Learning Fine-Grained Visual Understanding for Video Question Answering via Decoupling Spatial-Temporal Modeling



Hsin-Ying Lee Hung-Ting Su Bing-Chen Tsai Tsung-Han Wu Jia-Fong Yeh Winston H. Hsu

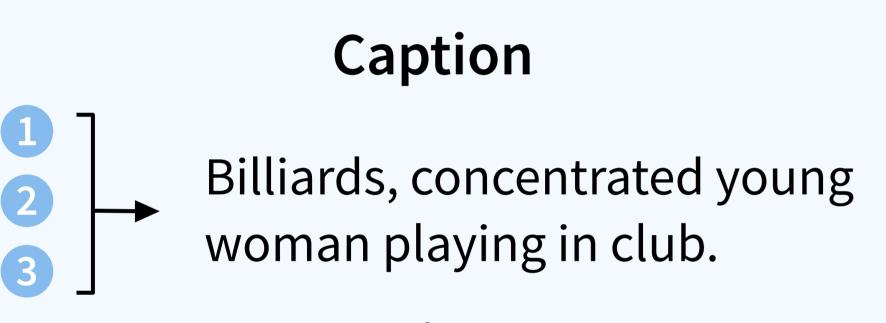
Decoupling spatial-temporal modeling into image and video-language models and pretraining to learn temporal relations between

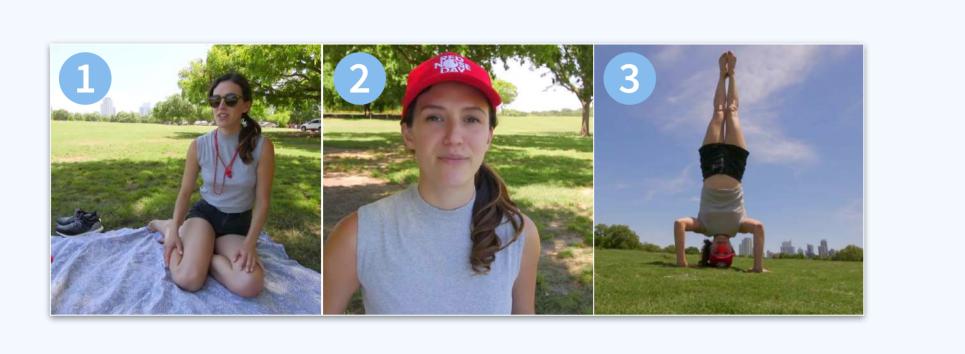
events in videos help video question answering.



Pre-train



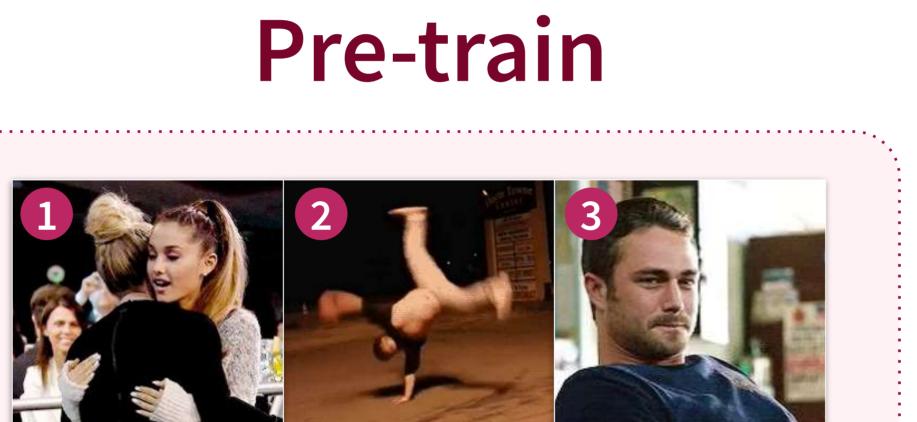




Transcripts
● "dancing with music …"
● "standing up nice and tall …"
● "connect your hands to the earth …"



Generated Video QA Dataset What type of animal do we see in this video? Fish



What happens after ① two blonde girls are hugging each other?

What was the person watching before holding a phone?

Ours

What was the person watching before holding a phone?

Image-Language Encoder

Video-Language E	Encoder
------------------	---------

2 A guy is dancing on the road.

▼ mirror

Fine-tune

Lack of Details

Video-Language Encoder

when taking shuffled input videos.

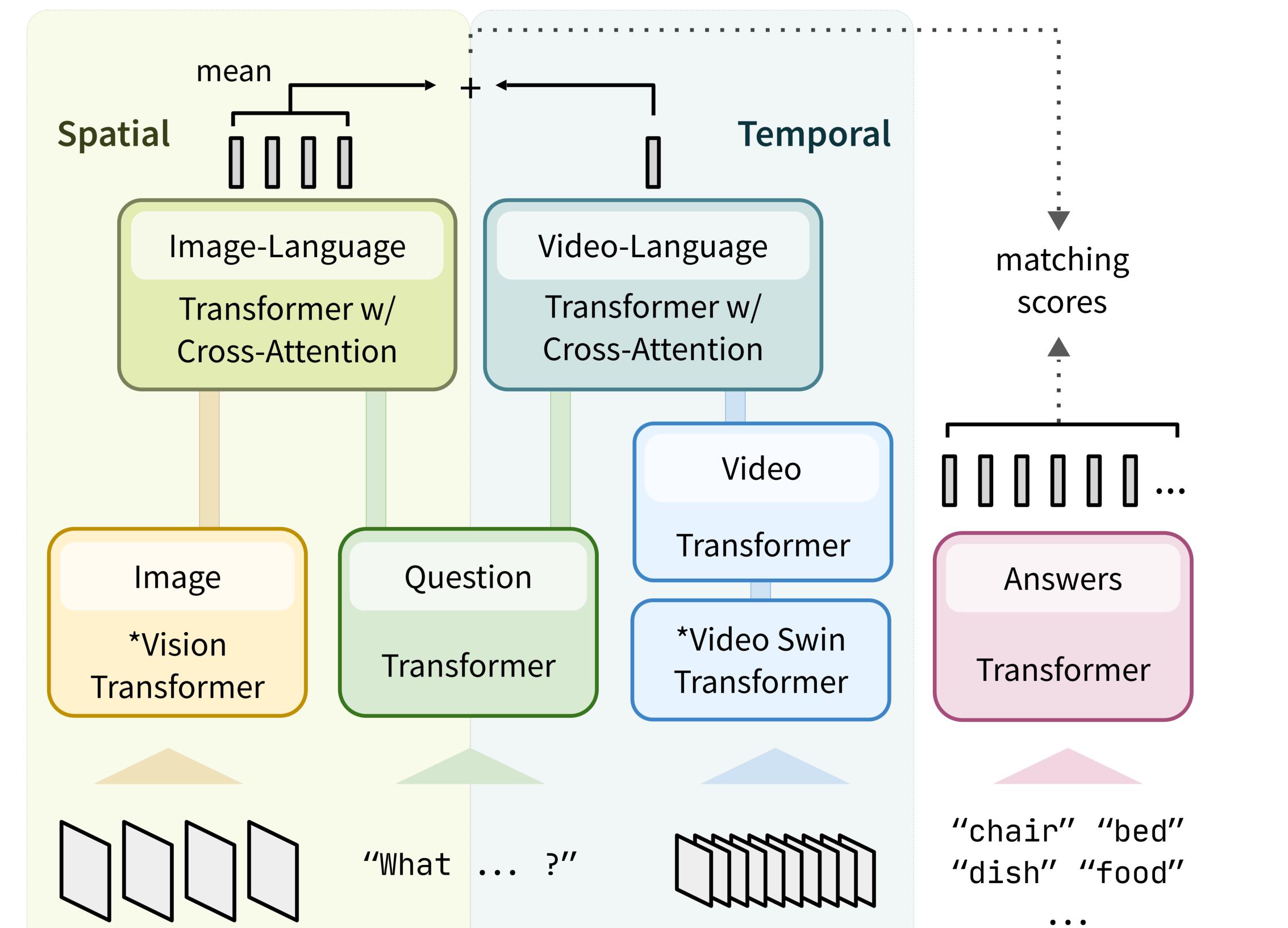
Introduction

- State-of-the-art approaches to video question answering mostly perform coarse-grained spatial-temporal modeling.
- Image-language (IL) models encode regions and grids, showing great potential for encoding fine-grained spatial semantics for video question answering.
- To answer questions about temporal relations, video-language (VL) models have to recognize events in videos and associate events with time conjunctions in questions.

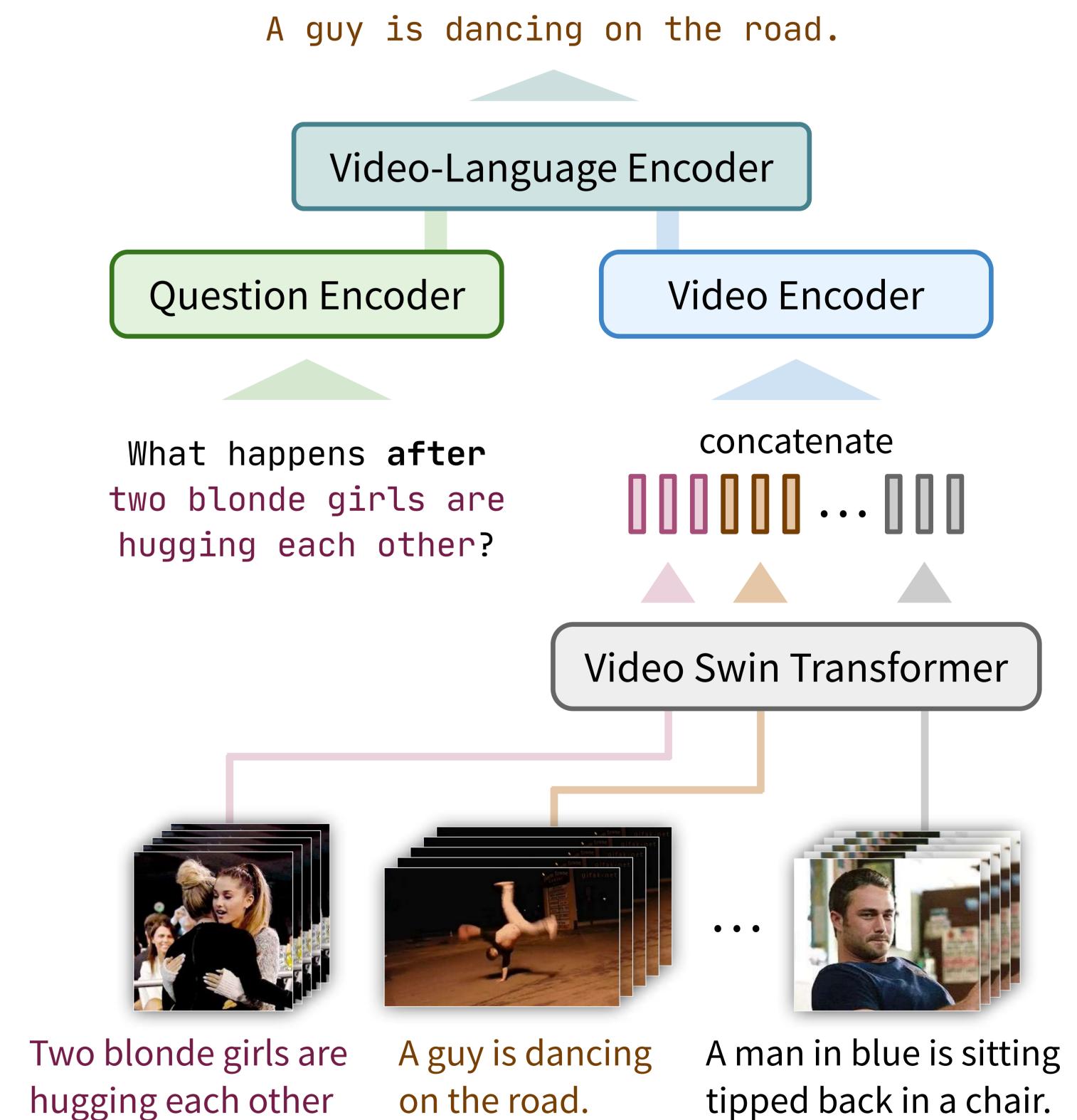
Method

- We propose **Decoupled Spatial-Temporal Encoders (DeST)**, decoupling spatial-temporal modeling into IL and VL encoders.
 - We incorporate a pre-trained IL encoder to encode static spatial semantics by averaging sparsely sampled frame-by-frame predictions at high spatial resolution.
- For questions requiring temporal relations, we train a VL encoder to model temporal dynamics, operating at high temporal but low spatial resolution.

Decoupled Spatial-Temporal Encoders



Temporal Referring Modeling



- The VL encoder is pre-trained with our proposed objective, Temporal Referring Modeling (TRM).
 - TRM queries absolute and relative positions of events in videos synthesized by concatenating clips sampled from video captioning datasets.

Experiments

- We conduct preliminary analyses to estimate the spatial and temporal modeling capability of previous work.
 - IL models perform better than VL models on questions about spatial understanding.
 - Some VL models perform similarly when taking normal and shuffled videos on questions about temporal understanding.
- DeST outperforms the previous state-of-the-art on two video QA benchmarks, ActivityNet-QA and AGQA. The ablation studies also demonstrate the efficacy of the proposed pipeline DeST and pre-training objective TRM.

Туре	Just-Ask	VIOLET	ALBE	F	Metho	od	P	re-training	g Data			Acc
Motion	28.00	18.25	32.50		CoMV	/T	1	00M				38.8
Spatial Rel.	17.50	15.00	24.38		Just-A			9M vid				38.9
Temporal Rel.	4.88	2.12	3.75		MV-G			00M				39.1
Yes / No	66.28	71.87	79.75		SiaSar			6M img				39.8
Color	34.29	31.28	57.39		MERI			80M vid				41.4
Object	26.73	22.33	31.45		VIOL				- 2.5M vid	+ 3M in		37.5
Location	35.75	30.57	36.01			BiLM		OM vid			0	43.2
Number	50.17	50.33	55.61		Singul				2.5M vid			44.1
Other	36.82	33.02	40.16	-		<i>.</i>	200			1 1 172		
	20.00	27.44	16.66		Ours		14	4M mg +	120K VQ/	A + 14K	vid 4	40.8
Overall Image-langu spatial mode	•	37.44 Iels are go		C					ethods or	ו Activi	tyNet	-QA
Image-langu	lage moc		ood at -		Compa Best		with pr		ethods or	ו Activi	tyNet	-QA
Image-langu	lage moc		ood at	C	Best				ethods or	ו Activi	tyNet	-QA
Image-langu	lage moc		ood at	Туре	Best 32.50	Ours	Diff (%					
Image-langu spatial mode	age moc eling.	lels are go	ood at	Type Motion Spatial Rel. Temporal Rel.	Best 32.50 24.38 4.88	Ours 35.75 23.88 5.25	Diff (% 10.00 -2.05 7.58		ethods or Question			Ac
Image-langu spatial mode	age moc eling. hmark Acc	els are go	ood at	Type Motion Spatial Rel. Temporal Rel. Yes / No	Best 32.50 24.38 4.88 79.75	Ours 35.75 23.88 5.25 78.61	Diff (% 10.00 -2.05 7.58 -1.43					Ac 41.3
Image-langu spatial mode Method Benc VIOLET AG	age mod eling. hmark Acc QA 49.	els are go uracy	ood at	Type Motion Spatial Rel. Temporal Rel. Yes / No Color	Best 32.50 24.38 4.88 79.75 57.39	Ours 35.75 23.88 5.25 78.61 59.11	Diff (% 10.00 -2.05 7.58 -1.43 3.00			Frames		Ac 41.3 50.0
Image-langu spatial mode Method Benc VIOLET AG AG	age mod eling. hmark Acc QA 49. QA* 49.	els are go uracy 15 22±.02	ood at	Type Motion Spatial Rel. Temporal Rel. Yes / No Color Object	Best 32.50 24.38 4.88 79.75 57.39 31.45	Ours 35.75 23.88 5.25 78.61 59.11 30.50	Diff (% 10.00 -2.05 7.58 -1.43 3.00 -3.02					Ac 41.3 50.0 51.0
Image-langu spatial mode Method Benc VIOLET AG AG	lage mod eling. hmark Acc QA 49. QA* 49. QA 51.	els are go curacy 15 $22\pm.02$ 27	ood at	Type Motion Spatial Rel. Temporal Rel. Yes / No Color Object Location	Best 32.50 24.38 4.88 79.75 57.39 31.45 36.01	Ours 35.75 23.88 5.25 78.61 59.11 30.50 36.27	Diff (% 10.00 -2.05 7.58 -1.43 3.00 -3.02 0.72			Frames		Ac 41.3 50.0 51.0 51.0
Image-langu spatial mode Method Benc VIOLET AG AG	lage mod eling. hmark Acc QA 49. QA* 49. QA 51.	els are go uracy 15 22±.02	ood at	Type Motion Spatial Rel. Temporal Rel. Temporal Rel. Yes / No Color Object Location Number	Best 32.50 24.38 4.88 79.75 57.39 31.45 36.01 55.61	Ours 35.75 23.88 5.25 78.61 59.11 30.50 36.27 55.28	Diff (% 10.00 -2.05 7.58 -1.43 3.00 -3.02 0.72 -0.59			Frames √ VQA VQA	Video ✓ TRM ✓	Ac 41.3 50.0 51.0 55.6 56.6
Image-langu spatial mode Method Benc VIOLET AG Just-Ask AG HERO VIO	lage modeling. hmark Acceling QA 49. QA* 49. QA* 49. QA* 49. QA* 49. QA* 69.	els are go uracy 15 $22\pm.02$ $73\pm.06$	ood at	Type Motion Spatial Rel. Temporal Rel. Yes / No Color Object Location	Best 32.50 24.38 4.88 79.75 57.39 31.45 36.01 55.61 40.16	Ours 35.75 23.88 5.25 78.61 59.11 30.50 36.27	Diff (% 10.00 -2.05 7.58 -1.43 3.00 -3.02 0.72			Frames √ VQA	Video	Ac 41.3 50.0 51.0 55.6 56.6 56.9

AcivityNet-QA by question type.

	Туре	Best w/o PT	Best w/ PT	Ours
	Object-Rel.	40.33	48.91	59.66
Reasoning	RelAction	49.95	66.55	72.98
	Object-Action	50.00	68.78	75.20
	Superlative	33.55	39.83	48.94
	Sequencing	49.78	67.01	73.53
	Exists	50.01	59.35	63.21
	Duration Compar.	47.03	50.49	60.39
	Activity Recog.	5.52	21.53	27.78
ntic	Object	40.40	49.31	61.27
Semantic	Rel.	49.99	59.60	63.93
	Action	47.58	58.03	65.96
Structure	Query	36.34	47.98	61.22
	Compare	49.71	65.11	72.04
	Choose	46.56	46.90	53.01
	Logic	50.02	56.20	59.18
	Verify	50.01	58.13	63.02
Overall	Binary	48.91	55.35	62.61
	Open	36.34	47.98	61.22
	All	42.11	51.27	61.91

Comparison with prior work on AGQA 2.0.

Training Stream	Acc
Image-Language Video-Language	49.91 16.56
Both	61.91

Ablation study of encoding streams.





and pre-training strategies.



