A Limitations and Societal Impacts

Limitations One limitation of our model is its potential for data bias. KOSMOS-1 is trained on a web-scale multimodal corpus, which means that it is likely to be biased towards the data that it was trained on. This could lead to the model generating text that is biased towards certain demographics or viewpoints.

Another limitation of KOSMOS-1 is its relatively small size compared to other large language models. This means that the model may not be able to learn as complex relationships between different modalities. This could lead to the model making mistakes when it is asked to perform tasks that require a deep understanding of multiple modalities.

Finally, KOSMOS-1 only supports vision modality. This means that the model cannot process other modalities such as speech. This could limit the applications of the model.

Societal Impacts The broader impact of this paper is that it introduces a new type of large language model that can perceive general modalities, follow instructions, and perform in-context learning. This has the potential to be used for a variety of beneficial applications, such as new educational tools and interactive dialogue assistants in video games. However, there are also potential negative impacts of MLLMs. MLLMs could be used to create fake news articles or social media posts. MLLMs could be used to generate text that reveals private information from web-scale pre-training data.

B Hyperparameters

B.1 Training

We report the detailed model hyperparameter settings of KOSMOS-1 in Table 1 and training hyperparameters in Table 2.

Hyperparameters	
Number of layers	24
Hidden size	2,048
FFN inner hidden size	8,192
Attention heads	32
Dropout	0.1
Attention dropout	0.1
Activation function	GeLU [1]
Vocabulary size	64,007
Soft tokens V size	64
Max length	2,048
Relative position embedding	xPos [2]
Initialization	Magneto [3]

Table 1: Hyperparameters of causal language model of KOSMOS-1

B.2 Language-Only Instruction Tuning

The detailed instruction tuning hyperparameters are listed in Table 3.

C Datasets

C.1 Pretraining

The models are trained on web-scale multimodal corpora. The training datasets consist of text corpora, image-caption pairs, and interleaved data of images and texts.

Hyperparameters	
Training steps	300,000
Warmup steps	375
Batch size of text corpora	256
Max length of text corpora	2,048
Batch size of image-caption pairs	6,144
Batch size of interleaved data	128
Optimizer	Adam
Learning rate	2e-4
Learning Rate Decay	Linear
Adam ϵ	1e-6
Adam β	(0.9, 0.98)
Weight decay	0.01

Table 2: Training hyperparameters of KOSMOS-1

Hyperparameters	
Training steps	10,000
Warmup steps	375
Batch size of instruction data	256
Batch size of text corpora	32
Batch size of image-caption pairs	768
Batch size of interleaved data	16
Learning rate	2e-5

Table 3: Instruction tuning hyperparameters of KOSMOS-1

Text Corpora We train our model with The Pile [4] and Common Crawl (CC). The Pile is a massive English text dataset built for training large-scale language models, which is produced from a variety of data sources. We exclude data splits from GitHub, arXiv, Stack Exchange, and PubMed Central. We also include the Common Crawl snapshots (2020-50 and 2021-04) datasets, CC-Stories, and RealNews datasets [5, 6]. The entire datasets have been purged of duplicate and near-duplicate documents, as well as filtered to exclude downstream task data.

Table 4 provides a full overview of the language datasets that were used in the training of KOSMOS-1 model. These data sources can be divided into the following three categories:

- **Academic**: NIH Exporter
- **Internet**: Pile-CC, OpenWebText2, Wikipedia (English), CC-2020-50, CC-2021-04, Realnews
- **Prose**: BookCorpus2, Books3, Gutenberg [7], CC-Stories

Image-Caption Pairs The image-caption pairs are constructed from several datasets, including English LAION-2B [8], LAION-400M [9], COYO-700M [10], and Conceptual Captions [11, 12]. English LAION-2B, LAION-400M, and COYO-700M are collected from web pages of the Common Crawl web data by extracting image sources and the corresponding alt-text. Conceptual Captions are also from internet web pages.

LAION-2B contains about 2B English image-caption pairs, LAION-400M consists of 400M English image-caption pairs, and COYO-700M has 700M English image-caption pairs. Conceptual Captions contains 15M English image-caption pairs and consists of two datasets: CC3M and CC12M, which are also collected from internet webpages using a Flume pipeline. For Conceptual Captions, we discard pairs whose captions contain special tags such as “<PERSON>”.

Interleaved Image-Text Data We collect interleaved multimodal data from the Common Crawl snapshot, which is a publicly available archive of web pages. We use a filtering process to select about 71M web pages from the original 2B web pages in the snapshot. We then extract the text and

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

Table 4: Language datasets used to train the KOSMOS-1 model.

images from the HTML of each selected web page. For each document, we limit the number of images to five to reduce noise and redundancy. We also randomly discard half of the documents that only have one image to increase the diversity. By using this corpus, we enable KOSMOS-1 to handle interleaved text and image and improve its few-shot ability.

To ensure quality and relevance, we apply several filtering criteria. First, we discard any pages that are not written in English. Second, we discard any pages that do not have images interspersed in the text. Third, we discard any images that have a resolution lower than 64 by 64 pixels or that are single-colored. Fourth, we discard any text that is not meaningful or coherent, such as spam or gibberish. We use some heuristics to identify and remove gibberish text containing emoji symbols, hashtags, and URL links. After applying these filters, we end up with about 71 million documents for training.

C.2 Data Format

The training data is organized in the format as follows:

Datasets	Format Examples
Text	<s> KOSMOS-1 can perceive multimodal input, learn in context, and generate output. </s>
Image-Caption	<s> <image> Image Embedding </image> WALL-E giving potted plant to EVE. </s>
Multimodal	<s> <image> Image Embedding </image> This is WALL-E. <image> Image Embedding </image> This is EVE. </s>

Table 5: The examples of the data format to train the KOSMOS-1 model.

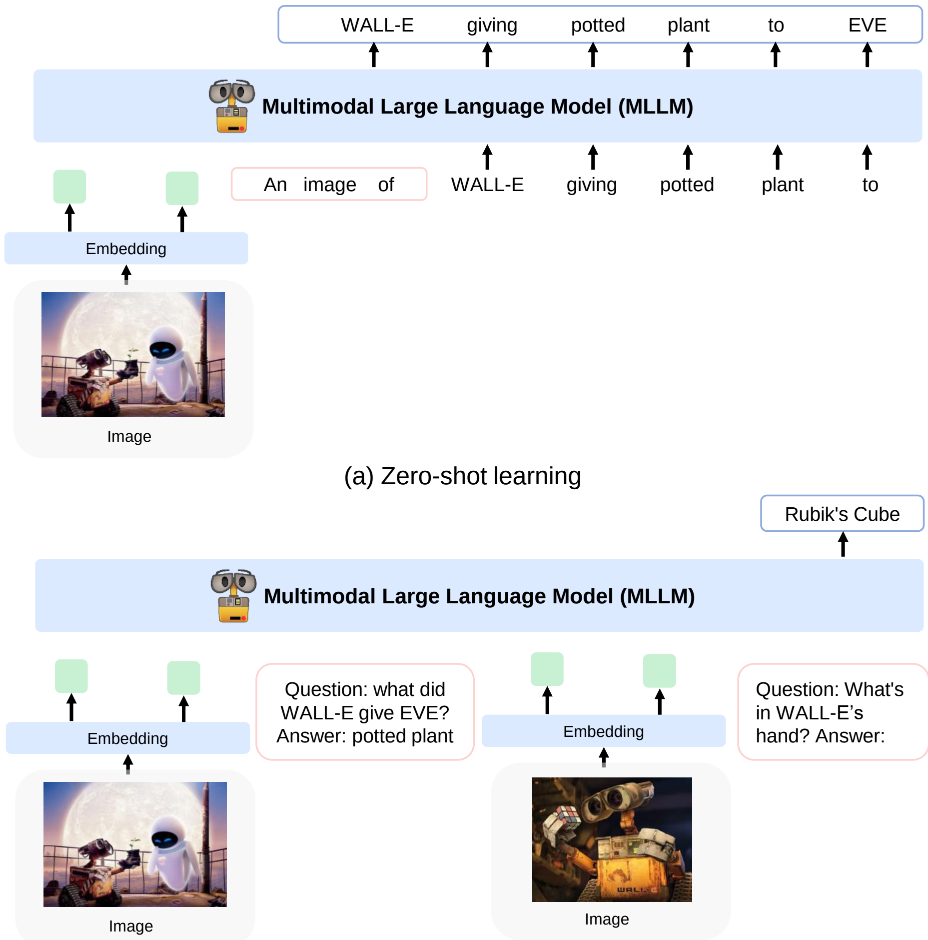
D Evaluation

D.1 Input Format Used for Perception-Language Tasks

Figure 1 shows how we conduct zero-shot and few-shot evaluations on perception-language tasks.

D.2 Perception-Language Tasks

We evaluate the caption generation on MS COCO Caption [13], and Flickr30k [14]. We use the test set of COCO *Karpathy split* [15], which re-partitions the train2014 and val2014 images [13] into 113,287, 5,000, and 5,000 for the training set, validation set, and test set, respectively. We conduct an evaluation on Flickr30k's *Karpathy split* test set. The image resolution is 224×224. We use beam search to generate the captions, and the beam size is 5. In the few-shot settings, we randomly



(a) Zero-shot learning

(b) Few-shot learning

Figure 1: We evaluate KOSMOS-1 on the perception-language tasks in zero- and few-shot settings. (a) Zero-shot learning, e.g., zero-shot image captioning with language prompts. (b) Few-shot learning, e.g., visual question answering with in-context learning.

sample demonstrations from the training set. We use COCOEvalCap[1] to compute CIDEr [16] and SPICE [17] scores as the evaluation metrics. We prompt KOSMOS-1 with *"An image of"* for zero-shot and few-shot caption generation experiments.

For visual question-answering tasks, we evaluate zero-shot and few-shot results on test-dev set of VQAv2 [18] and test-dev set of VizWiz [19], respectively. The resolution of images is 224×224. We use greedy search for the decoding. We follow the normalization rules of the VQAv2 evaluation code[2] when computing the VQA accuracy. We evaluate the performance of VQA in an open-ended setting that KOSMOS-1 generates answers and stops at the </s> ("end of sequence") token. The prompt is *"Question: {question} Answer: {answer}"* for visual question answering tasks.

[1] https://github.com/salaniz/pycocoevalcap

[2] https://github.com/GT-Vision-Lab/VQA

D.3 IQ Test Tasks

To evaluate the KOSMOS-1 on zero-shot nonverbal reasoning, we construct a dataset of the Raven IQ test. It consists of 50 examples collected from different websites[3][4][5][6]. Each example has three (i.e., 2×2 matrix), four, or eight (i.e., 3×3 matrix) given images. The goal is to predict the next one. Each instance has six candidate images with a unique correct completion. We measure accuracy scores to evaluate the models. The evaluation dataset is available at `https://aka.ms/kosmos-iq50`.

The matrix-style images are flattened and fed into the models one-by-one. To enable the model to better understand the desired task, we also use a textual instruction *"Here are three/four/eight images:"*, *"The following image is:"*, and *"Is it correct?"* for conditioning. We append each possible candidate to the context separately and compare the probability that the model outputs "Yes" in a close-ended setting. The candidate that yields the largest probability is regarded as the prediction.

D.4 OCR-Free Tasks

We evaluate OCR-free language understanding on the Rendered SST-2 [20] test set and HatefulMemes [21] validation set. We use accuracy as the metric for the Rendered SST-2 and report ROC AUC for the HatefulMemes dataset. We use the prompt *"Question: what is the sentiment of the opinion? Answer: {answer}"*, where the answer is either positive or negative for the Rendered SST-2. For the HatefulMemes task, the prompt is *"Question: does this picture contain real hate speech? Answer: {answer}"*, where the answer is either yes or no.

D.5 Web Page Tasks

We compare the performance on the Web-based Structural Reading Comprehension (WebSRC) dataset [22]. For comparisons, we train a language model (LLM) on the same text corpora with the same training setup as in KOSMOS-1. The LLM takes the text extracted from the web page as input. Its template of the prompt is *"Given the context below from web page, extract the answer from the given text like this: Qusestion: Who is the publisher of this book? Answer: Penguin Books Ltd. Context: {WebText} Q: {question} A: {answer} "*, where the *{WebText}* presents the text extracted from the web page. Besides using the same prompt, KOSMOS-1 prepends the image before the prompt. Two example images from WebSRC are shown in Appendix D.11. Following the original paper [22], we use exact match (EM) and F1 scores as our evaluation metrics.

D.6 Multimodal CoT Tasks

We evaluate the ability of multimodal chain-of-thought prompting on the Rendered SST-2. We use the prompt *"Introduce this picture in detail:"* to generate the content in the picture as the rationale. Then, we use the prompt *"{rationale} Question: what is the sentiment of the opinion? Answer: {answer}"* to predict the sentiment, where the answer is either positive or negative.

D.7 Zero-shot image classification Tasks

Given an input image, we concatenate the image with the prompt *"The photo of the"*. The input is then fed into the model to obtain the category name of the image. We evaluate the model on ImageNet [23], which contains 1.28M training images and 50k validation images in 1k object categories. The prediction is classified as correct if it is exactly the same as the ground-truth category name. The image resolution used for evaluation is 224×224. We use beam search to generate the category names and the beam size is 2.

D.8 Zero-Shot Image Classification with Descriptions

Following CUB [24], we construct a bird classification dataset that contains images and natural-language descriptions of categories. The dataset has three groups of binary image classification. Each

[3] `https://en.testometrika.com/intellectual/iq-test/`

[4] `https://en.testometrika.com/intellectual/iq-test-for-kids-7-to-16-year-old/`

[5] `https://iqpro.org/`

[6] `https://iqhaven.com/matrix-g`

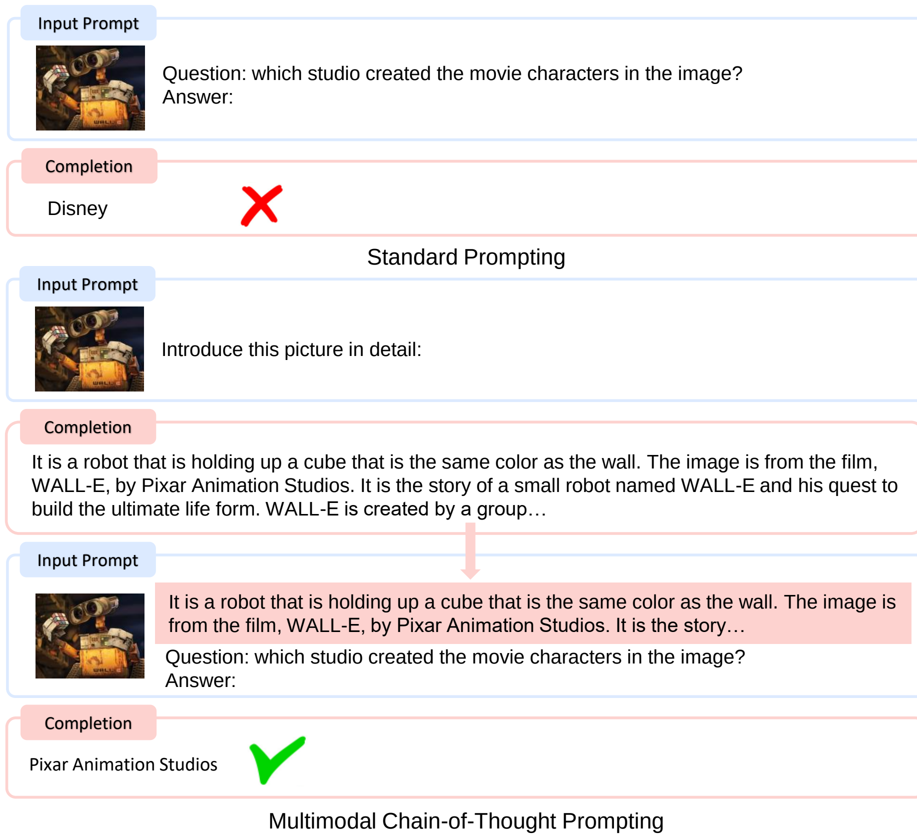


Figure 2: Multimodal Chain-of-Thought prompting enables KOSMOS-1 to generate a rationale first, then to tackle complex question-answering and reasoning tasks.

group contains two animal categories with similar appearances. Our goal is to classify images given the categories' descriptions. Table 6 presents the data samples. The first group is from [24], while the other two groups are collected from the website. Each category contains twenty images.

The evaluation procedure is illustrated in Figure **??**. For the zero-shot setting, we provide detailed descriptions of two specific categories and use the template *"Question:what is the name of {general category} in the picture? Answer:"* to prompt the model for the specific category name in an open-ended manner. To evaluate the effect of providing verbal descriptions in context, we also implement a zero-shot baseline without prompting descriptions. Instead, we provide the corresponding specific names in the prompt.

D.9 Cross-modal Transfer task

We compare KOSMOS-1 and the LLM baseline on three object commonsense reasoning datasets, RELATIVESIZE [25], MEMORYCOLOR [26] and COLORTERMS [27] datasets. Table 7 shows some examples of object size and color reasoning tasks. RELATIVESIZE contains 486 object pairs from 41 physical objects. The model is required to predict the size relation between two objects in a binary question-answering format with "Yes"/"No" answers. MEMORYCOLOR and COLORTERMS require the model to predict the color of objects from a set of 11 color labels in a multiple-choice format. We use only text as our input and do not include any images. We measure the accuracy of our model on these three datasets.

D.10 Language Tasks

We train a language model (LLM) baseline with the same text corpora and training setup. We evaluate KOSMOS-1 and the LLM baseline on eight language tasks, including cloze and completion tasks (i.e,

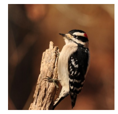
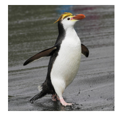
Category 1		Category 2	
three toed woodpecker		downy woodpecker	
	It has black and white stripes throughout the body and a yellow crown.		It has white spots on its black wings and some red on its crown.
Gentoo penguin		royal penguin	
	It has a black head and white patch above its eyes.		It has a white face and a yellow crown.
black throated sparrow		fox sparrow	
	It has white underparts and a distinctive black bib on the throat.		It has a reddish-brown plumage and a streaked breast.

Table 6: The detailed descriptions of different categories for in-context image classification.

Task	Example Prompt	Object / Pair	Answer
Object Size Reasoning	*Is {Item1} larger than {Item2}? {Answer}*	(*sofa*, *cat*)	*Yes*
Object Color Reasoning	*The color of {Object} is? {Answer}*	*the sky*	*blue*

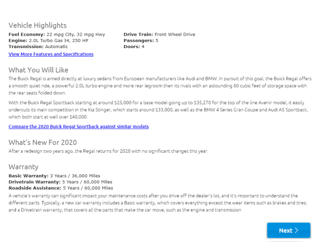
Table 7: Evaluation examples of object size and color reasoning.

StoryCloze, HellaSwag), Winograd-style tasks (i.e, Winograd, Winogrande), commonsense reasoning (i.e, PIQA), and three datasets BoolQ, CB, and COPA from the SuperGLUE benchmark [28]. The detailed descriptions of these datasets are provided in Appendix D.10. We conduct experiments under zero-shot and few-shot settings. We evaluate each test example by randomly sampling examples from the training set as demonstrations. We set the number of shots to 0, 1, and 4 in our experiments.

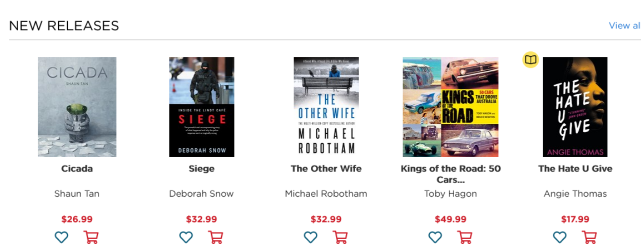
We conduct experiments on language tasks in four categories:

- Cloze and completion tasks: StoryCloze [29], HellaSwag [30]
- Winograd-style tasks: Winograd [31], Winogrande [32]
- Commonsense reasoning: PIQA [33]
- Three datasets from SuperGLUE benchmark [28]: BoolQ [34], CB [35], COPA [36]

D.11 WebSRC Task Examples



(a) Question is "*What is the type of this drive?*"



(b) Question is "*Who is the author of "Cicada"?*"

Figure 3: Examples form WebSRC [22].

D.12 Selected Examples

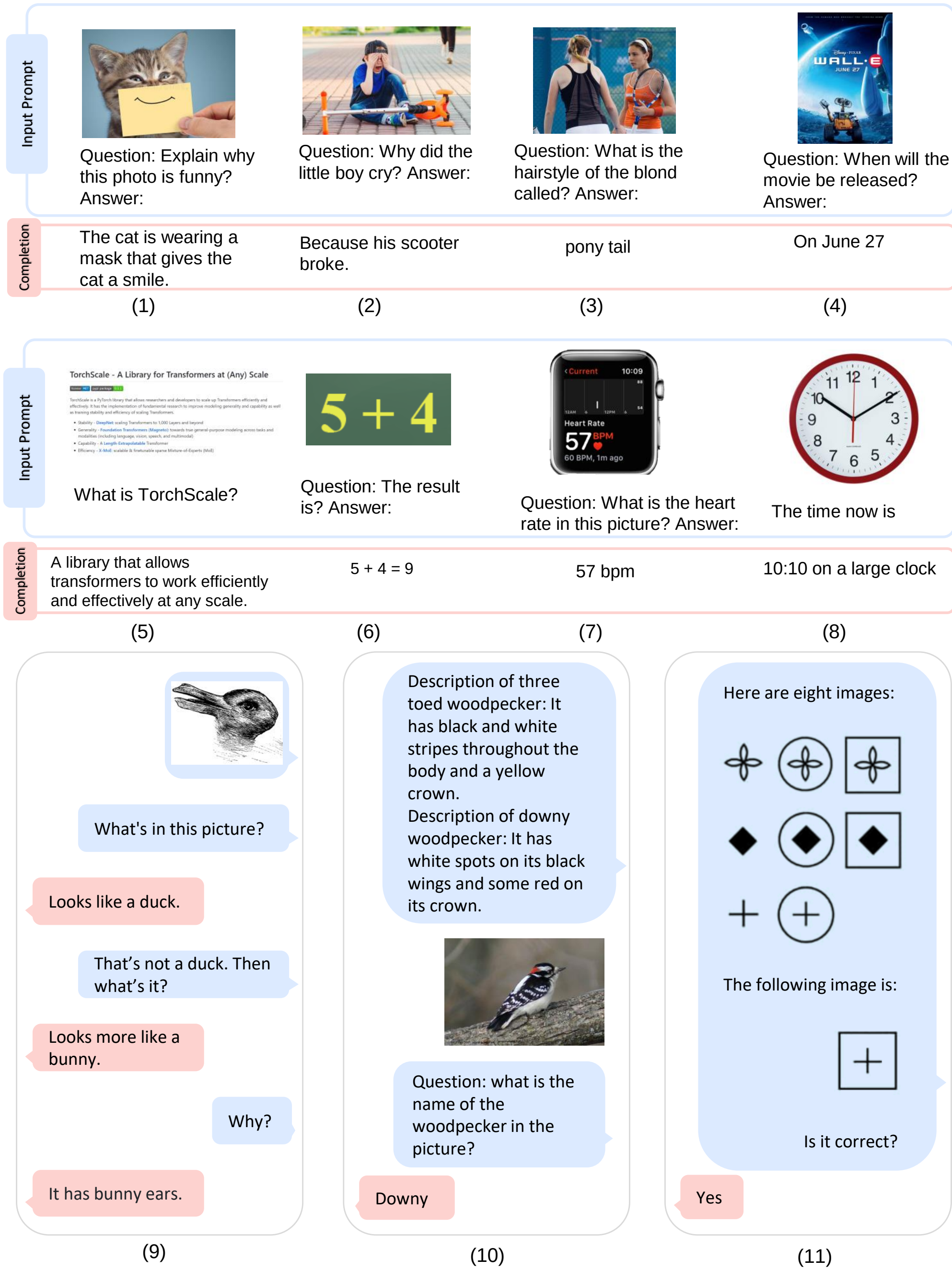


Figure 4: Selected examples generated from KOSMOS-1. Blue boxes are input prompt and pink boxes are KOSMOS-1 output. The examples include (1)-(2) visual explanation, (3)-(4) visual question answering, (5) web page question answering, (6) simple math equation, and (7)-(8) number recognition, and (9)-(11) visual dialogue.

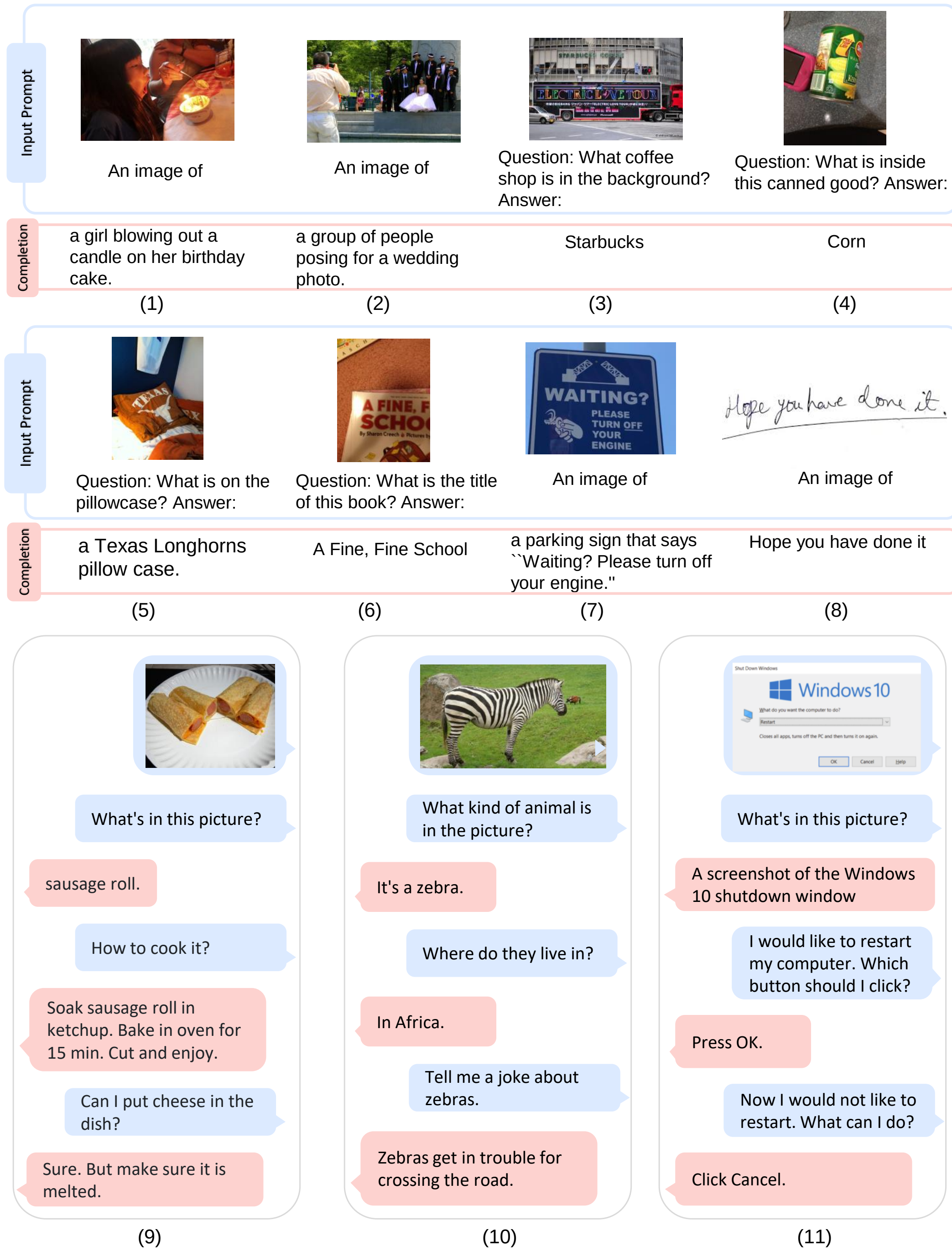


Figure 5: Selected examples generated from KOSMOS-1. Blue boxes are input prompt and pink boxes are KOSMOS-1 output. The examples include (1)-(2) image captioning, (3)-(6) visual question answering, (7)-(8) OCR, and (9)-(11) visual dialogue.

References

[1] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[2] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022.

[3] Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Foundation transformers. *CoRR*, abs/2210.06423, 2022.

[4] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[5] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[6] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model, 2022.

[7] Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *ICLR*, 2020.

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[9] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[10] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018.

[12] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[14] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[16] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[17] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334, 2017.

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, 2020.

[22] Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. WebSRC: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4173–4185, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.343. URL https://aclanthology.org/2021.emnlp-main.343.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[25] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. *ArXiv*, abs/1602.00753, 2016.

[26] Tobias Norlund, Lovisa Hagström, and Richard Johansson. Transferring knowledge from vision to language: How to achieve it and how to measure it? *ArXiv*, abs/2109.11321, 2021.

[27] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in technicolor. In *ACL*, 2012.

[28] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

[29] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, 2017.

[30] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[31] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning*, 2012.

[32] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI*, pages 8732–8740, 2020.

[33] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[34] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

[35] Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124, Jul. 2019.

[36] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium*, 2011.