System Report for CCL24-Eval Task 9:Chinese Vision-Language Understanding Evaluation

Jiangkuo Wang, Linwei Zheng, Kehai Chen*, Xuefeng Bai, Min Zhang

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

{220110927,220110604}@stu.hit.edu.cn

{chenkehai,baixuefeng,zhangmin2021}@hit.edu.cn

Abstract

This paper introduces our systems submitted for the Chinese Vision-Language Understanding Evaluation task at the 23rd Chinese Computational Linguistics Conference. In this competition, we utilized X^2-VLM and CCLM models to participate in various subtasks such as image-text retrieval, visual grounding, visual dialogue, and visual question answering. Additionally, we employed other models to assess performance on certain subtasks. We optimized our models and successfully applied them to these different tasks.

1 Introduction

In today's era of information explosion, multimodal understanding and interaction between vision and language have become increasingly important. With the rapid development of deep learning technology, vision-language pre-training models can establish associations between images and text, demonstrating powerful performance.

This competition encompasses various subtasks, spanning image-text retrieval, visual grounding, visual dialogue, and visual question answering. These tasks not only assess the model's comprehension and processing abilities with multimodal data but also evaluate its adaptability and robustness across diverse application scenarios. Through participation in these tasks, our goal is to validate the effectiveness of different vision-language pre-training models and delve into their potential in practical deployments.

The image-text retrieval task requires the model to retrieve relevant images based on textual descriptions or find matching text based on images. This demands the model to efficiently encode and align image and text features. The visual grounding task requires the model to accurately locate target objects in images, testing its fine-grained feature extraction capabilities. The visual dialogue task involves understanding the content of images and generating natural dialogues based on context. The visual question answering task requires the model to answer questions based on image content, assessing its visual understanding and language generation capabilities.

In this competition, we successfully completed the above tasks by optimizing and adjusting multiple models, including the X^2-VLM model. This paper provides a detailed introduction to our research methods, experimental process, and results, and summarizes the performance and achievements of the models in each task. Additionally, we will share the challenges and solutions encountered during the competition, aiming to offer references for future research and applications.

2 Methodology

The main models discussed in this paper include X^2-VLM (Zeng Y et al.), CCLM (Zeng Y et al.), Chinese CLIP (Redford et al.), and OFA (Wang P et al.). X^2-VLM is a general pre-training model capable of handling tasks that combine vision and language. CLIP, proposed

*Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 353–363, Taiyuan, China, July 25 – 28, 2024.
Volume3: Evaluations

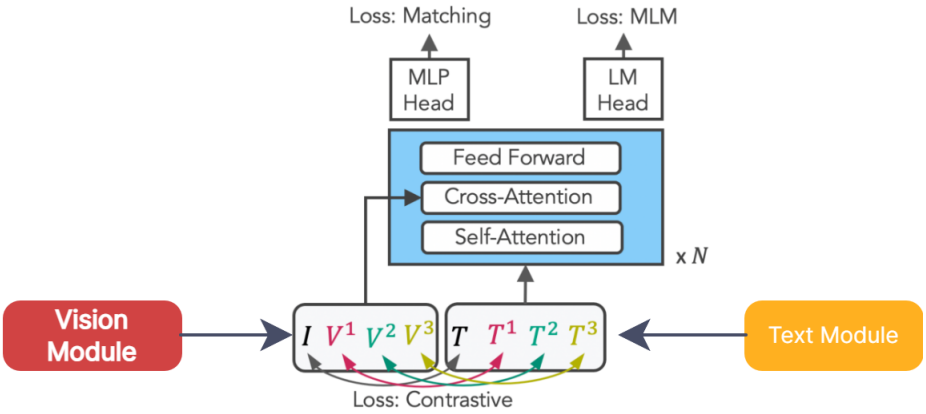

Figure 1: The overall structure of the X^2-VLM model

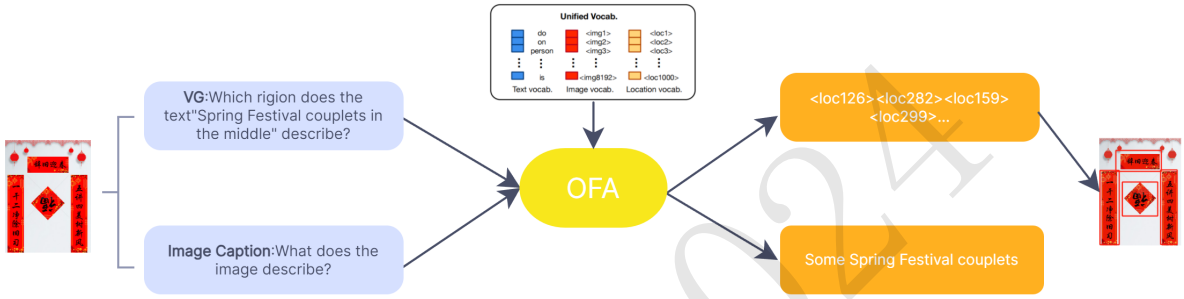

Figure 2: The overall structure of the OFA model

by OpenAI, is a multimodal model that can embed images and text into the same semantic space.

2.1 Model Architecture

Existing research on vision-language alignment generally falls into two categories: coarse-grained and fine-grained. Coarse-grained approaches use convolutional neural networks (He et al.) or vision transformers (Alexey et al.) to encode overall image features (Huang, Kim,Li rt al.). However, these methods struggle with fine-grained vision-language alignments, such as at the object level, from noisy image-text pairs that are typically weakly correlated (Huo et al.). To achieve fine-grained alignment, many methods use pre-trained object detectors as image encoders (Hao et al. Lu et al. Gan et al. Chen et al.). However, object detectors produce object-centric features that cannot encode relationships among multiple objects and can only recognize a limited number of object categories.

X^2-VLM is a unified model with vision, text, and multimodal fusion modules, all based on the Transformer architecture (as shown in Figure 1). We use three types of data for vision-language pre-training: object labels on images (Lin,Shao et al.) such as "man" or "backpack," region annotations on images (Kuznetsova et al.) such as "boy wearing backpack," and text descriptions for images such as "The first day of school gives a mixed feeling to both students and parents." The fusion module integrates vision and text features through cross-attention mechanisms. During pre-training, the modules act as encoders, and the text and fusion modules are adaptable for generative tasks. The model handles various data types, including image-text pairs, video-text pairs, and image annotations. It aligns visual concepts with textual descriptions and localizes them within images. This architecture facilitates unified encoding for both images and videos, leveraging pre-training to enhance understanding across modalities.

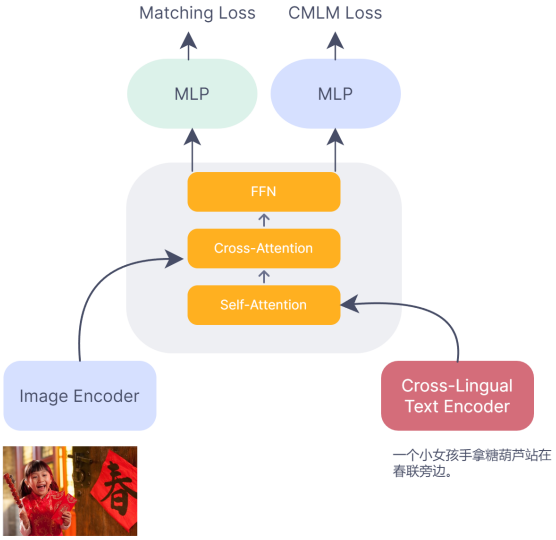

Figure 3: The overall structure of the CCLM model

The Cross-view Language Modeling (CCLM) framework combines cross-lingual and cross-modal pre-training using shared architecture and objectives (as shown in Figure 2). It consists of a Transformer-based image encoder, a cross-lingual text encoder, and a fusion model. The image encoder (Dovitskiy et al.) splits images into patches and embeds them, while the text encoder processes text inputs. The fusion model integrates text and image features through cross-attention. CCLM aligns representations of paired inputs in a common semantic space, sharing input-output formats, architectures, and training objectives. It uses contrastive, matching, and conditional masked language modeling losses to maximize sequence and token-level mutual information between inputs.

OFA is a unified Seq2Seq framework designed to integrate input/output modalities, architectures, and tasks. The model uses ResNet for visual feature extraction and byte-pair encoding for text processing. It employs a unified vocabulary for text, images, and objects (as shown in Figure 3). The architecture is based on the Transformer encoder-decoder framework, incorporating self-attention, feed-forward networks, and cross-attention layers. OFA supports multi-task and multimodal learning, including tasks such as visual grounding, image captioning, and visual question answering. It leverages large-scale pre-training datasets (Wei, Sanh et al.) and optimizes performance using cross-entropy loss and a Trie-based search strategy. Compared to models that rely on much larger paired datasets (Wang et al. Yuan et al.), OFA achieves better performance in various vision and language downstream tasks.

2.2 Innovative Text Data Augmentation

To further enhance the performance of our models in the image-text retrieval (ITR), visual dialogue (VD), and visual question answering (VQA) subtasks, we introduced an innovative text data augmentation strategy. Leveraging the advanced capabilities of the ChatGPT large language model, we performed various augmentation techniques on the textual data in our dataset. These techniques included synonym replacement, random insertion, random deletion, and random swapping of words within the text descriptions, showcasing our novel approach to enhancing textual data diversity.

Synonym Replacement Our method involved using ChatGPT to identify and replace words in the text with their synonyms. This innovative approach generated diverse versions of the same text, improving the model's ability to generalize across different expressions of the same concept. For example, the sentence "The boy is playing with a ball" could be augmented to "The child

is playing with a sphere."

Random Insertion For random insertion, we utilized ChatGPT to insert contextually relevant words at various positions within the text. This technique, innovative in its contextual relevance, augmented sentences like "The boy is playing with a ball" to "The energetic boy is playing with a ball happily," thus increasing data variety.

Random Deletion In random deletion, we employed ChatGPT to randomly remove words from the text while maintaining the overall meaning. This method enhances the model's robustness by forcing it to infer missing information. For example, "The boy is playing with a ball" could be augmented to "The boy playing a ball."

Random Swapping Random swapping involved using ChatGPT to randomly exchange the positions of words within the text, creating syntactically varied sentences that convey the same meaning. This technique improved the model's flexibility in understanding different word orders, such as augmenting "The boy is playing with a ball" to "Playing with a ball, the boy is."

Implementation Details The text augmentation was automated using ChatGPT's API. We systematically applied these innovative techniques to the entire dataset, ensuring contextually appropriate synonyms, semantically related insertions, comprehensible deletions, and structured swaps. This novel integration significantly increased the diversity of our textual data, improving the training process and enhancing the model's performance across tasks.

Overall, our innovative use of ChatGPT for text data augmentation demonstrated a unique and effective strategy, significantly boosting the robustness and generalization capabilities of our models.

2.3 Data Preprocessing

To enhance the model's performance across various tasks, we conducted the following preprocessing operations:

- **Visual Question Answering Task:** We augmented the text input of the model by incorporating historical question-answer pairs and special tokens. This addition provides additional contextual information, aiding the model in better understanding the input text and generating accurate answers.

- **Visual Dialogue Task:** Similarly, we enriched the text input by including historical dialogue pairs and special tokens. This ensures that the model considers the context comprehensively during the dialogue process, leading to more informed predictions.

- **Data Augmentation:** For all tasks, we applied diverse data augmentation techniques to the images, including random cropping, flipping, rotation, and color jittering. These operations increase data diversity, mitigate overfitting, and improve the model's generalization capability.

- **Answer List Processing:** In visual question answering and visual dialogue tasks, we modified the answer list content without removing duplicates. This approach prioritizes frequently occurring answers in the list, thereby enhancing response accuracy and diversity.

2.4 Multi-Task Learning Approach

To enhance our models' performance and robustness, we employed an innovative multi-task learning (MTL) strategy. This allows our models to learn from multiple related tasks simultaneously, improving generalization across image-text retrieval (ITR), visual dialogue (VD), and visual question answering (VQA).

Multi-Task Learning Framework Our MTL framework trains a single model on multiple tasks concurrently. Shared layers learn common representations, while task-specific layers capture nuances for each task. This reduces overfitting and improves overall performance.

Joint Loss Function We designed a joint loss function that combines losses from each task:

$$L = \lambda_1 L_{ITR} + \lambda_2 L_{VD} + \lambda_3 L_{VQA}$$

where λ_i are weights for each task's loss. These weights were tuned to balance the contributions during training.

Training Procedure Our training procedure involved:

- **Data Preparation:** We prepared a unified dataset with samples for ITR, VD, and VQA tasks.
- **Model Initialization:** Models were initialized with pre-trained weights.
- **Joint Training:** We trained the model using the joint loss function, updating shared and task-specific layers based on combined gradients.
- **Hyperparameter Tuning:** Extensive tuning was conducted to balance task losses.

Innovative Impact Our MTL approach leverages shared knowledge to enhance model performance, setting a new standard for integrating multiple tasks in a unified framework. This demonstrates substantial advancements in vision-language understanding.

By employing this MTL strategy, we achieved significant improvements in developing robust and efficient multimodal models.

2.5 Model Fusion

In our image-text retrieval task, we used an innovative model fusion approach to enhance accuracy and robustness. We combined four models: X^2-VLM, CCLM, Chinese CLIP, and OFA, leveraging their unique strengths.

First, we independently trained and evaluated each model on our dataset. We then extracted retrieval results from each model for a given image and combined these results using a weighted voting method. Each model's contribution was weighted based on its performance metrics (e.g., accuracy and recall) from the validation phase. This ensured that the most reliable models had a greater influence on the final results.

For each image, each model generated a ranked list of text descriptions. We calculated a weighted score for each text description across all models, selecting the descriptions with the highest aggregated scores as the best matches. This comprehensive approach improved retrieval performance and robustness.

The fusion process involved:

- **Model Training and Evaluation:** Independently training and evaluating each model on the image-text retrieval task.
- **Result Extraction:** Generating ranked lists of text descriptions for each image from each model.
- **Weighted Voting:** Applying a weighted voting mechanism based on individual model performance.
- **Score Aggregation:** Aggregating scores for each text description across all models.
- **Final Selection:** Choosing the top-ranked text descriptions as the most relevant matches.

This model fusion approach significantly improved retrieval performance and ensured robust results by mitigating individual model weaknesses.

Model Usage Strategy In addition to utilizing the X^2-VLM and CCLM models to tackle all five subtasks, we also leveraged the Chinese CLIP model and the OFA model for certain tasks to evaluate their performance in multimodal scenarios. Specifically, we employed a model fusion approach for the visual image-text retrieval subtasks. By combining the strengths of X^2-VLM, CCLM, Chinese CLIP, and OFA through a weighted voting mechanism, we enhanced the accuracy and robustness of our results. This fusion strategy allowed us to leverage the unique capabilities of each model, leading to improved performance across these key subtasks.

3 Experiments

3.1 Dataset Description

Our experiments utilized the Chinese image-text multimodal understanding evaluation dataset provided by the organizers. The dataset includes images from 15 main categories and 92 subcategories, manually curated to reflect elements commonly found in Chinese cultural contexts or everyday life.

3.2 Experimental Setup

We structured our experiments into several key steps, all conducted on four NVIDIA RTX A6000 GPUs:

1. **Data Preprocessing:** We enhanced text data diversity using innovative text data augmentation techniques, including synonym replacement, random insertion, random deletion, and random swapping with ChatGPT. Additionally, we applied diverse image augmentation techniques like random cropping, flipping, rotation, and color jittering to improve data diversity and model generalization.

2. **Model Selection and Architecture Optimization:** We selected four models: X^2-VLM, CCLM, Chinese CLIP, and OFA, optimizing their architectures for performance and efficiency. This included integrating multi-task learning frameworks.

3. **Model Training:** Each model was independently trained using the prepared training and validation sets. Multi-task learning strategies were employed to improve generalization and robustness by allowing models to learn from multiple related tasks simultaneously.

4. **Hyperparameter Tuning:** Extensive hyperparameter tuning was conducted to optimize model performance, adjusting learning rates, batch sizes, and other critical parameters.

5. **Model Fusion:** Post-training, we combined the strengths of X^2-VLM, CCLM, Chinese CLIP, and OFA using a weighted voting mechanism. This ensured that the most reliable models had a greater influence on the final results, enhancing accuracy and robustness.

6. **Result Analysis:** We analyzed the experimental results, comparing the performance of different models across each subtask. The analysis highlighted the strengths and weaknesses of each model and assessed the impact of our innovative techniques, providing insights for further optimization and future research.

The experimental results are shown in Tables 1 through 5.

Table 1: Experimental Results for Text Retrieval

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Model Fusion	66.5	88.9	92.3
X^2-VLM	66.8	88.2	93.3
CCLM	59.9	85.4	91.3
OFA	58.2	80.3	87.9
CLIP	54.3	77.6	84.1

Table 2: Experimental Results for Image Retrieval

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Model Fusion	48.5	78.9	87.3
X^2-VLM	48.7	77.4	87.0
CCLM	43.6	73.5	83.4
OFA	42.3	74.1	80.6
CLIP	45.1	73.9	86.3

Table 3: Experimental Results for Visual Question Answering

Model	**Accuracy (%)**
X^2-VLM	54.4
CCLM	58.3

Table 4: Experimental Results for Visual Grounding

Model	**IoU (%)**
X^2-VLM	55.7
CCLM	44.6
OFA	47.6

Table 5: Experimental Results for Visual Dialog

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
X^2-VLM	29.3	42.5	49.4
CCLM	34.3	48.5	54.7

3.3 Experiment Analysis

The X2VLM model excels in the image-text retrieval and visual grounding subtasks due to its unified architecture integrating vision, text, and fusion modules with Transformers. This enables effective cross-attention between text and vision features, associating visual concepts with text across images, videos, and annotations for comprehensive understanding. Its multi-grained training optimizes alignments and localizations, improving comprehension and localization. Additionally, the use of various loss functions enhances multimodal understanding and performance.

The Cross-View Language Modeling (CCLM) model performs strongly in VQA and VD subtasks due to its cross-lingual and cross-modal pre-training framework. Using a shared Transformer-based architecture, CCLM aligns image-text and text-translation pairs in a common semantic space. This approach maximizes mutual information between different data views through contrastive loss, matching loss, and conditional masked language modeling loss, enhancing sequence-level and token-level understanding. By sharing input-output formats and optimizing mutual information, CCLM effectively integrates visual and linguistic information, leading to superior performance in multimodal tasks.

Additionally, our text data augmentation strategy using ChatGPT further boosts model generalization and robustness by creating diverse textual inputs.

4 Conclusion

In the Chinese Vision-Language Understanding Evaluation task of the 23rd Chinese Computational Linguistics Conference, we designed and submitted a multi-model system to participate in several subtasks, including image-text retrieval, text retrieval, visual question answering, visual localization, and visual dialogue. We primarily utilized the X^2-VLM, CCLM, Chinese CLIP, and OFA models, leveraging their strengths to enhance performance in each subtask.

The experimental results demonstrate that the X^2-VLM model excelled in image-text retrieval and visual grounding tasks, particularly showcasing strong capabilities in aligning image and text features. The CCLM and OFA models also exhibited good performance in specific tasks. Through model optimization and data preprocessing, we successfully enhanced the system's performance in each subtask.

In future work, we plan to further optimize these model architectures and explore additional data augmentation and preprocessing methods to further improve model performance. Additionally, we aim to apply these models to a broader range of multimodal tasks to verify their applicability in diverse scenarios.

Overall, the outcomes of this competition underscore the potential and advantages of these multi-model systems in tackling complex multimodal tasks, offering valuable insights and guidance for future research and applications.

Acknowledgements

Thank you to all the reviewers for their valuable suggestions, which have greatly improved the content of this paper.The work was supported by the National Natural Science Foundation of China under Grant 62276077, Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), and Shenzhen College Stability Support Plan under Grants GXWD20220811170358002 and GXWD20220817123150002

Reference

Zeng Y, Zhang X, Li H, et al. 2023. *X^2-VLM: All-in-one pre-trained model for vision-language,* tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zeng Yan, et al. 2022. *Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training.* arXiv preprint arXiv:2206.00621 .

Radford A, Kim J W, Hallacy C, et al. 2021.*Learning transferable visual models from natural language supervision*International conference on machine learning. PMLR, 2021: 8748-8763.

Wang P, Yang A, Men R, et al. 2022 *Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework*International Conference on Machine Learning. PMLR, 2022: 23318-23340.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.2016. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. *An image is worth 16x16 words: Transformers for image recognition at scale. In International* Conference on Learning Representations, 2020.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu.2020. *Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.* arXiv preprint arXiv:2004.00849

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. *Vilt: Vision-and-language transformer without convolution or region supervision.* In International Conference on Machine Learning, pages 5583–5594. PMLR, 2021.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021.*Align before fuse: Vision and language representation learning with momentum distillation.* Advances in Neural Information Processing Systems, 34, 2021.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. *Wenlan: Bridging vision and language by large-scale multi-modal pre-training.* arXiv preprint arXiv:2103.06561, 2021.

Hao Tan and Mohit Bansal.2019. *LXMERT: Learning cross-modality encoder representations from transformers.* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee.2019. *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. In Advances in Neural Information.Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019,NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13–23, 2019.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. *Visualbert: A simple and performant baseline for vision and language*. arXiv preprint arXiv:1908.03557, 2019.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu.2020. *Large-scale adversarial training for vision-and-language representation learning*. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020.*Uniter: Universal image-text representation learning. In European conference on computer vision*, pages 104–120. Springer, 2020.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020.*Oscar: Object-semantics aligned pre-training for vision-language tasks*. In European Conference on Computer Vision, pages 121–137. Springer, 2020.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao.2021. *Vinvl: Revisiting visual representations in vision-language models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5579–5588, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014.*Microsoft coco: Common objects in context. In European conference on computer vision*, pages 740–755. Springer, 2014.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019.*Objects365: A large-scale, high-quality dataset for object detection*. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset,Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al.2018. *The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale*. arXiv preprint arXiv:1811.00982, 2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.2017. *Visual genome: Connecting language and vision using crowdsourced dense image annotations*. International journal of computer vision, 123(1):32–73, 2017.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai,and Quoc V Le. 2021.*Finetuned language models are zero-shot learners*. arXiv preprint arXiv:2109.01652, 2021.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021.*Multitask prompted training enables zero-shot task generalization*. arXiv preprint arXiv:2110.08207, 2021.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021.*Simvlm: Simple visual language model pretraining with weak supervision.* ArXiv, abs/2108.10904, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021.*Florence: A new foundation model for computer vision.* ArXiv, abs/2111.11432, 2021.