

---

# How2: A Large-scale Dataset for Multimodal Language Understanding

---

**Ramon Sanabria**  
Carnegie Mellon University  
`ramons@cs.cmu.edu`

**Ozan Caglayan**  
Le Mans University

**Shruti Palaskar**  
Carnegie Mellon University

**Desmond Elliott**  
University of Copenhagen

**Loïc Barrault**  
Le Mans University

**Lucia Specia**  
University of Sheffield

**Florian Metze**  
Carnegie Mellon University

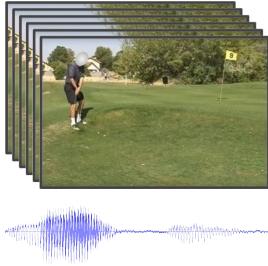
## Abstract

Human information processing is inherently multimodal, and language is best understood in a situated context. In order to achieve human-like language processing capabilities, machines should be able to jointly process multimodal data, and not just text, images, or speech in isolation. Nevertheless, there are very few multimodal datasets to support such research, resulting in a limited interaction among different research communities. In this paper, we introduce How2, a large-scale dataset of instructional videos covering a wide variety of topics across 80,000 clips (about 2,000 hours), with word-level time alignments to the ground-truth English subtitles. In addition to being multimodal, How2 is multilingual: we crowdsourced Portuguese translations of the subtitles. We present results for monomodal and multimodal baselines on several language processing tasks with interesting insights on the utility of different modalities. We hope that by making the How2 dataset and baselines available we will encourage collaboration across language, speech and vision communities.

## 1 Introduction

Multimodal sensory integration is an important aspect of human concept representation, language processing and reasoning [1]. From a computational perspective, major breakthroughs in natural language processing (NLP), computer vision (CV), and automatic speech recognition (ASR) have resulted in improvements in a wide range of multimodal tasks, including visual question-answering [2], multimodal machine translation [3], visual dialogue [4], and grounded ASR [5]. Despite these advances, state-of-the-art computational models are nowhere near integrating multiple modalities as effectively as humans. This can be partially attributed to a lack of resources that are *pervasively* multimodal: existing datasets are typically focused on a single task, *e.g.* images and text for image captioning [6], images and text for visual-question answering [2], or speech and text for ASR [7]. These datasets play a crucial role in the development of their fields, but their single-task nature limits the collective ability to develop general purpose artificial intelligence.

We introduce How2, a dataset of instructional videos paired with spoken utterances, English subtitles and their crowdsourced Portuguese translations, as well as English video summaries. The pervasive multimodality of How2 makes it an ideal resource for developing new models for multimodal understanding. In comparison to other multimodal resources, How2 is a naturally occurring dataset:



**I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.**

*Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.*

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

Figure 1: `How2` contains a large variety of instructional videos with utterance-level English subtitles (in bold), aligned Portuguese translations (in italics), and video-level English summaries (in the box). Multimodality helps resolve ambiguities and improves understanding.

Table 1: Statistics of `How2` dataset.

		Videos	Hours	Clips/Sentences	Per Clip Statistics
300h	train	13,168	298.2	184,949	5.8 seconds & 20 words
	val	150	3.2	2,022	5.8 seconds & 20 words
	test	175	3.7	2,305	5.8 seconds & 20 words
	held	169	3.0	2,021	5.4 seconds & 20 words
2000h	train	73,993	1,766.6	-	-
	val	2,965	71.3	-	-
	test	2,156	51.7	-	-

neither the subtitles, nor the summaries have been crowdsourced. Furthermore, the visual content is inherently related to the spoken utterances. Figure 1 shows an example in which the presenter is explaining how to play a golf shot. If one only has access to the text, it is unclear whether the “green” in the subtitles refers to the colour green (“verde” in Portuguese), or the surface type (“green” in Portuguese). The textual context alone is not enough to disambiguate the meaning of the subtitles, and at the time of writing, both Google Translate and Microsoft Translator incorrectly translate “green” as “verde”. However, given additional visual context (green grass with a flag pole), or the audio context (outside with the sound of chipping a golf ball), our multimodal models can correctly interpret this utterance. See Appendix A.1 for more examples.

The value of additional modalities can also be demonstrated for ASR. Object and motion level visual cues can filter out systematic noise that co-occurs with activities. Scene information from an image can be used to learn a common auditory representational space for different environmental characteristics such as indoor vs. outdoor settings [8]. Entities in an image can also be used to adapt a language model towards a domain [9].

Together with the dataset, we also release code to perform baseline experiments on several tasks covering different subsets of `How2`. We find that action-level visual features improve automatic speech recognition, video summarization and speech-to-text translation. These results demonstrate the potential of the `How2` dataset for future multimodal research.

## 2 How2 Dataset

The `How2` dataset consists of 79,114 instructional videos (2,000 hours in total, with an average length of 90 seconds) with English subtitles. The corpus can be (re-)created using the scripts and metadata we made available at <https://github.com/srvk/how2-dataset>. The website also contains information on how to obtain pre-computed features for validation or saving computation, and how to reproduce the experimental results we present using `nmtpy` [10].

**Collection** We downloaded the videos from *YouTube*, along with various types of metadata, including ground-truth subtitles and summaries (referred to as “descriptions”) in English, written by the

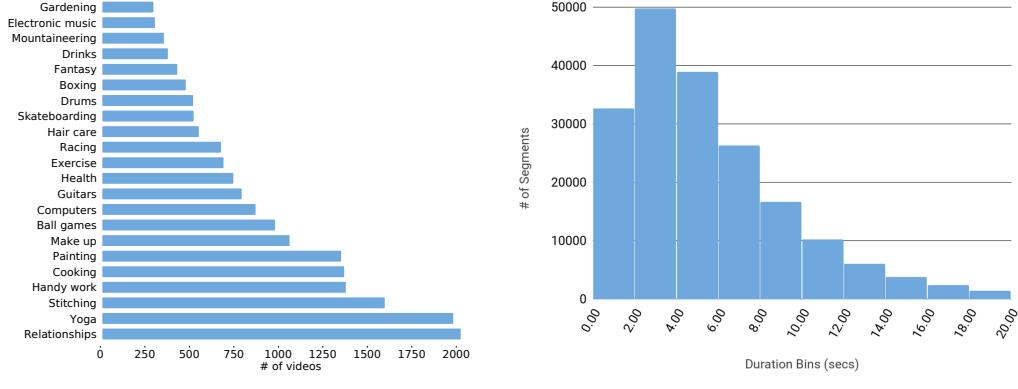


Figure 2: LDA topic distributions and segment duration for the  $300h$  subset. The  $2000h$  overall corpus exhibits very similar characteristics.

video creators. Videos were scraped from the YouTube platform using a keyword based spider as described in [11]. In order to produce a multilingual and multimodal dataset, the English subtitles were first re-segmented into full sentences, which were then aligned to the speech at the word level. The visual features were extracted from the video *clips* that correspond to these sentence-level alignments. The distribution of the duration of segments can be seen in Figure 2b. See Appendix A.2 for more details on the alignment process.

To generate translations, we used the *Figure Eight* crowdsourcing platform. After conducting a pilot experiment with a small set of languages, we chose Portuguese as target language because of the availability of workers and the quality and reliability of the annotations performed by them. In order to reduce the amount of time it would take to annotate the dataset, we posed translation as a post-editing task: in another pilot experiment, we instructed crowd workers to “choose the best translation” from English to Portuguese among candidate translations provided by three state-of-the-art online neural machine translation systems. We then selected the system that was preferred most often, and had crowdworkers post-edit the candidate translations. This process is still ongoing.

During annotation on *Figure Eight*, the worker population was restricted to those living in Portugal or Brazil. Workers were paid US\$ 0.05 to watch a short video and post-edit the automatically translated Portuguese segment into correct Portuguese. Workers thus performed the annotation (and the pilots) in a multimodal setting. To ascertain worker reliability, a content word of each 5 sentences of the candidate translations was replaced by another random content word that was not part of the translation. If the word inserted was still present in the final translation, the annotations from that worker were discarded and the worker was banned from further contributing to the project.

At the time of writing, we had completed the collection of Portuguese translations for a  $300h$  subset of the entire dataset from 200 workers (each was limited to 5,000 segments to post-edit but none of them reached this limit). We discarded and re-annotated 18% of the  $300h$ . The total cost for data collection thus far was US\$ 8,771.

In a **verification experiment**, we determined found that training an English-Portuguese neural MT system on  $300h$  of machine generated training data degrades performance by about 1 BLEU point, when compared to a system that has been trained on the post-edited translations, when evaluated against expert-validated post-edited translations, showing that the approach is justified.

**Topic distribution** We clustered the English subtitles using Latent Dirichlet Allocation (LDA) [12]. Upon analyzing the clusters with top words in each topic, inter-topic and intra-topic distances, we found that a good representation for the  $300h$  subset consists of 22 topics. We hand-labeled these topics based on top words in each cluster, as shown in Figure 2a.

**Splits** Table 1 presents summary statistics of the  $2000h$  set and  $300h$  subset: the *val* and *test* sets can be used for early-stopping, model selection and evaluation; the *held* set is reserved for future evaluations or challenges. The total set (*i.e.*  $2000h$ ) contains around 22.5M words. The tokenized

Table 2: Results of the automatic speech recognition, machine translation, speech-to-text translation, and summarization experiments on *test* set. The arrows indicate direction of improvement.

	ASR (% WER ↓)	MT (BLEU ↑)	STT (BLEU ↑)	SUM (ROUGE-L ↑)
Baseline	19.4	54.4	36.0	53.9
Multimodal	18.0	54.4	37.2	54.9

training set of *300h* subset contains around 3.8M (43K unique) and 3.6M (60K unique) words for English and Portuguese respectively.

### 3 Experiments

To demonstrate and explore the potential of the *How2* dataset, we propose several tasks and developed systems for them using a sequence-to-sequence (S2S) approach. Table 2 summarizes the baseline results on the *300h* training set for all tasks; only the summarization task uses the entire *2000h* set. More details can be found in the Appendix A.4.

1. **Automatic speech recognition.** We use an S2S model with a deep bi-directional LSTM encoder [13]. For multimodal ASR, we apply visual adaptive training [5, 9] where we re-train an ASR model by adding a linear adaptation layer which learns a video-specific bias to additively shift the speech features. All parameters of the network are jointly trained in this re-training step. The adaptation layer increases the model size by less than 1%.
2. **Machine translation.** We train an S2S MT model for English→Portuguese using a bidirectional GRU [14]. For multimodal MT (MMT), we apply the same adaptive approach as we did for ASR but the inputs to be shifted are now word embeddings instead of speech features. The adaptation layer increases the model size by 8%.
3. **Speech-to-text translation.** We directly translate from English speech to Portuguese using the same ASR architecture but with a different target vocabulary, which is similar to previous approaches [15, 16]. For multimodal STT, we apply the same adaptive scheme as in ASR.
4. **Summarization.** The baseline is again the same S2S MT model. For multimodal summarization, we follow the hierarchical attention approach [17, 18] to combine textual and visual modalities by using a sequence of action-level features instead of an average-pooled one as in the other experiments. This latter increases the model size by 14%.

### 4 Related work

Lying at the intersection of NLP and CV [39], image captioning (IC) is the multimodal task with the largest number of datasets available. The most widely used datasets in this field are the ones with human crowdsourced descriptions, such as Flickr8K [19], Flickr30K [40], MSCOCO [6] and their extensions to other languages. A closely related task to IC is multimodal machine translation. So far, MMT has been addressed using captioning datasets extended with translations in different languages such as IAPR-TC12 [41] and Multi30K which is an extension of Flickr30K into German [23], French [25], and Czech [26]. One major pitfall of these datasets is that they lack syntactic and semantic diversity.

A similar task to IC is that of automatically describing videos (VD). The most popular datasets for VD are MSR-VTT [29], LSMDc [28], and Microsoft Research Video Description (MSVD) corpus [27] which is the only multilingual resource of this type providing 122K crowdsourced descriptions. However, two-thirds of the descriptions are in English and the ones in other languages are not parallel. *How2* offers a larger amount of data, all of which is in two languages.

Lipreading can be seen as a form of multimodal ASR, albeit not fusing information at the semantic level. Popular and large-scale datasets include Grid [30] and Lip Reading in The Wild [32]. *How2* is the first dataset that allows to perform multimodal ASR, using images as acoustic and linguistic context to improve accuracy. *How2* is also a valuable resource for speech-to-text translation, which

Table 3: Comparison with previous datasets: (IC) and (VD) stand for image and video captioning. Language names are encoded in ISO-639-1.

Task	Dataset	Languages	Audio	Visual	Size
IC	Flickr8K [19]	EN, TR [20], ZH [21]	✗	✓ (I)	8K
IC	Flickr30K [22]	150K EN, DE [23]	✗	✓ (I)	30K
IC	MSCOCO [6]	414K EN	✗	✓ (I)	82K
IC		820K JA [24]	✗	✓ (I)	164K
MMT	Multi30K [23]	EN, DE, FR [25], CZ [26]	✗	✓ (I)	30K
VD	MSVD [27]	122K total in many	✓	✓ (V)	5.3 hours
VD	LSMDC [28]	EN	✓	✓ (V)	150 hours
VD	MSRVTT [29]	EN	✓	✓ (V)	41 hours
AV-ASR	Grid [30]	EN	✓	✓ (V)	50 hours
AV-ASR	ViaVoice [31]	EN	✓	✓ (V)	34.9 hours
AV-ASR	LRW [32]	EN	✓	✓ (V)	800 hours
STT	Fisher [33]	EN, ES	✓	✗	150 hours
STT	Audiobooks [34]	EN, FR	✓	✗	236 hours
SUM	CNN/DMC [35, 36]	EN	✗	✗	286,817 pairs
SUM	DUC [37]	EN	✗	✗	500 pairs
SUM	[38]	EN CZ	✓	✓ (V)	492,402 pairs

is otherwise often performed using Fisher-Callhome [33] and Audiobooks [34]. `How2` is the only corpus for multimodal STT currently available.

Multimodal neural abstractive summarization is an emerging field for which there are no well established benchmarking datasets yet. Li et al. [38] collected a multimodal corpus of news articles containing 500 videos of English news articles paired with human annotated summaries. UzZaman et al. [42] collected a corpus of images, structured text and simplified compressed text for summarization of complex sentences. More traditional text-based summarization is commonly based on CNN/Daily Mail [35, 36], Gigaword [43] and the Document Understanding Conference challenge data [37]. An older version and non-released version of `How2` was used for experiments on learning action examples in videos in [11].

## 5 Conclusions

We have introduced `How2`, a multimodal collection of instructional videos with English subtitles and crowdsourced Portuguese translations. We have also presented sequence-to-sequence baselines for machine translation, automatic speech recognition, spoken language translation, and multimodal summarization. By making available data and code for several multimodal natural language tasks, we hope to stimulate more research on these and similar challenges to obtain a deeper understanding of multimodality in language processing.

## Acknowledgements

This work was mostly conducted at the 2018 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies,<sup>1</sup> hosted and sponsored by Johns Hopkins University. Lucia Specia received funding from the MultiMT (H2020 ERC Starting Grant No. 678017) and MMVC (Newton Fund Institutional Links Grant, ID 352343575) projects. Loïc Barrault and Ozan Caglayan received funding from the CHISTERA M2CR (No. ANR-15-CHR2-0006-01).

<sup>1</sup><https://www.clsp.jhu.edu/workshops/18-workshop/>

## References

- [1] Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 2003.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2015.
- [3] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description (WMT). In *Proceedings of the First Conference on Machine Translation (WMT)*. ACL, 2016.
- [4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.
- [5] Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multimodal speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *Computing Research Repository (CoRR)*, 2015.
- [7] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1992.
- [8] Yajie Miao and Florian Metze. Open-domain audio-visual speech recognition: A deep learning approach. In *Proceedings of Interspeech*. ISCA, 2016.
- [9] Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [10] Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. NMTPY: A flexible toolkit for advanced neural machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 2017.
- [11] Shou-I Yu, Lu Jiang, and Alexander Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *Proceedings of the International Multimedia Conference (ACMM)*. ACM, 2014.
- [12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 2003.
- [13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*.
- [14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014.
- [15] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech*. ISCA, 2017.
- [16] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. End-to-end automatic speech translation of audiobooks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [17] Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2017.

- [18] Jindřich Libovický, Shruti Palaskar, Spandana Gella, and Florian Metze. Multimodal abstractive summarization of open-domain videos. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NIPS, 2018.
- [19] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research (JAIR)*, 2013.
- [20] Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. Tasviret: Görüntülerden otomatik türkçe açıklama oluşturma İçin bir denektaçılı veri kümesi (TasvirEt: A benchmark dataset for automatic Turkish description generation from images). In *Proceesdings of the Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2016)*. IEEE, 2016.
- [21] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. Adding Chinese captions to images. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*. ACM, 2016.
- [22] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 2017.
- [23] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the Workshop on Vision and Language*. ACL, 2016.
- [24] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2017.
- [25] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation (WMT)*. ACL, 2018.
- [26] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella. Frank. Findings of the shared task on multimodal machine translation (WMT). In *Proceedings of Conference on Machine Translation (WMT)*. ACL, 2018.
- [27] David L. Chen and William B. Dolan. Building a persistent workforce on Mechanical Turk for multilingual data collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP)*. AAAI, 2011.
- [28] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017.
- [29] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: a large video description dataset for bridging video and language. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.
- [30] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 2006.
- [31] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.
- [32] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2016.
- [33] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved Speech-to-Text translation with the Fisher and Callhome Spanish-English speech translation corpus. In *Proceedings International Workshop on Spoken Language Translation (IWSLT)*. ACL, 2013.

- [34] Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In Proceedings of the International Conference on Language Resources and Evaluation (LREC). ELRA, 2018.
- [35] Karl Moritz Hermann, Tomas Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS). NIPS, 2015.
- [36] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In Proceedings of the Conference on Computational Natural Language Learning (CoNLL). ACL, 2016.
- [37] Paul Over, Hoa Dang, and Donna Harman. Duc in context. Information Processing & Management, 2007.
- [38] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2017.
- [39] Raffaella Bernardi, Ruken Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research (JAIR), 2016.
- [40] Peter Young, Alice Lai, Micha Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL), 2014.
- [41] Michael Grubinger, Paul D. Clough, Henning Muller, and Thomas Desealiers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In Proceedings of International Conference on Language Resources and Evaluation (LREC). ELRA, 2006.
- [42] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. Multimodal summarization of complex sentences. In Proceedings International Conference on Intelligent User Interfaces (IUI). ACM, 2011.
- [43] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. ACL, 2012.
- [44] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi speech recognition toolkit. In Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2011.
- [45] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d CNNs retrace the history of 2d cnns and imagenet? In Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2018.
- [46] Taku Kudo. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL.
- [47] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
- [48] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch-Mayne, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a toolkit for neural machine translation. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). Software Demonstrations. ACL, 2017.

- [49] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Computing Research Repository (CoRR)*, 2014.
- [50] Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, 2017.
- [51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.
- [52] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 2014.
- [53] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the International Conference on International Conference on Machine Learning (ICML)*. IMIS, 2013.
- [54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*. ACL, 2002.
- [55] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2004.



**Actually I use this beautiful moroccan oil and it's really wonderful on the hair.**

*Na verdade eu uso este belo óleo marroquino e é realmente maravilhoso no cabelo.*

Relaxed African-American hair should be moisturized and washed to maintain good and healthy hair. Care for African-American, biracial or ethnic hair with tips from a professional hairstylist in this free video on hair care.



**Like said you can cook this pretty quickly with your family.**

*Como disse você pode cozinhar isso muito rapidamente com sua família.*

Learn how to cook and serve picadillo con arroz with expert cooking tips in this free classic Cuban recipe video clip.



**When your wide receivers get into their stance they want to have one foot forward, they want to have good position and be ready to fire off the line.**

*Quando seus receptores largos entram em sua posição eles devem ter um pé para a frente, e eles devem ter uma boa posição e estar pronto para disparar a linha.*

Learn some great tips on how to line up as a receiver in this free video clip on how to play football.

Figure 3: Examples of the `How2` dataset that reflect its sample variability and modality correlation. The image is a randomly selected frame from a concrete segment of the video, the text in **bold** is the utterance pronounced during this segment, the Portuguese text in *italic* are the translations corresponding to the utterance and finally, the summary of the whole video is placed inside the rectangle.

## A Appendix

### A.1 How2 Examples

In Figure 3, we list three typical instances from the `How2` dataset. In these examples, we can see the correspondence between the content of the video frame, the summary, and the utterance. This multimodal correspondence is what systems can exploit by using the `How2` dataset. In the first example, a hairdresser is explaining how to use a specific hair product. In that case, the visual elements (*i.e.*, hair product, hairdresser) and the scene, a hairdressing salon, are a rich source of context. In the second example, a woman is cooking in a kitchen with many cooking devices. In the third example, we can infer the acoustic environment (*i.e.*, outdoors) by the scene.

### A.2 Modality Alignment and Data Checks

To combine all modalities (*i.e.*, speech, video, transcriptions, and translations) successfully, we need to establish and validate their correspondence in time. While the audio transcription is generated from subtitles, these do not always correspond to the actual audio. We thus decided to generate token-level (*e.g.* word-level) time stamps that link text, audio, and video modality. From these, utterance-level start and end times were also calculated.

To align text and audio, we perform a Viterbi alignment between the transcriptions, which were provided by the users who uploaded the videos, and the audio track of `How2`, using Kaldi’s [44] Wall Street Journal (WSJ) GMM/HMM acoustic model. This alignment process estimates the start and end times of each sentence in the audio track. Finally, by using the estimated alignments, we can segment the audio and video track according to the utterances.

To make sure the data will be suitable for the proposed use, we validated two properties: **First**, we verified that the word alignment is indeed accurate by manually inspecting randomly chosen utterances. The 300 h sub-set was selected to give good alignment scores, and the WSJ model seemed to perform best in that respect: “good” (according to the score) utterances were indeed accurately aligned, when compared to alignments generated with other models, including those developed for speech synthesis. **Second**, the “transcription” data has been generated from video subtitles, which were not meant to be verbatim and highly accurate “transliterations” of the spoken content. Rather, the “transcription” text is a somewhat canonical form of the spoken word, which is fine for our proposed uses, although it may lead to slightly higher overall word error rates for the speech-to-text tasks.

### A.3 Feature Extraction and Processing

**Speech Features** We used Kaldi [44] to extract 40-dimensional filter bank features from *16kHz* raw speech using a time window of *25ms* with *10ms* frame shift and concatenated 3-dimensional pitch features to obtain the final 43-dimensional speech features. A *per-video* Cepstral Mean and Variance Normalization (CMVN) is further applied to account for speaker variability.

**Visual Features** A 2048-dimensional feature vector per 16 frames is extracted from the videos using a CNN trained to recognize 400 different actions [45]. It should be noted that this results in a sequence of feature vectors per video rather than a single/global one. In order to obtain the latter, we average pooled the extracted features into a single 2048-dimensional feature vector which will represent all sentences segmented out of a single video.

**Text Features** All texts are normalized, lowercased and filtered from punctuation. A SentencePiece model [46] is learned separately for English and Portuguese resulting into vocabularies of 5K each except for summarization which uses word-level tokens.

### A.4 Architecture Details

**Automatic Speech Recognition** We use a 6-layer pyramidal encoder [47] (with interleaved *tanh* projection layers) where the middle two layers skip every other input resulting into a subsampling rate of 4. The decoder is a 2-layer conditional GRU (CGRU) decoder [48], a transitive decoder where the hidden state of the second GRU is fed back to the first GRU. The first GRU is initialized with the mean encoder hidden state transformed using a *tanh* layer. A feed-forward attention mechanism [49] is used inside the decoder and the input and output embeddings are tied [50].

**Multimodal Automatic Speech Recognition** For multimodal ASR, we apply video adaptive training [5, 9] with a learned linear transformation of the visual feature vector  $v$  that will act as a *visual bias* over a given speech feature  $x_t$  at time  $t$ . The shifted speech feature  $\hat{x}_t$  is thus computed as follows:

$$\hat{x}_t = x_t + (\mathbf{W}^T v + b)$$

To train this model, we first initialize the model parameters using a previously trained ASR and jointly optimize all parameters including  $\mathbf{W}$  and  $b$  above.

**Machine Translation** We train a sequence-to-sequence neural MT model with a 2-layer bidirectional GRU [14] encoder and a 2-layer conditional GRU decoder[48]. A dropout [51] with 0.3 drop probability is used after source embeddings, source encodings and before the final *softmax* operation.

**Multimodal Summarization** We follow the hierarchical attention approach [17] to combine textual and visual modalities. Unlike previous multimodal ASR, MT and STT architectures, the visual

features described in Section 2 are now used as-is, i.e. as a sequence of 2048-dimensional vectors rather than being average pooled into a single vector.

### A.5 Training Details

Unless otherwise specified, we use ADAM [52] as the optimizer with an initial learning rate of 0.0004. The gradients are clipped to have unit norm [53]. The training is early stopped if the task performance on validation set does not improve for 10 consecutive epochs. Task performance is assessed using Word Error Rate (WER) for speech recognition, BLEU [54] for translation tasks and ROUGE-L [55] for the summarization task. The learning rate is halved whenever the task performance does not improve for two consecutive epochs. All systems are trained three times with different random initializations. The hypotheses are decoded using beam search with a beam size of 10. We report the average results of the three runs. We use *nmtptorch* [10] to train the models.