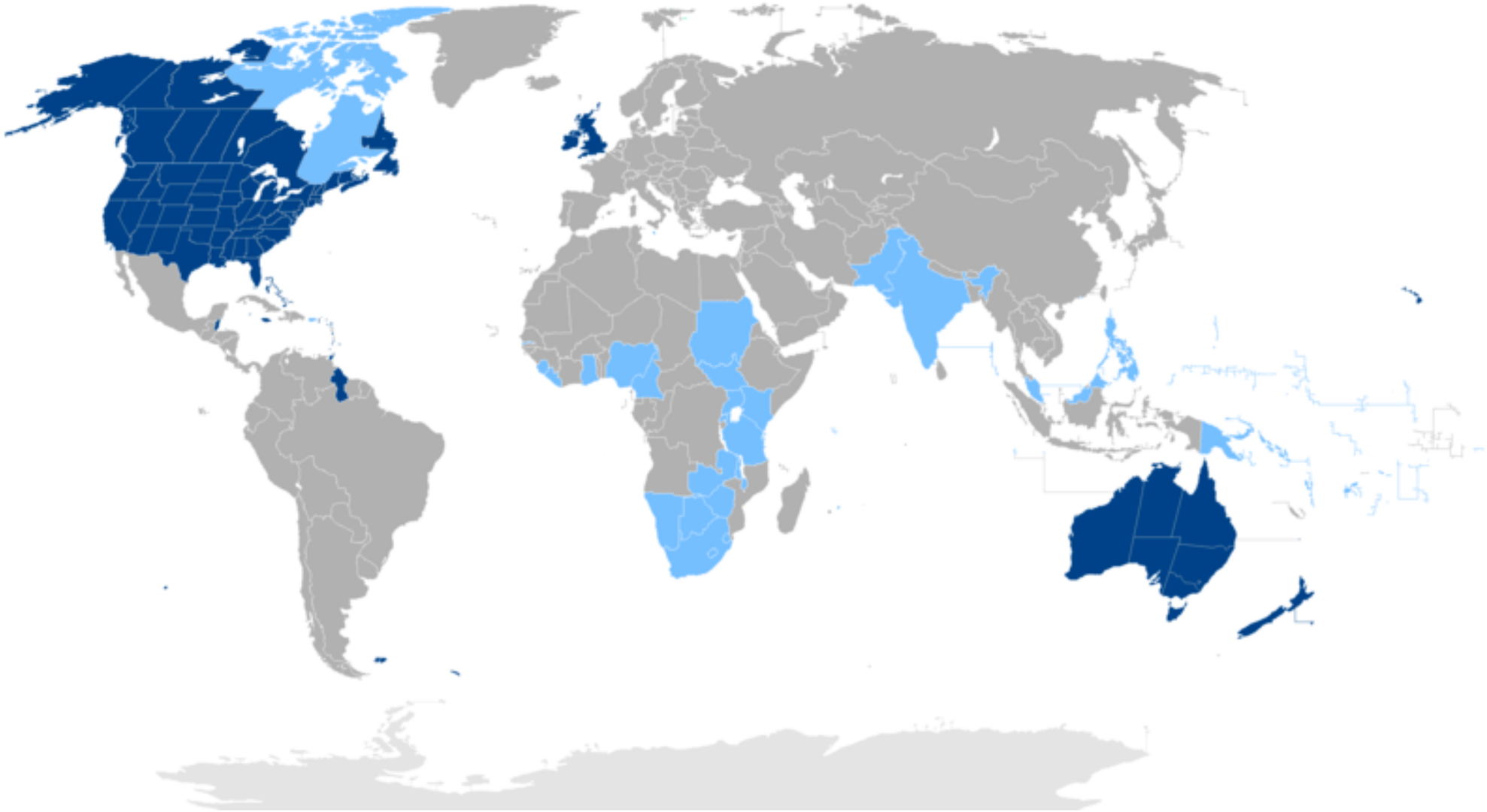
VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research

Xin (Eric) Wang
UC Santa Cruz
@ACL 2020 ALVR Workshop

Why VaTeX?


ActyNet-Cap
M-VAD
TACoS
Charades
VideoStory
VideoStory
VideoStory
Charades
MSR-VTT
MSR-VTT
MSVD
VideoStory
Charades
MPII
ActyNet-Cap
TACoS
MSR-VTT
TGIF
Youcook
LSMDC
Youcook
Youcook
Charades
MSVD
VTW
VideoStory
TGIF
M-VAD
TACoS
VideoStory
TACoS
TACoS
MSVD
MPII
MSR-VTT
ActyNet-Cap
VTW
Youcook
ActyNet-Cap
Youcook
TACoS
MSR-VTT
TGIF
ActyNet-Cap
Youcook
MSVD
Youcook
LSMDC
TACoS-MLevel
M-VAD
M-VAD
TACoS-MLevel
MSVD
Youcook
MPII
TGIF
MPII
Charades
VTW
Youcook
MPII
Charades
MSVD
M-VAD
Youcook
MPII
Charades
TACoS
VTW
TGIF
VTW
Charades
M-VAD
LSMDC
Youcook



Source: Wikipedia

There are thousands of languages on earth!

Why VaTeX?

Unique and Fine-grained

Why VaTeX?

257 Classes → 600 Classes

Why VaTeX?

200K Captions → 826K Captions

41.3K Unique Video Clips

826K Unique Captions in English & Chinese

600 Human Activities

Comparison with other Video Description Datasets

Dataset	MLingual	Domain	#classes	#videos:clips	#sent	#sent/clip
TACoS[45]	-	cooking	26	127:3.5k	11.8k	-
TACoS-MLevel[46]	-	cooking	67	185:25k	75k	3
Youcook[16]	-	cooking	6	88:-	2.7k	-
Youcook II[72]	-	cooking	89	2k:15.4k	15.4k	1
MPII MD[47]	-	movie	-	94:68k	68.3k	1
M-VAD[56]	-	movie	-	92:46k	55.9k	-
LSMDC[48]	-	movie	-	200:128k	128k	1
Charades[52]	-	indoor	157	10k:10k	27.8k	2-3
VideoStory[22]	-	social media	-	20k:123k	123k	1
ActyNet-Cap[30]	-	open	200	20k:100k	100k	1
MSVD[14]	✓	open	-	2k:2k	70k	35
TGIF[34]	-	open	-	-:100k	128k	1
VTW[69]	-	open	-	18k:18k	18k	1
MSR-VTT[66]	-	open	257	7k:10k	200k	20
VATEX (ours)	✓	open	600	41.3k:41.3k	826k	20

Meet VaTeX!

- **A person wearing a bear costume is inside an inflatable play area as they lose their balance and fall over.**
- **A person in a bear costumer stands in a bounce house and falls down as people talk in the background.**
- **A person dressed in a cartoon bear costume attempts to walk in a bounce house.**
- **A person in a mascot uniform trying to maneuver a bouncy house.**
- **A person in a comic bear suit falls and rolls around in a moon bounce.**

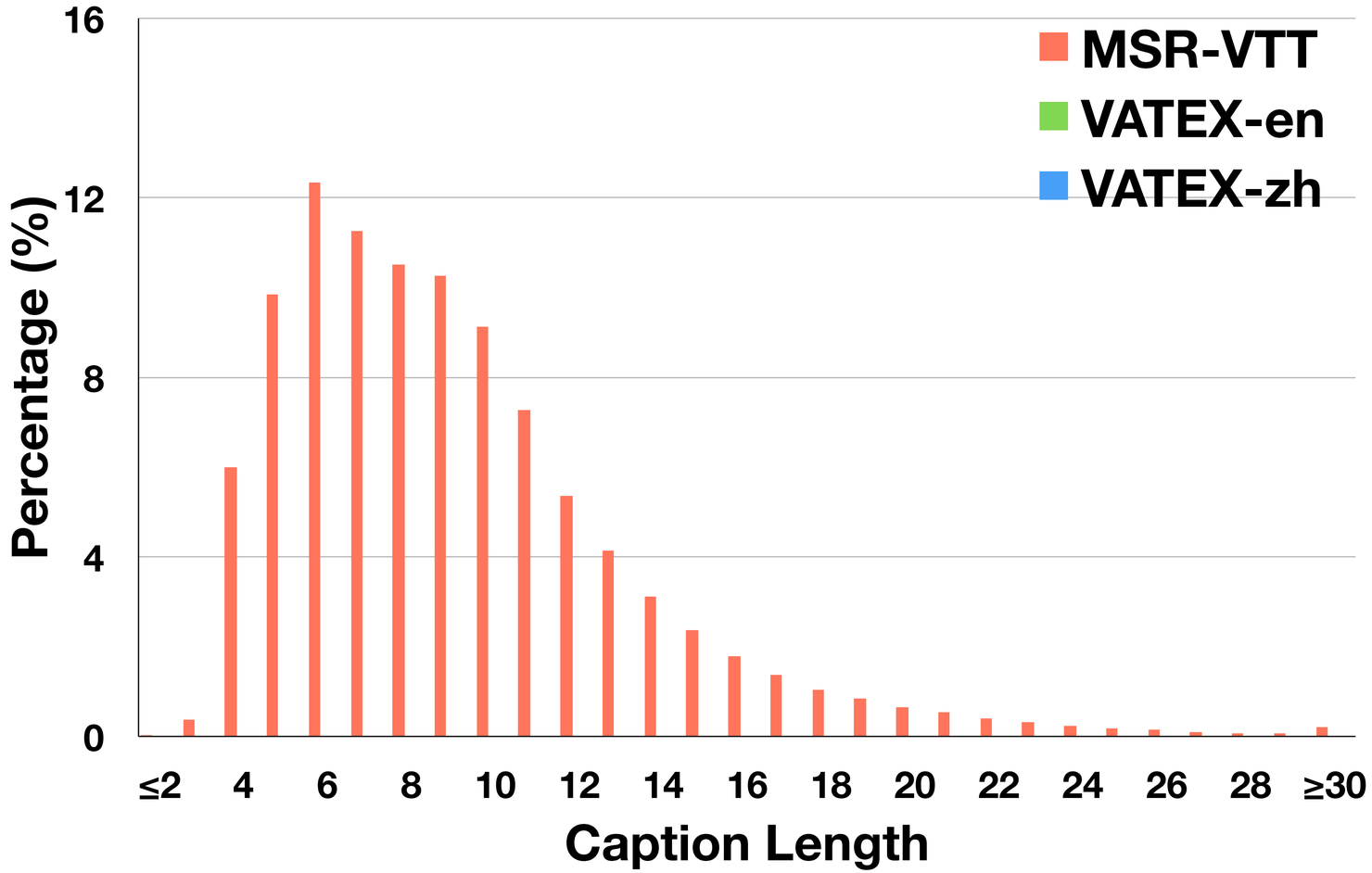
- 一个人穿着熊的布偶外套倒在了蹦床上。
- 一个人穿着一套小熊服装在充气蹦蹦床上摔倒了。
- 一个穿着熊外衣的人在充气垫子上摔倒了。
- 一个穿着深色衣服的人正在蹦蹦床上。
- 在一个充气大型玩具里,有一个人穿着熊的衣服站了一下之后就摔倒了。

- **A person dressed as a teddy bear stands in a bouncy house and then falls over.** ⟷
- **Someone dressed in a bear costume falling over in a bouncy castle.** ⟷
- **A person dressed up as a bear is standing in a bouncy castle and falls down.** ⟷
- **A man in a bear costume is balancing in a bouncy castle before they tumble to the floor.** ⟷
- **A man in costume was trying to stand straight on a bouncy castle but fell.** ⟷

- 一个打扮成泰迪熊的人站在充气房上，然后摔倒了。
- 有个穿着熊装的人在充气城堡摔倒了。
- 一个装扮成熊的人站在充气蹦床里，然后摔倒了。
- 一个穿着熊服装的人在一个有弹性的城堡里平衡, 然后他们就倒在了地板上。
- 一个穿着布偶熊的人试图站在一个充气城堡上，但却摔倒了。

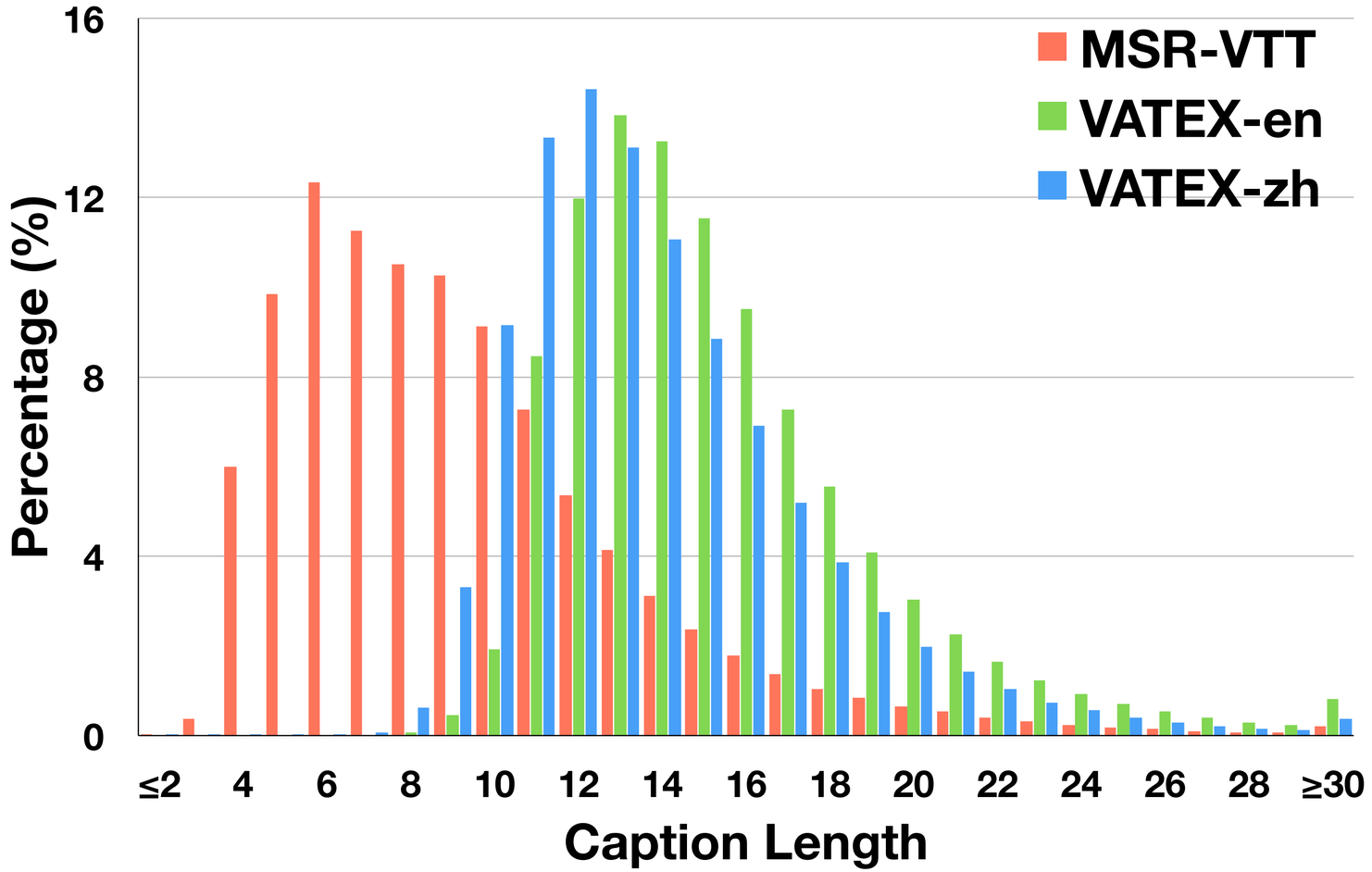
VaTeX *vs.* MSR-VTT

- **Distributions of Caption Lengths.**



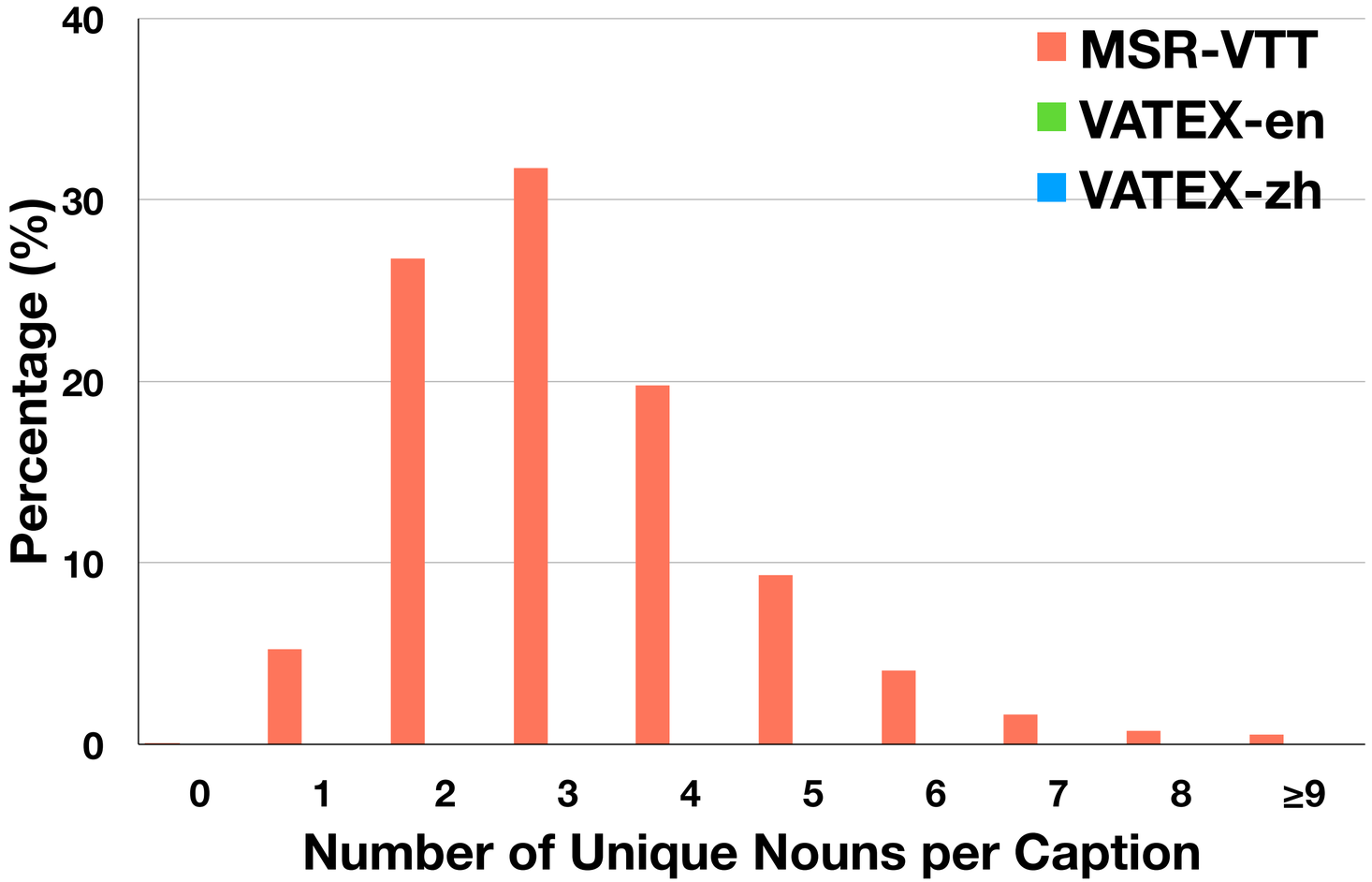
VaTeX *vs.* MSR-VTT

- **Distributions of Caption Lengths.**



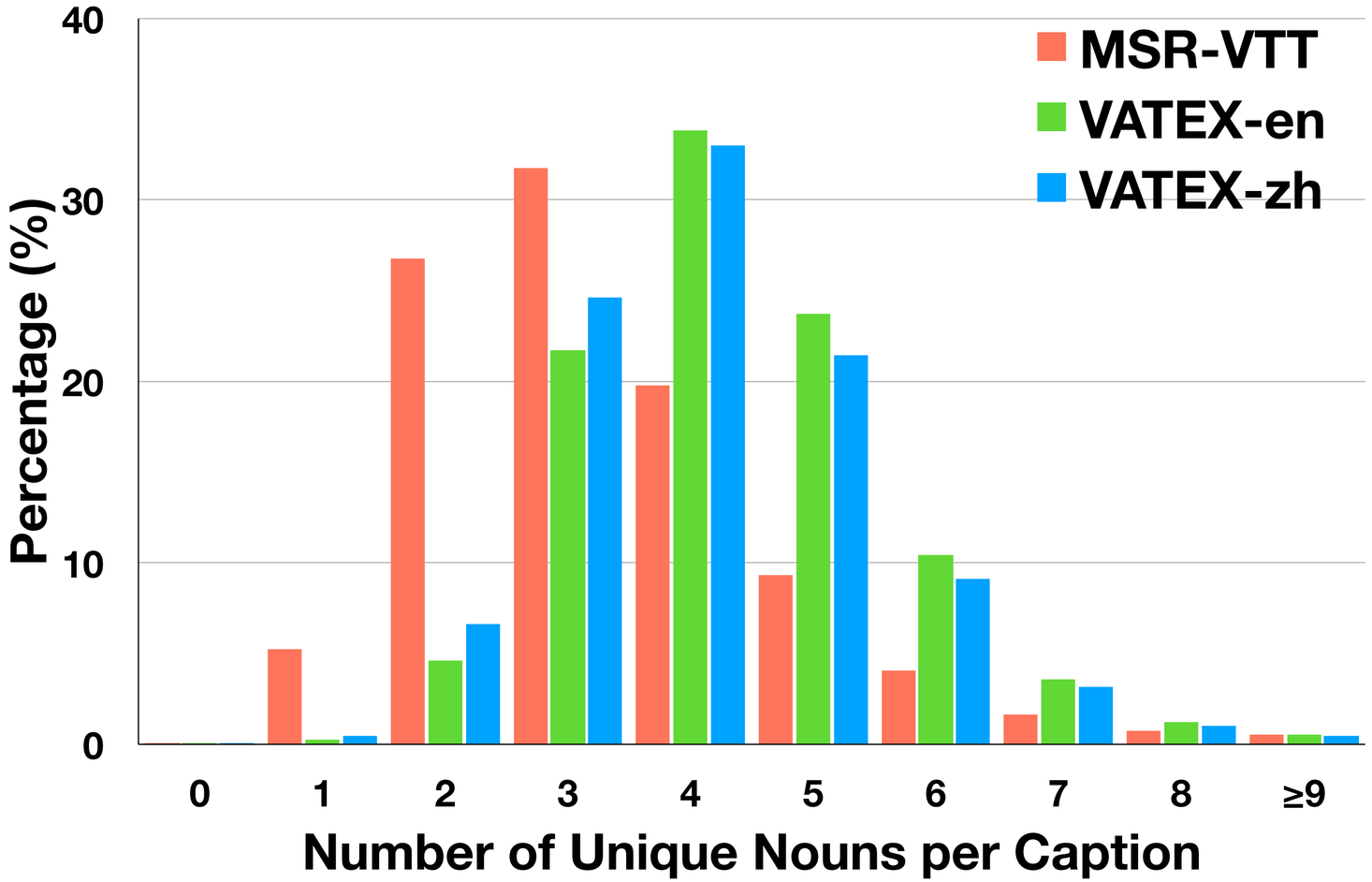
VaTeX *vs.* MSR-VTT

- **Distributions of Unique Nouns Per Caption.**



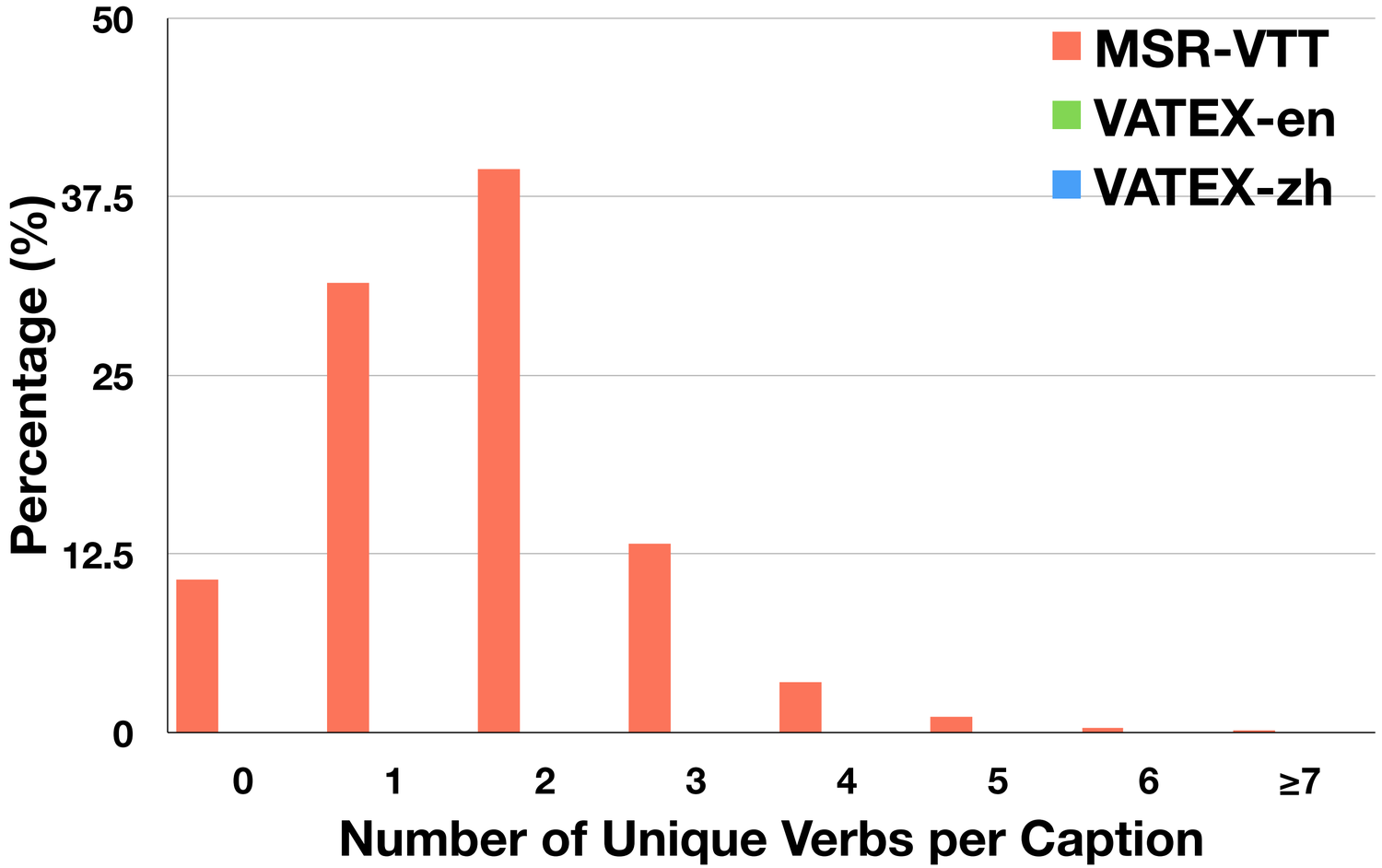
VaTeX *vs.* MSR-VTT

- **Distributions of Unique Nouns Per Caption.**



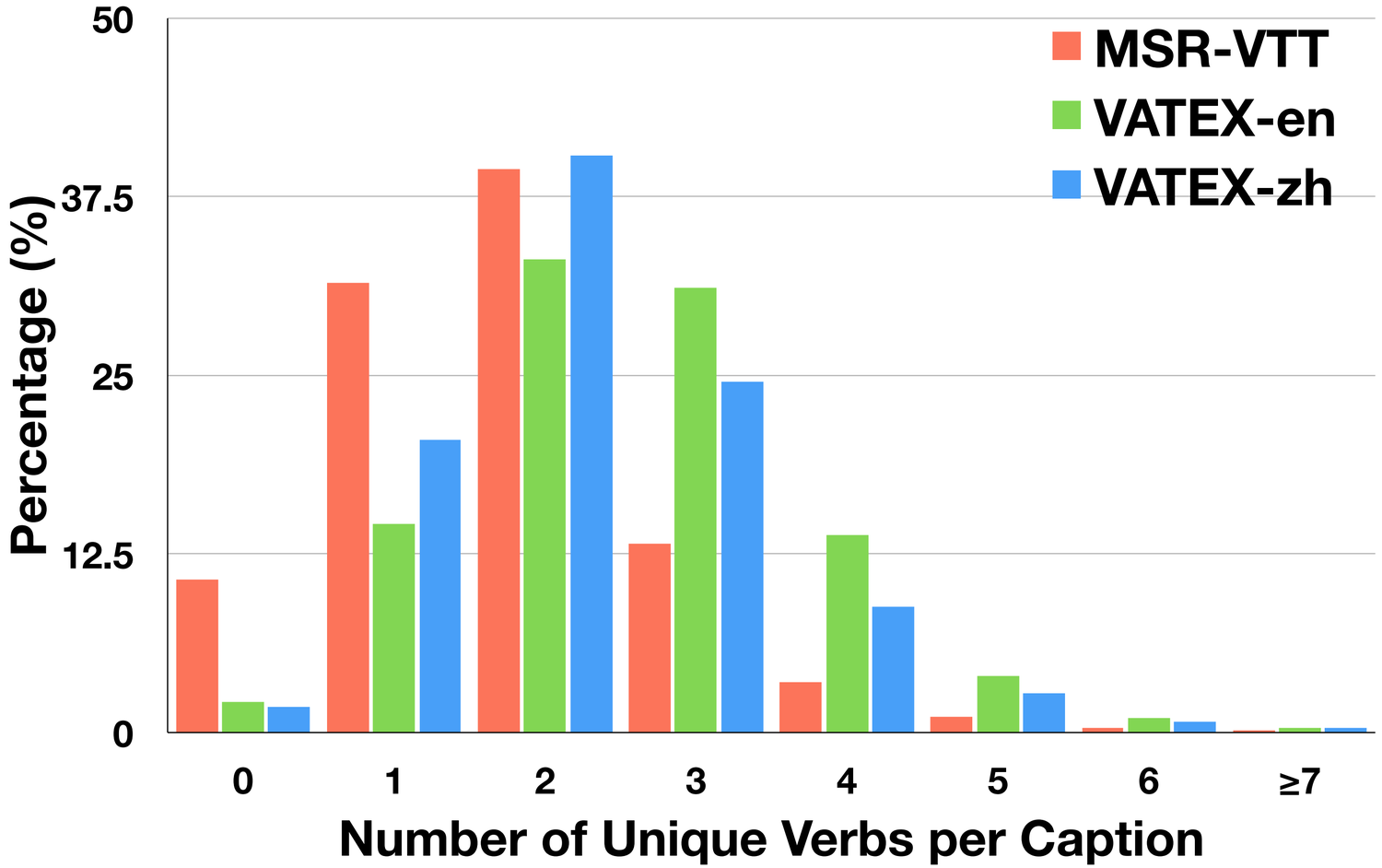
VaTeX *vs.* MSR-VTT

- **Distributions of Unique Verbs Per Caption.**



VaTeX *vs.* MSR-VTT

- **Distributions of Unique Verbs Per Caption.**

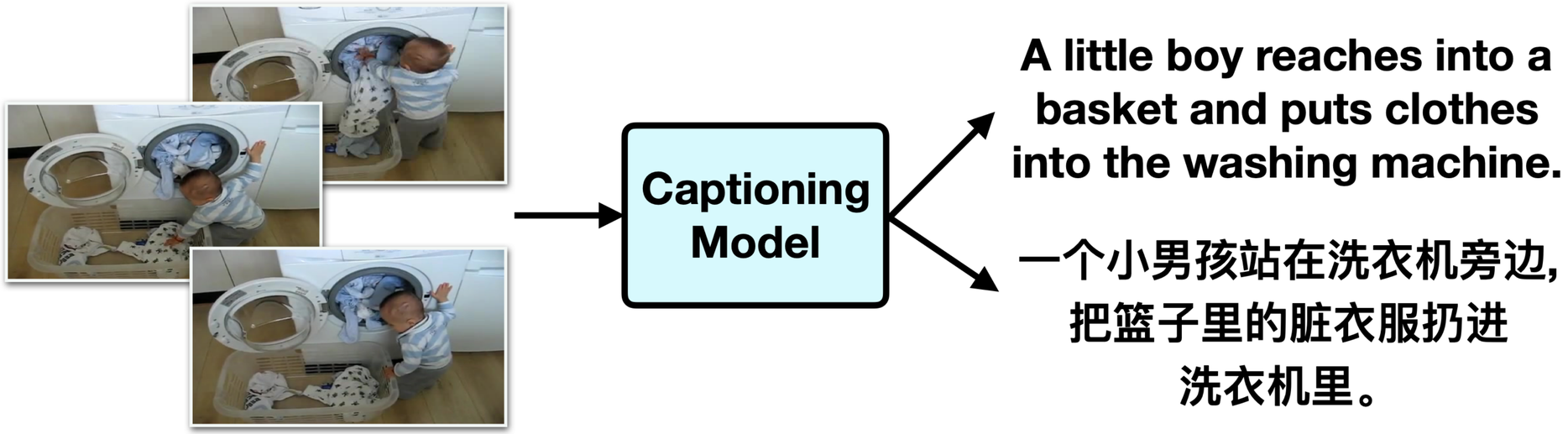


VaTeX *vs.* MSR-VTT

Dataset	sent length	duplicated sent rate		#unique n-grams				#unique POS tags			
		intra-video	inter-video	1-gram	2-gram	3-gram	4-gram	verb	noun	adjective	adverb
MSR-VTT	9.28	66.0%	16.5%	29,004	274,000	614,449	811,903	8,862	19,703	7,329	1,195
VaTeX-en	15.23	0	0	35,589	538,517	1,660,015	2,773,211	12,796	23,288	10,639	1,924
VaTeX-zh	13.95	0	0	47,065	626,031	1,752,085	2,687,166	20,299	30,797	4,703	3,086

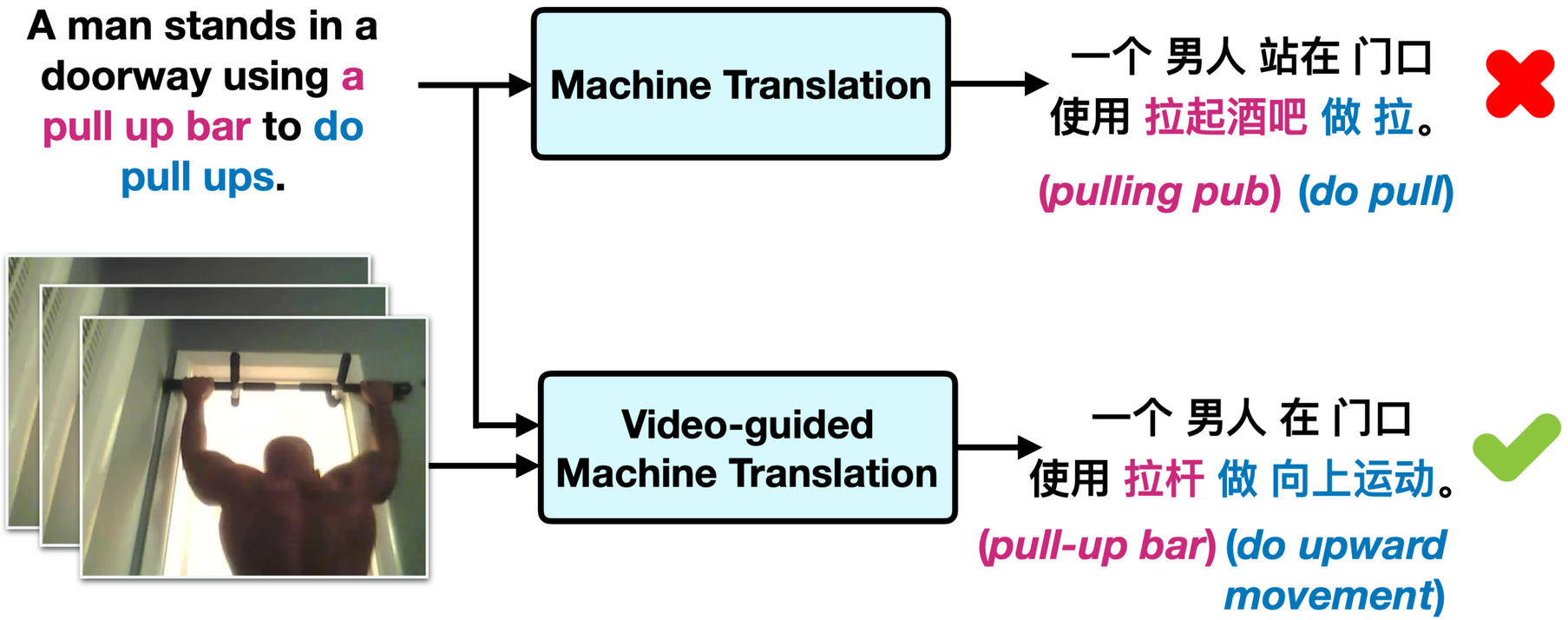
VaTeX Tasks

- **Multilingual Video Captioning**



VaTeX Tasks

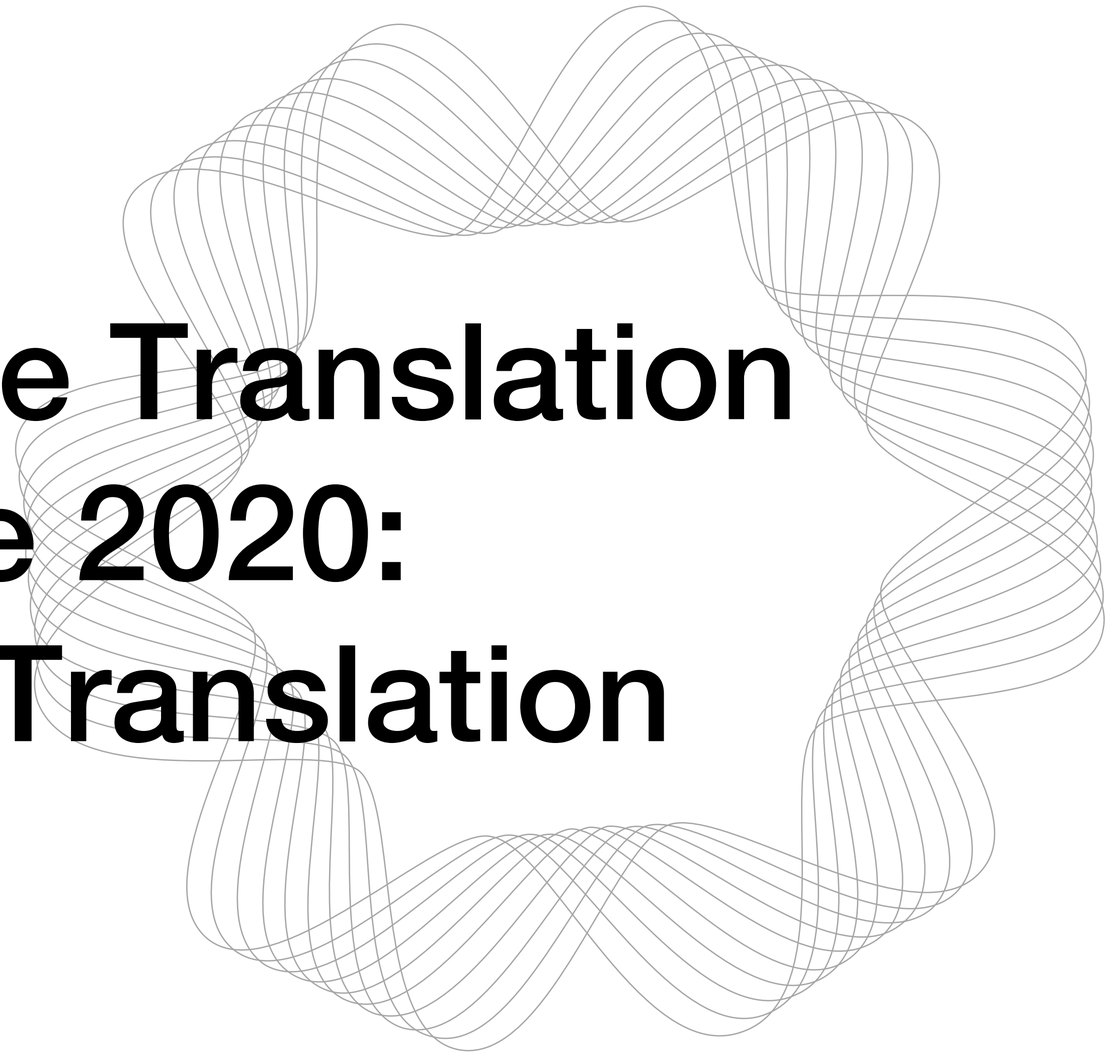
- **Video-Guided Machine Translation**



VaTeX Potentials

- **Video-text retrieval**
- **Query-based moment retrieval in untrimmed videos**
- **Zero-/few- shot video captioning**
- **Cultural and linguistic differences**
- **… …**

Video-guided Machine Translation (VMT) Challenge 2020: English-to-Chinese Translation



Requirements

	Train	Validation	Test
Videos	25,991	3,000	6,000
English Captions	259,910	30,000	60,000
Chinese Translations	259,910	30,000	60,000
Activity Label?	yes	yes	no
Released?	yes	yes	English only

- **Do NOT use any external corpora or pre-trained MT models.** The participants may not build upon any existing pre-trained machine translation models for this challenge. The VMT model must be trained on our VATEX dataset from scratch.

4/12/2020 -> 6/15/2020: 21 participants

VMT Leaderboard

Results

#	User	Entries	Date of Last Entry	Team Name	Corpus Bleu-4 ▲	Bleu-1 ▲	Bleu-2 ▲	Bleu-3 ▲	Bleu-4 ▲
1	tosho	3	06/15/20		0.366 (1)	0.631 (3)	0.419 (1)	0.302 (1)	0.225 (1)
2	zsyzsx1823	4	06/15/20		0.358 (2)	0.633 (1)	0.413 (2)	0.292 (2)	0.215 (2)
3	syuqing	6	06/15/20		0.353 (3)	0.632 (2)	0.409 (3)	0.287 (3)	0.209 (3)
4	acdart	2	05/11/20		0.314 (4)	0.598 (5)	0.371 (4)	0.250 (4)	0.175 (4)
5	zzxslp	1	04/15/20		0.311 (5)	0.599 (4)	0.368 (5)	0.247 (5)	0.172 (5)
6	Tcat	2	05/10/20		0.282 (6)	0.559 (6)	0.337 (6)	0.221 (6)	0.152 (6)

VMT Challenge 2020 Winner

VMT Challenge 2020 Second Place

VMT Challenge 2020 Third Place

Thanks!

vatex-challenge.org

Now VMT Challenge on CodaLab is open forever!

Organizers

Xin (Eric) Wang
UC Santa Cruz

An Yan
UC San Diego

Lei Li
ByteDance AI Lab

William Yang Wang
UC Santa Barbara