中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

TaiSu: A 166M Large-scale High-Quality Dataset for Chinese Vision-Language Pre-training

Yulong Liu[1,2], Guibo Zhu[1,4,5]*, Bin Zhu[3], Qi Song[3], Guojing Ge[1], Haoran Chen[1,4], Guanhui Qiao[1,4], Ru Peng[2], Lingxiang Wu[1], and Jinqiao Wang[1,4,5]

[1]Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China
[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
[3]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[5]Wuhan AI Research

Introduction to Vision-Language Pretraining (VLP)

- **Vision-Language pretraining**

  Learning visual and linguistic representation from large-scale data pairs with weak correlation.

- **Downstream tasks**
  - Image Captioning,
  - Visual Question Answering,
  - Cross-modal Retrieval
  - Text to Image Generation
  - ……

- **VLP model architectures**
  - Dual-stream model
  - Single-stream model

VLP Datasets

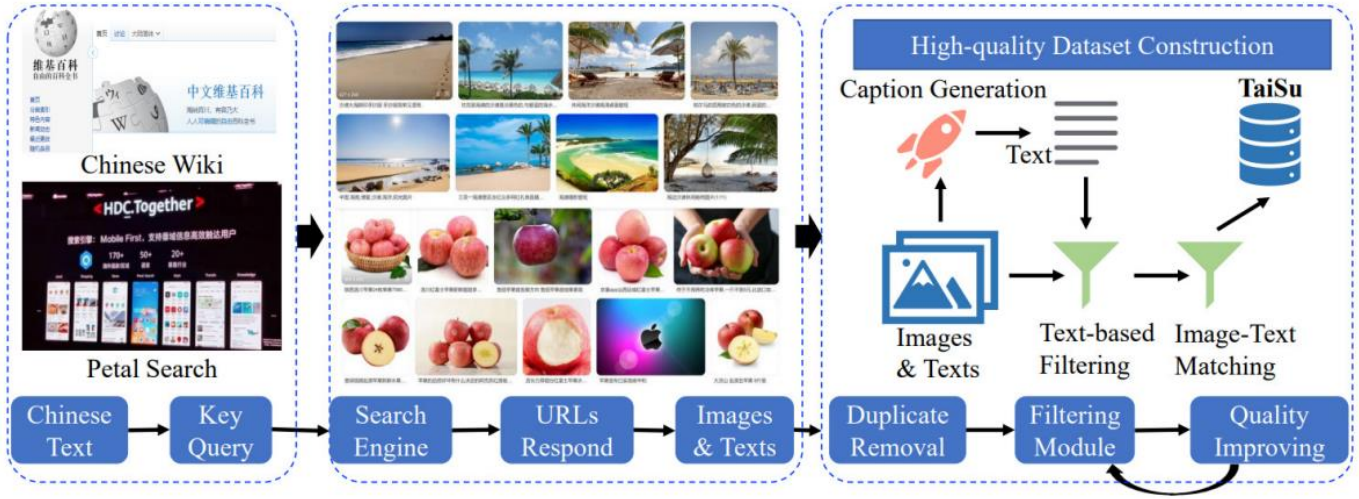
- Most of the public large-scale multi-modal datasets are based on English corpus.
- Almost all the large-scale multi-modal datasets only provide one caption for each image.

Dataset	Language	Available	Size
CC3M[9]	English	Yes	3,000,000
M5Product[10]	English	Yes	5,000,000
JFT-300M[11]	English	No	300,000,000
WIT[12]	multilingual	Yes	11,500,000
CC12M[13]	English	Yes	12,000,000
YFCC100M[14]	English	Yes	99,200,000
LAION-400M[15]	English	Yes	400,000,000
JFT-3B[16]	English	No	3,000,000,000
IG-3.5B-17K[17]	English	No	3,500,000,000
LAION-5B[18]	multilungual	Yes	5,850,000,000
Product1M[19]	Chinese	Yes	1,000,000
WudaoMM[8]	Chinese	Yes	5,000,000
M6-Corpus[6]	Chinese	No	60,500,000
WuKong[7]	Chinese	Yes	101,483,885
TaiSu(Ours)	Chinese	Yes	**166,000,000**

TaiSu is the largest public Chinese multi-modal dataset where one image may have more than one captions.

Main Contributions

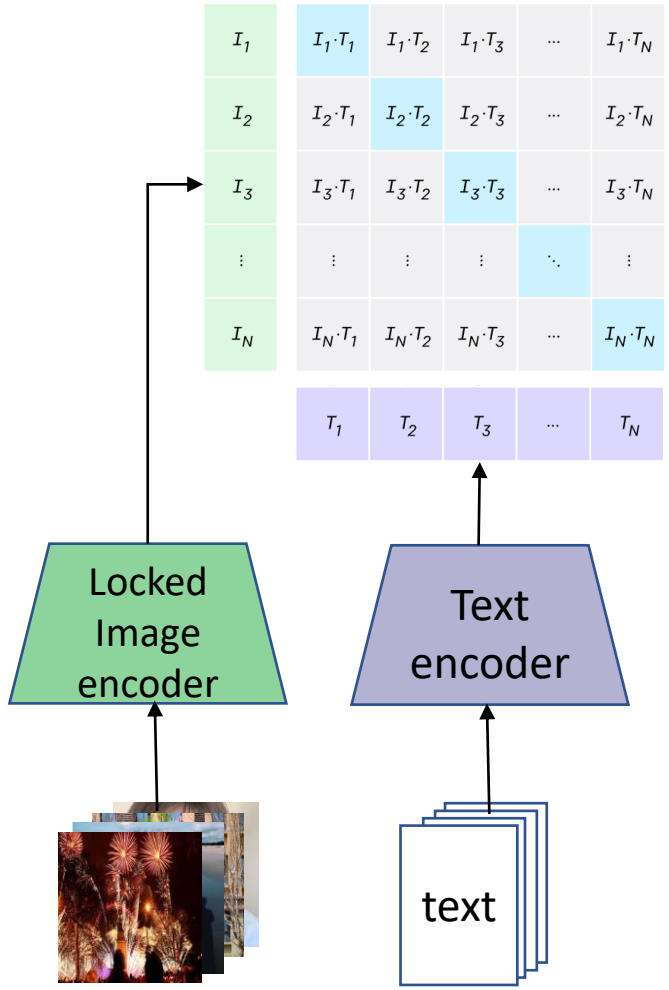
- A pipeline for the construction of a large-scale multi-modal dataset.
  text-based filtering,
  image-text matching,
  text augmentation
- The largest public multi-modal dataset for Chinese VLP.
  166M images
  219M Chinese captions



- Our experiments show that the pretrained models can benefit from the filtering process and the combination of the web data and the generated captions.

Summary of the pretraining methods

- **Dual-stream architecture**
  image encoder + text encoder.
- **Contrastive loss with global similarity**
  Follow the settings of CLIP
- **LiT-tuning**
  The pretrained image encoder → **frozen**,
  The text encoder → **need to be trained**.
- **Three kinds of data**
  Raw web data
  Filtered web data
  Filtered web data + generated data



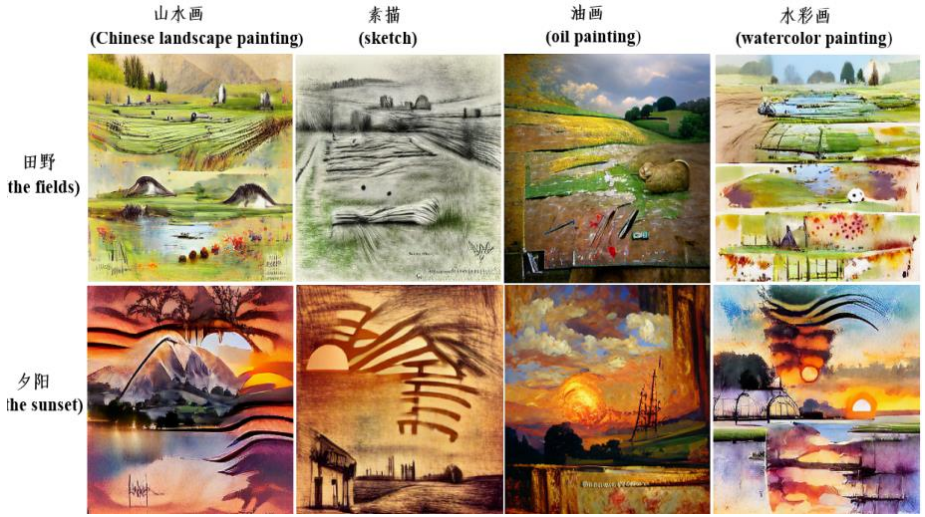
Zero-shot Image-text Retrieval & Zero-shot Image Classification & Text-to-image Generation

Zero-shot Image-text Retrieval

Dataset	Method	Image-to-Text			Text-to-Image			MR
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr8K-CN	BriVL[44]	13.4	31.2	40.7	8.0	20.7	29.5	23.9
	$\text{Wukong}_{\text{ViT-B}}$[7]	55.4	82.3	90.0	43.2	71.3	81.3	70.6
	$\text{Wukong}_{\text{Swin-L}}$[7]	47.2	78.8	87.6	36.6	64.8	76.2	65.2
	$\text{Ours}_{\text{RN101}}$	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	BriVL[44]	17.7	42.3	54.3	10.3	27.5	37.9	31.7
	$\text{Wukong}_{\text{ViT-B}}$[7]	**66.2**	88.7	94.3	45.7	73.8	82.2	75.1
	$\text{Wukong}_{\text{Swin-L}}$[7]	58.7	86.7	92.7	40.9	68.0	78.4	70.9
	$\text{Ours}_{\text{RN101}}$	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	BriVL[44]	17.1	41.7	57.5	14.8	39.0	54.2	37.4
	$\text{Wukong}_{\text{ViT-B}}$[7]	48.3	77.8	88.8	49.2	79.4	87.9	71.9
	$\text{Wukong}_{\text{Swin-L}}$[7]	47.3	78.0	88.3	46.4	77.0	87.6	70.8
	$\text{Ours}_{\text{RN101}}$	**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	BriVL[44]	-	-	-	12.7	30.9	41.8	28.5
	$\text{Wukong}_{\text{ViT-B}}$[7]	-	-	-	33.4	59.3	69.7	54.1
	$\text{Wukong}_{\text{Swin-L}}$[7]	-	-	-	**34.5**	**60.6**	**71.2**	**55.5**
	$\text{Ours}_{\text{RN101}}$	-	-	-	27.5	53.9	64.8	48.7
	$\text{Ours}_{\text{ViT-B}}$	-	-	-	29.7	57.0	67.4	51.4

Zero-shot Image classification

Dataset	Metric	BriVL[44]	RN101			ViT-B		
			RD	TS_{Web}	TS_{all}	RD	TS_{Web}	TS_{all}
CIFAR10	R@1	72.3	79.8	81.0	74.1	87.9	**89.2**	85.5
CIFAR100	R@1	35.9	43.3	43.0	43.0	55.5	**56.7**	55.4
Caltech101	R@1	72.0	71.4	71.8	73.5	72.6	73.2	**74.2**
Caltech256	R@1	58.0	68.5	69.9	70.2	69.1	71.0	**72.0**
DTD	R@1	18.8	24.9	26.2	**29.8**	26.5	29.3	28.7
Flowers	R@1	18.4	24.3	23.3	21.3	23.8	**30.6**	22.4
EuroSAT	R@1	25.5	29.9	23.6	39.7	36.0	29.5	**52.6**
ImageNet	R@1	24.3	33.3	34.2	33.4	33.7	**35.4**	34.0
AVG		40.65	46.93	46.63	48.13	50.64	51.86	**53.10**



Ablation study

Dataset	Model	Image-to-Text			Text-to-Image			MR
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr8k	RD-RN101	41.0	69.0	80.0	28.3	54.4	66.8	56.6
	TS_{Web}-RN101	44.2	69.2	80.2	29.4	56.0	68.3	57.9
	TS_{all}-RN101	55.1	82.6	**90.9**	44.9	74.2	**84.3**	72.0
	RD-ViT	35.5	63.3	76.8	26.9	52.4	64.7	53.8
	TS_{Web}-ViT	40.8	70.1	80.6	31.3	57.4	69.0	58.2
	TS_{all}-ViT	**57.6**	**83.4**	90.6	**45.4**	**74.4**	84.1	**72.6**
Flickr30k	RD-RN101	47.0	75.5	85.3	31.9	61.5	72.9	62.4
	TS_{Web}-RN101	48.5	77.4	86.2	35.2	63.5	73.8	64.1
	TS_{all}-RN101	65.3	88.6	94.1	**51.2**	**79.1**	**89.5**	77.6
	RD-ViT	42.3	71.6	82.8	29.7	57.1	68.6	58.7
	TS_{Web}-ViT	49.8	77.3	87.2	34.7	63.0	73.0	64.2
	TS_{all}-ViT	**65.6**	**90.1**	**94.9**	49.9	78.9	87.0	**77.7**
MUGE	RD-RN101	-	-	-	26.7	52.4	63.8	47.6
	TS_{Web}-RN101	-	-	-	**30.1**	56.4	**67.4**	51.3
	TS_{all}-RN101	-	-	-	27.5	53.9	64.8	48.7
	RD-ViT	-	-	-	27.6	53.6	65.1	48.8
	TS_{Web}-ViT	-	-	-	29.7	**57.0**	**67.4**	**51.4**
	TS_{all}-ViT	-	-	-	28.5	54.2	65.2	49.3
COCO-CN	RD-RN101	41.8	71.6	84.2	39.8	69.9	81.5	64.8
	TS_{Web}-RN101	43.4	75.1	85.7	41.8	70.6	84.8	66.9
	TS_{all}-RN101	**54.1**	**82.8**	**91.8**	**54.3**	82.7	**92.4**	**76.4**
	RD-ViT	40.8	70.3	83.3	37.2	68.6	82.2	63.7
	TS_{Web}-ViT	44.4	73.3	84.8	39.8	72.3	83.9	66.4
	TS_{all}-ViT	52.5	81.5	91.4	53.6	**83.7**	**92.4**	75.9

Dataset	Metric	BriVL[44]	RN101			ViT-B		
			RD	TS_{Web}	TS_{all}	RD	TS_{Web}	TS_{all}
CIFAR10	R@1	72.3	79.8	81.0	74.1	87.9	**89.2**	85.5
CIFAR100	R@1	35.9	43.3	43.0	43.0	55.5	**56.7**	55.4
Caltech101	R@1	72.0	71.4	71.8	73.5	72.6	73.2	**74.2**
Caltech256	R@1	58.0	68.5	69.9	70.2	69.1	71.0	**72.0**
DTD	R@1	18.8	24.9	26.2	**29.8**	26.5	29.3	28.7
Flowers	R@1	18.4	24.3	23.3	21.3	23.8	**30.6**	22.4
EuroSAT	R@1	25.5	29.9	23.6	39.7	36.0	29.5	**52.6**
ImageNet	R@1	24.3	33.3	34.2	33.4	33.7	**35.4**	34.0
AVG		40.65	46.93	46.63	48.13	50.64	51.86	**53.10**

- The filtering based on image-text matching can improve the performances of the pretrained models both on the zero-shot image-text retrieval task and on the zero-shot image classification task.
- The addition of the generated captions can significantly improve the performance on the zero-shot image-text retrieval task.

Comparison with CLIP + translator

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

CLIP's training data: 400M Englsih image-text pairs

On the Chinese image-text retrieval task, TaiSu's models obtain better performance than CLIP's models. This can reveal the necessity of the construction of a large-scale Chinese multi-modal dataset