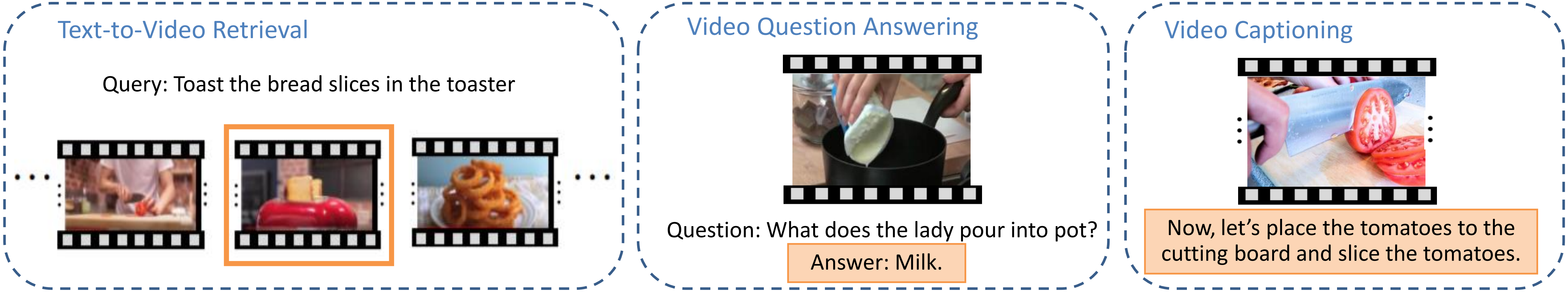
VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation

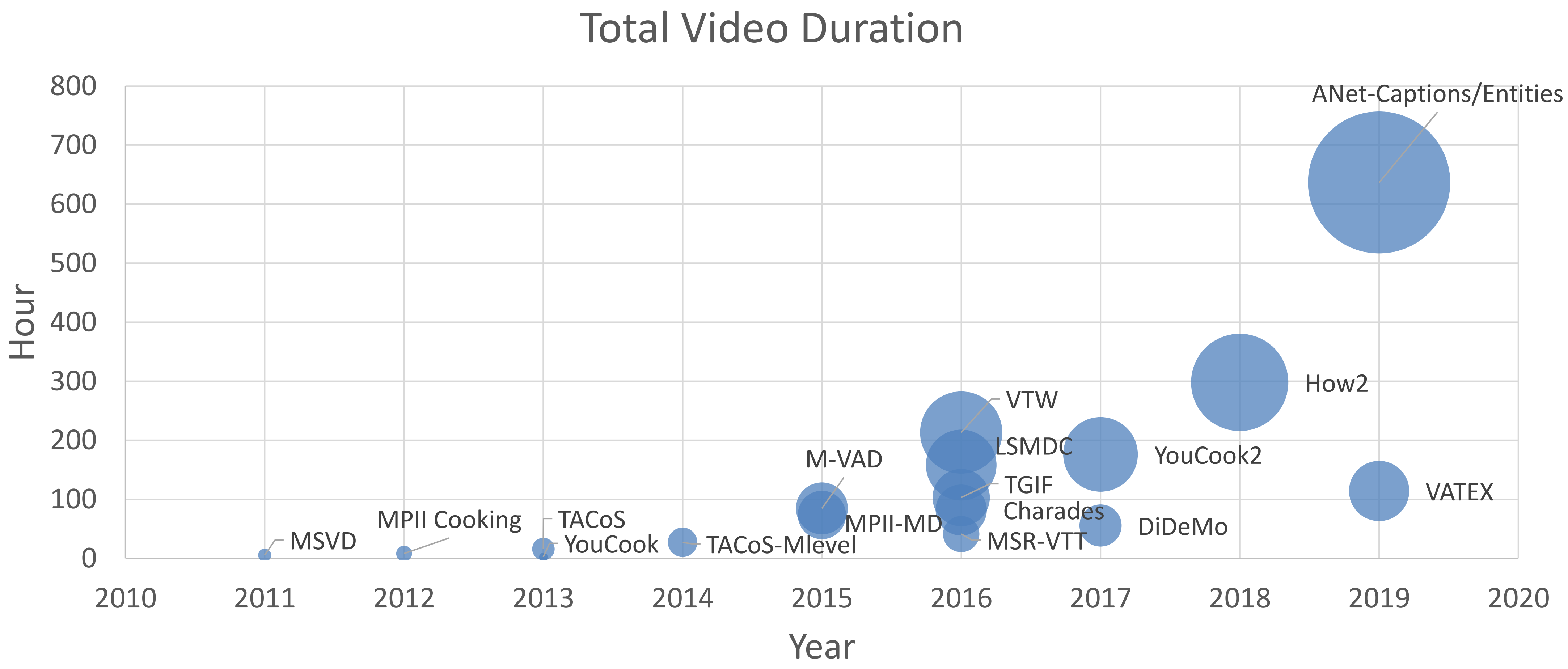
Linjie Li*, Jie Lei*, Zhe Gan, Licheng Yu, Yen-Chun Chen,
Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang,
Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang and Zicheng Liu

Microsoft
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
UNIVERSITY OF CALIFORNIA SANTA CRUZ
UC SANTA BARBARA
清華大學 Tsinghua University

Video-and-Language Tasks



Video-and-Language Datasets



Credit to: Luowei Zhou

Motivation

Single-Channel Video

Video Frames

Video+Language Datasets on Single-Channel Videos

- *Video Retrieval*: YouCook2, ActivityNet, …
- *Video QA*: MSRVTT-QA, MSVD-QA, …
- *Video Captioning*: YouCook2, ActivityNet, …

Motivation

Single-Channel Video

Video Frames

Video+Language Datasets on Single-Channel Videos

- *Video Retrieval*: YouCook2, ActivityNet, …
- *Video QA*: MSRVTT-QA, MSVD-QA, …
- *Video Captioning*: YouCook2, ActivityNet, …

Video+Language Models

Model	PT Dataset	Retrieval Tasks	QA Tasks	Captioning Tasks
HowTo100M [Miech et al.]	HowTo100M	MSRVTT, YouCook2	-	-
ActBERT [Zhu and Yang]		MSRVTT, YouCook2	MSRVTT-QA, LMSDC	YouCook2
ClipBERT [Lei et al.]	VG + COCO	MSRVTT, ActivityNet, DiDeMo	MSRVTT-QA, TGIF-QA	-
Frozen in Time [Bain et al.]	CC+WebVid-2M	MSRVTT, MSVD, DiDeMo, LSMDC	-	-

Motivation

Single-Channel Video

Video Frames

Video+Language Datasets on Single-Channel Videos

- *Video Retrieval*: YouCook2, ActivityNet, …
- *Video QA*: MSRVTT-QA, MSVD-QA, …
- *Video Captioning*: YouCook2, ActivityNet, …

Video+Language Models

Model	PT Dataset	Retrieval Tasks	QA Tasks	Captioning Tasks
HowTo100M [Miech et al.]	HowTo100M	MSRVTT, YouCook2	-	-
ActBERT [Zhu and Yang]		MSRVTT, YouCook2	MSRVTT-QA, LMSDC	YouCook2
ClipBERT [Lei et al.]	VG + COCO	MSRVTT, ActivityNet, DiDeMo	MSRVTT-QA, TGIF-QA	-
Frozen in Time [Bain et al.]	CC+WebVid-2M	MSRVTT, MSVD, DiDeMo, LSMDC	-	-

A general video-and-language system should do well on diverse tasks/domains/datasets.

Motivation

Multi-Channel Video

Video Frames ASR/Subtitles Audio

Video+Language Datasets on Multi-Channel Videos

- *Video Retrieval*: TVR, How2R, …
- *Video QA*: TVQA, How2QA, VIOLIN, …
- *Video Captioning*: TVC, …

Motivation

Multi-Channel Video

Video Frames ASR/Subtitles Audio

Video+Language Datasets on Multi-Channel Videos

- *Video Retrieval*: TVR, How2R, …
- *Video QA*: TVQA, How2QA, VIOLIN, …
- *Video Captioning*: TVC, …

Video+Language Models

Model	PT Dataset	Retrieval Tasks	QA Tasks	Captioning Tasks
HERO [Li et al.]	HowTo100M	TVR, How2R (Multi-channel) DiDeMo, MSRVTT (Single-channel)	TVQA, How2QA, VIOLIN	TVC

Motivation

Multi-Channel Video

Video Frames

ASR/Subtitles

Audio

Video+Language Datasets on Multi-Channel Videos

- *Video Retrieval*: TVR, How2R, …
- *Video QA*: TVQA, How2QA, VIOLIN, …
- *Video Captioning*: TVC, …

Video+Language Models

Model	PT Dataset	Retrieval Tasks	QA Tasks	Captioning Tasks
HERO [Li et al.]	HowTo100M	TVR, How2R (Multi-channel) DiDeMo, MSRVTT (Single-channel)	TVQA, How2QA, VIOLIN	TVC

A smart video-and-language system should be able to leverage information from different modalities.

Motivation

NLP Benchmarks

GLUE

SuperGLUE

XGLUE

XTREME

Publicly accessible large-scale multi-task benchmarks can facilitate advances in modeling.

VALUE Benchmark

- A comprehensive benchmark for **V**ideo-**A**nd-**L**anguage **U**nderstanding **E**valuation

Multi-channel Video

With both **Video Frames** and **Subtitle/ASR**

Diverse Video Domain

Diverse video content from **YouTube**, **TV Episodes** and **Movie Clips**

Various Datasets over Representative Tasks

11 datasets over **3** tasks: **Retrieval**, **Question Answering** and **Captioning**.

Leaderboard!

To track the advances in Video-and-Langauge research.

Videos in VALUE

- Diverse video domains
- Varying video lengths
- High multimodal ratios

Table 1: Statistics of video data used in VALUE benchmark. Multi-channel ratio refers to percentage of videos with subtitles. Video lengths are measured in terms of seconds (s) on average.

Video Data	Source	#Video	Multi-channel Ratio	Length
TV (TVQA, TVR, TVC)	TV episodes	21.8K	100%	76s
How2 (How2R, How2QA)	Instructional Videos on Youtube	31.7K	99.36%	59s
VIOLIN	TV episodes, Movie Clips	15.9K	99.33%	40s
VLEP	TV episodes, Vlog on Youtube	10.2K	98.11%	32s
YouCook2 (YC2C, YC2R)	Cooking Videos on Youtube	15.4K	94.40%	20s
VATEX (VATEX-EN-R/C)	Various Youtube Videos	41.3K	50.93%	10s

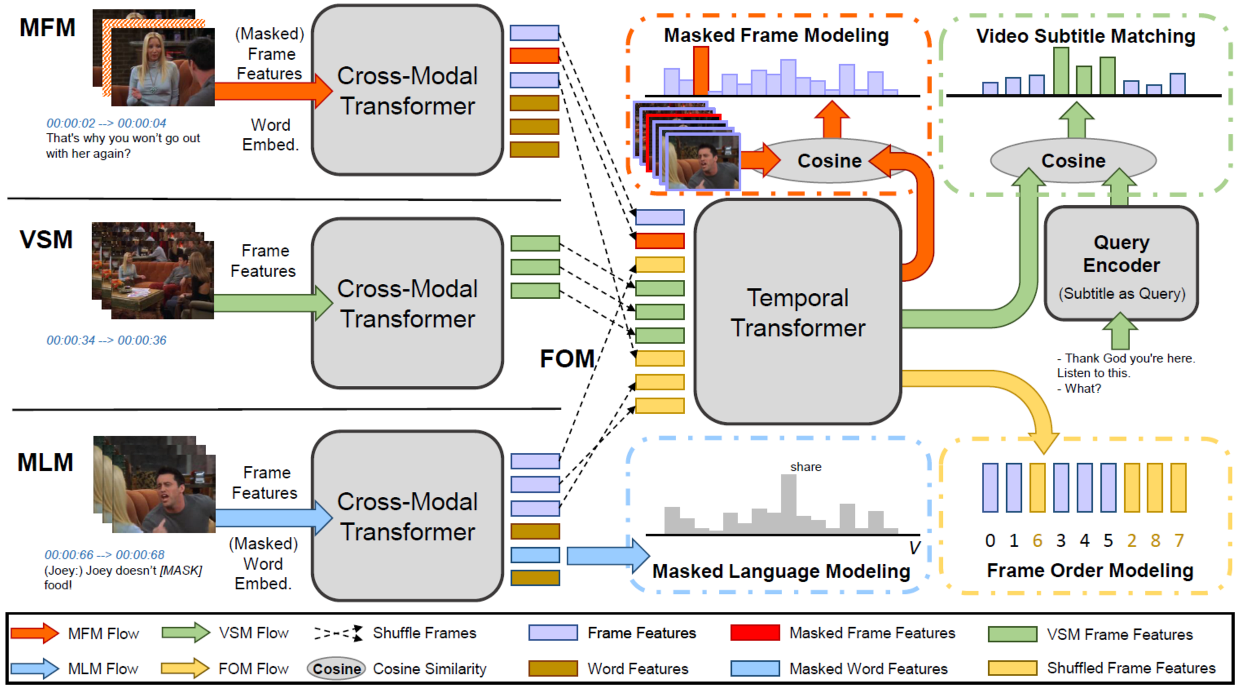
Task Name	Video Source	More info	Metric
		Retrieval Tasks	
TVR	TV episodes		Average(R@1, 5, 10) with tIoU >= 0.7
How2R	YouTube (HowTo100M)		Average(R@1, 5, 10) with tIoU >= 0.7
YC2R	YouTube		Average(R@1, 5, 10)
VATEX-EN-R	YouTube		Average(R@1, 5, 10)
		QA Tasks	
TVQA	TV episodes		Accuracy
How2QA	YouTube (HowTo100M)		Accuracy
VIOLIN	TV episodes, Movie clips		Accuracy
VLEP	TV episodes, YouTube		Accuracy
		Captioning Tasks	
TVC	TV episodes		CIDEr-D
YC2C	YouTube		CIDEr-D
VATEX-EN-C	YouTube		CIDEr-D

Analysis on VALUE Benchmark

- Impact of Input Channels and Video-Subtitle Fusion Methods
- Impact of Visual Representations
- Task Transferability Evaluation
- Multi-Task Learning Evaluation

VALUE Baseline: HERO

- **H**ierarchical **E**ncode**R** for **O**mni-representation Pre-training
  - Model Architecture
    - Cross-Modal Transformer
    - Temporal Transformer
  - Pre-training
    - Masked Language Modeling (MLM)
    - Masked Frame Modeling (MFM)
    - Video-Subtitle Matching (VSM)
    - Frame Order Modeling (FOM)
  - Achieving competitive results on multi-channel video-and-language tasks



Analysis on VALUE Benchmark

- Q1: Is video channel alone sufficient to achieve good performance?

Table 3: Impact of **input channels**. For video-only experiments, we replace all subtitle texts with empty strings. For sub-only experiments, the visual features are replaced with zero vectors. All results are reported on Val/Test (public) split without pre-training.

Input Channel	TVR	How2R	YC2R	VATEX-EN-R	TVQA	How2-QA	VIO-LIN	VLEP	TVC	YC2C	VATEX-EN-C	Meta-Ave
	AveR	AveR	AveR	AveR	Acc.	Acc.	Acc.	Acc.	C	C	C	
Video-only	4.49	1.70	9.74	57.50	44.17	60.42	58.53	57.56	37.52	53.61	51.14	39.67
Sub-only	1.95	0.98	32.31	5.21	70.15	68.15	66.26	58.06	38.74	93.33	9.28	40.40
Video+Sub	**7.72**	**1.91**	**33.91**	**58.99**	**71.08**	**69.44**	**66.83**	**58.79**	**48.48**	**108.46**	**52.15**	**52.52**

- Meta-Ave: Average of scores across all tasks

Subtitle/ASR contains rich information that are helpful for solving the tasks

Analysis on VALUE Benchmark

- Q2: What is the effective way to fuse video and subtitle embeddings?

Fusion Method	TVR	How2R	YC2R	VATEX-EN-R	TVQA	How2-QA	VIO-LIN	VLEP	TVC	YC2C	VATEX-EN-C	Meta-Ave
	AveR	AveR	AveR	AveR	Acc.	Acc.	Acc.	Acc.	C	C	C	
1 two-stream	5.66	1.90	32.60	48.19	**71.15**	**69.63**	66.61	58.49	42.67	99.35	39.04	48.66
2 sequence concat	5.60	2.73	**35.55**	**60.24**	69.61	68.99	66.09	**60.91**	44.73	99.78	**52.65**	51.53
3 temp. align + sum	6.75	2.44	31.84	58.11	70.23	69.44	66.33	57.72	47.80	104.97	52.07	51.61
4 temp. align + concat	7.10	**3.19**	32.59	57.33	69.81	69.31	66.16	58.54	47.12	100.90	52.09	51.29
5 HERO	**7.72**	1.91	33.91	58.99	71.08	69.44	**66.83**	58.79	**48.48**	**108.46**	52.15	**52.52**

Early fusion of temporally-aligned frames and subtitles can help boost model performance

Analysis on VALUE Benchmark

- Q3: How VALUE tasks relate to each other?
  - There are large differences between tasks
    - Domain gaps
    - Different video lengths
    - Different task formalization

(a) Retrieval Tasks.

Train Data	TVR	How2R	YC2R	VATEX-R
TVR	**7.72**	0.00	0.35	2.79
How2R	0.03	**1.91**	10.30	10.31
YC2R	-	-	**33.91**	1.01
VATEX-R	-	-	3.82	**58.99**

(b) QA Tasks.

Train Data	TVQA	How2-QA	VIO-LIN	VLEP
TVQA	**71.08**	36.89	50.01	53.23
How2QA	21.75	**69.44**	53.85	55.65
VIOLIN	20.12	40.55	**66.83**	44.26
VLEP	22.16	26.04	50.00	**58.79**

(c) Captioning Tasks.

Train Data	TVC	YC2C	VATEX-C
TVC	**48.48**	1.35	1.72
YC2C	0.43	**108.46**	0.74
VATEX-C	4.25	7.09	**52.15**

Analysis on VALUE Benchmark

- Q4: Can we have one model to conquer them all?

Training Setting	TVR	How2R	YC2R	VATEX-EN-R	TVQA	How2-QA	VIO-LIN	VLEP	TVC	YC2C	VATEX-EN-C	Meta-Ave
	AveR	AveR	AveR	AveR	Acc.	Acc.	Acc.	Acc.	C	C	C	
1 Human	-	-	-	-	89.41	90.32	91.39	90.50	62.89	-	62.66	-
Finetune-only												
2 ST	7.70	1.74	40.69	38.34	70.54	69.00	63.75	57.94	<u>46.76</u>	<u>106.24</u>	<u>52.16</u>	50.44
3 MT by Task	7.75	1.90	46.38	38.17	71.26	71.43	64.74	68.01	46.01	105.22	51.07	52.00
4 MT by Domain	10.01	<u>2.69</u>	44.58	36.10	73.94	70.01	<u>65.93</u>	67.37	46.53	100.74	50.46	51.97
5 AT	9.76	2.42	47.91	37.33	<u>73.98</u>	71.14	65.80	<u>68.03</u>	46.46	101.72	51.07	52.33
6 AT→ST	<u>10.43</u>	2.68	<u>49.48</u>	<u>38.58</u>	73.46	<u>71.88</u>	65.73	67.80	46.12	103.73	51.87	<u>52.89</u>

- ST: single-task training
- MT by Task: jointly train tasks within the same task type (e.g.: retrieval tasks)
- MT by Domain: jointly train tasks within the same domain (e.g.: YouTube domain)
- AT: a single model trained on all 11 datasets
- AT -> ST: all-task training as pre-training, then finetune on each single task

Analysis on VALUE Benchmark

- Q4: Can we have one model to conquer them all?

Training Setting	TVR AveR	How2R AveR	YC2R AveR	VATEX-EN-R AveR	TVQA Acc.	How2-QA Acc.	VIO-LIN Acc.	VLEP Acc.	TVC C	YC2C C	VATEX-EN-C C	Meta-Ave
1 Human	-	-	-	-	89.41	90.32	91.39	90.50	62.89	-	62.66	-
Finetune-only												
2 ST	7.70	1.74	40.69	38.34	70.54	69.00	63.75	57.94	46.76	106.24	52.16	50.44
3 MT by Task	7.75	1.90	46.38	38.17	71.26	71.43	64.74	68.01	46.01	105.22	51.07	52.00
4 MT by Domain	10.01	2.69	44.58	36.10	73.94	70.01	65.93	67.37	46.53	100.74	50.46	51.97
5 AT	9.76	2.42	47.91	37.33	73.98	71.14	65.80	68.03	46.46	101.72	51.07	52.33
6 AT→ST	10.43	2.68	49.48	38.58	73.46	71.88	65.73	67.80	46.12	103.73	51.87	52.89
Pre-train + Finetune												
7 ST	12.04	4.09	57.88	**40.63**	74.36	**74.76**	65.31	**68.46**	**48.97**	**127.94**	**52.57**	**57.00**
8 MT by Task	**12.63**	**4.66**	**59.20**	39.97	74.56	74.40	66.34	68.11	48.02	123.40	50.49	56.53
9 MT by Domain	11.53	4.03	52.14	36.97	74.54	74.08	65.92	68.06	47.23	100.29	45.95	52.79
10 AT	11.61	4.03	52.20	38.01	**75.12**	73.66	66.60	68.27	46.04	109.11	49.74	54.04
11 AT→ST	12.17	4.51	54.16	38.86	75.05	74.24	**66.93**	67.96	46.38	120.86	50.59	55.61

VALUE Challenge

Welcome to VALUE Challenge 2021!

Overview

We are pleased to announce VALUE Challenge 2021! The challenge will be hosted at the Forth Workshop on Closing the Loop Between Vision and Language, ICCV 2021.

Please stay tuned for more information!

Important Dates

- Challenge Launch: **June 7th, 2021.**
- Results Submission Deadline: **23:59:59 (AoE), September 13th, 2021.**
- Decision to participants: **September 27th, 2021.**
- The winners will be announced at the CLVL workshop, ICCV 2021 on October 17th, 2021 .

https://value-benchmark.github.io

VALUE Leaderboard

VALUE | Retrieval | QA | Captioning

The models are ranked by the Mean-Rank, the average of model ranks over 11 tasks. We break ties using the Meta-Ave, the average of model performance across 11 tasks. AveR, accuracy and CiDER are used as evaluation metrics for Retrieval, QA and Captioning tasks, respectively.

Rank	Model	Mean-Rank	Meta-Ave	TVR	How2R	YC2R	VATEX-EN-R	TVQA	How2QA	VIOLIN	VLEP	TVC	YC2C	VATEX-EN-C
- 06/07/2021	Human *VALUE baseline*	-	-	-	-	-	-	89.41	90.32	91.39	90.50	62.89	-	62.66
1 09/14/2021	craig.starr *Kakao Brain*	1.18	62.87	15.41	6.75	66.04	53.07	77.52	78.31	67.52	69.04	54.16	143.17	60.53
2 09/14/2021	lgh	2.45	60.00	13.12	4.64	62.68	49.86	75.45	73.92	67.47	68.37	53.34	128.87	62.30
3 06/07/2021	HERO (AT->ST, PT+FT) *VALUE baseline*	3.27	57.58	13.56	3.95	54.28	49.09	74.83	74.60	67.18	69.37	48.13	121.89	56.54
4 06/07/2021	HERO (AT->ST, FT-only) *VALUE baseline*	4.18	56.07	12.40	3.61	50.93	49.91	74.38	71.88	66.80	68.68	49.41	110.63	58.09

https://value-benchmark.github.io

Thank you