Video Graph Transformer for Video Question Answering

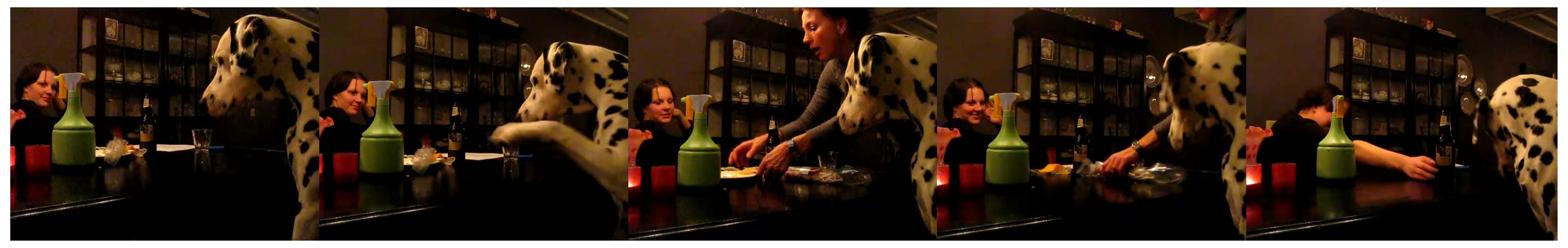
Junbin Xiao, Pan Zhou, Tat-Seng Chua and Shuicheng Yan
Sea AI Lab (SAIL) & National University of Singapore (NUS)

October 23-27, 2022, Tel Aviv

Read on your phone | Watch our video | Data & Code

Introduction:

Existing transformer-style models only demonstrate their success in answering questions that involve the coarse recognition or description of video contents. Their performances are either unknown or weak in answering questions that challenge real-world **visual relation reasoning**, especially the **causal** and **temporal** relations that feature **video dynamics at action and event level**. Cross-modal pretraining seems promising, yet it requires the handling of million-scale *video*-text data.



MSRVTT-QA & MSVD-QA [Xu et al, MM'17]:

Who is looking at the dog? Lady.

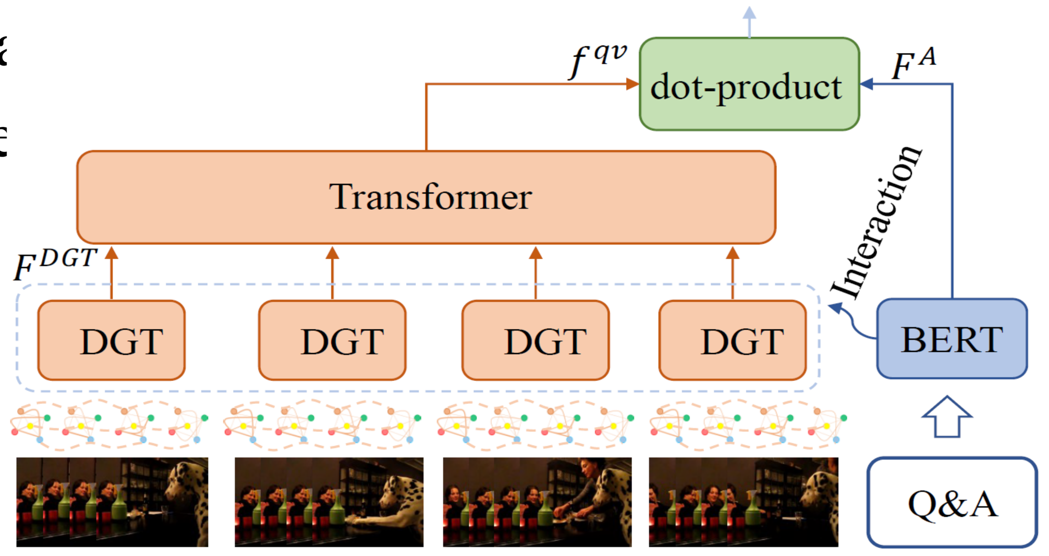
What is the dog doing? Sitting.

NExT-QA[Xiao et al, CVPR'21]:

Why did the woman walk towards the table in the middle of the video? Clean the table.

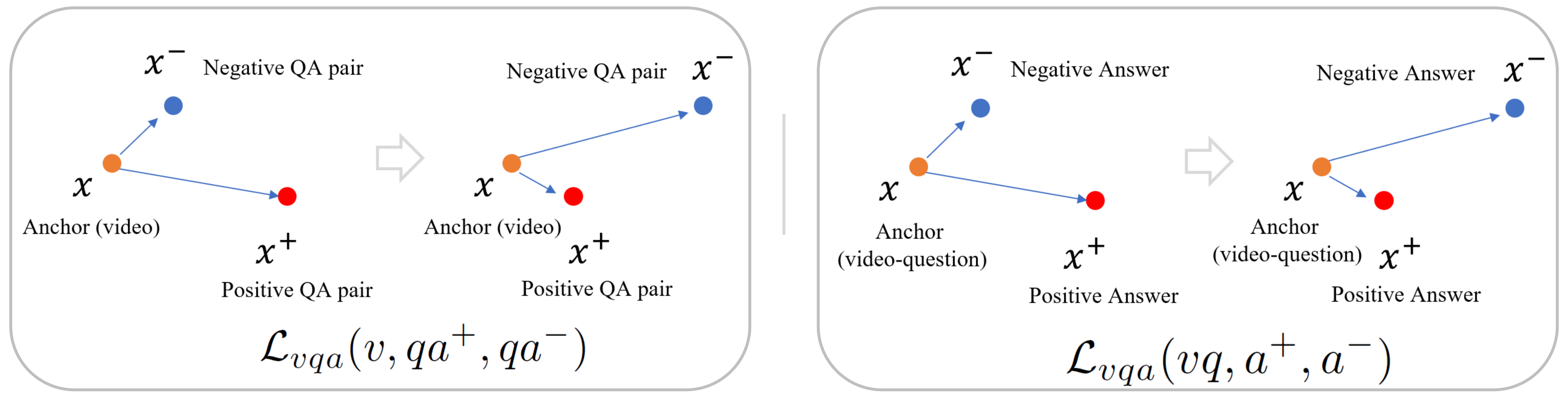
Method:

We propose Video Gra (**VGT**) to improve pre vious arts in answer- ing relation-type questions from two major aspects:



- **Video Encoding:** Dynamic Graph Transformer (DGT) that explicitly encodes the visual objects, their relations and dynamics, for spatial and temporal relation reasoning.
- **Contrastive Learning**: we design ***separate*** video and text transformers to encode video and QA information respectively for contrastive learning, instead of cross-modal transformer for answer classification. Fine-grained vision-text information communication is done by additional light-weight cross-modal interaction modules.

➢ Contrastive Learning



(a) Multi-choice QA (b) Open-ended QA

$$\mathcal{L}_*(x, x^+, x^-) = -\mathbb{E}_i[\log(\frac{e^{s_{\mathrm{VGT}}(x_i, x_i^+)}}{e^{s_{\mathrm{VGT}}(x_i, x_i^+)} + \sum_{(x_i, x_j^-)\in\mathcal{N}_i} e^{s_{\mathrm{VGT}}(x_i, x_j^-)}})]$$

➢ Dynamic Graph Transformer

- **Spatial-temporal**:

  Consider contextual graphs to improve the graphs obtained at static frames.

- **Compositional**:

  Summarize local/atomic interactions to global activities.

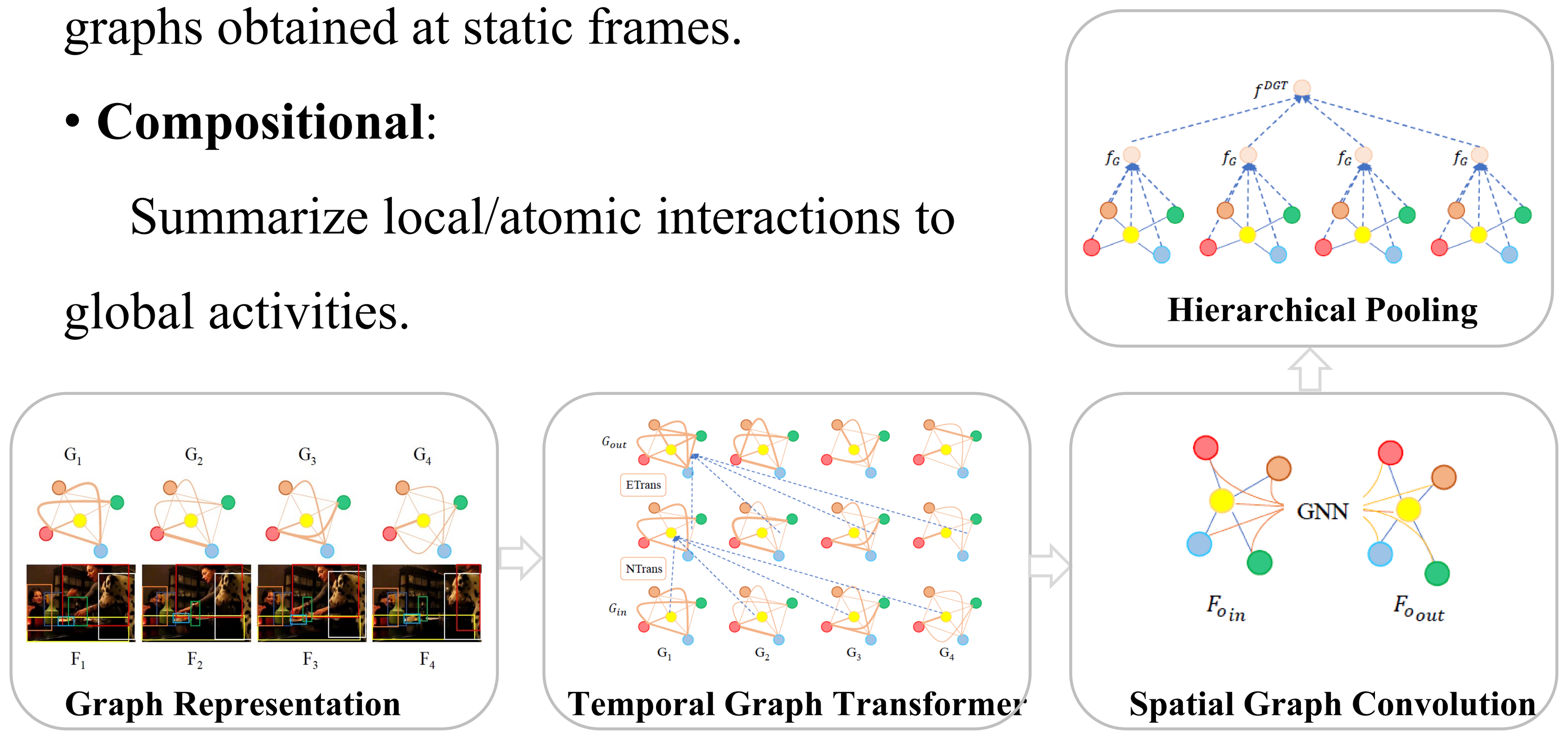


Illustration of the 4 stages to encode a video clip.

➢ Cross-modal Interaction $x^{qv} = x^v + \sum_{m=1}^{M} \beta_m x^q_m,$ where $\beta = \sigma(x^v (X^q)^\top)$

x^v: visual representations, e.g., F^{DGT}

x^q: textual representations, e.g., Outputs from BERT.

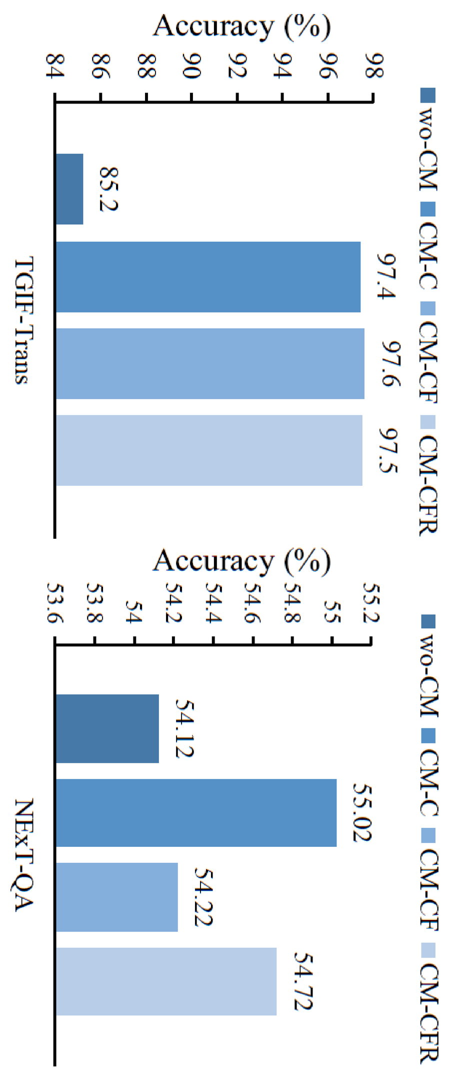
Experiment:

➢ SoTA Comparison.

Methods	NExT-Val	NExT-Test
VQA-T*[ICCV'21]	45.30	44.54
HQGA[AAAI'22]	51.42	51.75
VQA-T* (PT)	52.32	50.83
P3D-G[AAAI'22]	53.40	-
VGT (Ours)	<u>55.02</u>	<u>53.68</u>
VGT(PT)	**56.89**	**55.70**

Methods	TGIF-FQA	MSRVTT-QA
HCRN[CVPR'20]	55.9	35.6
ClipBERT[CVPR'21]	60.3	37.4
HQGA[AAA'22]	61.3	38.6
MERRLOT(PT)	**69.5**	**43.1**
VGT (Ours)	<u>61.6</u>	39.7

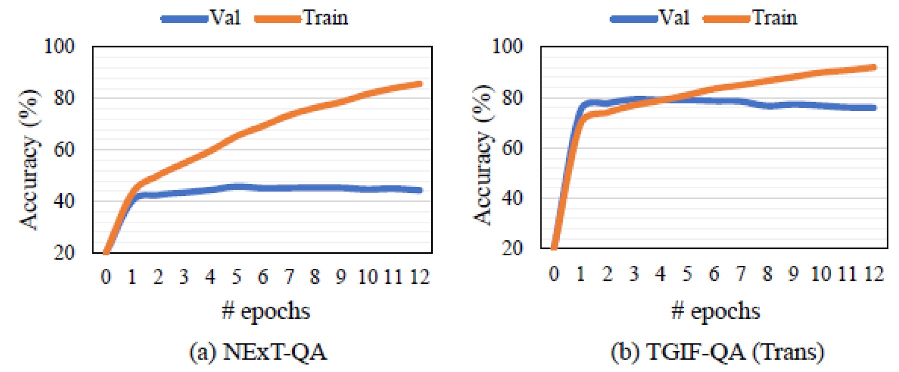
Methods	TGIF-QA Act	TGIF-QA Trans	TGIF-QA-R* Act	TGIF-QA-R* Trans
PGAT[MM'21]	80.6	85.7	58.7	65.9
ClipBERT[CVPR'21]	82.8	87.8	-	-
MERLOT[NeurIPS'21]	<u>94.0</u>	<u>96.2</u>	-	-
VGT (Ours)	**95.0**	**97.6**	<u>59.9</u>	<u>70.5</u>
VGT(PT)	-	-	**60.5**	**71.5**



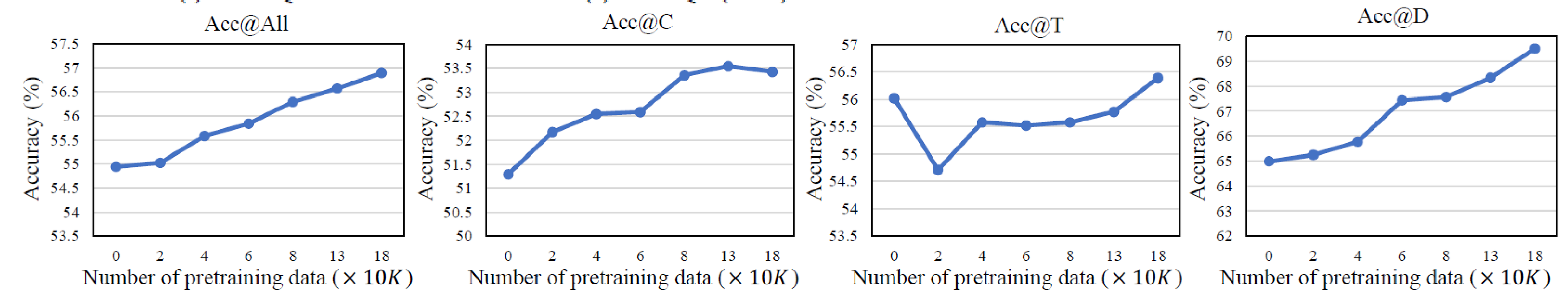
Cross-modal interaction at different levels.

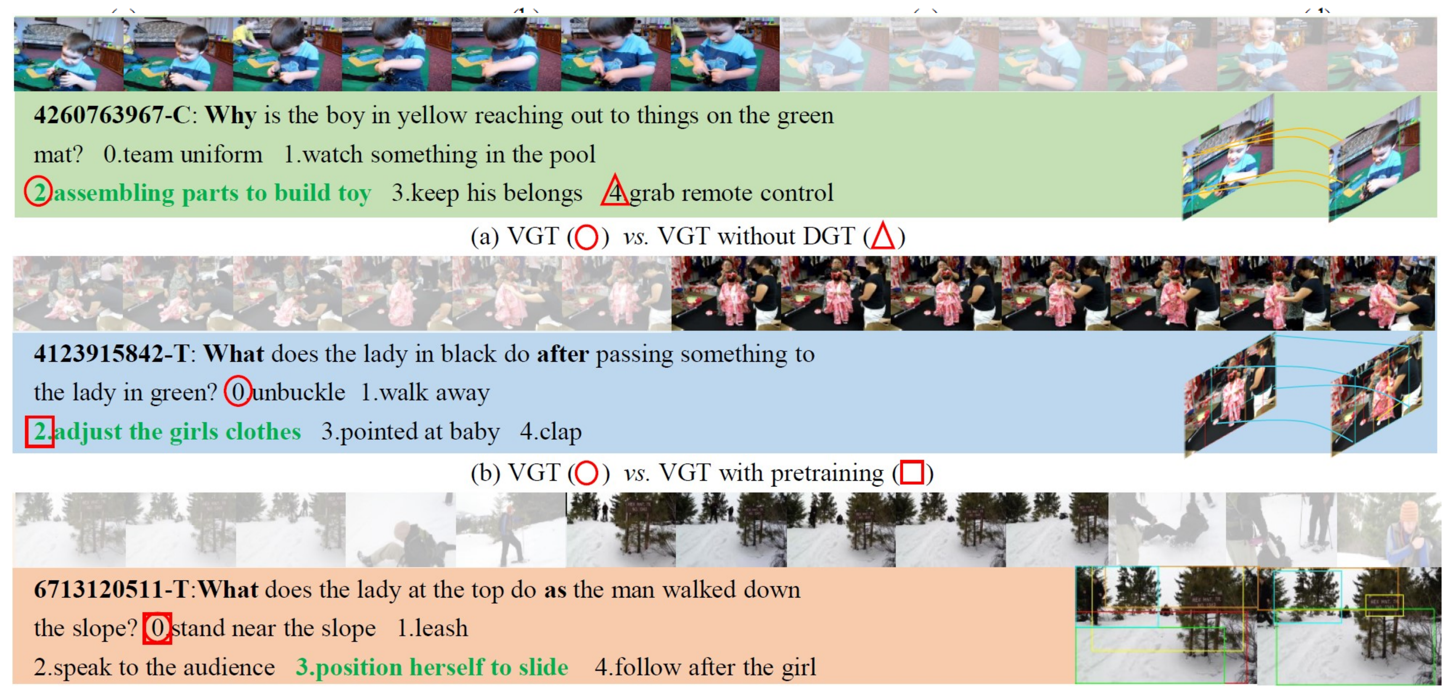
➢ Ablation Study

Models	TGIF-QA Action	TGIF-QA Trans	NExT-QA Val Acc@C	NExT-QA Val Acc@T	NExT-QA Val Acc@D	NExT-QA Val Acc@All
VGT	**95.0**	**97.6**	**52.28**	**55.09**	64.09	**55.02**
w/o DGT	89.6	95.4	50.10	52.85	64.48	53.22
w/o TTrans	94.0	97.6	50.86	53.04	64.86	53.74
w/o NTrans	94.5	97.4	50.79	54.22	63.32	53.84
w/o ETrans	94.8	97.4	51.25	54.34	64.48	54.30
w/o F_I	93.5	97.0	50.44	53.97	63.32	53.58
Comp→CLS	70.1	79.9	42.96	46.96	53.02	45.82



Classification model variant suffers from over-fitting.





Conclusion:

- We propose **video graph transformer** to advance VideoQA from coarse recognition and description to fine-gained visual reasoning in dynamic scenarios, and we achieve **SOTA** results on related benchmarks.
- We propose **dynamic graph transformer** to encode visual graph dynamics for relation reasoning in space-time. Most importantly, we demonstrate that **contrastive learning** significantly outperforms classification for multi-choice cross-modal video reasoning.
- We are the 1st to shown that **pretraining visual graph transformer** can benefit video-language understanding towards a more data-efficient and fine-grained direction.