INK: Intensive-Neural-Knowledge Aligned Image Text Retrieval

Jae Sung (James) Park[§‡*] **Qiuyuan Huang**[‡] **Yonatan Bisk**[†‡]
Subhojit Som[‡] **Ali Farhadi**[§] **Yejin Choi**[§]
Jianfeng Gao[‡]
[‡]Microsoft Research, Redmond [§] University of Washington [†] Carnegie Mellon University

Abstract

Knowledge-based vision language systems are increasingly ubiquitous in our everyday lives. However, despite the introduction of numerous benchmarks, the community has siloed models of different types of knowledge rather than building general knowledge-intensive models that encompass both commonsense and factoid knowledge. We introduce ***INK*** – ***I**ntensive **N**eural **K**nowledge* – a new task that involves extracting the necessary *knowledge* to accurately perform image and text retrieval[1]. In particular, ***INK*** leverages existing resources to require understanding of factoid, object-commonsense, or social-consciousness knowledge to successfully perform retrieval. Finally, we provide a set of competitive baseline models whose weak performance motivates the need to develop new knowledge understanding models and systems.

1 Introduction

Large-scale pre-trained neural models have dramatically improved the AI systems' performance on natural language and vision tasks. However, most state of the art models still cannot encode factoid and commonsense knowledge in a way that would make them suitable for many real-world knowledge-based tasks that require sophisticated reasoning and generating explainable answers (Bommasani et al., 2021; Marcus, 2020). Specifically, most existing models are pre-trained on raw text and/or image data and are evaluated on classification and generation tasks that do not require using external knowledge. Thus, deep learning has not yet produced deep knowledge (Marcus and Davis, 2019; Gao et al., 2020). To address these challenges, we argue that AI models should not only be trained using raw data of multiple modalities (i.e., vision and language) but also incorporate

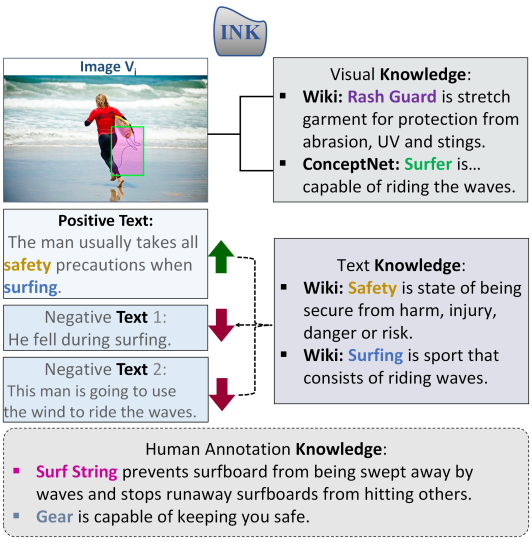


Figure 1: Example of INK task that uses knowledge to identify text relevant to the image from a set of text candidates. Our task involves leveraging visual and text knowledge retrieved from web and human-annotated knowledge.

factoid and commonsense knowledge (Choi, 2022) from various sources such as pre-compiled knowledge bases and wiki documents. To this end, we have developed a new knowledge-based text-image retrieval task, paired with a benchmark, INK to foster the development of deep learning models that can incorporate external knowledge, generate interpretable answers, and be easily generalize to new domains. Fig.1 shows an image-text retrieval example that requires external knowledge. In INK, the retrieval is grounded not only in visible objects, but also in commonsense/factoid knowledge and the knowledge implicitly conveyed by the image and text. To human users, it is the use of knowledge that demonstrates the role of human intelligence in performing various complex retrieval tasks such as exploratory search, investigation and learning (Gao et al., 2022). To the best of our knowledge, INK is the first knowledge-based text-image retrieval task with a public benchmark for evaluation.

[*]Work done when interning at Microsoft Research.
[1]We define *Neural Knowledge* as approach of injecting knowledge to neural networks.

2 Related Work

Vision and Language Task. There are several large datasets that address various other tasks across the language and vision space. Some of most popular ones are image captioning (Lin et al., 2014; Sharma et al., 2018a; Young et al., 2014), Visual Genome (Krishna et al., 2016), Visual Question Answering (VQA 1.0 and 2.0) (Antol et al., 2015), Question Answering on Image Scene Graphs for Compositional Question Answering (CQA) (Ren et al., 2015) and TextVQA (reasoning based on the text in images) (Singh et al., 2019) to name a few. However, these datasets do not emphasize on commonsense nature of factual knowledge mentioned above.

Knowledge-Intensive Tasks and Approaches. Numerous papers have injected knowledge into language pretraining models (Yu et al., 2020; Xu et al., 2021; Rosset et al., 2021; Zhou et al., 2020; He et al., 2020a; Xiong et al., 2019; He et al., 2020b; Agarwal et al., 2021) with an emphasis on knowledge-intensive NLP tasks (Petroni et al., 2021). More recently, knowledge-intensive Visual Question Answering tasks such as OKVQA (Marino et al., 2019), KB-VQA (Wang et al., 2017b), FVQA (Wang et al., 2017a), and WebQA (Chang et al., 2021) have been introduced. Sherlock (Hessel et al., 2022) contains image, bounding box regions, and grounded inference texts that require commonsense reasoning to perform the retrieval task. KAT (Gui et al., 2022) combined explicit Knowledge (e.g., wiki/web search) and implicit Knowledge (e.g., GPT-3) on OK-VQA dataset. KRISP (Marino et al., 2021) was proposed to retrieve knowledge stored in pretrained language models, MAVEx (Wu et al., 2021) make better use of the noisy retrieved knowledge. In this paper, we introduce a intensive knowledge retrieval task `INK`, the knowledge representations in our `INK` can be enhanced current deep learning models and improve their depth of knowledge, generalization, and interpretability.

3 Dataset Collection

We describe dataset collection process for `INK`, which includes knowledge-intenisve image-text pairs and human annotation of their relevant knowledge. Comparison with the prior dataset is shown in Table 1.

3.1 Human Evaluation and Knowledge Category

Collecting Knowledge-Intensive Image-Text Pairs. We use existing image and caption dataset as backbone to generate the INK benchmark of image-text retrieval-based comparison task. In general, the evaluation data should contain *knowledge-intensive* instances (Petroni et al., 2019), and require leveraging explicit knowledge to accurately perform the retrieval. Using the Mechanical Turk platform[2], we pre-judge the image-text pairs in parts of COCO (Lin et al., 2014), VizWiz (Gurari et al., 2018), and Sherlock (Hessel et al., 2022) datasets, and found 3%, 4%, 50% required external knowledge respectively. We reason that instances in COCO depend upon low-level semantics, and in VizWiz include captions for blind people rely on information from image itself without the needed of explicit knowledge to understand the alignment. After the survey testing, we use Sherlock images with regions and the relevant inference text pairs for human evaluation of knowledge category, and human annotation of golden knowledge. We acquire 2.5K knowledge-intensive image-text pairs from the setting.

Knowledge Category. We first categorized knowledge into three dimensions for each image-text pair that involve: 1) looking up factual information from encyclopedia (factoid knowledge), 2) reasoning about object properties and relations among different objects (object-commonsense knowledge), and 3) understanding social norms in everyday situations (social-consciousness knowledge). Then, given an image and its corresponding text, we asked MTurk workers to evaluate if one of the three knowledge category, or no knowledge is needed to understand the image-text pair. The distribution of knowledge category for *object-commonsense*, *social-consciousness*, and *factoid* which required are 63%, 30% and 7% respectively after M-Turk worker evaluated.

3.2 Human Annotation and Knowledge Resources

The knowledge resource pool of our INK is organized by 1) explicit retrieval-based web knowledge, and 2) implicit human annotated golden knowledge. The statistics are shown in Table 3 in Appendix C. We next describe their collection process.

[2] `https://mturk.com/`

Dataset	# Image-Text	Task	Knowledge Intensive	Aligned Knowledge Source
COCO (Lin et al., 2014)	5K	Retrieval	✗	✗
Flickr30K (Plummer et al., 2015)	1K	Retrieval	✗	✗
VQA (Goyal et al., 2017)	453K	QA	✗	✗
FVQA (Wang et al., 2017a)	700	QA	✓	DBPedia
KB-VQA (Wang et al., 2017b)	2.9K	QA	✓	ConceptNet, DBPedia, WebChild
OK-VQA (Marino et al., 2019)	5K	QA	✓	✗
Sherlock (Hessel et al., 2022)	22K	Retrieval	✓	✗
INK (ours)	2.5K	Retrieval	✓	Human

Table 1: Comparison between INK and previous image-text dataset on the *test* set.

Retrieval-based Web Knowledge. We first collect knowledge from the knowledge base in the web. In particular, we accumulated 180K entity with definition provided in Wikidata[3]. The entities are selected based on the categories of factual objects used in (Gui et al., 2022). Knowledge from ConceptNet (Speer and Havasi, 2013) is extracted to consider object common sense in our knowledge resource. We build the concept list from (Zhong et al., 2022), which includes common concepts found in Conceptual Caption dataset (Sharma et al., 2018b), and append verbs and nouns in the Sherlock inference text. The statistics of knowledge source and their examples are shown in Table 3 and 4 in Appendix.

Human Annotated Knowledge. While web knowledge can be acquired automatically, the pool may not cover all the relevant knowledge for the image-text instances. We thus obtained the gold knowledge for image and text pairs on the *evaluation data* by asking humans to identify three entities and annotate useful knowledge description information about these entities. More details of annotation process is described in Appendix D. We show examples of human annotated knowledge in Figure 5 in Appendix.

4 Task Setup

Knowledge Setting of INK. INK focuses on the task of knowledge aligned image *to* text retrieval. For each image V_i, we are given a consistent set of N text candidates that include the paired text T_i. We additionally provide unified knowledge source $\mathbf{K}$ across train/val/test split, which consists of external auto-retrieval web knowledge and human annotation knowledge described in Section 3.2. The top sample in Fig. 1 shows an example of this task setting with knowledge.

Task Definition of INK. In our task, we extract a set of relevant knowledge automatically $\tilde{\mathbf{K}}_\mathbf{i} = \{\tilde{K}_{i1}, ..., \tilde{K}_{ij}, ...\}$ from knowledge pool $\mathbf{K}$ to help understand the alignment b.w. V_i and T_j, which we define as *automatic knowledge extraction.* As our approach, we extract a set of knowledge separately for image and text, and their union is considered as $\tilde{K}_{ij}$ (Eq. 1). We leave it as future work that considers both modalities to extract the relevant knowledge.

$$\begin{aligned}\tilde{K}_{ij} &= \text{Know}(V_i, T_j, \mathbf{K}) \\ &\approx \text{Know}(V_i, \mathbf{K}) \cup \text{Know}(T_j, \mathbf{K}) \end{aligned} \quad (1)$$

Then, our INK task is based on image-text retrieval. With V_i and $\tilde{\mathbf{K}}_\mathbf{i}$ as context, we acquire the similarity score with all T_j candidates, and evaluate if paired text T_i gets a high score.

5 Experiments

5.1 Automatic Knowledge Extraction

Our automatic knowledge extraction system extracts knowledge separately from image and text. For image knowledge, we acquire the CLIP (Radford et al., 2021) similarity score between image and all the knowledge snippets in the resource, and retrieve the k-nearest neighbors using the FAISS library (Johnson et al., 2019). On the text side, we use entity-linker[4] to detect entities in the Wikidata and acquire their one sentence description. More details of our knowledge extraction are shown in Appendix B.

5.2 Model Details

Our first baseline is zeroshot and finetuned CLIP (Radford et al., 2021), a strong contrastive learning based model does not leverage knowledge

[3] https://www.wikidata.org

[4] https://github.com/egerber/spaCy-entity-linker

Approach	*Object* Rank↓	*Object* R@5(%)↑	*Social* Rank↓	*Social* R@5(%)↑	*Factoid* Rank↓	*Factoid* R@5(%)↑	**Overall** Rank↓	**Overall** R@5(%)↑
w/o knowledge								
Random	1265	0.1	1265	0.1	1265	0.1	1265	0.1
CLIP zero-shot	322	17.0	309	14.2	251	20.6	313	16.4
CLIP fine-tuned	42	48.5	69	34.1	54	48.9	51	44.1
w/ extracted knowledge								
KAT-Retrieval	135	26.3	200	14.8	191	27.6	159	22.8
INK-CLIP zero-shot	299	18.0	313	13.0	237	22.2	299	16.8
INK-CLIP fine-tuned	42	48.4	69	34.0	54	49.0	51	44.0
w/ gold human knowledge (oracle)								
INK-CLIP zero-shot (*oracle*)	232	22.5	264	17.1	209	22.2	240	20.8
INK-CLIP fine-tuned (*oracle*)	34	53.1	57	37.7	55	53.3	43	48.9

Table 2: Results of image *to* text retrieval. We report the average rank of ground truth text and Recall@5 (R@5) measuring if ground truth text is retrieved in top k retrieved sentences.

and achieves the good performance on INK retrieval task. We use the cosine similarity score of image-text pairs to perform the ranking.

We additionally introduce an intensive neural knowledge approach that injects extracted knowledge to the contrastive learning based model at test time, which we define as INK-CLIP. For each visual knowledge snippet, we measure its semantic similarity with the corresponding text based on the embeddings acquired by the text encoder. We pick the highest knowledge-inference similarity score and get the weighted sum with the image-text score to rank the image-text pairs.

KAT (Gui et al., 2022) is the state of the art knowledge augmented model in OKVQA (Marino et al., 2019) but is not designed for image-text retrieval task. We formulate the model to perform retrieval (KAT-retrieval) with knowledge snippets extracted in Section 5.1 and corresponding text as input, and have it to generate answer "yes/no". The probability of "yes" is used to score the alignment between knowledge and text, and rank retrieval sentences based on this score. More implementation details of each model are provided in the Appendix A.

5.3 Results and Ablation Study

Evaluation and Results. As shown in Table 2, we present the performance of random chance and above models on image to text retrieval task. We report the average rank of ground truth text, and Recall@5 (R@5) for each knowledge type and the entire dataset. Overall, our INK provide a progressive improvement retrieval task than existing tasks: SOTA models in COCO achieve 90% (Geigle et al., 2022), while our best model, CLIP-finetuned, achieves 48.5% on Recall@5. KAT-Retrieval outperforms the CLIP zero-shot model but not the finetuned version. We find all models achieve the lower performance on *social* knowledge. In fact, incorporating extracted knowledge to the INK-CLIP (zero-shot) model gives an overall improvement (313 → 299 in avg rank), but not on the social domain.

Does Extracting Better Automatic Knowledge Improve Retrieval Performance? We notice that automatic knowledge extraction show comparable results on INK task with the vanilla baseline. We additionally consider when *gold human knowledge* has been perfectly retrieved from our knowledge pool. In the last two rows of Table 2, we provide INK-CLIP (zero-shot oracle) and INK-CLIP (fine-tuned oracle) models which achieve significant improvement by leveraging the gold knowledge in every knowledge dimension. This implies that our knowledge pool contains relevant information to enhance the image-text alignment and developing a more advanced automatic knowledge extraction system is crucial to improve performance in our retrieval task.

6 Conclusion

We present INK, a new image-text retrieval task that requires extracting relevant knowledge to achieve good performance. To accelerate research in this domain, we release a new dataset for human annotation of gold knowledge with image-text pairs, and a set of baseline models with the benchmark, encouraging researchers to develop new models and systems, and explore ways of evaluating the INK performance. We show that integrating relevant knowledge is the key to achieve good performance in image-text retrieval task, and leave as future work to develop more advanced methods of knowledge integration to vision-language models.

Ethical Considerations

To push the frontier of this important vision and language area, our dataset INK aims at bringing researchers and practitioners in relevant fields together, to share ideas and insights. INK, while in some ways similar to other grounded-retrieval tasks, is focused more on involve in more knowledge information. This is an emerging research area that poses new challenges for AI systems and there is still significant room for improvement. Such a deeper understanding between vision and language has also started to play a key role in human-machine interaction systems. This will greatly advance computer vision technologies including visual entity and object recognition, knowledge analysis and aesthetic evaluation.

The creation of models that leverage external sources makes them more controllable. While this may benefit factuality and reduction of bias, the inverse also holds. Such a model assumes its sources are true and makes compelling arguments that leverage them. If said sources were false or misinformation, a controllable knowledge based system could be leveraged for harmful goals.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv:2010.12688*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *CVPR*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. WebQA: Multihop and Multimodal QA. *arXiv preprint arXiv:2109.00590*.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus 151*, pages 139–155.

Jianfeng Gao, Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Heung-Yeung Shum. 2020. Robust conversational ai with grounded text generation. *arXiv preprint arXiv:2009.03457*.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, and Iryna Gurevych. 2022. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521.

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

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL 2022. Long paper, Oral.* arXiv:2112.08614.

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

Bin He, Xin Jiang, Jinghui Xiao, and Qun Liu. 2020a. Kgplm: Knowledge-guided language model pre-training via generative and discriminative learning. *arXiv:2012.03551*.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020b. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models. *Proceedings of ACL*.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. *arXiv preprint arXiv:2202.04800*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. *Proceedings of ECCV*.

Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *The 34th Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Question answering about images using visual semantic embeddings. In *ICML Deep Learning Workshop*.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2021. Knowledge-aware language model pretraining. *arXiv:2007.00655*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018a. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018b. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Robyn Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017a. Fvqa: Fact-based visual question answering. *TPAMI*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017b. Explicit knowledge-based reasoning for visual question answering. *ArXiv*, abs/1511.02570.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2021. Multi-modal answer validation for knowledge-based vqa. In *arXiv preprint, arXiv:2103.12248*.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv:1912.09637*.

Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. K-plug: Knowledge-injected pre-trained language model for natural language understanding and generation in e-commerce. *arXiv:2104.06960*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv:2010.00796*.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Bill Yuchen Lin, and Xiang Ren. 2020. Pre-training text-to-text transformers for concept-centric common sense. *arXiv:2011.07956*.

Appendix

A Implementation Details

- CLIP: We use `RN50x64` version of CLIP model and finetune it following (Hessel et al., 2022) by drawing the bounding box region on image in pixel space. We use batch size of 64 trained with 8 Nvidia RTX6000 GPUs with 48GB of memory each. Learning rate of $1e^{-5}$ with AdamW (Kingma and Ba, 2015) is used and the model is trained for 5 epochs.

- INK-CLIP: We retrieve the best visual knowledge $\tilde{k}$ from knowledge pool $\mathbf{K}$ using the best region-based knowledge extractor in Appendix 5.1. We then take account of the relevance of visual knowledge when ranking the text candidates. The CLIP text transformer encodes knowledge and text with a size of 1024, and cosine similarity is measured to get the relevance score. Lastly, we perform the weighted sum: $\text{score}(V, K, T) = \text{score}(V, T) + \alpha \cdot \text{score}(\tilde{k}, T)$ to score the image and text pairs. We perform hyperparameter search for $\alpha = (0.1, 0.5)$ with step of 0.1 to get the INK-CLIP retrieval scores. We found $\alpha = 0.1$ and $\alpha = 0.2$ to be the best for extracted and gold knowledge performance.

- KAT (Gui et al., 2022) separately encodes explicit and implicit knowledge from image and leverages T5-base encoder-decoder (Raffel et al., 2020) with reasoning module (Izacard and Grave, 2020) to attend over the encoded relevant knowledge to answer the question (see Figure 2 for the model overview). Following Gui et al. (2022), knowledge for image patches are acquired, which we define as global knowledge. We additionally extract information for the specified region in the Sherlock dataset, which we define as local knowledge. The performance of different image preprocessing techniques are described in Appendix B. We use the best global and local visual knowledge method and acquire $n = 5$ knowledge snippets. Text knowledge is additionally acquired as described in Section 5.1. To train the retrieval version, negative knowledge instances are drawn from batch inspired by contrastive training (Radford et al., 2021), which are given the label "no". Drawing hard negatives to train the model is left as future work. On 8 GPUs with 32GB of memory each, we use batch size of 32, learning rate of 3e-5, and is trained for 5 epochs.

B Automatic Knowledge Extraction

We provide more details of automatic knowledge extraction system and investigate the performance of different visual extraction strategies. The overall model for the details of automatic visual knowledge extraction is shown in Figure 3. Global visual knowledge is extracted for each image patch following (Gui et al., 2022). We then extract local visual knowledge by either cropping the specified region or using finetuned model with the drawn region in (Hessel et al., 2022). The global and local extraction step is both performed on the unified knowledge pool $\mathbf{K}$ defined in Section 4 to get the visual knowledge. Lastly, the overview of the text knowledge extraction system is shown in Figure 4.

	WikiData	ConceptNet	Human
Unique Entities	187308	8863	4836
Unique Sentence	187308	22207	7284
Vocab Size	172842	15316	13172
Avg Sent Length	9.74	4.39	11.95

Table 3: Statistics of our knowledge resource. Note human knowledge source is collected on val and test set only.

To evaluate the visual knowledge extraction system, we consider the ground truth image-text pairs and measure if top 5 knowledge contains at least one gold (1 Recall @5), or if top 100 contains all three gold knowledge (3 Recall @100). The retrieval results only on the human annotated pool are shown in Table 5. We find that region finetuned performs the best to extract relevant knowledge. The global and best local knowledge is used to acquire the visual knowledge for knowledge augmented models.

C Knowledge Source Details and Examples

Table 3 presents statistics of different knowledge resources used for our task. Table 4 shows example of knowledge text extracted in the Wikidata and Conceptnet knowledge base. One sentence definition is retrieved for each entity detected in Wikidata and 7 types of relation knowledge (e.g. IsCapableOf, HasProperty, Causes, AtLocation,

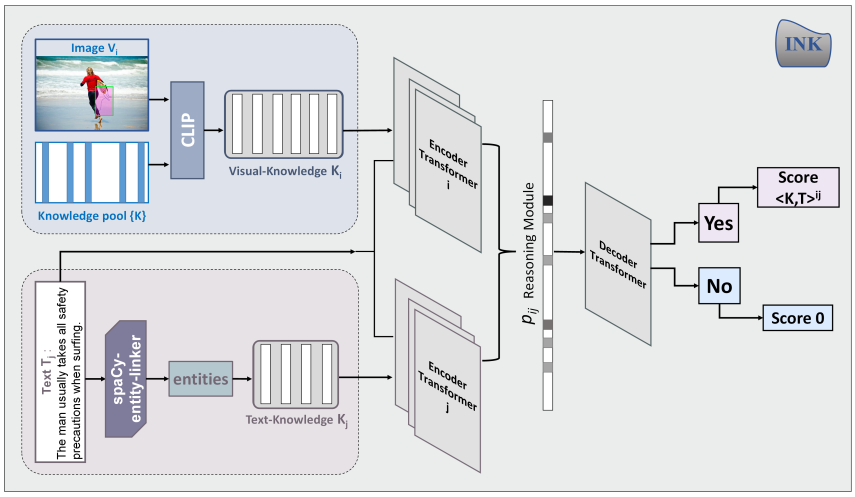


Figure 2: Our version of *KAT-retrieval* model uses a contrastive-learning-based module (CLIP) to retrieve visual-knowledge entries from wiki and concept explicit knowledge base, and uses a named entity recognition (spaCy-entity-linker) to retrieve text-knowledge from the explicit knowledge base (see Section 5.1). The integration of knowledge is processed by the respective encoder transformer, and jointly with reasoning module and the decoder transformer as an end-to-end training with the retrieval based comparison task generation.

Source	Knowledge Type	Knowledge Sentence Example
Wiki	Definition	Musette Waltz: is type of dance System Administrator: is person who maintains and operates a computer system and/or network Phillies: is baseball team in Philadelphia, Pennsylvania, United States
Conceptnet	<IsCapableOf>	Lifeguard *is capable of* saving a drowning person
	<HasProperty>	Biking *has property* good for your health.
	<Causes>	Knitting *generally causes* relaxation.
	<AtLocation>	Intersection *is usually at* the place where two streets meet.
	<PartOf>	Bills *is part of* finance.
	<MadeOf>	Sword *is made of* steel.
	<UsedFor>	Wig *is used for* changing one's appearance

Table 4: Examples of knowledge extracted automatically from the two web knowledge base, Wikidata and Conceptnet. We combine the entity/concept with the knowledge information as our knowledge sentence.

Approach	1 R@5	3 R@100
Global: CLIP zero-shot (Patch)	8.1	2.1
Local: CLIP zero-shot (Cropped)	9.6	2.7
Local: CLIP finetune (Region Drawn)	18.4	6.8

Table 5: Multimodal Knowledge Retrieval with CLIP and different image preprocessing methods. See Section5.1 for more details.

PartOf, MadeOf, UsedFor) are extracted for Conceptnet. Figure 5 shows qualitative examples of extracted knowledge from the web and human annotated knowledge for the image-text pairs.

D Additional Human Evaluation and Annotation Details

Using the template in Figure 6, we first collect image-text pairs evaluated as knowledge-intensive by humans. In Figure 7, we then ask annotators to write relevant entities and knowledge information to understand the pairs together. To capture concepts and entities that are not explicitly shown in the data, we asked them not to include words that are mentioned in the text description. Two words are annotated with one line description following the format in Wikidata, and other followed the triplet format in ConceptNet, in order to align the text domain with explicit knowledge.

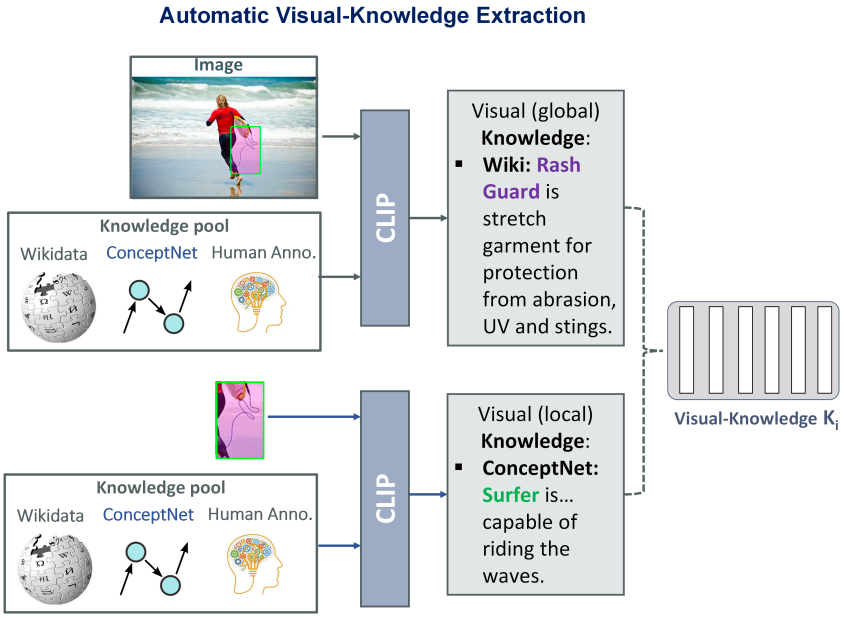


Figure 3: Overview of automatic visual-knowledge extraction.

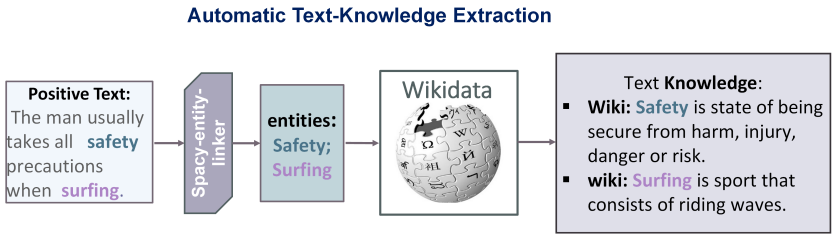
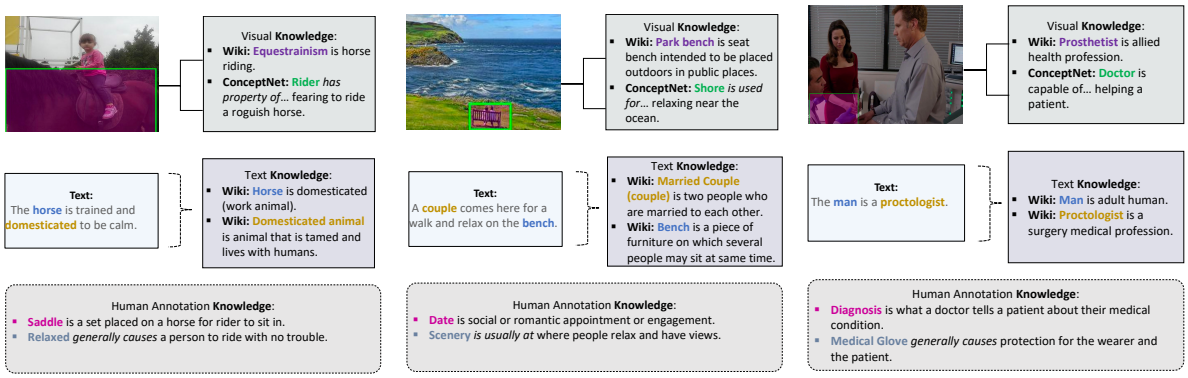


Figure 4: Overview of automatic text-knowledge extraction.



(a) Object-commonsense knowledge (b) Social-consciousness knowledge (c) Factoid knowledge

Figure 5: Qualitative examples of knowledge extracted from web knowledge base such as Wikidata and Conceptnet, and human annotation knowledge for the three knowledge dimensions. We extract visual and text knowledge separately for each image and text.

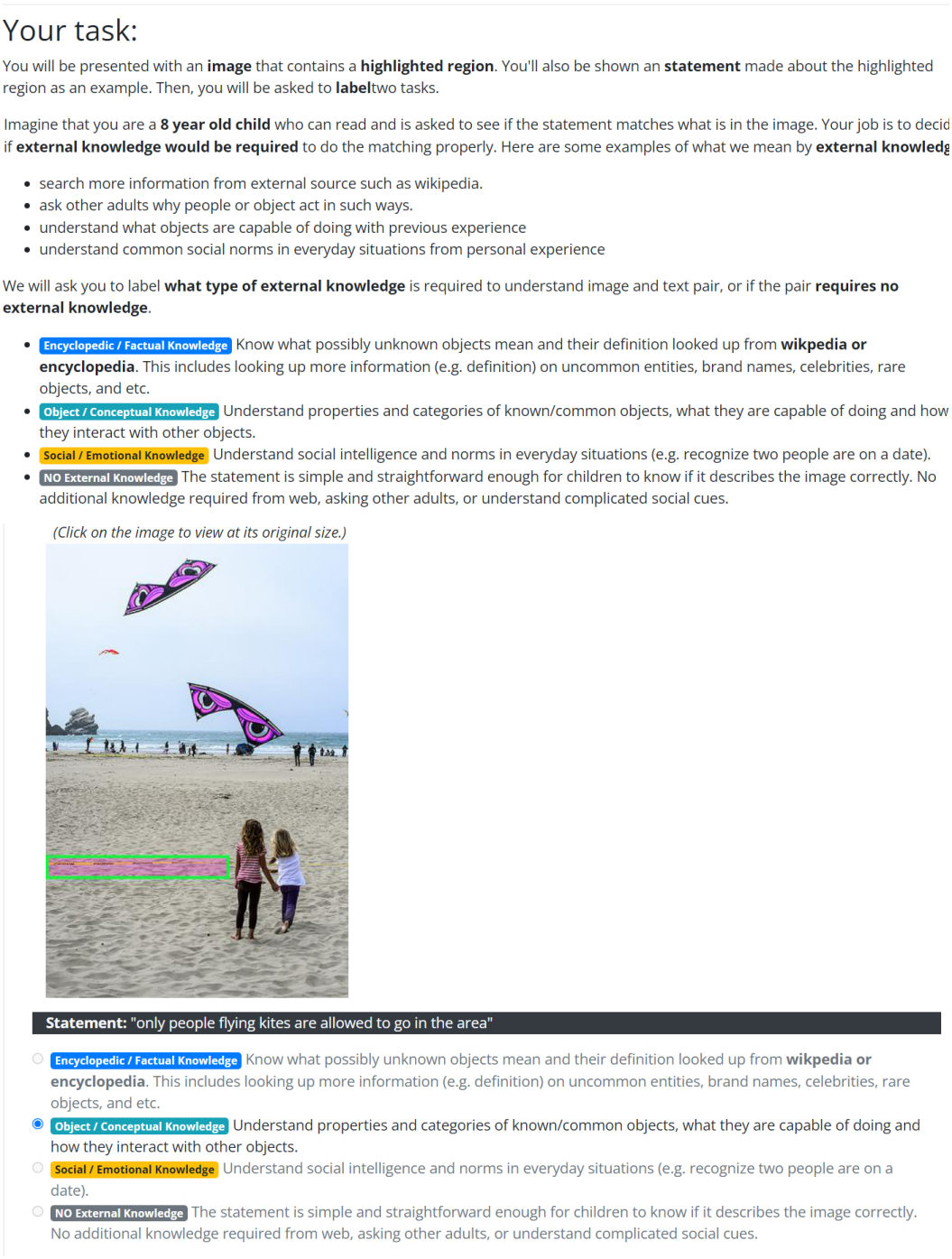


Figure 6: Example template of identifying knowledge-intensive image and text pairs. We define three types of knowledge dimension, ask humans to select if such knowledge, or external knowledge would be required to understand the image and text together. We specifically asked workers if children would be able to comprehend the alignment, and select no external knowledge is required if not.

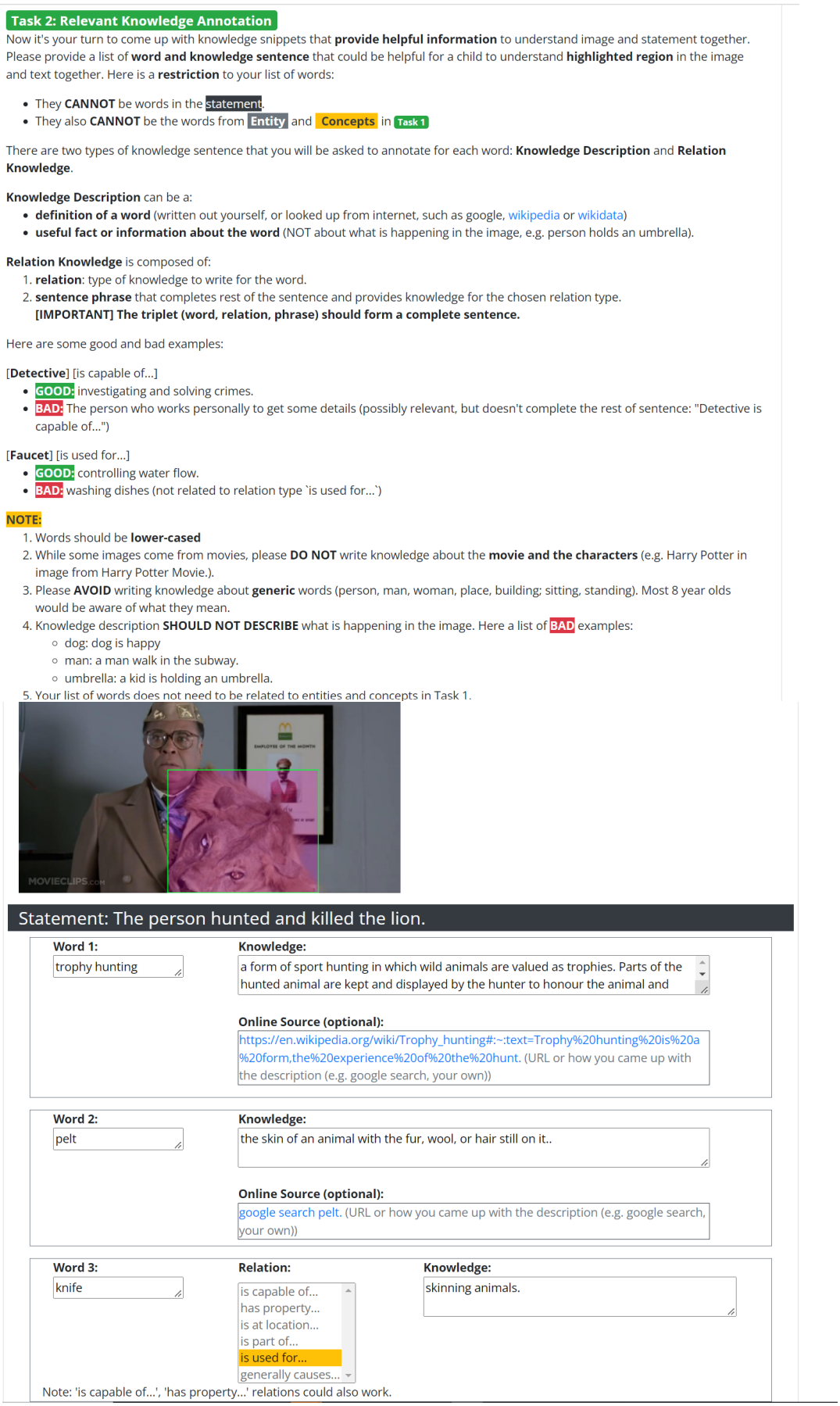


Figure 7: Example template of collecting human annotation of relevant knowledge for image and text pairs. We ask humans to define entities and one sentence description that could aid the understanding of the multimodal content.