Appendix

A The necessity of the construction of a large-scale Chinese cross-modal dataset.

In this section, we will demonstrate the necessity of the construction of a large-scale Chinese cross-modal dataset by comparing TaiSu with CLIP[1] on the zero-shot Chinese image-text retrieval task. CLIP's models are trained with 400M English image-text pairs and have shown great generalization ability on different downstream tasks. However, they can not directly process Chinese captions. We thus use an offline machine translator[1] to translate Chinese captions into English. The results are shown in Table 6.

Even though CLIP's models are trained with much more data, they show a significantly poor performance on this task. We believe that this is due to the limited capacity of the translator. Because of the huge difference between Chinese and English, current machine translation models can not bridge the semantic gap. Therefore, simply attaching a machine translator to a model pretrained on a large-scale English corpus does not yield the results we expect. We also notice that CLIP's models obtain relatively low scores on MUGE, which is a Chinese E-commercial retrieval dataset. In fact, the E-commercial captions have very different characteristics compared with the sentences used in our daily life and are more difficult to translate. Hence, it's necessary to construct a large-scale Chinese cross-modal dataset to facilitate the deployment of cross-modal applications in a Chinese or multilingual environment.

Table 6: Comparison with CLIP on zero-shot Chinese image-text retrieval task. The Chinese captions are translated into English for CLIP by machine translation. In fact, this comparison is unfair, but the aim is to demonstrate the necessity of the construction of a large-scale Chinese cross-modal dataset.

Dataset	Method	Image-to-Text			Text-to-Image			MR
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr8K-CN	$\text{CLIP}_{\text{RN101}}$[1]	50.2	77.4	86.9	32.2	59.0	70.0	62.6
	$\text{CLIP}_{\text{ViT-B/32}}$	51.1	77.1	85.9	34.3	60.5	71.2	63.4
	$\text{Ours}_{\text{RN101}}$	55.1	82.6	**90.9**	44.9	74.2	**84.3**	72.0
	$\text{Ours}_{\text{ViT-B}}$	**57.6**	**83.4**	90.6	**45.4**	**74.4**	84.1	**72.6**
Flickr30K-CN	$\text{CLIP}_{\text{RN101}}$[1]	57.2	83.8	91.1	34.2	58.8	68.5	65.6
	$\text{CLIP}_{\text{ViT-B/32}}$	58.5	83.9	90.2	34.4	60.1	69.1	66.0
	$\text{Ours}_{\text{RN101}}$	65.3	88.6	94.1	**51.2**	**79.1**	**89.5**	77.6
	$\text{Ours}_{\text{ViT-B}}$	**65.6**	**90.1**	**94.9**	49.9	78.9	87.0	**77.7**
COCO-CN	CLIP[1]	37.8	66.2	77.6	33.2	62.6	75.8	58.9
	$\text{CLIP}_{\text{ViT-B/32}}$	37.5	65.0	76.8	34.3	63.1	75.3	58.7
	$\text{Ours}_{\text{RN101}}$	**54.1**	**82.8**	**91.8**	**54.3**	82.7	**92.4**	**76.4**
	$\text{Ours}_{\text{ViT-B}}$	52.5	81.5	91.4	53.6	**83.7**	**92.4**	75.9
MUGE	$\text{CLIP}_{\text{RN101}}$[1]	-	-	-	8.8	18.2	24.1	17.0
	$\text{CLIP}_{\text{ViT-B/32}}$	-	-	-	8.4	18.3	23.5	16.7
	$\text{Ours}_{\text{RN101}}$	-	-	-	27.5	53.9	64.8	48.7
	$\text{Ours}_{\text{ViT-B}}$	-	-	-	**29.7**	**57.0**	**67.4**	**51.4**

B Details of the filtering based on image-text retrieval

In our proposed pipeline, a filtering process based on image-text retrieval is applied to the collected Web image-text pairs. As shown in Figure 3, We conduct image-text retrieval in a small window to filter the image-text pairs. If a diagonal element in the window is the largest one of that line or that row within the window, we regard the corresponding image-text pairs as qualified data. For TaiSu,the window size is set to 120.

[1] `https://github.com/benywon/en-ch-NMT/`

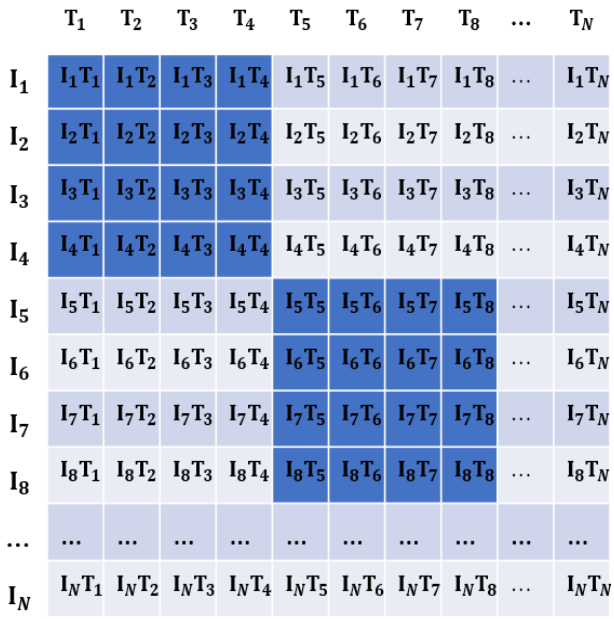

Figure 3: The similarity matrix of data in a batch. I_i is the embedding of the i-th image and T_i presents the embedding of the i-th text. $I_iT_j$is the cosine similarity between the i-th image and the j-th text. We conduct image-text retrieval in a small window(the dark blue areas) to filter the image-text pairs. If a diagonal element in the window is the largest one of that line or that row within the window, we regard the corresponding image-text pairs as qualified data.

C Hyperparameters for pretrained models

We show the hyperparameters of our pretrained models in Table 7. The backbone of the text encoder used in our experiments is a transformer encoder with causal mask. And the backbone of the image encoder is either a pretained RN101 or a pretrained ViT-B/32.

Table 7: Hyperparameters for two variants trained in our experiments.

Model Variant	Embedding dimension	Input image resolution	Image encoder	Input text token length	Text encoder layers	width	heads
RN101	512	224x224	RN101	52	12	512	8
ViT-B/32	512	224x224	ViT-B/32	52	12	512	8

ViT-B-32				**RN101**	
patch size	width	layers	heads	layers	width
32	768	12	8	[3,4,23,3]	256

D Text-to-image generation on MSCOCO

We translate the captions sampled from MSCOCO[2] into Chinese and use them as text prompts. We use TS_{all}-ViT-B/32 to guide the VQGAN to generate images according to the given text prompts. The generated samples are shown in Figure 4.

Figure 4: Some generated samples. The text prompts are from MSCOCO.

E Statistics of TaiSu

Figure 5 shows the world cloud of TaiSu generated with the Chinese text segmentation module jieba[2]. Figure 6 shows the distribution of sentence length in TaiSu. And Figure 7 shows the 50 most common nouns in TaiSu.

[2]https://github.com/fxsjy/jieba

Figure 5: The word cloud generated with texts in TaiSu dataset. For instance, “一个” means “one”; “这是” means “this is”; “女孩” means “girl”.

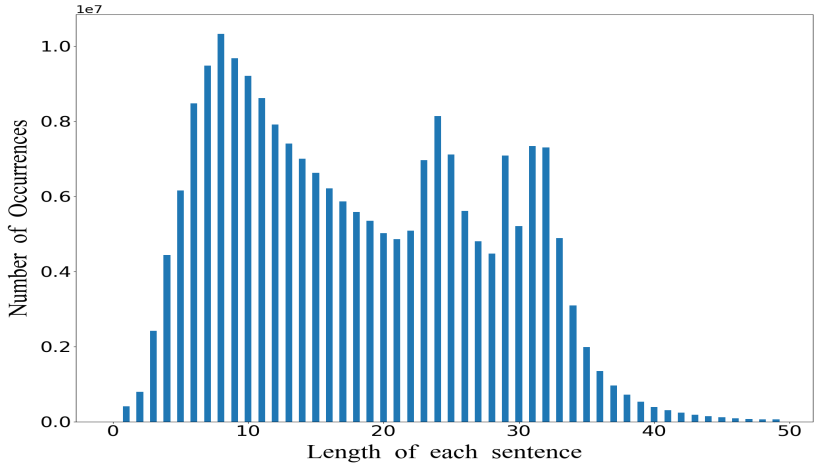

Figure 6: The distribution of the length of sentences in TaiSu. The minimum length is set to 1 and the maximum length is set to 50.

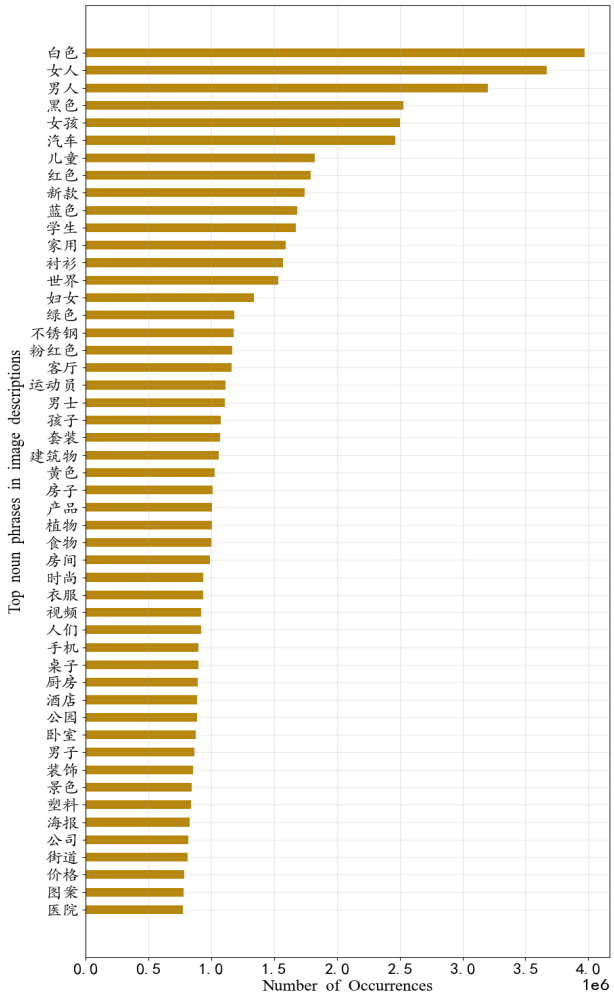

Figure 7: The 50 most common nouns in TaiSu and their number of occurrences. They are "白色"(white), "女人"(woman),"男人"(man),"黑色"(black),"女孩"(girl),"汽车"(car),"儿童"(child),"红色"(red),"新款"(new style),"蓝色"(blue),"学生"(student),"家用"(household),"衬衫"(shirt),"世界"(world),"妇女"(lady),"绿色"(green),"不锈钢"(stainless steel),"粉红色"(pink),"客厅"(living room),"运动员"(athlete),"男士"(gentleman),"孩子"(kid),"套装"(suit),"建筑物"(building),"黄色"(yellow),"房子"(house),"产品"(product),"植物"(plant),"食物"(food),"房间"(room),"时尚"(fashion),"衣服"(clothes),"视频"(video),"人们"(people),"手机"(cell phone),"桌子"(table),"厨房"(kitchen),"酒店"(hotel),"公园"(park),"卧室"(bed room),"男子"(man),"装饰"(decorate),"景色"(scenery),"塑料"(plastic),"海报"(poster),"公司"(company),"街道"(street),"价格"(price),"图案"(pattern), and "医院"(hospital).

F Potential negative societal impacts

Since the data are collected from Web, there exists inevitably some personal information such as name and job. Although we have performed person-name substitutions by substituting person names with a special token, it's still possible that some person names are left in TaiSu dataset. This may lead to the privacy disclosure. In addition, the web sites where we crawl images have strict content censorship, so the risk of pornography and violence is relatively low. In order not to infringe the image copyright, we will not release the raw images. Instead, we will release the image URLs and the corresponding captions. And TaiSu dataset is not allowed for commercial use.

G License

This Dataset are provided to You under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("CC BY-NC-SA 4.0"), with the additional terms included herein. You can access the CC BY-NC-SA 4.0 at `https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode`. When You download or use the Dataset from the Website or elsewhere, You are agreeing to comply with the terms of CC BY-NC-SA 4.0, and also agreeing to the Dataset Terms. This dataset is only for non-commercial purposes such as academic research, teaching, or scientific publications.

H Maintenance plan

For future maintenance and easy download, TaiSu Dataset and the models evaluated in our paper are hosted with websites and cloud disks. We use a Github Page as the homepage of TaiSu, where you can find the download links of our dataset and the pretrained models. For evaluation contents, we host the pytorch code on Github repo and model checkpoints and image-text datasets on Google Drive and Baidu Cloud Disk, with the latter more convenient in China.

I Datasheet

Motivation

Q1. For what purpose was the dataset created?

Answer: The goal of this dataset is to provide a large publicly accessible Chinese cross-modal dataset for vision-language pretraining and to facilitate the development of Chinese VLP community.

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

Answer: This dataset belongs to Institute of Automation, Chinese Academy of Sciences. Reasearchers: Yulong Liu, Guibo Zhu, Guojing Ge, Guanhui Qiao, Haoran Chen, JinQiao Wang, Lingxiang Wu, Bin Zhu, Qi Song, Ru Peng

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.

Answer: Our dataset is funded by CASIA.

Q4. Any other comments?

Answer: No.

Composition

Q5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Answer: Each instance is a image-text pair containing the visual image and at least one caption.

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

Answer: There are in total approximately 166 million image-text pairs in our dataset.

Q7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Answer: No. It's impossible to contain all possible instances for VLP. And our dataset is not a somple of instances from a larger set.

Q8. What data does each instance consist of? (e.g., unprocessed text or images) or features?

Answer: The raw data consists of images crawled from the internet and the corresponding captions.

Q9. Is there a label or target associated with each instance?

Answer: Yes, each image in one instance is labeled with at least one caption.

Q10. Is any information missing from individual instances?

Answer: No.

Q11. Are relationships between individual instances made explicit (e.g., users?movie ratings, social network links)?

Answer: Yes

Q12. Are there recommended data splits (e.g., training, development/validation, testing)?

Answer: No.

Q13. Are there any errors, sources of noise, or redundancies in the dataset?

Answer: Since the data are crawled from the internet, there exist noise and redundancies.

Q14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Answer: We only provide the image URLs and corresponding captions. The raw images are contained on corresponding servers.

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)?

Answer: No. All the data are publicly accessible from webpages, so this risk is minimum.

Q17. Does the dataset relate to people?

Answer: No.

Collection Process

Q18. How was the data associated with each instance acquired?

Answer: Each instance is crawled from internet according to a query item.

Q19. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Answer: Python web crawler.

Q20. If the dataset is a sample from a larger set, what was the sampling strategy?

Answer: No, the dataset is not a sample from a larger set.

Q21. Who was involved in data collection process (e.g., students, crowd-workers, contractors) and how were they compensated (e.g., how much were crowd-workers paid)?

Answer: Our dataset is collected by authors of this paper. They got their salary monthly from the institute.

Q22. 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)?

Answer: The dataset is crawled from July 2021 to march 2022. This timeframe does not match the creation timeframe of the data associated with the instances.

Q23. Were any ethical review processes conducted (e.g., by an institutional review board)?

Answer: No.

Q24. Does the dataset relate to people?

Answer: No.

Preprocessing, Cleaning, and/or Labeling

Q25. 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)?

Answer: Yes. After the image-text pairs were collected, we cleaned the dataset by a image-text retrieval process and we augment the text data by image-captioning.

Q26. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Answer: No. We only provide the cleaned dataset.

Q27. Is the software used to preprocess/clean/label the instances available?

Answer: Yes. The source code to clean the image-text pairs is available on our webpage.

Q28. Any other comments?

Answer: No.

Uses

Q29. Has the dataset been used for any tasks already?

Answer: Yes. As illustrated in the main text, we apply our dataset to a variety of multimodal downstream tasks. Results demonstrate that our dataset can serve as a promising VLP dataset, both for understanding and generation tasks.

Q30. Is there a repository that links to any or all papers or systems that use the dataset?

Answer: We do not have a repository to record all papers using our dataset. However, we can track these papers via Google Scholar.

Q31. What (other) tasks could the dataset be used for?

Answer: Our dataset is potentially suitable for multimodal downstream tasks, such as multimodal unstanding, text-to-image generation, zero-shot classification, etc.

Q32. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Answer: No.

Q33. Are there any tasks for which the dataset should not be used?

Answer: Our dataset is designed for VLP community. It may be not appropriate for tasks outside this domain. And it is not allowed for commercial purposes.

Q34. Any other comments?

Answer: No.

Q35. 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?

Answer: Yes. Our dataset is publicly available through Github.

Q36. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)

Answer: The dataset is available at `https://github.com/ksOAn6g5/TaiSu`

Q37. When will the dataset be distributed?

Answer: The dataset has been publicly released.

Q38. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Answer: Yes. License: CC BY-NC-SA 4.0

Q39. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

Answer: No.

Q40. Any other comments?

Answer: No.

Maintenance

Q41. Who will be supporting/hosting/maintaining the dataset?

Answer: Institute of Automation,Chinese Academy of Sciences.

Q42. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Answer: The email of authors of our dataset is available on the project webpage and our paper.

Q43. Is there an erratum?

Answer: No. If we notice errors in the future, we will put them in an erratum.

Q44. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Answer: Yes, we will further improve the effectiveness of this dataset.

Q45. 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)?

Answer: No. Our dataset is not related to people.

46. Will older versions of the dataset continue to be supported/hosted/maintained?

Answer: Yes. We will not delete the older version if a new version is available.

47. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Answer: We have provided information about how the data was collected and cleaned. Thus, those who want to collect similar data can easily do so.

References

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

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