Reasoning with Heterogeneous Graph Alignment for Video Question Answering

Ahjeong Seo

Seoul National University

1. Introduction: Why this research is meaningful?

AS - IS → **TO - BE**

<Previous Dominant VQA Methods>

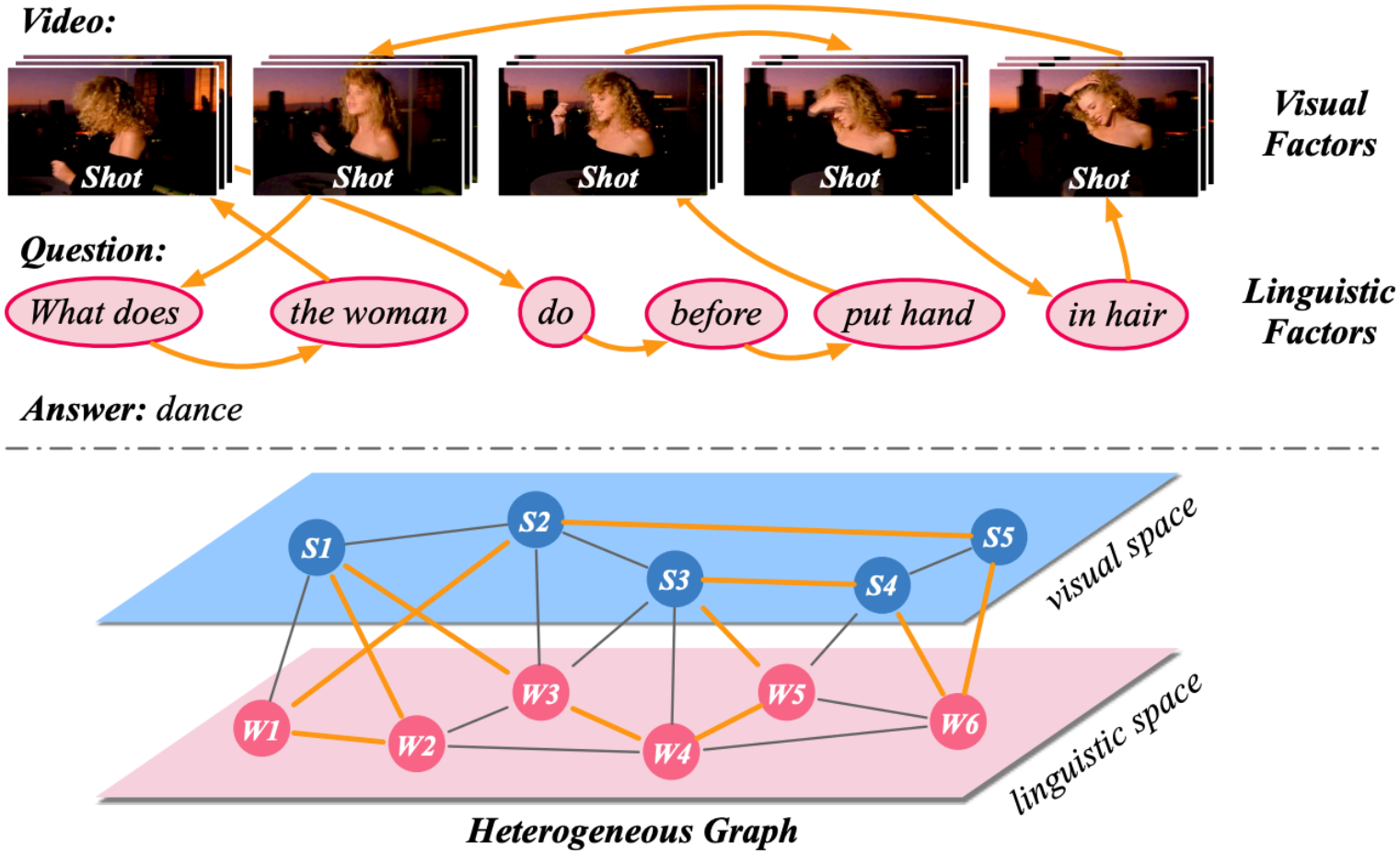
- process video and question separately
- Feed representations of different modalities and late fusion

Neglect correlations of both inter- and intra- modality!!

Heterogenous Graph Alignment

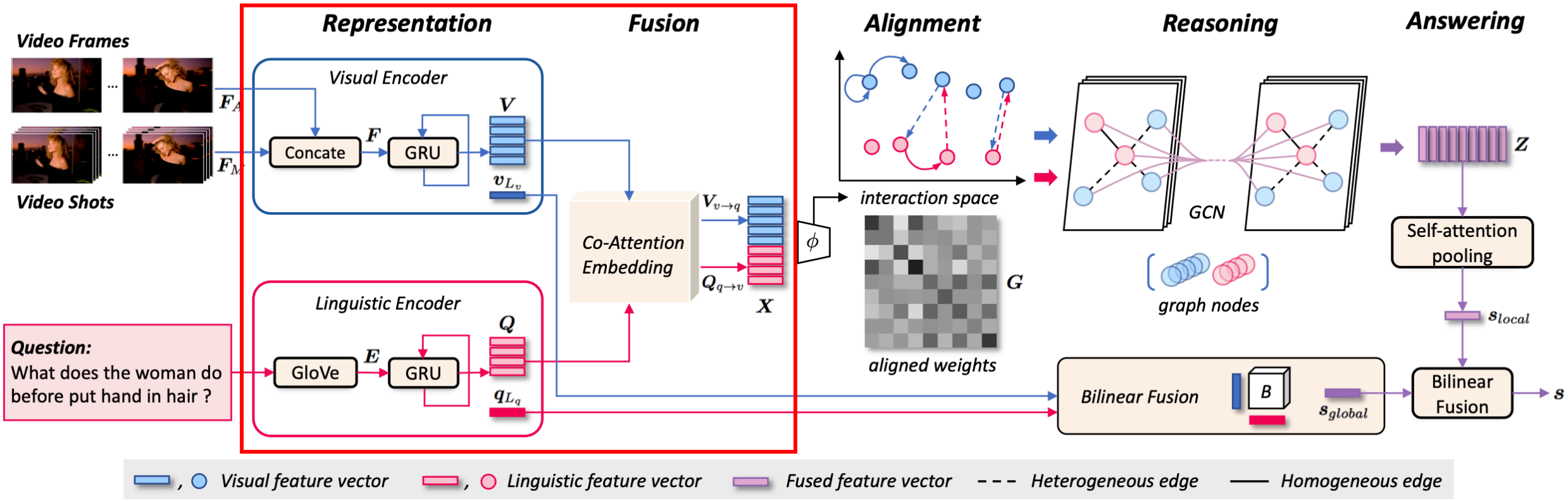
- Inter- and intra- modality information can be aligned and interacted simultaneously

1. Introduction: Why this research is meaningful?



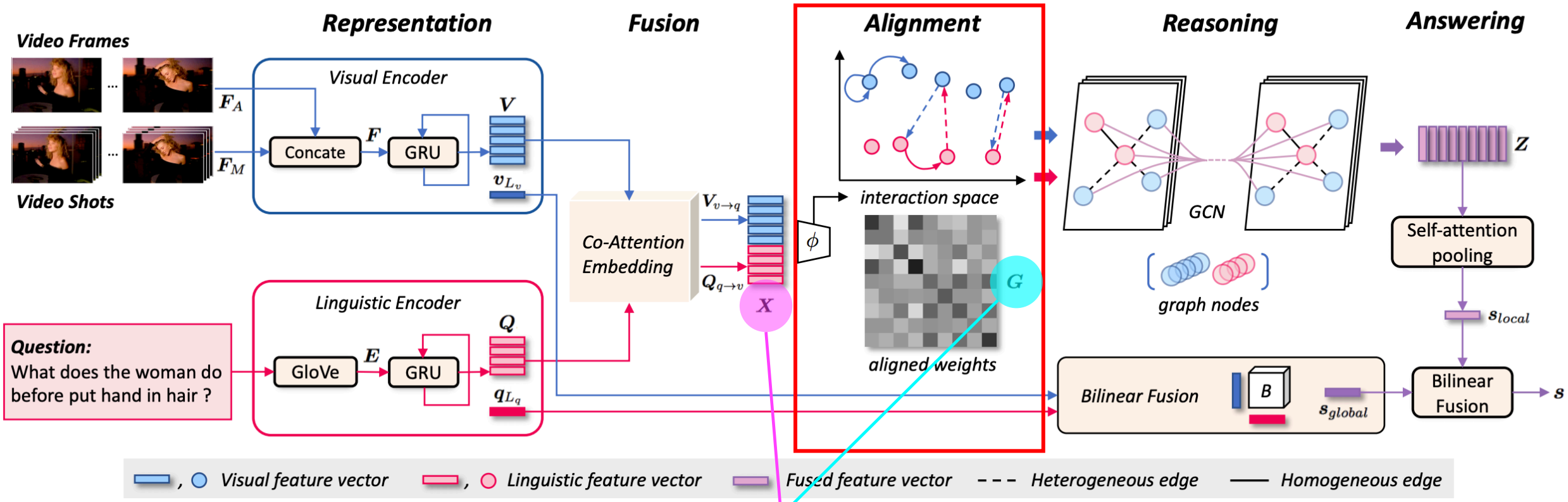
- Shot level video and question
- cross-modal reasoning over a heterogeneous graph betw een factors, shots and words
- Heterogeneous means both inter- and intra-modality

2. Method



1. Representation: get sequence representation of each modality
2. Fusion: Transform from linguistic to visual, and visual to linguistic space by co-attention

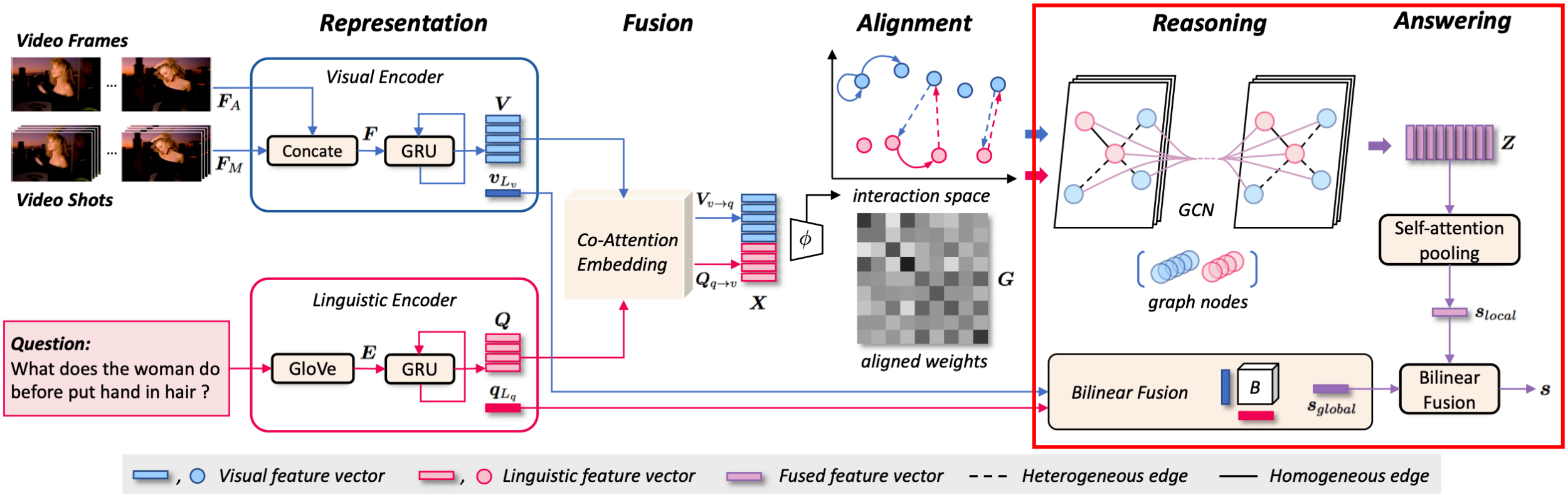
2. Method



3. Alignment: make graph input feature X by concatenating visual and linguistic embedding vector, and get adjacency matrix by learning correlation scores

$$G = \phi(X)\,\phi(X)^T$$

2. Method



4. Reasoning: get node representation by GCN
5. Answering: get answer by fusing global and local representation

3. Experiments

- Dataset : TGIF-QA
  - 165K Q&A pairs from 72K animated GIFs
  - Has four task types
  - SOTA on three task

- Also SOTA on other dataset: MSVD-QA, MSRVTT-QA

Table 1: State-of-the-art comparison on TGIF-QA dataset. Mean ℓ_2 loss for Count, and accuracy (%) for others.

Methods	Count	Action	Trans.	FrameQA
Random	19.62	20.00	20.00	0.06
ST-VQA-Sp.	4.28	57.3	63.7	45.5
ST-VQA-Tp.	4.40	60.8	67.1	49.3
ST-VQA-Sp.Tp.	4.56	57.0	59.6	47.8
CT-SAN	5.14	56.1	64.0	39.6
Co-Mem	<u>4.10</u>	68.2	74.3	51.5
PSAC	4.27	70.4	76.9	**55.7**
Fan et al.	<u>4.10</u>[1]	<u>73.9</u>	77.8	53.8
ST-VQA⋆	4.22	73.5	<u>79.7</u>	52.0
Ours HGA	**4.09**	**75.4**	**81.0**	<u>55.1</u>