGNMN	*NAACL'22*	—	—	—	—	—	—	—
COST	*ECCV'22*	69.5	55.9	46.5	3.82	27.8	57.4	105.1
MRLV	*NeurIPS'22*	—	—	—	—	—	—	—
THAM	*EMNLP'22*	<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	*TASLP'23*	75.6	63.3	53.2	44.9	29.3	60.1	125.3
ITR	*PAMI'23*	76.2	<u>64.1</u>	<u>54.3</u>	<u>46.0</u>	29.8	60.7	128.5
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**77.5**	**66.0**	**56.1**	**47.7**	**30.6**	**62.4**	**135.4**

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	AVSD-DSTC8						
		B-1	B-2	B-3	B-4	M	R	C
Baseline	*ICASSP'19*	61.4	46.7	36.5	28.9	21.0	48.0	65.1
MTN	*ACL'19*	—	—	—	—	—	—	—
JMAN	*AAAI'20*	64.5	50.4	40.2	32.4	23.2	52.1	87.5
VGD	*ACL'20*	—	—	—	—	—	—	—
BiST	*EMNLP'20*	68.4	54.8	45.7	37.6	27.3	56.3	101.7
SCGA	*AAAI'21*	71.1	59.3	49.7	41.6	27.6	56.6	112.3
RLM	*TASLP'21*	74.6	62.6	52.8	44.5	28.6	59.8	124.0
PDC	*ICLR'21*	74.9	62.9	52.8	43.9	28.5	59.2	120.1
AV-TRN	*ICASSP'22*	—	—	—	39.4	25.0	54.5	99.7
VGNMN	*NAACL'22*	—	—	—	—	—	—	—
COST	*ECCV'22*	69.5	55.9	46.5	3.82	27.8	57.4	105.1
MRLV	*NeurIPS'22*	—	—	—	—	—	—	—
THAM	*EMNLP'22*	76.4	64.1	53.8	45.5	30.1	61.0	130.4
DialogMCF	*TASLP'23*	75.6	63.3	53.2	44.9	29.3	60.1	125.3
ITR	*PAMI'23*	76.2	64.1	54.3	46.0	29.8	60.7	128.5
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**77.5**	**66.0**	**56.1**	**47.7**	**30.6**	**62.4**	**135.4**

Results

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	B-1	B-2	B-3	B-4	M	R	C
AV-TRN	*ICASSP'22*	–	–	–	24.7	19.1	43.7	56.6
+ Ext.	*ICASSP'22*	–	–	–	37.1	24.5	53.5	86.9
DSTC10	*AAAI'22*	67.3	54.5	44.8	<u>37.2</u>	24.3	53.0	<u>91.2</u>
DialogMCF	*TASLP'23*	<u>69.3</u>	<u>55.6</u>	<u>45.0</u>	36.9	<u>24.9</u>	<u>53.6</u>	<u>91.2</u>
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**70.0**	**57.4**	**47.6**	**40.0**	**25.7**	**54.5**	**99.8**

AVSD-DSTC10

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	B-1	B-2	B-3	B-4	M	R	C
AV-TRN	*ICASSP'22*	–	–	–	24.7	19.1	43.7	56.6
+ Ext.	*ICASSP'22*	–	–	–	37.1	24.5	53.5	86.9
DSTC10	*AAAI'22*	67.3	54.5	44.8	<u>37.2</u>	24.3	53.0	<u>91.2</u>
DialogMCF	*TASLP'23*	<u>69.3</u>	<u>55.6</u>	<u>45.0</u>	36.9	<u>24.9</u>	<u>53.6</u>	<u>91.2</u>
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**70.0**	**57.4**	**47.6**	**40.0**	**25.7**	**54.5**	**99.8**

AVSD-DSTC10

Model	Venue	B-4
MTN	*ACL'19*	21.7
GPT-2	*EMNLP'21*	19.2
BART	*NAACL'22*	33.1
PaCE	*ACL'23*	<u>34.1</u>
$\text{MST}_{\text{MIXER}}$	*ECCV'24*	**44.7**

SIMMC 2.0

Results

B-n = BLEU-n, **M** = METEOR, **R** = Rouge-L, **C** = CIDEr

Model	Venue	B-1	B-2	B-3	B-4	M	R	C
AV-TRN	*ICASSP'22*	–	–	–	24.7	19.1	43.7	56.6
+ Ext.	*ICASSP'22*	–	–	–	37.1	24.5	53.5	86.9
DSTC10	*AAAI'22*	67.3	54.5	44.8	37.2	24.3	53.0	91.2
DialogMCF	*TASLP'23*	69.3	55.6	45.0	36.9	24.9	53.6	91.2
$\mathrm{MST}_{\mathrm{MIXER}}$	*ECCV'24*	**70.0**	**57.4**	**47.6**	**40.0**	**25.7**	**54.5**	**99.8**

AVSD-DSTC10

Model	Venue	B-4
MTN	*ACL'19*	21.7
GPT-2	*EMNLP'21*	19.2
BART	*NAACL'22*	33.1
PaCE	*ACL'23*	34.1
$\mathrm{MST}_{\mathrm{MIXER}}$	*ECCV'24*	**44.7**

SIMMC 2.0

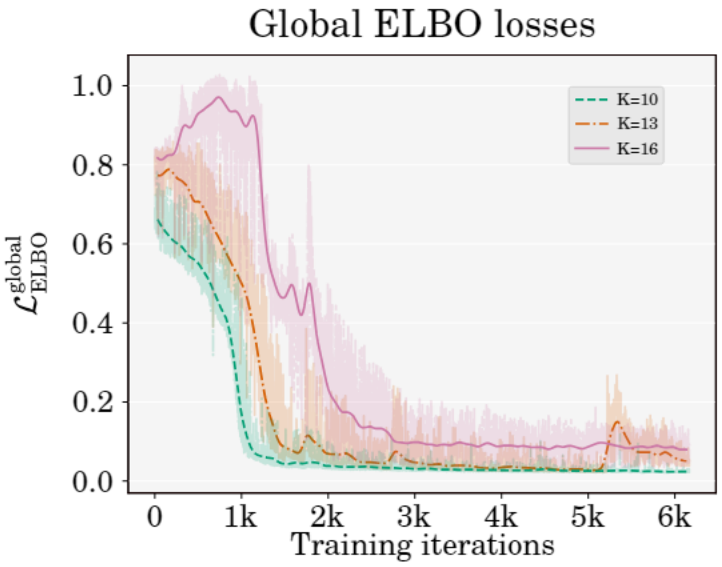
C = Causal, **T** = Temporal, **D** = Descriptive questions

Model	Venue	WUPS_C	WUPS_T	WUPS_D	WUPS
HCRN	*CVPR'20*	16.05	17.68	49.78	23.92
HGA	*AAAI'20*	17.98	17.95	50.84	24.06
Flamingo	*NeurIPS'22*	–	–	–	28.40
KcGA	*AAAI'23*	–	–	–	28.20
EMU	*arXiv'23*	–	–	–	23.40
$\mathrm{MST}_{\mathrm{MIXER}}$	*ECCV'24*	**22.12**	**22.20**	**55.64**	**29.50**

NExT-QA *(open-ended)*

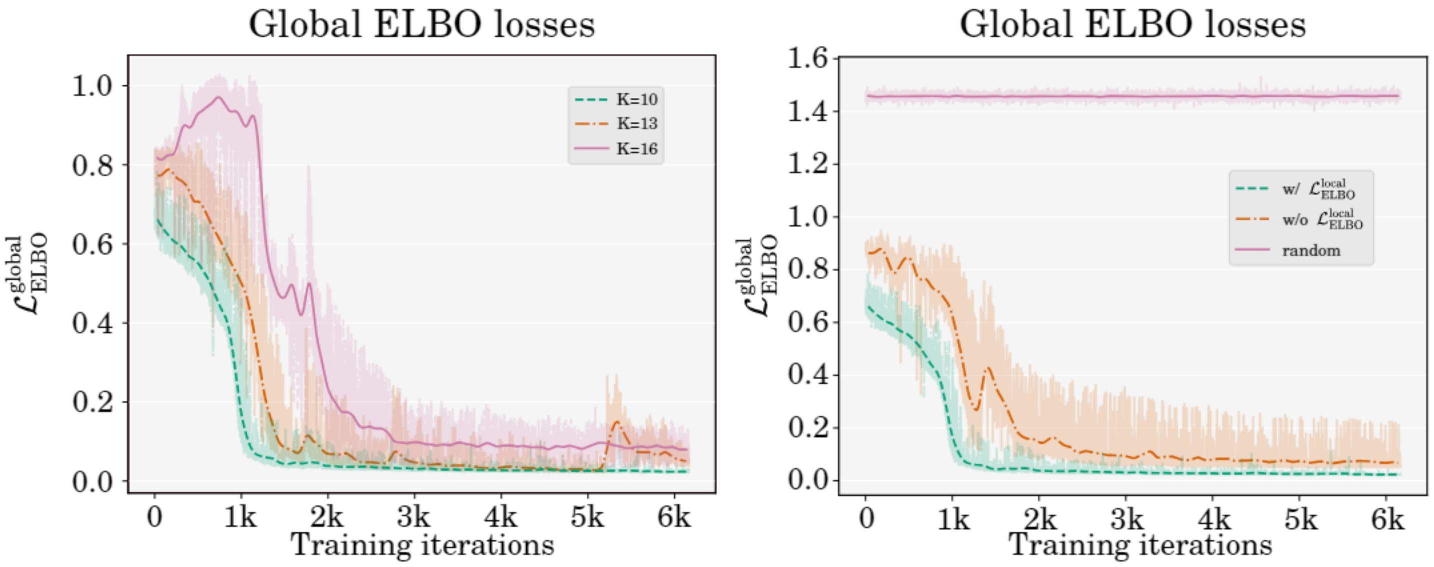
Ablation Study

Ablation Study



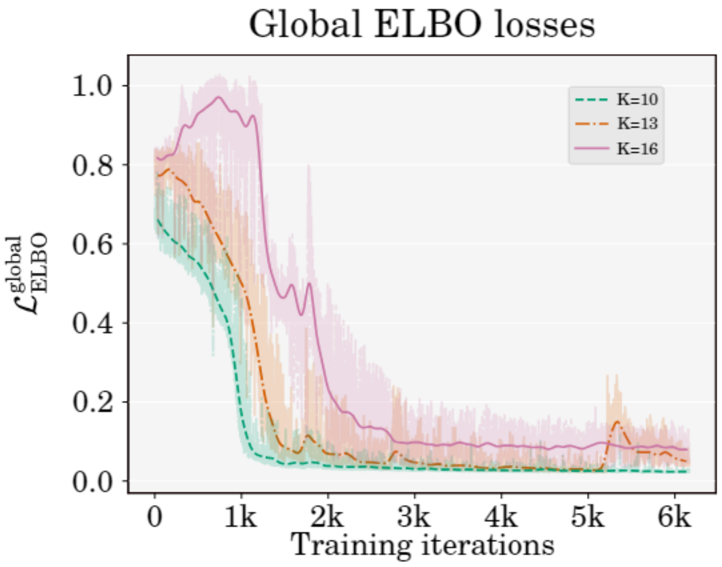
- Learning the latent graphs become more difficult when using a higher number of nodes *K*

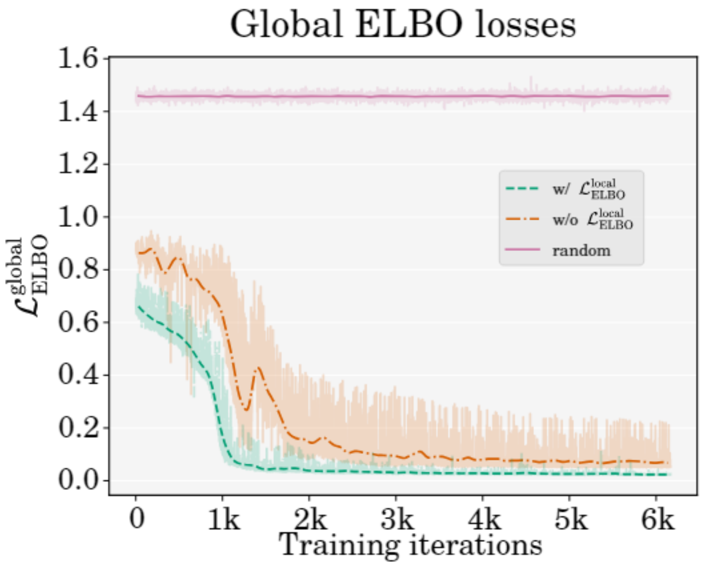
Ablation Study

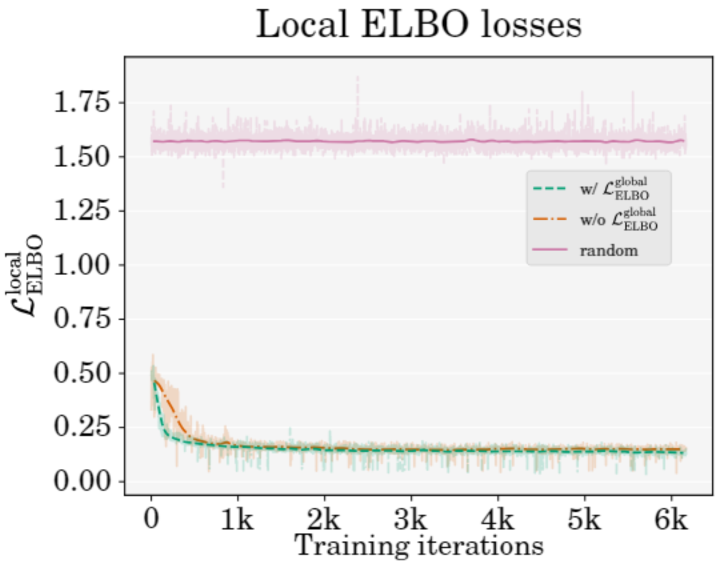


- Learning the latent graphs become more difficult when using a higher number of nodes *K*
- The *divide* stage alleviates the difficulty of learning the global latent graphs

Ablation Study

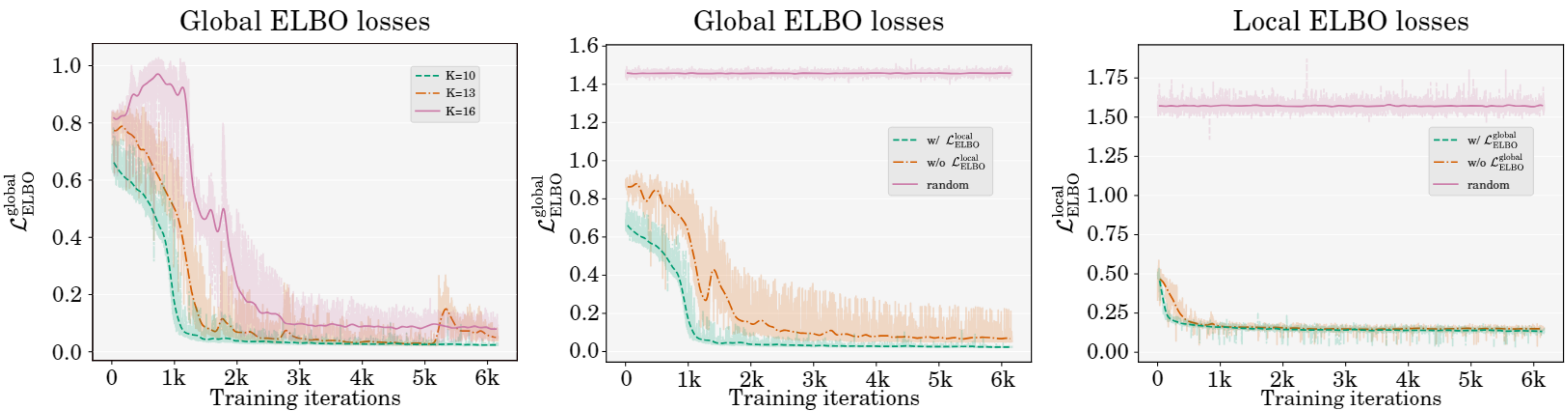






- Learning the latent graphs become more difficult when using a higher number of nodes *K*
- The *divide* stage alleviates the difficulty of learning the global latent graphs
- The *conquer* stage slightly improves the learning of the local latent graphs in the *divide* stage

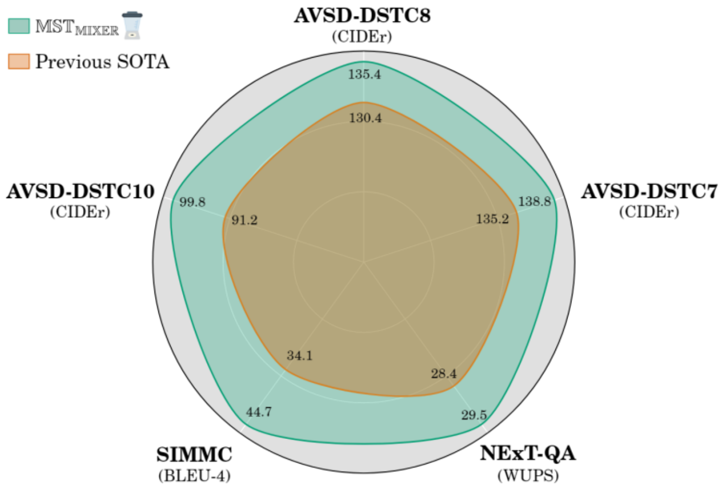
Ablation Study



- Learning the latent graphs become more difficult when using a higher number of nodes *K*
- The *divide* stage alleviates the difficulty of learning the global latent graphs
- The *conquer* stage slightly improves the learning of the local latent graphs in the *divide* stage
  - ➢ ***Conquer* stage benefits more from *divide* stage than vice-versa**

Summary

- We proposed $\text{MST}_{\text{MIXER}}$: A novel multi-modal state tracking model specifically geared towards video dialog
- It first identifies the most influential constituents at different semantic levels
- Then, it relies on a *divide-and-conquer* GNN-based approach to infer the missing underlying structure of the mix of all modalities
- Finally, it leverages these features to augment the hidden states of a backbone VLM
- $\text{MST}_{\text{MIXER}}$ achieves new SOTA results on a variety of challenging benchmarks



References

1. Alamri, H., Cartillier, V., Das, A., et al.: *Audio visual scene-aware dialog*. In **CVPR** (2019)
2. Antol, S., Agrawal, A., Lu, J., et al.: *VQA: Visual Question Answering*. In **ICCV** (2015)
3. Xu, D., Zhao, Z., et al.: *Video question answering via gradually refined attention over appearance and motion*. In **ACM MM** (2017)
4. Das, A., Kottur, S., Gupta, K., et al.: *Visual Dialog*. In **CVPR** (2017)
5. Lee, H., Lee, J., Kim, T.Y.: *SUMBT: Slot-utterance matching for universal and scalable belief tracking*. In **ACL** (2019)
6. Wu, C.S., Madotto, A., et al.: *Transferable multi-domain state generator for task-oriented dialogue systems*. In **ACL** (2019)
7. Le, H., Socher, R., Hoi, S.C.: *Non-autoregressive dialog state tracking*. In **ICLR** (2020)
8. Mrkšić, N., Ó Séaghdha, D., et al.: *Neural belief tracker: Data-driven dialogue state tracking*. In **ACL** (2017)
9. Le, H., Chen, N.F., Hoi, S.C.*: Multimodal Dialogue State Tracking*. In **NAACL** (2022)
10. Pang, W., Wang, X.: *Visual dialogue state tracking for question generation*. In **AAAI** (2020)
11. Abdessaied, A., Hochmeister, M., Bulling, A.: *OLViT: Multi-modal state trackingvia attention-based embeddings for video-grounded dialog*. In **LREC-COLING** (2024)
12. Kottur, S., Moon, S., et al.: *SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations*. In **EMNLP** (2021)
13. Moon, S., Kottur, S., Crook, P.A., et al.: *Situated and interactive multimodal conversations*. In **COLING** (2020)

University of Stuttgart
Germany

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

Thank You!

Adnen Abdessaied, Lei Shi, Andreas Bulling
University of Stuttgart, Germany
adnen.abdessaied@vis.uni-stuttgart.de

02.10.2024