SINC: Self-Supervised In-Context Learning for Vision-Language Tasks Appendix

Yi-Syuan Chen[1], Yun-Zhu Song[1], Cheng Yu Yeo[1], Bei Liu[2], Jianlong Fu[2], and Hong-Han Shuai[1]
[1]National Yang Ming Chiao Tung University, [2]Microsoft Research Asia
{yschen.ee09,yzsong.ee07,boyyeo123.ee08,hhshuai}@nycu.edu.tw
{Bei.Liu,jianf}@microsoft.com

1. Details of Implementations

The meta-model is a 12-layer decoder-only transformer, and the multi-source feature fuser comprises a single cross-attention layer. For data representation, METER [5], ViT [4], and RoBERTa [9] are considered as the vision-language, vision, and language knowledge sources. For ViT and RoBERTa, we use the pre-trained models from Huggingface library [18] with base configuration. During pre-training, we use 8 demonstrations, which can be easily expanded due to the low computation cost of the approach. For DID prompts, we leverage Faiss [7] to retrieve 100 related data based on VL representations, which are then randomly sampled for demonstrations. For LID prompts, we sample 1 class in addition to the query class. The model is pre-trained for 500k steps with 4k warm-up steps using AdamW [10] optimizer. Pre-training performance is monitored using LID and OD prompts from a separate validation set, and the best checkpoint is determined based on the summation of LID and OD validation performance. To avoid label leakage in demonstrations for downstream tasks, DID prompts are utilized as a selection method for ICL evaluation in downstream, as opposed to LID prompts. It should be noted that gradient-based methods are practically infeasible to benefit from the demonstration selection approach due to the high-cost re-training process.

1.1. Construction of Self-Supervised Datasets

We utilize spaCy library[1] to parse the texts from image-text datasets, including COCO [3], Visual Genome [8], Conceptual Captions [14], and SBU Captions [11]. We extract the tokens with part of speech belonging to nouns, verbs, adjectives, and adverbs. To construct a pre-training label space that facilitates the transfer to the downstream tasks, we select tokens that occur to be the labels in downstream, i.e., VQAv2, SNLI-VE, and NLVR2, and randomly sample additional 1000 tokens. As a result, our pre-training label set consists of 3126 classes, which can be readily extended to accommodate other tasks. We then retrieve image-text pairs that contain the identified classes in the text and replace the corresponding spans with special `[MASK]` tokens. This results in a manageable collection of 4,110,168 instances in our environment, with the possibility of further expansion.

1.2. Experiments on Learning Efficiency

We reimplement the architecture of both SINC and Frozen [17] under comparable settings to access their learning efficiency.[2] The architecture of Frozen consists of a language model and an image encoder, where the image encoder is utilized to extract features of text-related images. These features are then converted into multiple embedding vectors sharing the same dimension as the text token embeddings. The objective of Frozen is the conventional causal language modeling [13, 2], where the image embeddings are prepended to the sequence. During the learning process, the language model remains frozen while the image encoder is learnable.

For implementations, we utilize the Huggingface Transformer library [18] and deliberately choose configurations that maintain approximately equivalent learnable parameters across the different methods. For Frozen, we employ GPT-2 [13]

[1]https://spacy.io/
[2]As of this submission, Frozen [17] has not yet released the official implementation.

as the language model, which has 124M parameters, and ViT [4] with base and large configurations as the image encoder, containing 86.39M and 304.35M parameters, respectively. Note that Frozen actually uses a 7B language model, which is unfeasible for our devices (24GB VRAM), and thus we reduce the language model size. Therefore, the actual learning costs for Frozen should be higher than our estimates. The model is trained with embedding sequences of length 512, consisting of 2 image tokens and 510 text tokens. For SINC, we also employ two configurations of GPT-2 as the meta encoder, containing 81.9M and 354.82M parameters, respectively. To extract features, we use METER [5], RoBERTa-base [9], and ViT-base [4] for vision-language, language, and vision features. The meta model is trained with 8 demonstrations and 1 query data, resulting in a representation sequence length of 17 (16 for demonstration data and label, and 1 for query data). The GFLOPs and memory footprint are evaluated using the torch profiler library [3] as shown below.

```
from torch.profiler import profile

# Prepare data and targets
...

with torch.profiler.profile(
    activities=[
        torch.profiler.ProfilerActivity.CPU,
        torch.profiler.ProfilerActivity.CUDA,
    ],
    with flops=True, profile_memory=True) as prof:
        # Forward
        outputs = model(**inputs, output_hidden_states=True)
        loss = loss_function(outputs, targets)
        # Backward (if required)
        loss.backward()
        optimizer.step()

)
events = prof.events()

# Sum up the computation and memory consumption from events.
...
```

2. More Ablation Study

Tab. 1 presents additional ablation analyses for SINC. The results demonstrate the efficacy of the proposed correlation embeddings in enhancing the performance (row 2). This improvement can be attributed to the ability of correlation embeddings to capture the relationship between the demonstrations and the query data. This relationship acts as a condition that compels models to rely distinctively on the demonstrations for predictions, consequently leading to an improved performance.

We also explore alternative approaches to constructing self-supervision that could facilitate the transferability of SINC. Our initial approach involves employing K-means clustering to generate 100 clusters for the image-text datasets, as previously utilized. Subsequently, the data is labeled based on the corresponding cluster membership. We then randomly initialize 100 label embeddings to train models using SINC under the default setting. For downstream evaluation, since the pre-training label embeddings are not associated with any specific semantics, we randomly select a corresponding number of embeddings for downstream tasks. However, the evaluation results indicate that this approach fails to provide the ICL transferability to downstream tasks (row 3). We hypothesize that this is mainly due to the limited and non-generalizable nature of the pre-training label space. In comparison, the proposed methods effectively achieve transferability through the designs on self-supervision construction and architectures, thereby demonstrating the superiority.

3. Prior Methods for Few-Shot Visual Entailment

SNLI-VE [19] is a dataset for visual entailment that is derived from the text entailment dataset SNLI [1]. A datum in SNLI consists of a *text premise* and a *text hypothesis*, and the task is to determine whether the hypothesis entails, contradicts, or is neutral to the premise. The text premises in SNLI are sourced from the image-captioning dataset Flickr30k [20], and SNLI-VE is created by substituting these text premises with their corresponding images from Flickr30k. Therefore, the aim of SNLI-VE becomes to predict the entailment relations between an *image premise* and a *text hypothesis*.

[3] https://pytorch.org/docs/stable/profiler.html

Setting	VQAv2 val	SNLI-VE dev / test	NLVR2 dev / test
Default	44.42	53.35/53.23	54.97/56.39
Correlation Embeddings			
w/ → w/o	43.37 -1.05	51.45/51.15 -1.99	54.35/55.30 -0.86
Self-Supervision Construction			
Proposed → Clustering	1.02 -43.4	33.35/33.29 -19.97	50.00/50.02 -5.67

Table 1. Different settings of SINC. Across settings, the in-demo ratio is 0.2 and the number of demonstrations is 4 for fair comparisons, which may not yield optimal values for tasks.

[15] presents a zero-shot cross-modality transfer technique for SNLI-VE with additional usage of *text premises* from SNLI. The approach involves training the CLIP text encoder [12] to align the text premise and text hypothesis. Subsequently, the trained text encoder and the untrained CLIP image encoder are utilized to align the image premise and textual hypothesis, leading to a zero-shot cross-modality scenario. The technique yields an accuracy of 64.11% and 64.66% on the development and test sets of SNLI-VE, respectively. However, the method relies on text premises, which are not generally available. Moreover, due to the annotation process of SNLI, text premises can provide significant advantages to the prediction. As per [1], text hypotheses are generated by crowd-sourced annotators who were shown a text premise and asked to create entailing, contradicting, and neutral sentences. During the annotation process, annotators are not permitted to see the corresponding images of the text premises. Hence, the text premises should provide adequate information for predicting entailment relationships. Alternatively, if text premises are available, we can directly utilize language models for predictions. For example, RoBERTa [9] can achieve an accuracy of 82.5% and 83.5% on the development and test sets of SNLI. Therefore, the findings of [15] primarily indicate that cross-modality transfer is feasible with the additional annotation of accurate image description, and we do not consider it as the baseline for our experiments on SNLI-VE.

4. Visualization of Pre-training and Downstream Prompts

We provide the visualization for pre-training and downstream prompts, including OD (Fig. 1), LID (Fig. 2), DID (Fig. 3), fast concept binding (Fig. 4), VQAv2 (Fig. 5), SNLI-VE (Fig. 6), and NLVR2 (Fig. 7) prompts.

Figure 1. Examples of Out-Demo (OD) prompts in pre-training. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

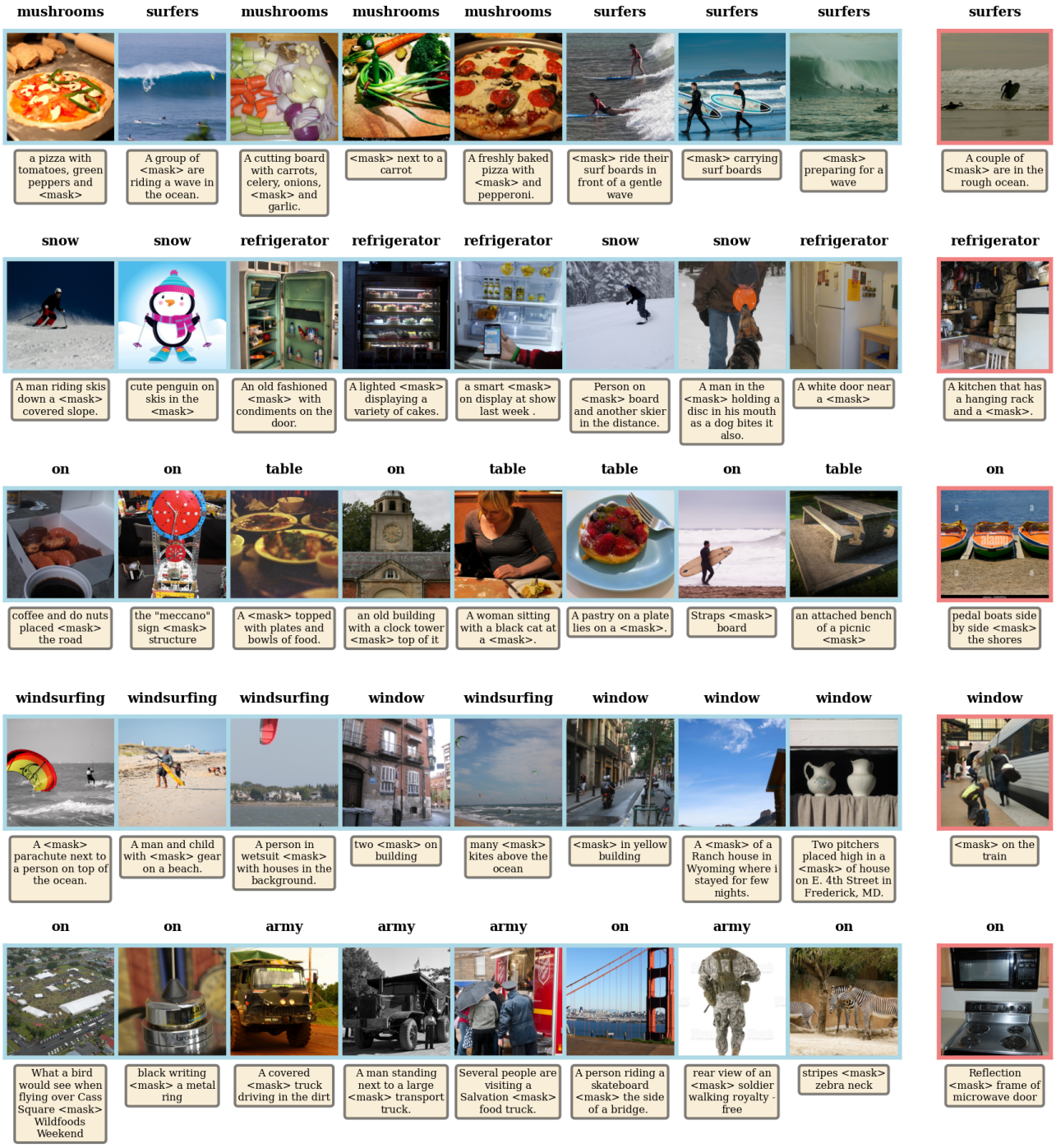

Figure 2. Examples of Label-In-Demo (LID) prompts in pre-training. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

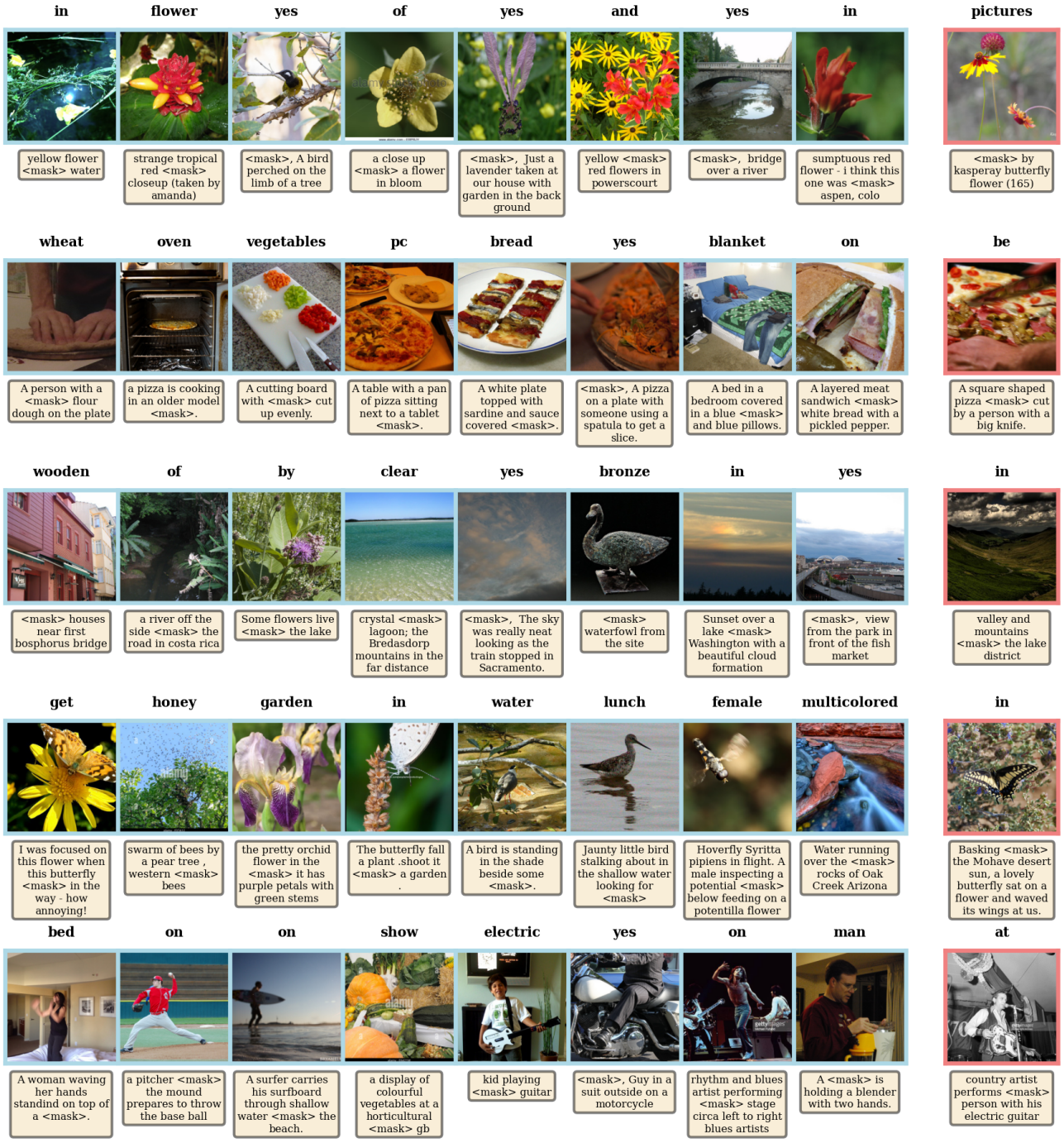

Figure 3. Examples of Data-In-Demo (DID) prompts in pre-training. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

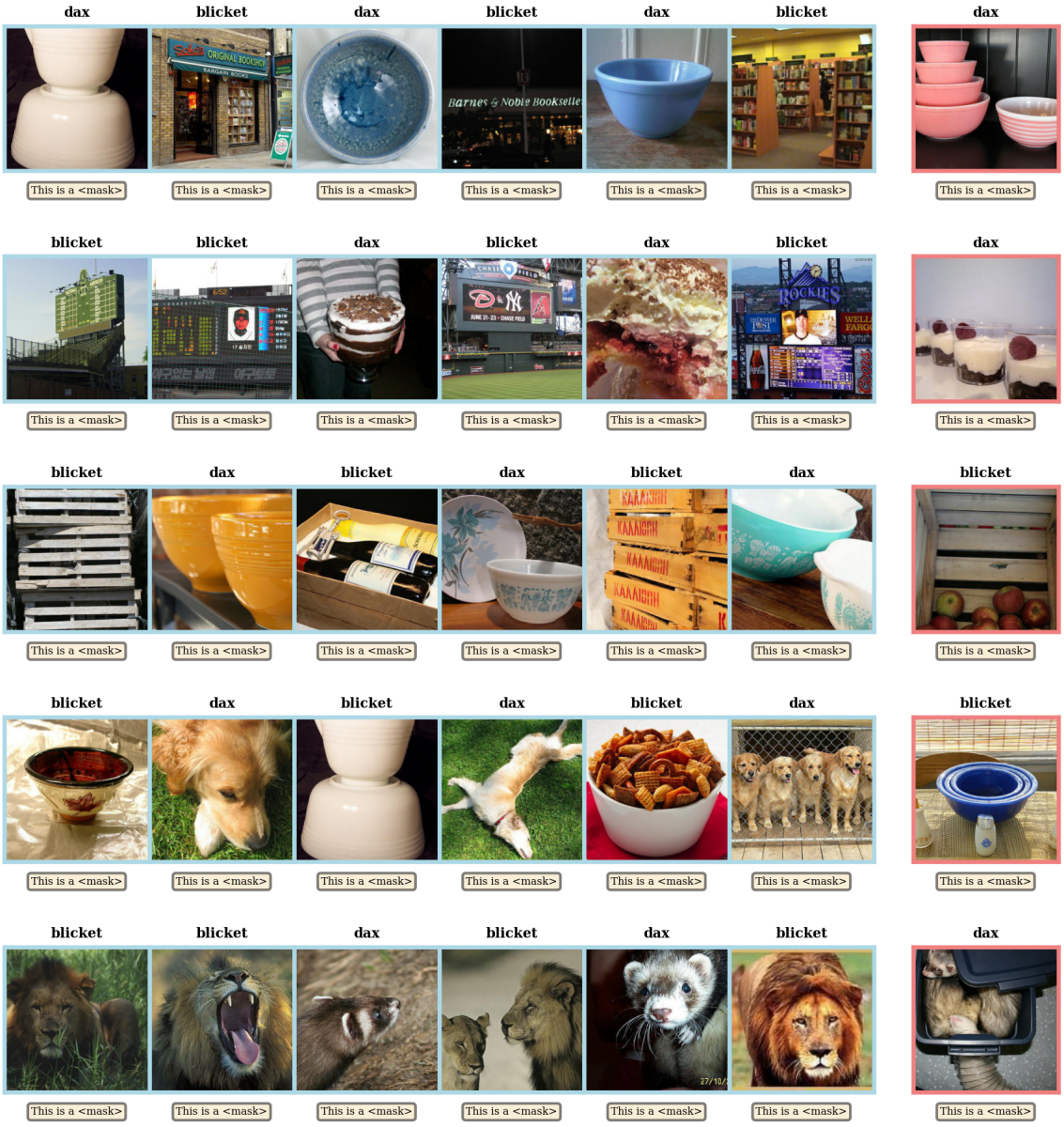

Figure 4. Examples of prompts in fast concept binding [17] for evaluation. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

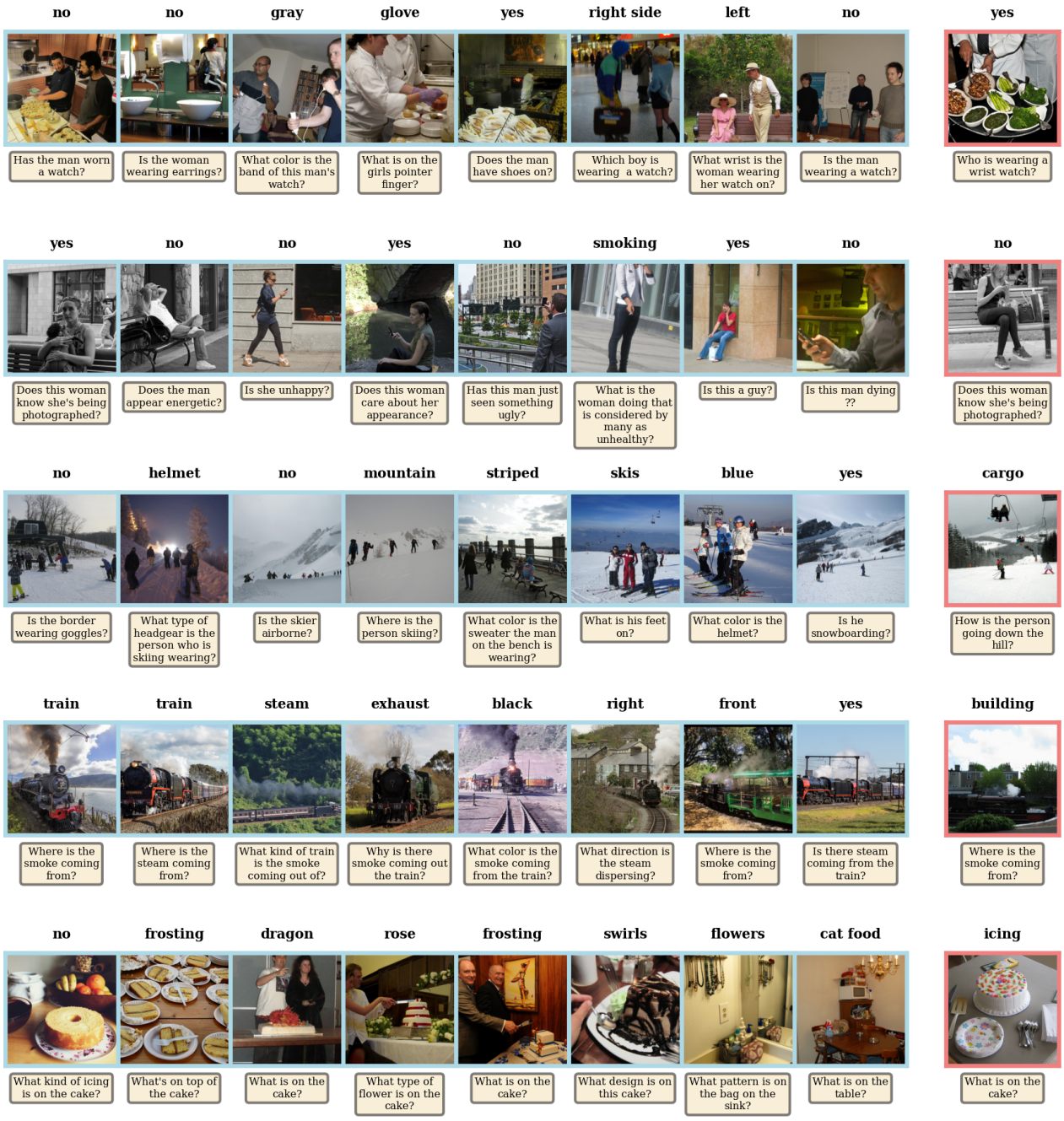

Figure 5. Examples of Data-In-Demo (DID) prompts in VQAv2 [6] for evaluation. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

Figure 6. Examples of Data-In-Demo (DID) prompts in SNLI-VE [19] for evaluation. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

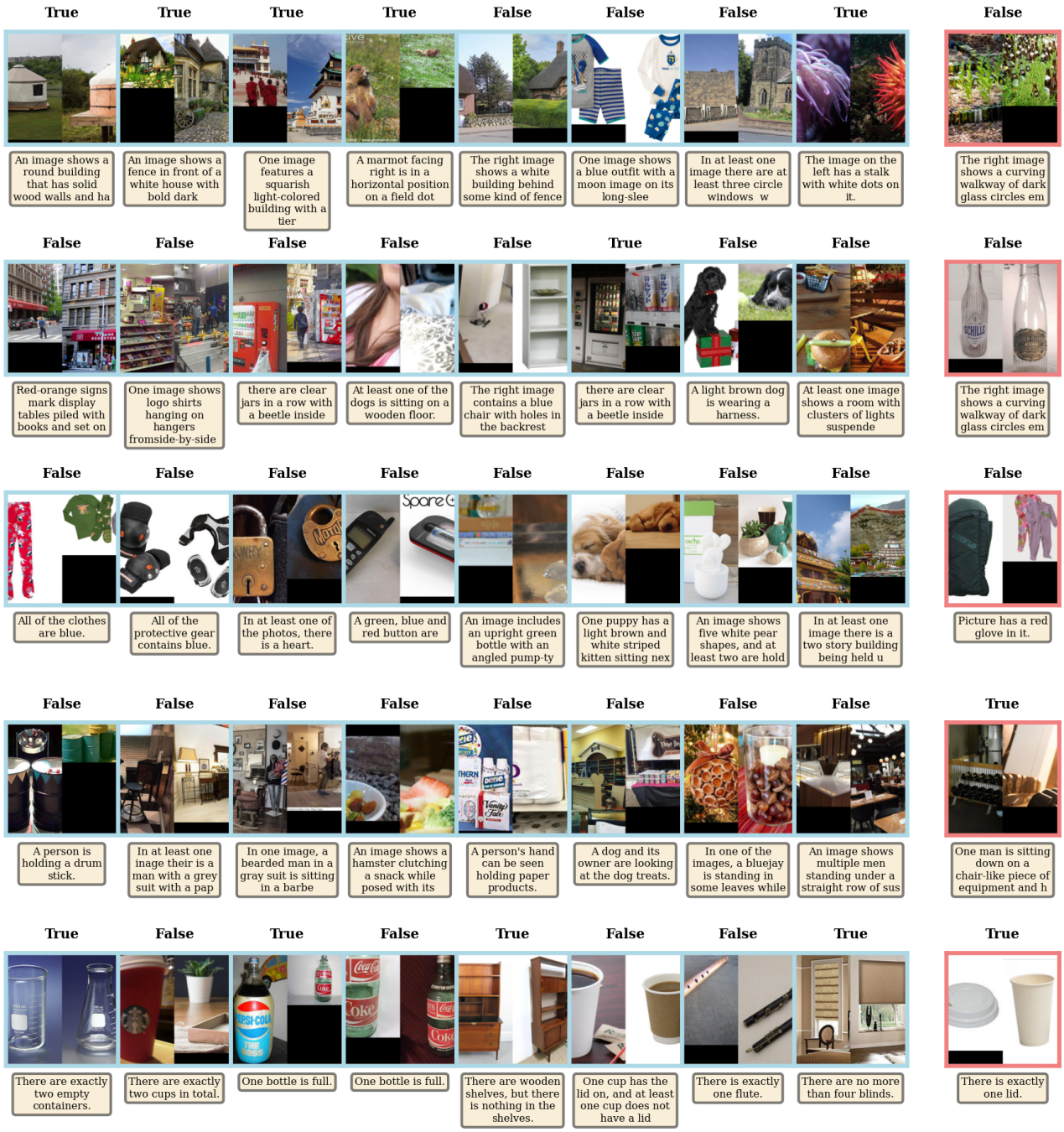

Figure 7. Examples of Data-In-Demo (DID) prompts in NLVR^2 [16] for evaluation. Demonstrations are outlined in blue and query data is outlined in red. The two images are horizontally concatenated and then center-cropped, and the text is truncated at 64 characters for better visualization.

References

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015. 2, 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 1

[3] 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. *arXiv preprint arXiv:1504.00325*, pages 1–8, 2015. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–8, 2021. 1, 2

[5] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, 2022. 1, 2

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2017. 8

[7] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, pages 535–547, 2019. 1

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, pages 32–73, 2017. 1

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, pages 1–8, 2019. 1, 2, 3

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[11] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1–8, 2011. 1

[12] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 3

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, pages 1–8, 2019. 1

[14] 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 (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1

[15] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, 2022. 3

[16] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019. 10

[17] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, pages 200–212, 2021. 1, 7

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 1

[19] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, pages 1–8, 2019. 2, 9

[20] Peter Young, Alice Lai, Micah 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*, pages 67–78, 2014. 2