SUGARCREPE++ Dataset: Vision-Language Model Sensitivity to Semantic and Lexical Alterations

Sri Harsha Dumpala* **Aman Jaiswal*** **Chandramouli Sastry**
Evangelos Milios **Sageev Oore** **Hassan Sajjad**
Dalhousie University, Canada.

1 Implementation Details

1.1 Hardware information

We performed all the experiments in this paper using a single 40G NVIDIA A100 GPU available in the Compute Canada Cluster.

1.2 Dataset sources

We obtain all existing datasets from their original sources released by the authors. We refer readers to these sources for questions regarding obtaining consent, dataset licenses and collection procedure.

- COCO [13]: We obtain COCO images from its official project website [2]. We use the images from the validation set [3]
- SUGARCREPE [4]: We obtain SUGARCREPE captions and hard negatives from its official website [4].

1.3 Model sources

Evaluation of VLMs. Source and links of the VLMs detailed in Table 8 (in the Appendix of our paper) is provided below.

- CLIP [16]: 'ViT-B/32' variant of CLIP available at HuggingFace Link
- RoBERTa-ViT-B-32 [18]: RoBERTa-ViT-B-32 trained on LAION dataset available at HuggingFace Link
- ALIGN [6]: Model available at HuggingFace Link
- ALIP [23]: Model available at Google Drive Link
- FLAVA [19]: Model available at HuggingFace Link
- ALBEF [8]: ALBEF base model available in LAVIS
- BLIP [9]: BLIP base model available in LAVIS
- BLIP2 [10]: BLIP2 pretrained model available in LAVIS
- ViLT [7]: Pre-trained ViLT model available at HuggingFace Link
- SegCLIP [14]: Model available at Google Drive Link
- XVLM-4M [25]: XVLM base model trained using 4 Million samples available at Google Drive Link

* The authors contribute equally to this work.
[2] https://cocodataset.org/
[3] http://images.cocodataset.org/zips/val2017.zip
[4] https://github.com/RAIVNLab/sugar-crepe

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- XVLM-16M [25]: XVLM base model trained using 16 Million samples available at Google Drive Link
- ViLT-ITR-COCO [7]: ViLT model finetuned for image-text retrieval task using MSCOCO dataset. This model is available at HuggingFace Link
- XVLM-16M-COCO [25]: XVLM 16 Million model fine-tuned for image-text-retrieval using MSCOCO dataset. This model is available at Google Drive Link
- XVLM-16M-Flickr [25]: XVLM 16 Million model fine-tuned for image-text-retrieval using Flickr dataset. This model is available at Google Drive Link
- NegCLIP [24]: NegCLIP is trained on top of CLIP and the model link is available in Github Link
- CLIP-SVLC [1]: Model is available at Google Drive Link
- BLIP-SGVL [3]: Model is available at Google Drive Link
- CyCLIP [2]: Model is available at Google Drive Link

Variants of CLIP: All the CLIP models' weights for the different variants of CLIP reported in Table 10 in the appendix are obtained from OpenCLIP [5] framework [5].

Evaluated ULMs: Source and links of the ULMs detailed in Table 11 (in the Appendix of out paper) is provided below.

- All-MiniLM-L6-v2 [21]: HuggingFace Link
- BGE-small-en-v1.5 [22]: HuggingFace Link
- All-MiniLM-L12-v2 [21]: HuggingFace Link
- GTE-small [12]: HuggingFace Link
- Angle-BERT-base-uncased-nli-en-v1 [11]: HuggingFace Link
- BGE-base-en-v1.5 [22]: HuggingFace Link
- Sentence-T5-base [15]: HuggingFace Link
- GTE-base [12]: HuggingFace Link
- Instructor-large [20]: HuggingFace Link
- Instructor-large (custom-ins)[20]: HuggingFace Link, we use *'Represent the sentence for spatial semantics'* as the custom instruction for Instructor-large (custom-ins) model.
- UAE-Large-V1 [11]: HuggingFace Link
- GTE-large [12]: HuggingFace Link
- All-RoBERTa-large-v1 [17]: HuggingFace Link
- Stsb-RoBERTa-large [17]: HuggingFace Link
- Sentence-T5-xl [15]: HuggingFace Link
- Angle-Llama-7b-nli-v2 [11]: HuggingFace Link

1.4 Reproducibility

We release SUGARCREPE++ dataset and the code to evaluate models on Github [6]. The datasheet for SUGARCREPE++ is provided in the Supplementary material and in the Appendix 2. The HuggingFace dataset **croissant metadata** is available here.

1.5 Author statement

In case of violation of rights, the authors will bear all responsibility. We publicly release SUGARCREPE++ dataset under the **CC-BY-4.0** license.

[5] https://github.com/mlfoundations/open_clip
[6] https://github.com/Sri-Harsha/scpp

1.6 License, Hosting and Maintenance Plan

We release the dataset publicly under the **CC-BY-4.0** license on Github. The authors of this paper are committed to support and maintain the dataset via our GitHub repository.

2 Datasheet

2.1 Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- The SUGARCREPE++ dataset was created to evaluate the sensitivity of vision language models (VLMs) and unimodal language models (ULMs) to semantic and lexical alterations. The SUGARCREPE dataset consists of (only) one positive and one hard negative caption for each image. Relative to the negative caption, a single positive caption can either have low or high lexical overlap. The original SUGARCREPE only captures the high overlap case. To evaluate the sensitivity of encoded semantics to lexical alteration, we require an additional positive caption with a different lexical composition. SUGARCREPE++ fills this gap by adding an additional positive caption enabling a more thorough assessment of models' abilities to handle semantic content and lexical variation.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- The SUGARCREPE++ dataset is created by the authors of this paper (affiliated to Faculty of Computer Science, Dalhousie University) to advance our understanding of language models through a new evaluation dataset/task.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- We acknowledge the support provided by the Faculty of Computer Science, Dalhousie University. Resources used in preparing this research were provided, in part, by the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Province of Ontario, the Government of Canada through Canadian Institute for Advanced Research (CIFAR), ACENET (ace-net.ca), the Digital Research Alliance of Canada (alliancecan.ca) and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners.

Q4 **Any other comments?**

- No.

2.2 Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The instances from SUGARCREPE++ dataset represent images from MS-COCO [13] and their associated text captions, negative captions from SUGARCREPE and newly introduced positive captions.

Q6 **How many instances are there in total (of each type, if appropriate)?**

- In total, SUGARCREPE++ dataset consists of 4757 instances. The detailed statistics of the subcategories are provided in https://github.com/Sri-Harsha/scpp.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

- We included all possible instances from the SUGARCREPE dataset, except those which are not suitable for our tasks.

Q8 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

- Each instance of SUGARCREPE++ dataset consists of an image associated with three captions, where two captions describe the image and one caption does not.

Q9 **Is there a label or target associated with each instance?** *If so, please provide a description.*

- Each instance in SUGARCREPE++ consists of an image and a triplet of captions. The label for a instance is whether each caption in the triplet correctly corresponds to the image or not.

Q10 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

- No.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

- To the best of our knowledge, there is no explicit relationship between the individual instances.

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

- No, this is only an evaluation dataset.

Q13 **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

- No, to the best of our knowledge there are no errors in SUGARCREPE++ dataset. We have done human validation as described in detail in the paper, to minimize any potential errors.

Q14 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The images used in our dataset are based on the MS-COCO [13] dataset, which is freely and publicly available. MS-COCO dataset is released under the Creative Commons Attribution 4.0 license as listed in their website https://cocodataset.org/#termsofuse.

Q15 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

- No, we source part of our dataset, such as image-caption pairs from MS-COCO [13] and negative captions from SUGARCREPE [4], both of which are open-source datasets.

Q16 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

- The authors did not create any content to be explicitly offensive. However, there may be instances that some users may find offensive. Since our SUGARCREPE++ dataset depends on the MS-COCO [13] and SUGARCREPE [4], we encourage the reader to refer to these datasets documentation for further details.

Q17 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- No, the dataset does not relate to people, and is not focused on people (although people may appear in the images and descriptions).

Q18 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

- We explicitly do not identify any sub-populations.

Q19 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

- Some images might contain identifiable individual faces.

Q20 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

- We do not provide any such data in our dataset that may be considered sensitive. All images in our datasets are taken from publicly available datasets.

Q21 **Any other comments?**

- No

2.3 Collection Process

Q22 **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data associated with each instance was acquired via our data generation process (see Section 2 in our paper for a detailed description).

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

- Please see Section 2 of our paper for a complete description of our data generation and extensive validation process.

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

- Not applicable.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

- The authors of this paper generated the textual content using generative AI as explained in Section 2 of the paper, and manually validated it.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

- The data was generated and evaluated over the course of approximately four months.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- We corresponded with the Research Ethics Board (REB) at Dalhousie University. After describing our project in detail, the REB confirmed that our project did not require ethics approval as it did not meet the regulatory definition of human subjects research. Therefore, we did not need to submit a formal application and were allowed to proceed with our research without additional REB review.

Q28 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- No, the dataset does not relate to people, and is not focused on people (although people may appear in the images and descriptions).

Q29 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

- Not applicable.

Q30 **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Not applicable.

Q31 **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- Not applicable.

Q32 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- Not applicable.

Q33 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Not applicable.

Q34 **Any other comments?**

- No.

2.4 Preprocessing, Cleaning, and/or Labeling

Q35 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

- No preprocessing or labelling was done for creating the scenarios.

Q36 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

- N/A.

Q37 **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

- Not applicable

Q38 **Any other comments?**

-

2.5 Uses

Q39 **Has the dataset been used for any tasks already?** *If so, please provide a description.*

- No. SUGARCREPE++ is a new benchmark.

Q40 **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

- To the best of our ability, we will try to maintain links to derivative papers and systems that use our dataset in the SUGARCREPE++ GitHub repository (https://github.com/Sri-Harsha/scpp).

Q41 **What (other) tasks could the dataset be used for?**

- The primary use case of our benchmark is to evaluate the sensitivity of VLMs and ULMs to semantic and lexical alterations. While we did not explore this direction in the present work, future work can use this dataset to evaluate any multi-modal system that uses VLMs and ULMs as foundation blocks such as text-to-image retrieval models, multi-modal chatbots, etc.

Q42 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

- Due to the reliance on the MS-COCO [13] and SUGARCREPE [4] datasets, SUGARCREPE++ may contain offensive material, or biases present in these source datasets. Users of SUGARCREPE++ should carefully consider how these limitations may impact their potential use case and exercise discretion in their application of the dataset.

Q43 **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

- The dataset should be avoided for a task if the limitations discussed above are unacceptable or potentially problematic for the intended use case.

Q44 **Any other comments?**

- No.

2.6 Distribution and License

Q45 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

- Yes, SUGARCREPE++ dataset will be open-sourced and freely available.

Q46 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

- Our datset and code will be made available at the following Github link: https://github.com/Sri-Harsha/scpp

Q47 **When will the dataset be distributed?**

- October 31, 2024 and onward.

Q48 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- We release data under the **CC-BY-4.0** license.
- Our code will be released under the **Apache-2.0** license

Q49 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- The dataset will be released under CC-BY-4.0 license.

Q50 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- No.

Q51 **Any other comments?**

- No.

2.7 Maintenance

Q52 **Who will be supporting/hosting/maintaining the dataset?**

- The authors will be supporting, hosting and maintaining the dataset and code through GitHub.

Q53 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

- The authors can be contacted through their email. Alternatively, an issue can be created on our GitHub repository.

Q54 **Is there an erratum?** *If so, please provide a link or other access point.*

- There is no erratum for our initial release. Errata will be documented as future releases on the benchmark website.

Q55 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

- SUGARCREPE++ will be updated. Updates can be monitored through Github.

Q56 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

- NA

Q57 **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

- We will host older versions in GitHub, in case we release newer versions.

Q58 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

- Users can extend and build on SUGARCREPE++ dataset as we did for SUGARCREPE. We do not take responsibility for validating any extension of our work.

Q59 **Any other comments?**

- No.

References

[1] S. Doveh, A. Arbelle, S. Harary, E. Schwartz, R. Herzig, R. Giryes, R. Feris, R. Panda, S. Ullman, and L. Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023.

[2] S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

[3] R. Herzig, A. Mendelson, L. Karlinsky, A. Arbelle, R. Feris, T. Darrell, and A. Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 14077–14098. Association for Computational Linguistics, 2023.

[4] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[5] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[7] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[8] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[9] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[10] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[11] X. Li and J. Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

[12] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] H. Luo, J. Bao, Y. Wu, X. He, and T. Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.

[15] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874. Association for Computational Linguistics, 2022.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[18] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

[19] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[20] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W. Yih, N. A. Smith, L. Zettlemoyer, and T. Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121. Association for Computational Linguistics, 2023.

[21] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[22] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[23] K. Yang, J. Deng, X. An, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023.

[24] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

[25] Y. Zeng, X. Zhang, and H. Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR, 2022.