3AM: An Ambiguity-Aware Multi-Modal Machine Translation Dataset

Xinyu Ma[1], Xuebo Liu[2], Derek F. Wong[1], Jun Rao[2], Bei Li[3], Liang Ding[4], Lidia S. Chao[1], Dacheng Tao[4], and Min Zhang[2]

[1]University of Macau, [2]Harbin Institute of Technology, [3]Northeastern University, [4]The University of Sydney

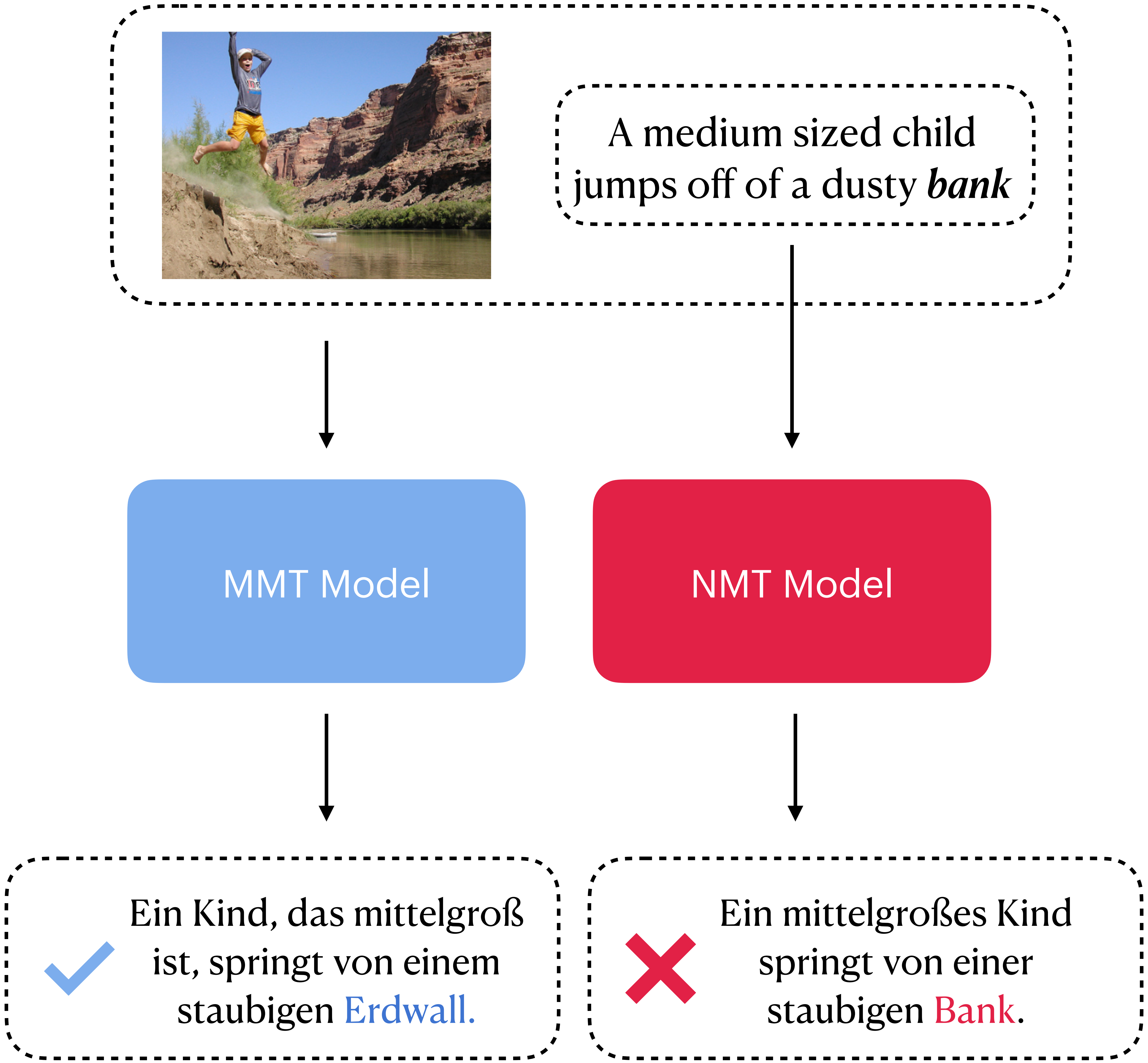
LREC-COLING 2024

Code and data available at https://github.com/MaxyLee/3AM

Multimodal Machine Translation

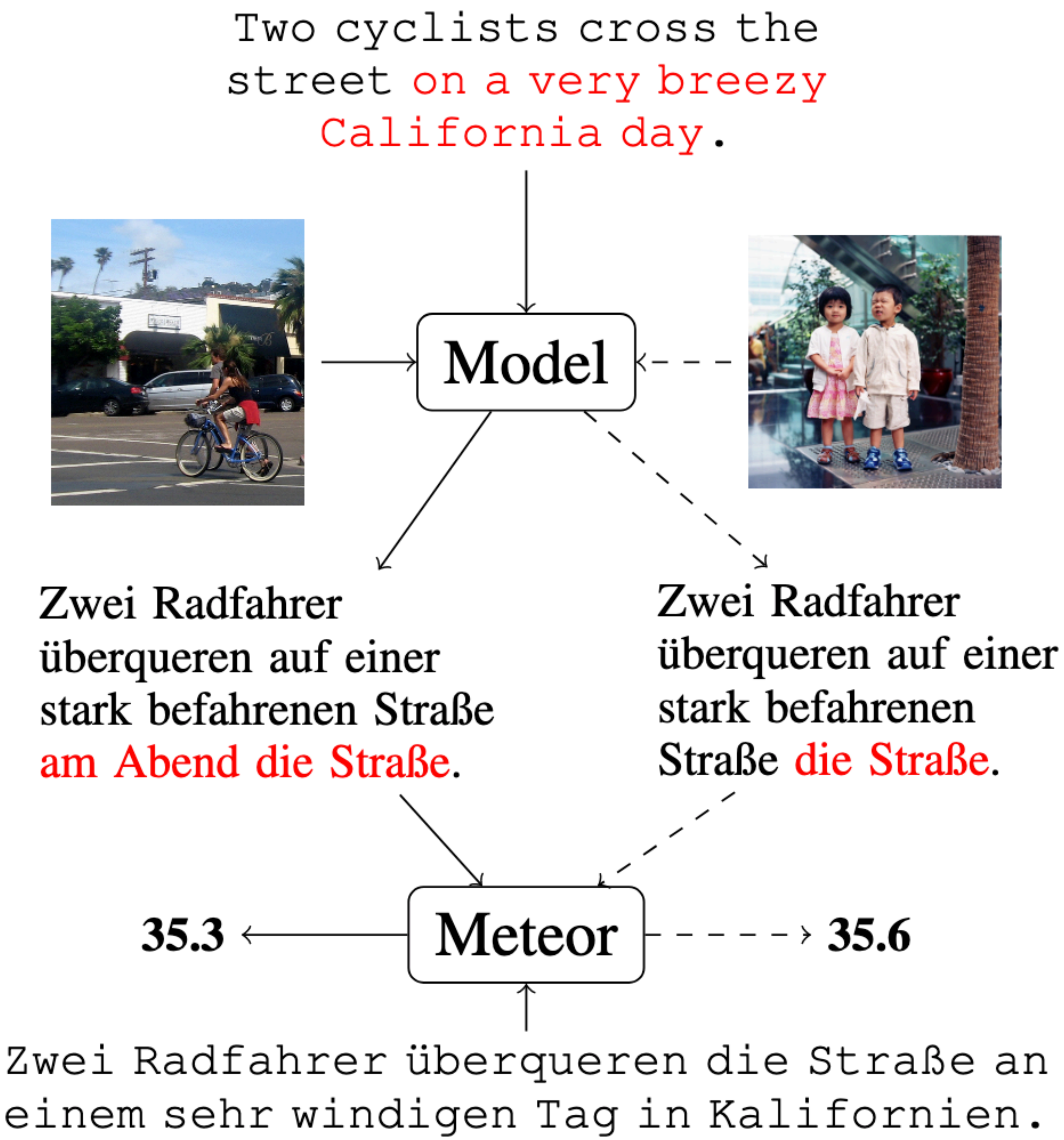
- Multimodal Machine Translation (MMT) aims at improving translation quality by utilizing additional visual information

- For example, visual information can help to remove ambiguity



Challenges

- Data scarcity
- Need for visual information
  - Text information is more important than visual information

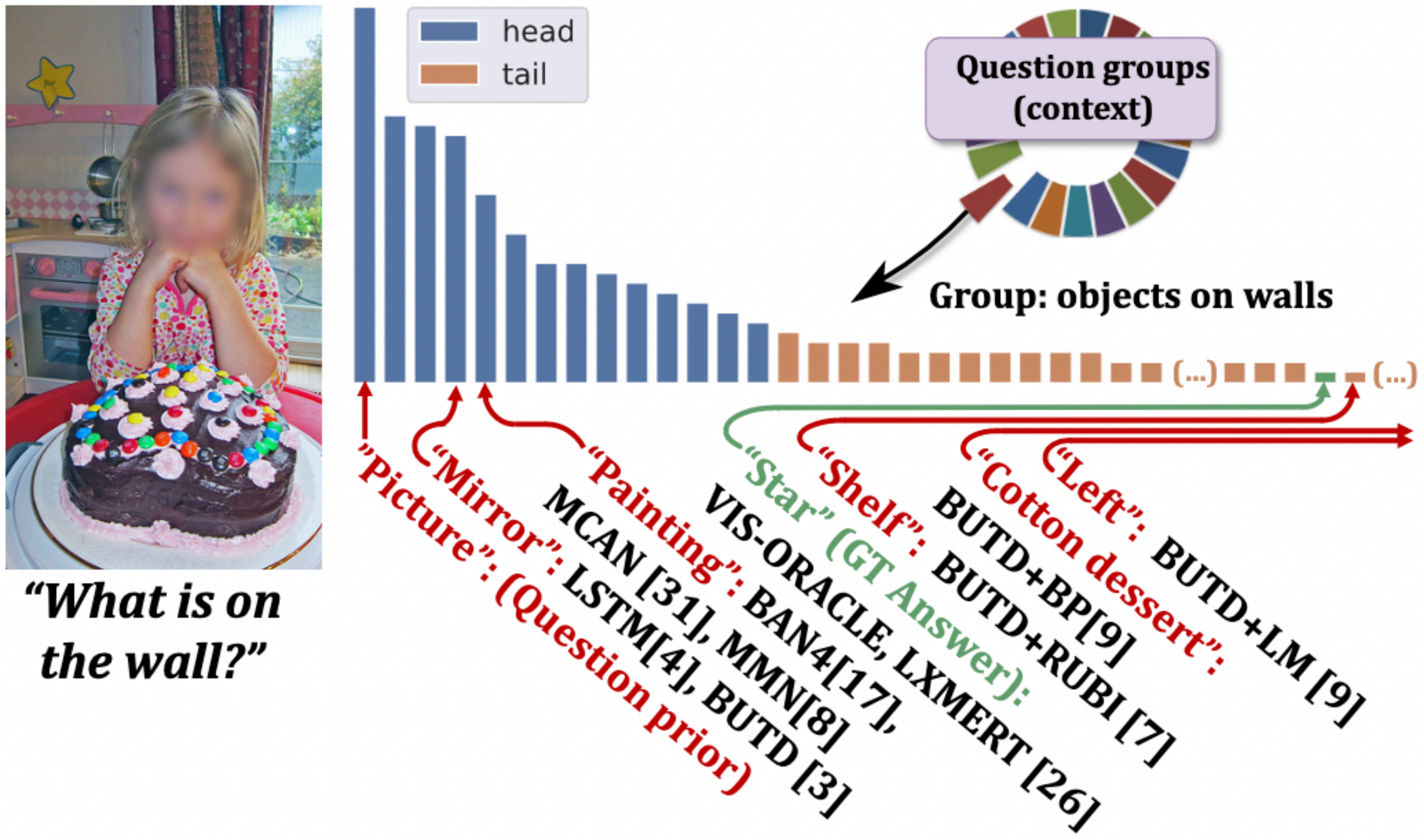


In some cases, the incongruent image performs better

[Elliott, EMNLP 2018]

Challenges

- Language Prior
  - VQA: an example
    - Q: 'What sport is'
      A: 'tennis' (41%)
    - Q: 'How many'
      A: '2' (39%)
- Hypothesis
  - Current MMT models rely on language prior and ignore the visual information



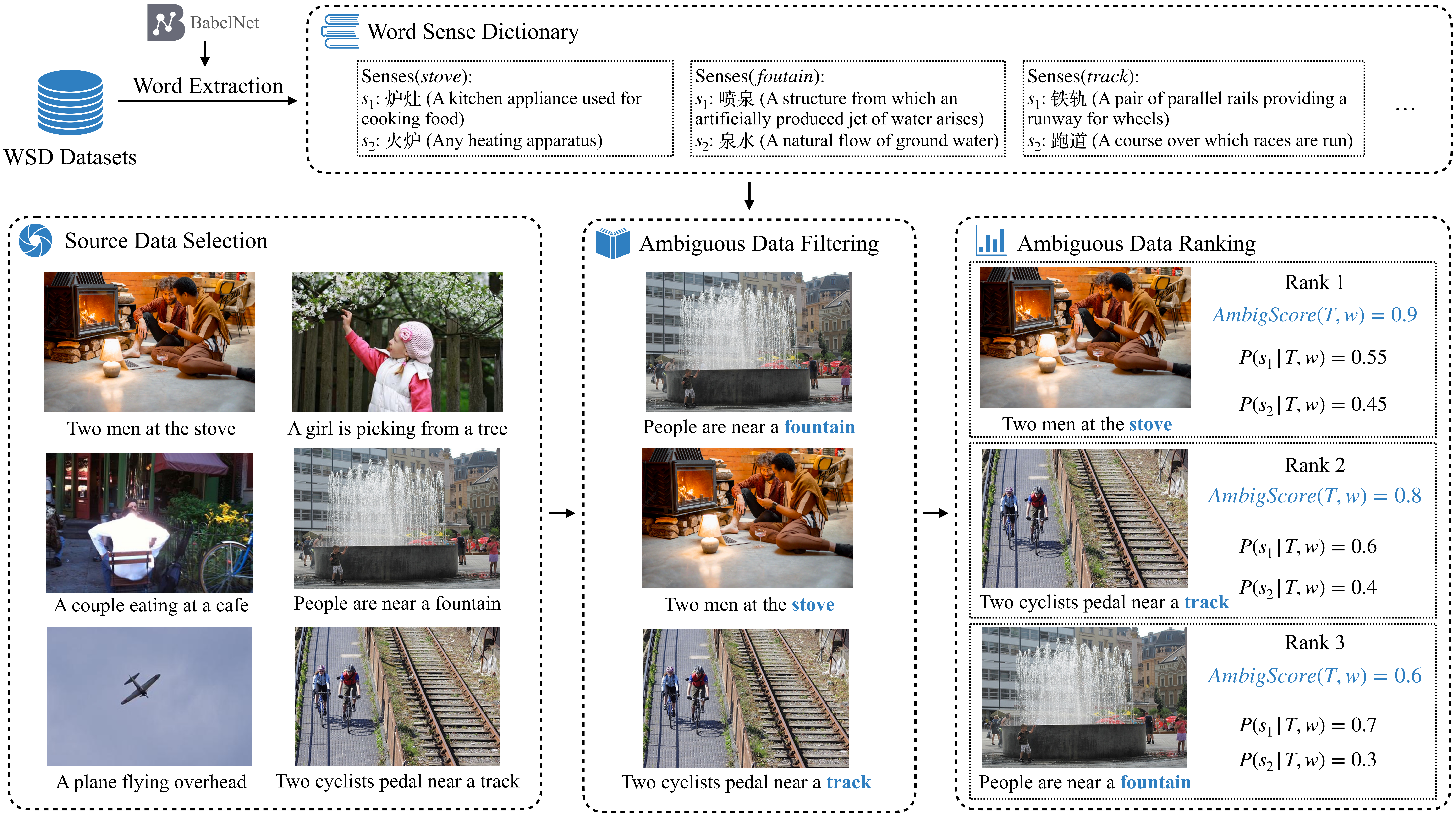
[Kervadec et al., CVPR 2021]

Motivation

- Select sentences with ambiguous words
- Force MMT models to utilize visual information

Image	English Sentence	Senses	Possible Chinese Translations
	A green gecko is seen on a **palm**.		✔ 在棕榈树上看到一只绿色壁虎。 ✘ 在手掌上看到一只绿色壁虎。
	A group of people on skis are being **taped**.		✔ 一群滑雪板上的人正在被录像。 ✘ 一群滑雪板上的人正在被录音。

Dataset construction

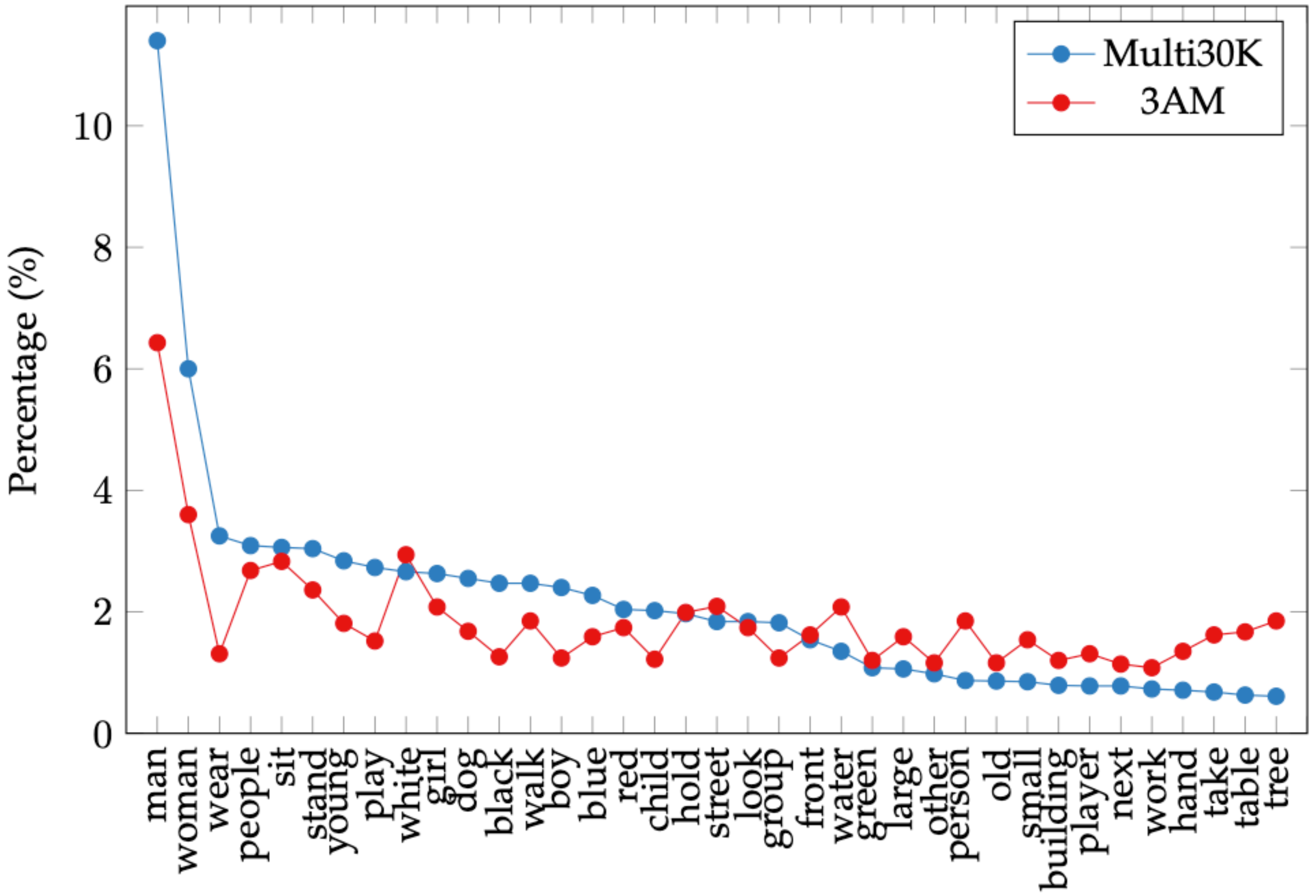


Dataset statistics

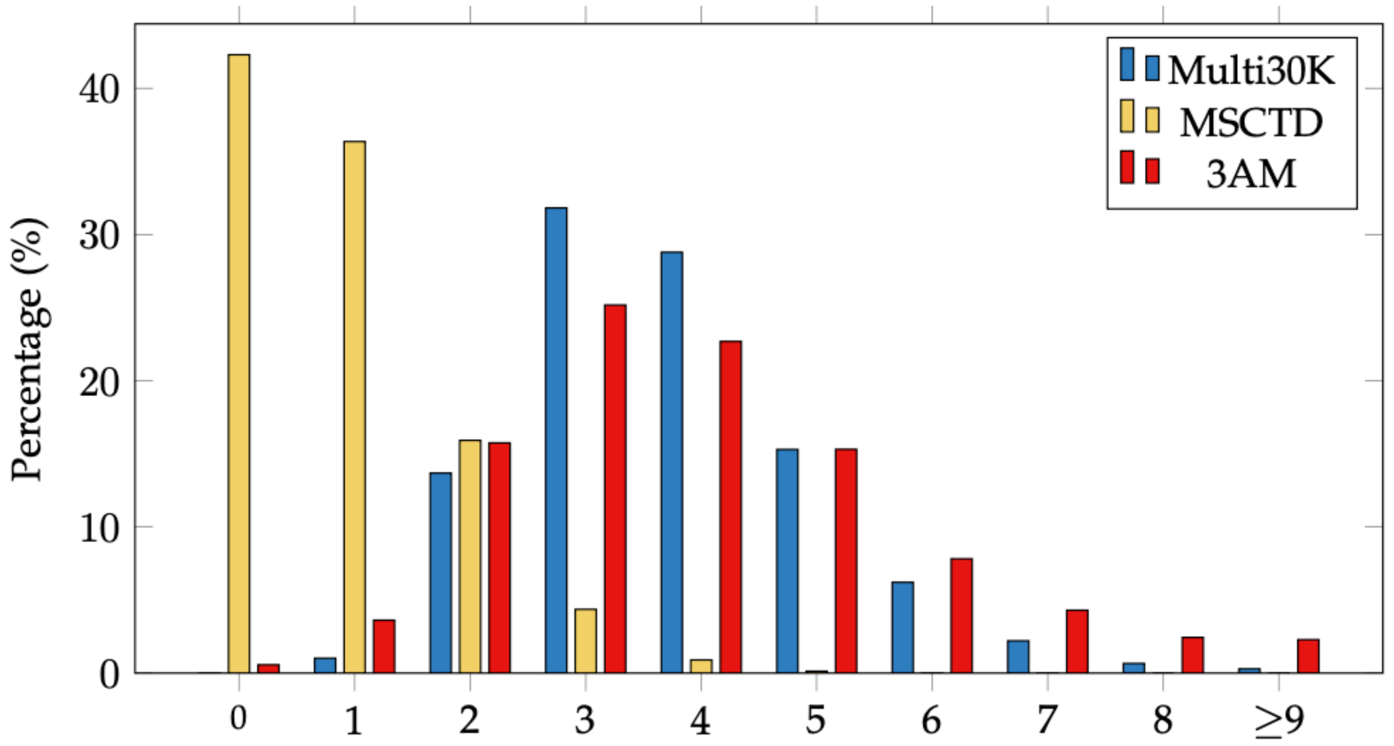
- Diversity
- Ambiguity

Dataset	Text					Image		
	Avg. length	Dist-1	Dist-2	Dist-3	Dist-4	LPIPS	IS	Ent-Obj
Multi30K	13.06	0.25	2.29	5.26	7.31	0.80584 ± 0.00010	23.25 ± 2.58	3.15
MSCTD	8.40	0.17	1.38	3.16	4.07	0.74149 ± 0.00011	7.85 ± 0.20	3.21
3AM	13.48	0.77	5.23	8.85	9.67	0.82975 ± 0.00011	29.94 ± 3.75	4.35

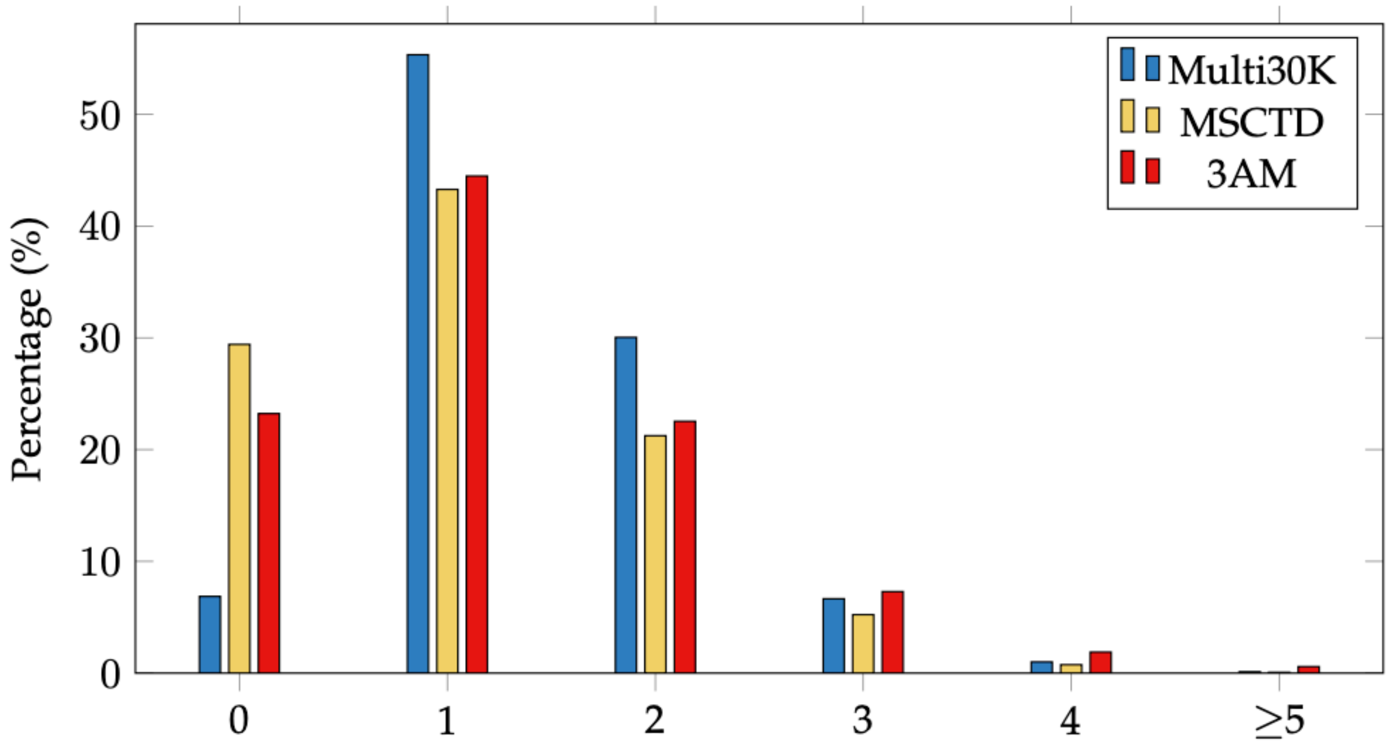
Detailed statistics of Multi30K, MSCTD, and 3AM



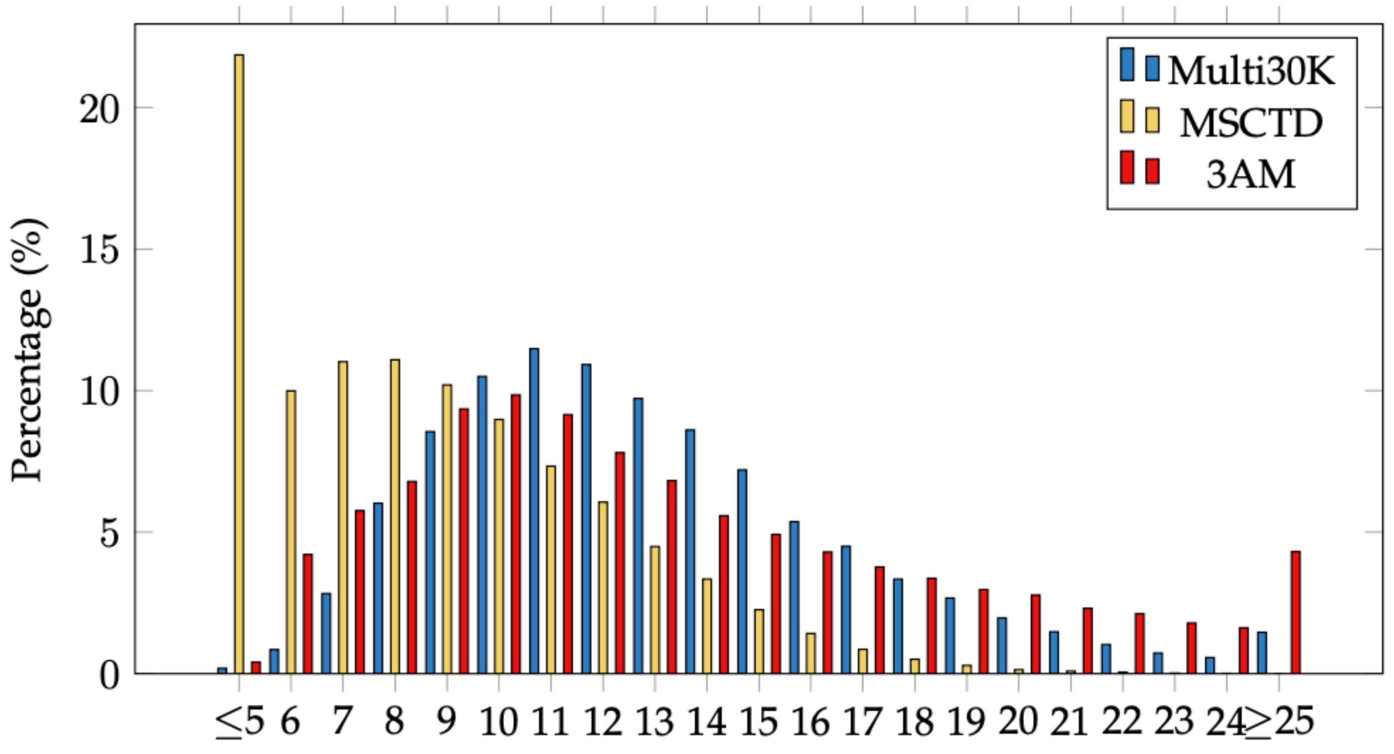
Plot of the most common words that occur in the captions of Multi30K and 3AM, the words in the 3AM dataset are more evenly distributed.



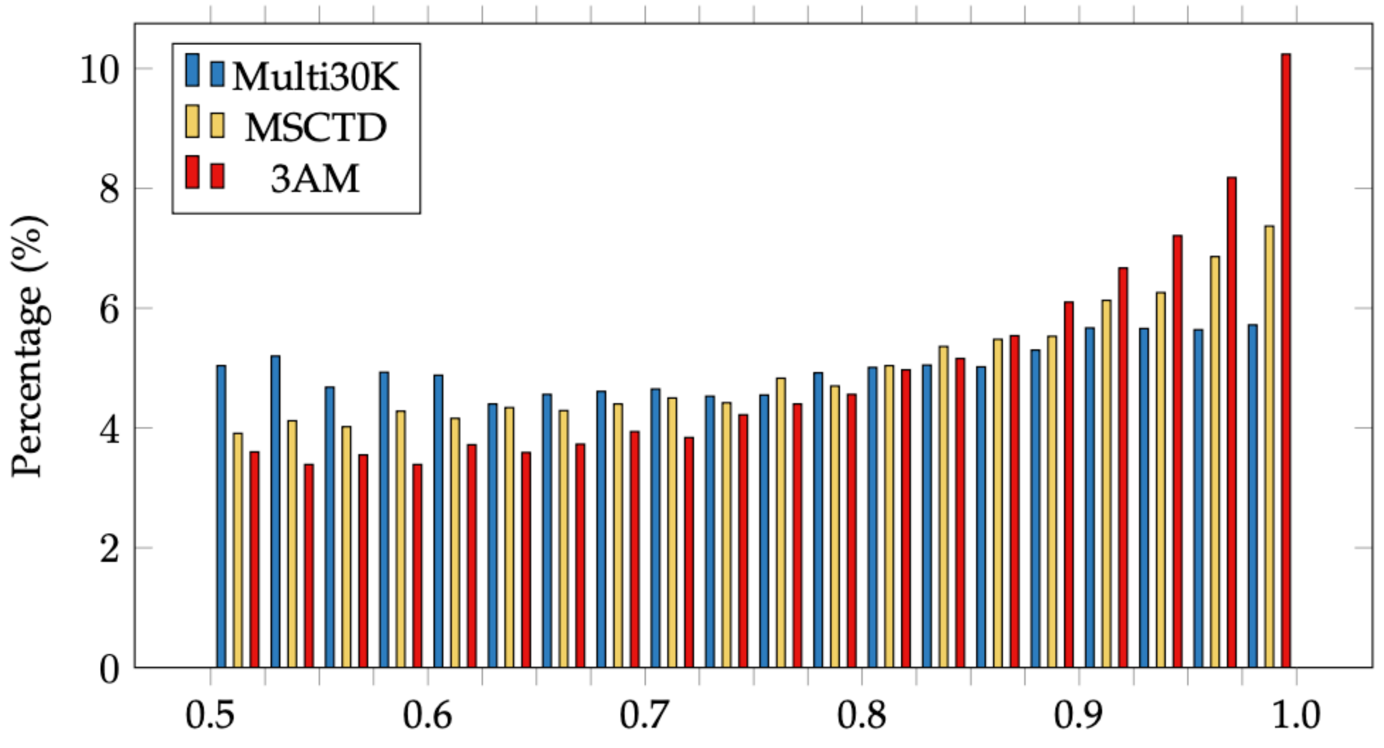
(a) Distributions of unique nouns per caption



(b) Distributions of unique verbs per caption



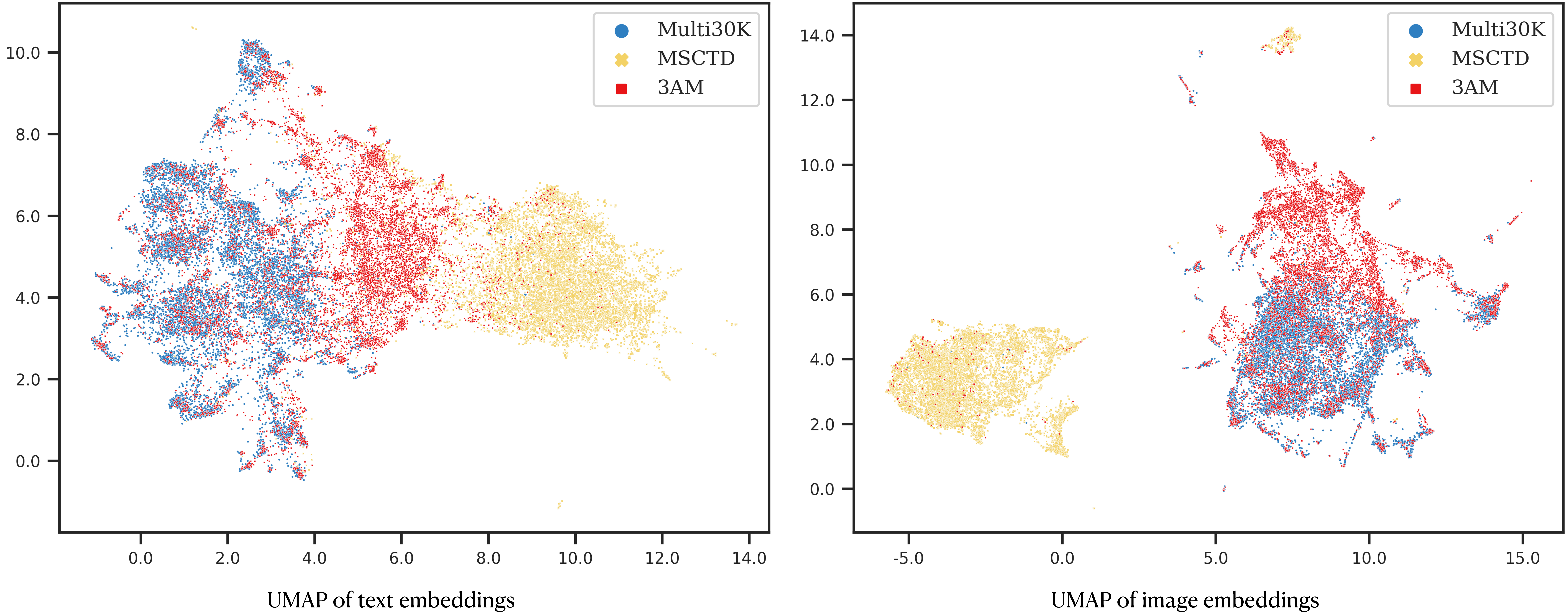
(c) Distributions of caption lengths



(d) Distributions of ambiguity scores

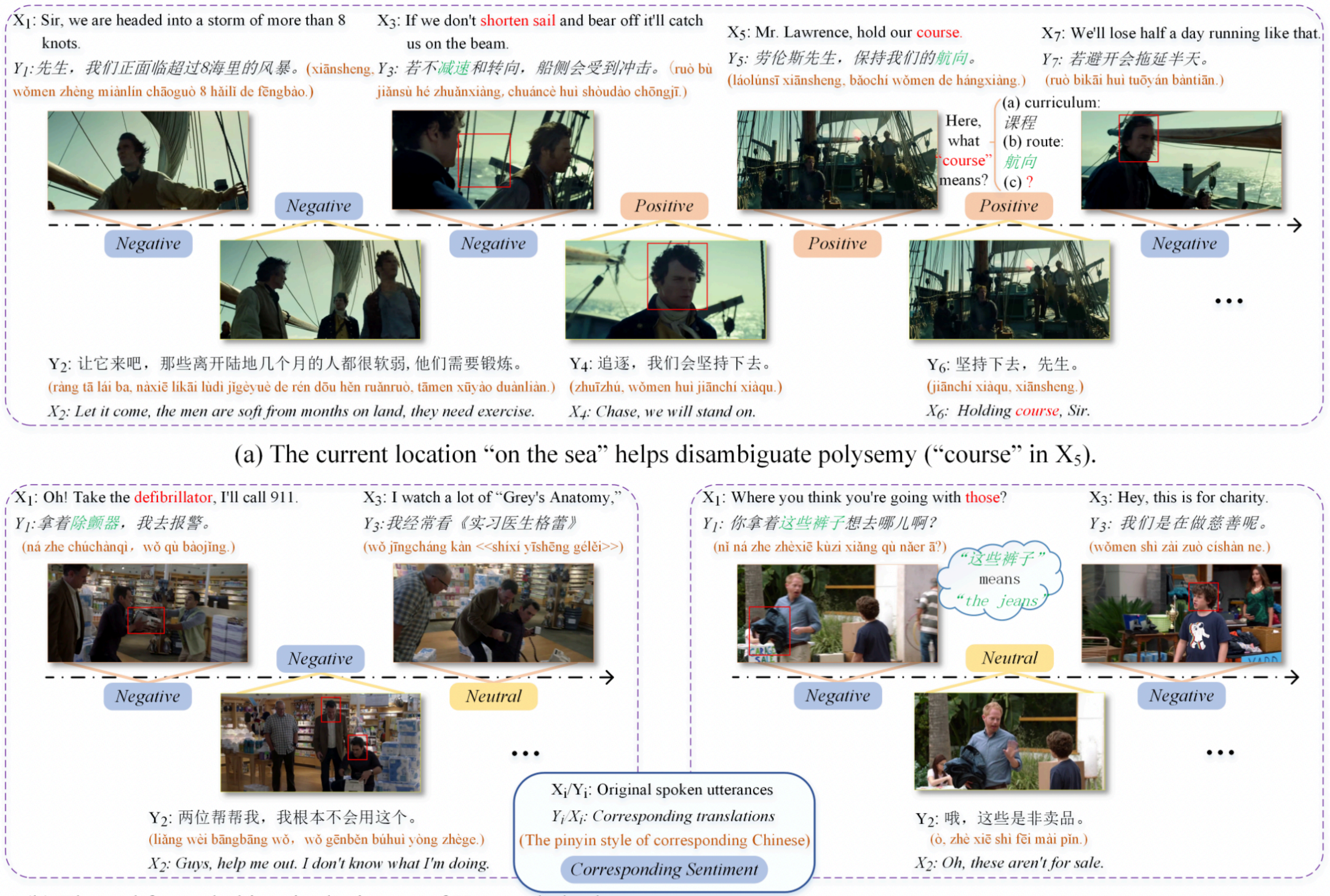
Dataset statistics

- Visualization
  - The 3AM dataset encompasses a greater diversity of caption styles and a wider range of visual concepts



Experiments

- Datasets
  - Multi30K
  - MSCTD
    - Multimodal sentiment chat translation dataset



(a) The current location "on the sea" helps disambiguate polysemy ("course" in X_5).

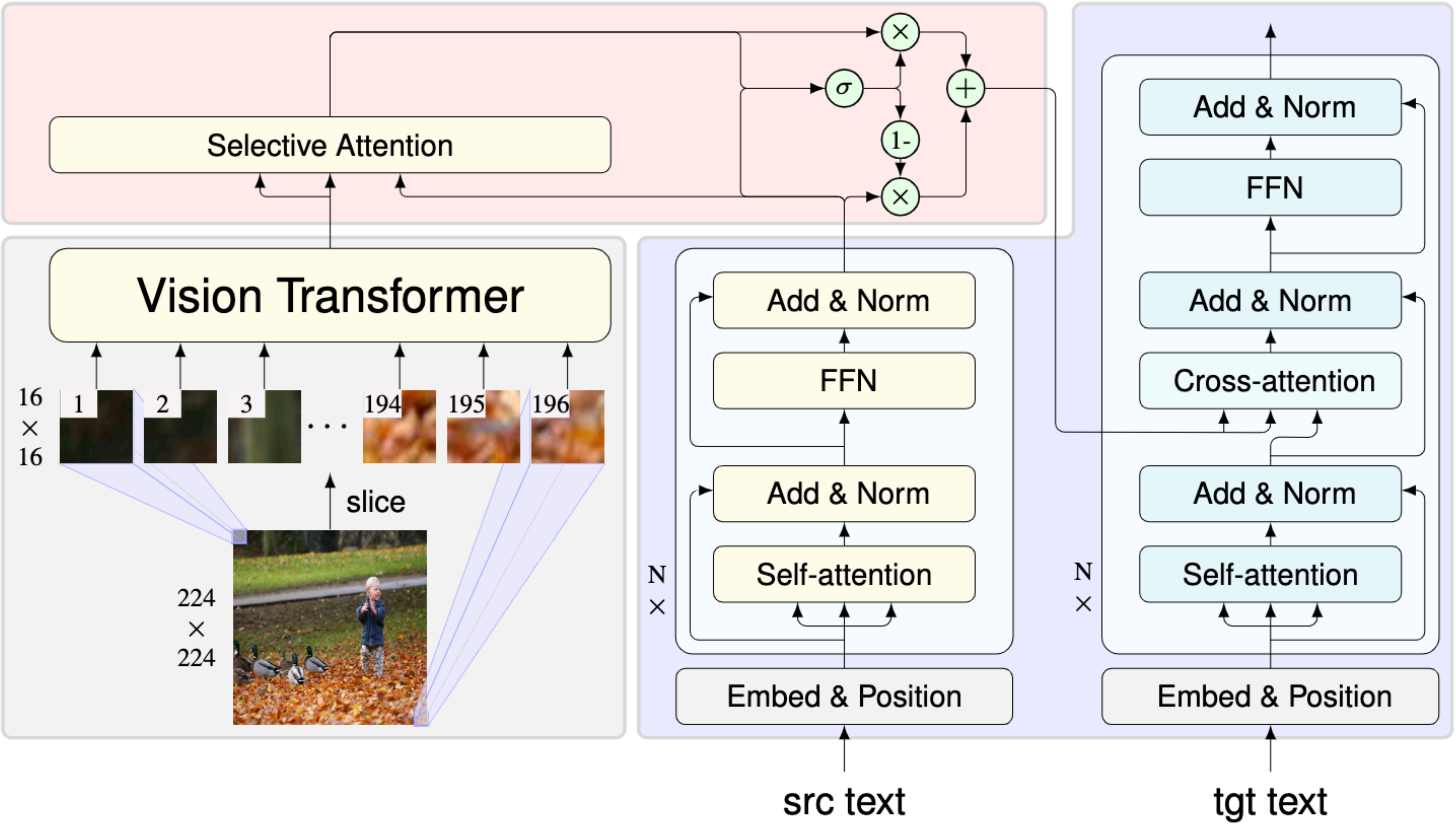
(b) The red framed object in the image of X_1 may help the translation of "defibrillator" in X_1.

(c) The red framed object (means "jeans") in the image of X_1 may help the translation of the pronoun "those" in X_1.

[Liang et al., ACL 2022]

Experiments

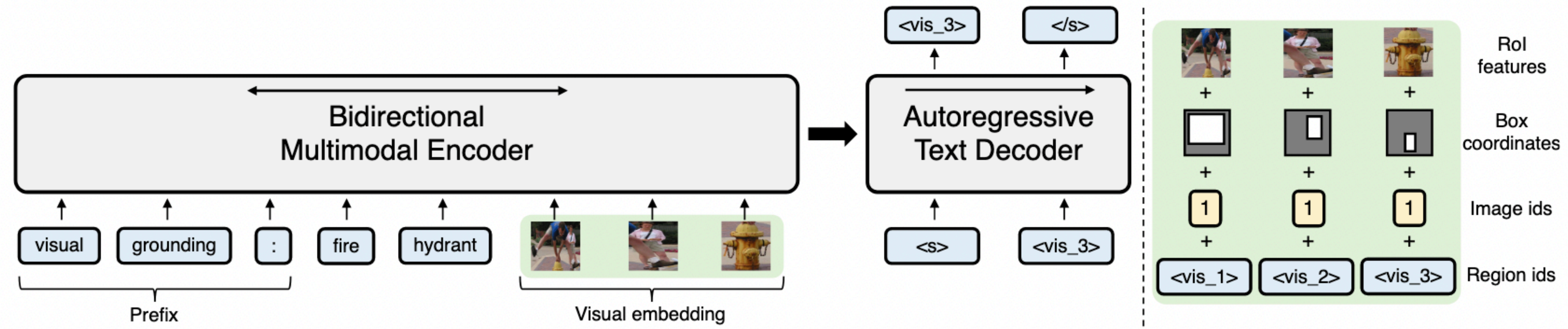
- Baseline models
  - Selective Attention



[Li et al., ACL 2022]

Experiments

- Baseline models
  - VL-Bart, VL-T5



(a) Our vision-and-language framework

(b) Visual embedding

[Cho et al., ICML 2021]

Experiment

- MMT models trained on 3AM outperform their text-only counterparts by a large margin
- While MMT model trained on other datasets perform close to or even worse than text-only models
- This result confirms our hypothesis that models trained on our dataset can better leverage visual information

Method	Multi30K (test)				MSCTD (test)				3AM (test)			
Multi30K (train)												
	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓
Trans	42.86	74.32	65.44	47.86	2.87	34.99	15.75	108.20	10.86	49.10	29.40	88.85
SelAttn	42.00	74.17	64.63	49.82	2.86	36.00	16.61	107.84	11.67	50.05	30.86	87.20
Bart	56.93	83.24	79.61	32.47	7.40	46.71	29.35	101.93	22.29	59.19	45.43	73.87
VL-Bart	56.70	82.93	77.89	32.00	8.12	46.29	27.22	86.40	23.20	60.20	45.75	70.95
T5	**60.59**	**85.69**	**82.85**	**27.61**	10.24	52.53	**38.78**	85.30	25.03	62.99	50.72	67.08
VL-T5	59.61	85.25	82.12	27.95	**11.10**	**52.96**	38.71	**77.71**	**25.34**	**63.25**	**50.89**	**66.35**
MSCTD (train)												
	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓
Trans	9.89	50.43	30.75	80.68	22.97	62.93	46.43	65.40	4.51	40.69	20.10	88.37
SelAttn	6.91	46.75	25.04	85.31	20.87	62.08	44.27	65.58	5.30	41.87	21.05	108.70
Bart	22.77	65.66	51.50	59.95	**32.68**	69.82	**56.68**	**52.60**	14.93	56.34	38.72	74.58
VL-Bart	18.10	60.34	44.81	65.29	30.81	68.96	55.63	54.03	13.61	54.24	36.46	77.53
T5	**29.17**	**72.04**	**59.82**	**51.32**	29.39	70.43	54.22	54.46	**18.49**	**59.68**	**44.13**	**70.26**
VL-T5	28.43	71.09	58.85	52.82	29.49	**70.63**	54.48	54.52	17.87	59.27	43.44	70.55
3AM (train)												
	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓	B ↑	BS ↑	M ↑	T ↓
Trans	25.95	64.51	49.88	63.92	3.53	39.23	19.02	102.93	11.33	49.51	31.34	89.68
SelAttn	27.81	67.06	52.13	59.77	4.25	40.34	19.84	100.19	13.33	51.54	33.47	87.05
Bart	48.13	80.16	76.07	39.19	13.45	54.61	38.30	84.94	31.47	65.87	55.62	63.65
VL-Bart	50.13	80.74	76.38	36.87	16.13	56.45	39.15	74.17	33.27	66.56	55.84	61.28
T5	50.16	81.84	79.18	35.92	15.56	59.18	48.04	77.79	33.09	68.15	57.26	60.09
VL-T5	**52.04**	**82.60**	**79.76**	**34.37**	**17.12**	**59.94**	**48.54**	**73.01**	**34.24**	**68.39**	**59.12**	**58.88**

Performance of MMT models on 3AM and other MMT datasets in terms of BLEU (B), BERT-Score (BS), METEOR (M), and TER (T)

Analysis

- Visual Awareness
  - The overall image awareness of a model $\mathscr{M}$ on dataset $\mathscr{D}$ can be defined as:

$$\Delta\text{-Awareness} = \frac{1}{|\mathscr{D}|}\sum_{i}^{|\mathscr{D}|} a_{\mathscr{M}}\left(x_i, y_i, v_i, \bar{v}_i\right)$$

where x is the source sentence, y is the target sentence, v is the congruent image, $\bar{v}$ is the incongruent image, and $a_{\mathscr{M}}(\cdot)$ is the image awareness of model M on a single instance:

$$a_{\mathscr{M}}\left(x_i, y_i, v_i, \bar{v}_i\right) = \varepsilon\left(x_i, y_i, v_i\right) - \varepsilon\left(x_i, y_i, \bar{v}_i\right)$$

Dataset	C	I	Δ-Awareness
Multi30K	74.16	74.11 ± 0.04	0.05 ± 0.04
MSCTD	62.08	62.08 ± 0.00	0.00 ± 0.00
3AM	51.54	50.17 ± 0.09	1.36 ± 0.09

BERT-Scores under Congruent (C) and Incongruent (I) settings, and the image awareness results.

Analysis

- Case Study
  - Tape→ S_1: 录像, S_2: 录音
  - MMT model (VL-T5) can correctly translate the ambiguous word

Source: A group of people on skis are being **taped**.

Target: 一群滑雪板上的人正在被录像。(record video)

T5: 一群踩着滑雪板的人正在被录音。(record audio)

VL-T5: 一群滑雪板上的人正在被录制视频。(record video)

Conclusion

- Contributions
  - Propose 3AM, a MMT dataset that is more challenging and contains a richer set of concepts
  - Evaluate SOTA MMT models and show that models that can leverage visual information outperform text-only models
- Limitations
  - The challenge of data scarcity remains: the size of 3AM is only 26K

Thank you