arXiv:2205.03860v5 [cs.CV] 17 Nov 2022

ZERO and R2D2: A Large-scale Chinese Cross-modal Benchmark and a Vision-Language Framework

Chunyu Xie[1], Heng Cai[1], Jincheng Li[1], Fanjing Kong[1], Xiaoyu Wu[1],Jianfei Song[1]
Henrique Morimitsu[1,2], Lin Yao[1], Dexin Wang[3], Xiangzheng Zhang[3], Dawei Leng[1]
Baochang Zhang[4], Xiangyang Ji[2], Yafeng Deng[1,2]
[1]360 AI Research, [2]Tsinghua University, [3]360 Search Department, [4]Beihang University
{xiechunyu, caiheng1, lijincheng, kongfanjing, wuxiaoyu1, songjianfei, yaolin}@360.cn
{wangdexin, zhangxiangzheng, lengdawei}@360.cn, henrique.morimitsu@mail.tsinghua.edu.cn
bczhang@buaa.edu.cn, xyji@tsinghua.edu.cn, dengyafeng@gmail.com

Abstract

*Vision-language pre-training (VLP) on large-scale datasets has shown premier performance on various downstream tasks. In contrast to plenty of available benchmarks with English corpus, large-scale pre-training datasets and downstream datasets with Chinese corpus remain largely unexplored. In this work, we build a large-scale high-quality Chinese cross-modal benchmark named ZERO for the research community, which contains the currently largest public pre-training dataset ZERO-Corpus and five human-annotated fine-tuning datasets for downstream tasks. ZERO-Corpus contains 250 million images paired with 750 million text descriptions, plus two of the five fine-tuning datasets are also currently the largest ones for Chinese cross-modal downstream tasks. Along with the ZERO benchmark, we also develop a VLP framework with pre-**R**anking + **R**anking mechanism, boosted with target-guided **D**istillation and feature-guided **D**istillation (R2D2) for large-scale cross-modal learning. A global contrastive pre-ranking is first introduced to learn the individual representations of images and texts. These primitive representations are then fused in a fine-grained ranking manner via an image-text cross encoder and a text-image cross encoder. The target-guided distillation and feature-guided distillation are further proposed to enhance the capability of R2D2. With the ZERO-Corpus and the R2D2 VLP framework, we achieve state-of-the-art performance on twelve downstream datasets from five broad categories of tasks including image-text retrieval, image-text matching, image caption, text-to-image generation, and zero-shot image classification. The datasets, models, and codes are available at https://github.com/yuxie11/R2D2*

1. Introduction

Vision-language pre-training (VLP) mainly learns the semantic correspondence between vision and natural language. Previous works [6, 23, 36, 41] explore the VLP model and achieve significant improvement on various vision-language (V+L) tasks. These methods are supported by massive data [34], excellent architectures such as Transformer [40], and cross-modal models such as CLIP [32].

There are plenty of available benchmarks with English corpus, such as Conceptual Captions [35], SBU Captions [30], and LAION [34]. Differently, large-scale pre-training datasets and downstream datasets with Chinese corpus are relatively few. M6-Corpus [26] is a multi-modal pre-training dataset in Chinese but not publicly available. BriVL [13] constructs a V+L dataset called WSCD, but only releases 5M image-text pairs. Wukong [14] is a newly published pre-training dataset with 100M image-text pairs. Most existing downstream Chinese datasets mainly focus on retrieval tasks, such as Flickr30k-CN [18] and COCO-CN [25], which are not sufficient for a complete evaluation of VLP models. Besides, Flickr30k-CN tries to translate English cross-modal downstream datasets into Chinese, however, fails to cover Chinese idioms and often causes translation errors.

In this paper, we introduce a large-scale Chinese cross-modal benchmark called ZERO, including a pre-training dataset (ZERO-Corpus) and five downstream datasets. Specifically, ZERO-Corpus consists of 250 million images and 750 million descriptive texts, which is the largest public Chinese V+L pre-training dataset. ZERO-Corpus are collected from the search engine with images and corresponding textual descriptions, by filtering from 5 billion image-text data by user click-through rate (CTR). Compared to existing pre-training datasets, ZERO-Corpus is high-quality due to the user CTR filtering method and the diverse textual information for each image. Table 1 shows an overview of

Dateset	Language	Availability	#Image	#Text
Visual Genome [17]	English	Yes	108K	5.4M
SBU Captions [30]	English	Yes	875K	875K
CC3M [35]	English	Yes	3.1M	3.1M
CC12M [2]	English	Yes	12M	12M
RedCaps [9]	English	Yes	12M	12M
WIT [37]	Multilingual	Yes	11.5M	37.6M
YFCC100M [39]	English	Yes	100M	200M
LAION-400M [34]	English	Yes	400M	400M
WSCD [13]	Chinese	Yes	5M	5M
M6-Corpus [26]	Chinese	No	60.5M	60.5M
Wukong [14]	Chinese	Yes	100M	100M
ZERO-Corpus	Chinese	Yes	250M	750M

Table 1. Statistics of the vision-language pre-training datasets. The details of ZERO-Corpus can refer to Section 2.1.

Dateset	Annotation	Image-Text Pairs		
		Train	Val	Test
Flickr30k-CN [18]	Machine Translation	29K	1K	1K
COCO-CN [25]	Human Annotation	18K	1K	1K
AIC-ICC [44]	Human Annotation	210K	30K	30K
MUGE [26]	-	129K	29K	30K
ECommerce-T2I [26]	-	9K	5K	5K
Flickr30k-CNA	Human Annotation	29K	1K	1K
ICR	Human Annotation	160K	20K	20K
IQR	Human Annotation	160K	20K	20K
ICM	Human Annotation	320K	40K	40K
IQM	Human Annotation	320K	40K	40K

Table 2. Statistics of the vision-language downstream datasets. Our proposed downstream datasets can refer to Section 2.2.

V+L pre-training datasets. Together with the pre-training dataset, we provide 5 high-quality human-annotated downstream datasets. Two of them are the largest Chinese V+L downstream datasets and first proposed for Chinese image-text matching task, which is also important for evaluating VLP models. For the image-text retrieval task, we provide 3 datasets, especially our Flickr30k-CNA, which is a more comprehensive and accurate human-annotated dataset than Flickr30k-CN [18]. The statistics of the public and our proposed downstream datasets are shown in Table 2. We build a leaderboard on the five downstream test datasets.

From the perspective of cross-modal learning, existing methods can be categorized as single-stream and dual-stream. Most single-stream methods (*e.g.*, [4, 24, 31]) employ an extra object detector to extract the patch embedding and then align patches and words. As illustrated in [21], object detectors are annotation-expensive and computing-expensive, because they require bounding box annotations during pre-training and high-resolution (*e.g.*, 600 × 1000) images during inference. On the other hand, for dual-stream architectures (*e.g.*, [13, 32, 43]), it is non-trivial to model the fine-grained associations between image and text, since the corresponding representations reside in their own semantic space.

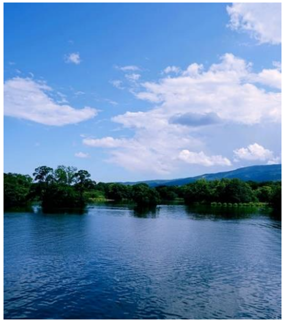
To address these limitations, we introduce two strategies in a cross-modal learning framework. We first omit the object detection module from our network to avoid expensive annotations and computation. We then combine dual-stream architecture with single-stream architecture, where the single-stream architecture consists of two cross encoders. The cross encoders are able to learn image-to-text and text-to-image interactions in a fine-grained manner. During pre-training, we design a global contrastive pre-**R**anking loss to obtain image-text representations and fine-grained **R**anking loss to further improve model performance, inspired by industrial technology such as recommend systems [5, 42] and online advertising [38]. We also introduce a two-way distillation method, consisting of target-guided **D**istillation and feature-guided **D**istillation. The target-guided distillation increases the robustness when learning from noisy labels, while feature-guided distillation aims to improve the generalization performance. We call our proposed pre-training framework **R2D2**. To summarize, our main contributions are as follows:

- We construct the largest public Chinese vision-language pre-training dataset, containing 250 million images and 750 million corresponding texts. It is high-quality due to the filtering method by user CTR and the diverse textual information for each image.
- We provide five human-annotated cross-modal downstream datasets, two of which are currently the largest Chinese V+L downstream datasets. We build a leaderboard on the five downstream test datasets.
- We introduce a VLP framework named **R2D2** for image-text cross-modal learning. Specifically, we propose a pre-**R**anking + **R**anking strategy to learn powerful vision-language representations and a two-way distillation method (*i.e.*, target-guided **D**istillation and feature-guided **D**istillation) to further enhance the learning capability.
- Our proposed method achieves state-of-the-art performance on twelve downstream datasets from five broad categories of V+L tasks, showing the superior ability of our pre-trained model.

2. ZERO Benchmark

2.1. Pre-training Dataset

Existing public pre-training datasets suffer from two major limitations. First, the image-text pairs are collected usually by their co-occurrence relationship coarsely from third-party search engines or websites. Thus, the collected pairs are inherently noisy. Second, the text corpus lacks diversity



Title: 大沼国立公园，这里水清白云蓝天，大沼、小沼、莼菜沼三个高山湖皆属于大沼国定公园
(Onuma National Park is with clear water, white cloud and blue sky. All three alpine lakes (i.e., Onuma, Konuma, and Uzbekistan) belong to Onuma National Park.)

Content: 大沼国定公园包含大沼、小沼和莼菜沼。风景最好的就在大沼的湖边附近。大沼是由驹岳火山喷发后生成的面积24平方公里的湖泊，有大小126个岛屿、32湖湾所组成，这些岛屿由18座桥梁连接的景象十分秀美，富有欧洲风味的风景。
(Onuma National Park consists of Onuma, Konuma, and Uzbekistan. The best scenery is near the lake of Onuma. Onuma is a lake with an area of 24 square kilometers formed after the eruption of Komagatake volcano. It consists of 126 islands and 32 bays. The view of these islands connected by 18 bridges is very beautiful, full of European-style scenery.)

ImageQuery: 水清白云蓝天 (clear water, white cloud and blue sky)

Figure 1. An example of ZERO-Corpus. More samples can be found in Appendix.

as each image usually has one corresponding text description. To overcome these drawbacks, we collect a new dataset for Chinese image-text pre-training, called ZERO-Corpus. Specifically, we extract 250 million images and 750 million corresponding texts from 5 billion image-text data collected by an image search engine. The key point is filtering the candidates by the higher user CTR, which means users have clicked more on an image searched by the same query. Moreover, we remove inappropriate images and harmful textual descriptions to keep only the most relevant and high-quality image-text pairs. We also provide diverse textual information for each image, *i.e.*, “Title”, “Content”, and “ImageQuery”. We show an example in Figure 1. More details about the pre-training datasets can be found in Appendix.

2.2. Downstream Dataset

We construct four Chinese image-text datasets from scratch. In these datasets, each image has one corresponding text. We divide the training set, validation set, and test set with a ratio of 8:1:1. 15 human annotators carefully label all the image-text pairs. We check the image-text pairs of the downstream datasets to ensure that these data do not appear in the pre-training dataset. We also translate all data of Flickr30k [47] by 6 professional linguists. The details of each dataset are as follows.

Image-Caption Matching Dataset (ICM). ICM is collected for the image-text matching task. Each image has a corresponding caption text. We first use CTR to select the image-text pairs. Then, human annotators manually perform a 2nd round manual correction, obtaining 400,000 image-text pairs, including 200,000 positive cases and 200,000 negative cases. We keep the ratio of positive and negative pairs consistent in each of the train/val/test sets.

Image-Query Matching Dataset (IQM). This is a dataset also for the image-text matching task. Different from ICM, we use the search query instead of detailed description text. Similarly, IQM contains 200,000 positive cases and 200,000 negative cases. ICM and IQM are currently the largest Chinese vision-language downstream datasets.

Image-Caption Retrieval Dataset (ICR). We collect 200,000 image-text pairs under the rules described in ICM. It contains image-to-text and text-to-image retrieval tasks.

Image-Query Retrieval Dataset (IQR). IQR is also proposed for the image-text retrieval task. We collect 200,000 queries and the corresponding images as the annotated image-query pairs similar to IQM. We show examples of the above four datasets in Appendix.

Flickr30k-CNA Dataset. Former Flickr30k-CN [18] translates the training and validation sets of Flickr30k [47] using machine translation, and manually translates the test set. We check the machine-translated results and find two kinds of problems. (1) Some sentences have language problems and translation errors. (2) Some sentences have poor semantics. In addition, the different translation ways prevent the model from achieving accurate performance. We gather 6 professional English and Chinese linguists to meticulously re-translate all data of Flickr30k and double-check each sentence. We name this dataset as Flickr30k-Chinese All (Flickr30k-CNA). We show some cases of the difference between Flickr30k-CN and Flickr30k-CNA in Appendix.

3. VLP Framework

Our proposed R2D2 VLP framework consists of the model architecture, pre-training strategies, and the two-way distillation method.

3.1. Model Architecture

From Figure 2, the architecture contains a text encoder, an image encoder, and two cross encoders. The text encoder is a BERT [10] using the tokenizer of RoBERTa-wwm-ext [7]. For the image encoder, we adopt the Vision Transformer (ViT) [11]. The two cross encoders are multi-layer transformers. The text encoder and image encoder transform texts and images into sequences of hidden states separately. Then the text and image hidden states interact in the two cross encoders through cross-attention.

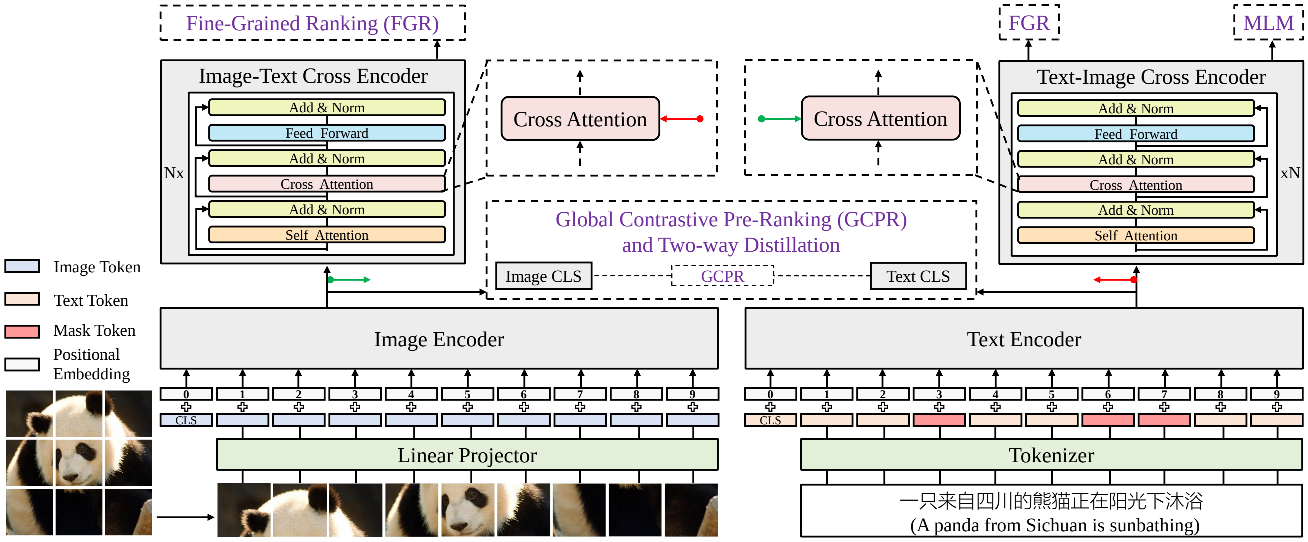


Figure 2. The overall architecture of the proposed framework. The image encoder and the text encoder aim to learn individual features of image and text, respectively. Then, the image features (green circled arrow) are fed into the text-image cross encoder. Similarly, the text features (red circled arrow) are fed into the image-text cross encoder. During pre-training, we apply global contrastive pre-ranking (GCPR) and fine-grained ranking (FGR) as pre-training objectives. Moreover, we introduce mask language modeling (MLM) with Enhanced Training (ET) and two-way distillation to obtain remarkable performance.

3.2. Pre-training Methods

To explore the relationship between image and text pairs, we design a mechanism of pre-ranking + ranking, named global contrastive pre-ranking (GCPR) and fine-grained ranking (FGR). We adopt masked language modeling (MLM) with Enhanced Training (ET) to learn the representation of cross-modal models.

Global Contrastive Pre-Ranking. Traditional contrastive learning aims to align the representation of multi-modal data (*e.g.*, paired image and text). It maximizes the similarity score of the positive pairs and minimizes the score of the negative pairs. In practice, we use global contrastive learning to accomplish the pre-ranking task. We perform full back-propagation across k GPUs. For each image I_i and the corresponding text T_i, the softmax-normalized similarity score of image-to-text and text-to-image can be defined as:

$$\begin{aligned}\mathbf{s}(I_i,T_i)&=\frac{\exp(\mathrm{sim}(I_i,T_i)/\tau)}{\sum_{j=1}^{n\times k}\exp(\mathrm{sim}(I_i,T_j)/\tau)},\\ \mathbf{s}(T_i,I_i)&=\frac{\exp(\mathrm{sim}(T_i,I_i)/\tau)}{\sum_{j=1}^{n\times k}\exp(\mathrm{sim}(T_i,I_j)/\tau)},\end{aligned}\quad(1)$$

where n is the batch size of one GPU, k is the number of GPUs, τ is a learnable temperature parameter, and $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity between a pair of image-text. Let $\mathcal{D}$ denote the training data and $\mathbf{y}(\cdot,\cdot)$ denote the ground-truth one-hot label. The global contrastive pre-ranking loss is calculated by the cross-entropy loss $\mathcal{L}_c(\cdot)$, as shown in Equation (2).

$$\begin{aligned}\mathcal{L}_{i2t}(I,T)&=\mathcal{L}_c(\mathbf{s}(I,T),\mathbf{y}(I,T)),\\ \mathcal{L}_{t2i}(T,I)&=\mathcal{L}_c(\mathbf{s}(T,I),\mathbf{y}(T,I)),\\ \mathcal{L}_{\text{GCPR}}&=\frac{1}{2}\mathbb{E}_{(I,T)\sim\mathcal{D}}[\mathcal{L}_{i2t}(I,T)+\mathcal{L}_{t2i}(T,I)]\,.\end{aligned}\quad(2)$$

Fine-Grained Ranking. As aforementioned, we apply global contrastive pre-ranking to obtain the individual representations of images and texts, respectively. Relying on these representations, we next perform Fine-Grained Ranking (FGR) loss to conduct a fine-grained ranking task. To be specific, this is a binary classification task, aiming to predict whether an image-text is matched. Formally, we denote $h_{I_{[CLS]}}$ and $h_{T_{[CLS]}}$ as the output representations of two cross encoders. Given an image representation $h_{I_{[CLS]}}$ and a text representation $h_{T_{[CLS]}}$, we feed the representations into a fully-connected layer $g(\cdot)$ to get the predicted probabilities respectively. Let $\mathbf{y}$ denote the ground-truth label of binary classification, we then compute the FGR loss as:

$$\mathcal{L}_{\text{FGR}}=\frac{1}{2}\mathop{\mathbb{E}}_{(I,T)\sim\mathcal{D}}[\mathcal{L}_c(g(h_{I_{[CLS]}}),\mathbf{y})+\mathcal{L}_c(g(h_{T_{[CLS]}}),\mathbf{y})]\quad(3)$$

The selection strategy of negative pairs is in Appendix.

Masked Language Modeling with Enhanced Training. We apply a masked language modeling loss to the text-image cross encoder to improve the ability to model the relationship between image and text at the token level. 15% of the text tokens are masked in the input. All of these tokens are replaced with the $[MASK]$ token. For the MLM task [10], the forward operations are executed individually in most VLP models [4, 21], increasing the computational cost of pre-training. In our model, the MLM task utilizes masked text and corresponding images together for denoising, which enhances the interaction between text and images. Since FGR relies heavily on this interaction ability, we propose enhanced training (ET), which integrates the MLM task into the FGR forward operations for positive image-text pairs. Experiments in Section 4.3 show that ET can reduce the computational cost of R2D2 while maintaining the accuracy of the model. For simplicity, $\mathcal{L}_{\text{MLM}}$ denotes the loss of the MLM task with enhanced training.

3.3. Two-way Distillation

Most image-text pre-training data are collected by a semi-automatic program, which may create noisy and inaccurate samples. Imprecise labels are problematic, since they may mislead the model. To address this, we propose target-guided distillation (TgD), a teacher-student paradigm with soft targets. To further improve the generalization performance of the pre-trained model, we introduce feature-guided distillation (FgD), another teacher-student based distillation. For convenience, we call the combination of these two distillations as two-way distillation (TwD).

Target-guided Distillation. To decrease the risk of learning from noisy labels, we propose to adopt soft targets generated by momentum-updated encoders. Here, the momentum-updated encoder is the teacher model of distillation, which contains the exponential-moving-average weights. We combine the similarity score $\mathbf{s}(\cdot,\cdot)$ with one-hot labels $\mathbf{y}(\cdot,\cdot)$ via coefficient α to generate the final soft targets. Let $\hat{\mathbf{y}}(I,T)$ and $\hat{\mathbf{y}}(T,I)$ denote the final soft targets. Taking $\hat{\mathbf{y}}(I,T)$ as the example, we define it as:

$$\hat{\mathbf{y}}(I,T) = \alpha \mathbf{s}(I_m,T) + (1-\alpha)\mathbf{y}(I,T), \tag{4}$$

where I_m represents that the images I are fed into the momentum-updated encoder. During training, we also introduce a queue mechanism and replace $\hat{\mathbf{y}}(I,T)$ with $\hat{\mathbf{y}}(I,T_q)$. In practice, the text queue with a fixed size aims to maintain the recent text representations. We then concatenate the text queue and the text representations of current mini-batch to compute $\mathbf{s}(I_m,T_q)$ and $\mathbf{y}(I,T_q)$. Similarly, we perform the same process when constructing $\hat{\mathbf{y}}(T,I_q)$.

Considering the effectiveness of features in the queue decreases with increasing time steps, we also maintain a weighted queue w to mark the reliability of the corresponding position features. Specifically, we decay each element in the queue by a factor of 0.99 per iteration, except for the new incoming item. Further, we replace $\mathcal{L}_c(\cdot)$ with weighted cross-entropy loss $\mathcal{L}_w(\cdot)$ in Equation 5. With the target-guided distillation, the $\mathcal{L}_{\text{GCPR}}^{\text{TgD}}$ is defined as follows.

$$\begin{aligned}\mathcal{L}_{i2t}^{w}(I,T) &= \mathcal{L}_w(\mathbf{s}(I,T_q), \hat{\mathbf{y}}(I,T_q); w),\\ \mathcal{L}_{t2i}^{w}(T,I) &= \mathcal{L}_w(\mathbf{s}(T,I_q), \hat{\mathbf{y}}(T,I_q); w),\end{aligned}$$

$$\mathcal{L}_{\text{GCPR}}^{\text{TgD}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim\mathcal{D}}[\mathcal{L}_{i2t}^{w}(I,T) + \mathcal{L}_{t2i}^{w}(T,I)]\,. \tag{5}$$

Feature-guided Distillation. Similar to TgD, we use a teacher-student paradigm to conduct feature-guided distillation. Taking the text encoder as the example below, the teacher character is the momentum-updated text encoder and the student is the text encoder. Here, the weights of the teacher are updated by all past text encoders via exponential-moving-average. To further improve the capability of the model, we apply a masking strategy to the inputs. In practice, we feed complete inputs into the teacher and masked inputs into the student. Relying on the momentum mechanism, we aim to make the features of the student closer to that of the teacher. Formally, the predicted distributions (*i.e.*, $\mathcal{P}_t(T)$, $\mathcal{P}_s(T)$) of the teacher and the student are defined as follows, respectively.

$$\begin{aligned}\mathcal{P}_t(T) &= \frac{\exp((f_t(T)-\mu)/\tau_t)}{\sum_{i=1}^{d}\exp((f_t(T)^{(i)}-\mu^{(i)})/\tau_t)},\\ \mathcal{P}_s(T) &= \frac{\exp(f_s(T)/\tau_s)}{\sum_{i=1}^{d}\exp(f_s(T)^{(i)}/\tau_s)},\end{aligned} \tag{6}$$

where $f_t(\cdot)$ and $f_s(\cdot)$ denote the networks of the teacher and the student, respectively. Moreover, μ is a momentum-updated mean of $f_t(\cdot)$, and d is the dimension of the features. τ_t and τ_s are the temperature parameters of the teacher and the student, respectively, which can sharpen the distribution of the features. Note that we do not use μ for $\mathcal{P}_s$ to avoid collapse in feature-guided distillation. We can obtain similar formulations for $\mathcal{P}_s(I)$ and $\mathcal{P}_t(I)$. We perform the feature-guided distillation by the cross-entropy loss, and the loss L_{FgD} is defined as:

$$\mathcal{L}_{\text{FgD}} = \frac{1}{2}\mathop{\mathbb{E}}_{(I,T)\sim\mathcal{D}}[\mathcal{L}_c(\mathcal{P}_s(I),\mathcal{P}_t(I)) + \mathcal{L}_c(\mathcal{P}_s(T),\mathcal{P}_t(T))]. \tag{7}$$

Our model is trained with the full objective:

$$\mathcal{L} = \mathcal{L}_{\text{GCPR}}^{\text{TgD}} + \mathcal{L}_{\text{FGR}} + \mathcal{L}_{\text{FgD}} + \mathcal{L}_{\text{MLM}}. \tag{8}$$

4. Experiments

4.1. Implementation Details

The number of transformer layers for the text encoder, and the two cross encoders are 12, 6, and 6, respectively. The text encoder is initialized from RoBERTa-wwm-ext [7] while the two cross encoders are randomly initialized. Following Wukong [14], we use the image encoder of 12-layers ViT-Base and 24-layers ViT-Large initialized from CLIP [32], and freeze it during pre-training. The resolution of the input image is 224×224 in pre-training and fine-tuning. The dimension of the feature vectors of both image and text is 768. We pre-train models with 15 epochs using a batch size of

4096 on 128 NVIDIA A100 GPUs. We set τ= 0.07 in Equation 1, α=0.4 in Equation 4, and τ_s= 0.1, τ_t= 0.04 in Equation 6. Moreover, the momentum is set as $m = 0.995$, and the queue size is 36,864. We adopt the Adam optimizer and the cosine learning rate schedule with a linear warmup [27]. The pre-trained model is adapted to five vision-language downstream tasks: image-text retrieval, image-text matching, image caption, text-to-image generation, and zero-shot image classification. More details about the downstream datasets and fine-tuning strategy can refer to Appendix.

4.2. Comparisons with State-of-the-art

For both image-to-text retrieval and text-to-image retrieval tasks, we report Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), and Mean Recall (R@M). The results of BriVL [13] and Wukong [14] are excerpted from their paper. Wukong reproduces the CLIP-style [32] and FILIP-style [46] models. Their results are also included. From Table 3, our models outperform state-of-the-art on all datasets. Moreover, R2D2$_{\text{ViT-L}}$ outperforms R2D2$_{\text{ViT-B}}$. These results indicate that our framework is able to learn better fine-grained associations between image and text. We report the results of Flickr30k-CNA on the test set of Flickr30k-CN for a fair comparison. R2D2 fine-tuned on Flickr30k-CNA outperforms that on Flickr30k-CN, since the quality of human-translated Flickr30k-CNA is much higher than that of machine-translated Flickr30k-CN.

Table 4 reports the comparison with existing methods on other V+L downstream tasks. Unlike the image-text retrieval task, there are few datasets for the Chinese image-text matching (ITM) task. Thus, we introduce image-caption matching dataset (ICM) and image-query matching dataset (IQM) for the Chinese ITM task and show the corresponding results. Also, we evaluate Wukong and BriVL on these datasets for the ITM task. We use Area Under Curve (AUC) as the metric. For the image captioning task, fine-tuning is conducted on the training split of AIC-ICC [44]. We adopt four widely-used evaluation metrics: BLEU, METEOR, ROUGE-L, and CIDEr following BriVL. Table 4 also presents text-to-image generation results on ECommerce-T2I dataset[1] [26]. The metric of Frechet Inception Distance (FID) is reported. We evaluate our pre-trained models on ImageNet [8] for the zero-shot image classification task. Class labels are translated from English. Top-1 and Top-5 accuracy are reported. Our model achieves state-of-the-art performance on these V+L downstream tasks, showing the superior capabilities.

Effect of the Proposed VLP Framework. To demonstrate the effectiveness of the proposed vision-language pre-training (VLP) framework, we conduct comparative experiments using the same pre-training data in a fairer way. From Table 5, we show the results (rows 1-2) of different VLP frameworks pre-trained on the Wukong dataset. Our R2D2 outperforms competitors on five downstream tasks when pre-trained with the same 100M dataset, which demonstrates the superiority of our VLP framework.

	Method	Image-to-Text Retrieval			Text-to-Image Retrieval			R@M
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr30k-CN	CLIP$_{\text{ViT-B}}$	87.1	97.7	98.8	69.0	90.3	95.0	89.7
	CLIP$_{\text{ViT-L}}$ [32]	91.6	99.1	99.7	77.3	94.4	97.2	93.2
	FILIP$_{\text{ViT-B}}$	72.1	91.3	95.8	57.5	84.3	90.6	81.9
	FILIP$_{\text{ViT-L}}$ [46]	90.6	98.8	99.6	76.9	94.9	97.4	93.0
	Wukong$_{\text{ViT-B}}$	83.9	97.6	99.0	67.6	89.6	94.2	88.7
	Wukong$_{\text{ViT-L}}$ [14]	92.7	99.1	99.6	77.4	94.5	97.0	93.4
	R2D2$_{\text{ViT-B}}$	93.2	99.2	99.8	79.2	95.2	97.3	94.0
	R2D2$_{\text{ViT-L}}$	**95.6**	**99.8**	**100.0**	**84.4**	**96.7**	**98.4**	**95.8**
COCO-CN	CLIP$_{\text{ViT-B}}$	68.7	93.6	97.5	68.9	93.3	97.3	86.6
	CLIP$_{\text{ViT-L}}$ [32]	68.3	93.0	97.3	70.1	92.2	96.4	86.2
	FILIP$_{\text{ViT-B}}$	52.7	81.3	88.3	56.2	86.8	94.3	76.6
	FILIP$_{\text{ViT-L}}$ [46]	69.1	91.3	96.9	72.2	92.4	97.2	86.5
	Wukong$_{\text{ViT-B}}$	65.8	90.3	96.6	67.0	91.4	96.7	84.6
	Wukong$_{\text{ViT-L}}$ [14]	73.3	94.0	98.0	74.0	94.4	98.1	88.6
	R2D2$_{\text{ViT-B}}$	78.1	96.2	98.6	76.0	94.9	98.3	90.3
	R2D2$_{\text{ViT-L}}$	**79.3**	**97.1**	**98.7**	**79.1**	**96.5**	**98.9**	**91.6**
AIC-ICC	BriVL [13]	45.6	68.0	76.3	34.1	58.9	69.1	58.7
	CLIP$_{\text{ViT-B}}$	50.5	73.0	80.2	38.1	63.7	73.3	63.1
	CLIP$_{\text{ViT-L}}$ [32]	59.1	79.5	85.2	46.2	70.7	78.6	69.9
	FILIP$_{\text{ViT-B}}$	42.5	67.2	76.0	32.9	58.4	68.8	57.6
	FILIP$_{\text{ViT-L}}$ [46]	54.1	75.8	82.8	44.9	69.0	77.5	67.4
	Wukong$_{\text{ViT-B}}$	47.5	70.6	78.6	36.7	36.7	71.7	57.0
	Wukong$_{\text{ViT-L}}$ [14]	61.6	**80.5**	**86.1**	48.6	72.5	80.2	71.6
	R2D2$_{\text{ViT-B}}$	56.8	76.2	82.1	47.6	72.8	80.2	69.3
	R2D2$_{\text{ViT-L}}$	**65.4**	80.3	84.7	**57.3**	**78.1**	**83.0**	**74.8**
MUGE	CLIP$_{\text{ViT-B}}$	-	-	-	43.5	71.7	80.6	65.3
	CLIP$_{\text{ViT-L}}$ [32]	-	-	-	50.1	76.9	84.9	70.6
	FILIP$_{\text{ViT-B}}$	-	-	-	30.6	58.2	70.2	53.0
	FILIP$_{\text{ViT-L}}$ [46]	-	-	-	43.5	71.5	80.9	65.3
	Wukong$_{\text{ViT-B}}$	-	-	-	39.2	66.9	77.4	61.2
	Wukong$_{\text{ViT-L}}$ [14]	-	-	-	52.7	77.9	85.6	72.1
	R2D2$_{\text{ViT-B}}$	-	-	-	53.4	78.1	86.0	72.5
	R2D2$_{\text{ViT-L}}$	-	-	-	**60.1**	**82.9**	**89.4**	**77.5**
CNA	R2D2$_{\text{ViT-B}}$	93.6	99.5	99.8	80.5	95.6	97.7	94.5
	R2D2$_{\text{ViT-L}}$	**96.9**	**99.8**	**100.0**	**84.9**	**97.0**	**98.6**	**96.2**
ICR	R2D2$_{\text{ViT-B}}$	53.4	75.4	83.4	52.1	73.3	82.0	69.9
	R2D2$_{\text{ViT-L}}$	**61.5**	**82.9**	**87.7**	**60.7**	**82.0**	**86.9**	**77.0**
IQR	R2D2$_{\text{ViT-B}}$	37.0	62.1	70.9	35.8	61.2	70.5	56.3
	R2D2$_{\text{ViT-L}}$	**41.9**	**67.8**	**75.9**	**41.3**	**67.6**	**75.4**	**61.7**

Table 3. Comparisons with state-of-the-art models on image-text retrieval task. CNA represents our proposed Flickr30k-CNA.

Effect of the Proposed Pre-training Dataset. Similarly, we provide comparison results (rows 2-4 of Table 5) of our R2D2 framework pre-trained on the 100M Wukong dataset and the proposed ZERO-corpus. R2D2 pre-trained on the 23M pre-training dataset (a subset of ZERO-corpus) achieves better results than on the much larger 100M Wukong dataset. This improvement comes from the data quality of our ZERO-corpus dataset, which is filtered by user click-through rate and provides diverse text descriptions along with each image. We achieve the best results on the whole pre-training dataset, *i.e.*, ZERO-corpus with 250M high-quality image-text pairs.

4.3. Ablation Study

Effect of Fine-Grained Ranking (FGR). We conduct ablation studies on the first 1% of ZERO-Corpus. For sim-

[1] https://tianchi.aliyun.com/muge

Method	Image-Text Matching		Image Caption				Text-to-Image Genearation	Zero-shot Image Classification	
	AUC (ICM)	AUC (IQM)	BLEU	METEOR	ROUGE-L	CIDEr	FID	Top-1 Acc.	Top-5 Acc.
BriVL [13]	61.9	57.6	66.1	41.1	71.9	220.7	-	24.3	56.8
Wukong$_{\text{ViT-B}}$	79.2	75.1	66.7	71.2	72.2	224.2	23.7	49.1	74.2
Wukong$_{\text{ViT-L}}$ [14]	81.8	78.1	68.9	74.5	72.3	243.1	18.8	55.0	80.5
R2D2$_{\text{ViT-B}}$	88.6	84.9	68.3	76.3	73.2	230.2	18.9	50.6	78.1
R2D2$_{\text{ViT-L}}$	**90.6**	**86.7**	**71.8**	**78.2**	**75.3**	**247.9**	**14.4**	**56.9**	**83.3**

Table 4. Comparison with state-of-the-art models on downstream vision-language tasks.

Method	Pre-training Dataset	Image-Text Retrieval		Image-Text Matching		Image Caption	Text-to-Image Generation	Classification
		Flick30k-CN	COCO-CN	ICM	IQM	AIC-ICC	ECommerce-T2I	ImageNet
Wukong$_{\text{ViT-L}}$ [14]	Wukong (100M)	93.4	88.6	81.8	78.1	243.1	18.8	55.0
R2D2$_{\text{ViT-L}}$	Wukong (100M)	95.2	90.1	86.5	81.5	245.8	16.4	55.6
R2D2$_{\text{ViT-L}}$	ZERO-Corpus (23M)	95.4	90.7	88.1	83.6	246.5	15.7	55.7
R2D2$_{\text{ViT-L}}$	ZERO-Corpus (250M)	**95.8**	**91.6**	**90.6**	**86.7**	**247.9**	**14.4**	**56.9**

Table 5. Effect of the proposed VLP framework and pre-training dataset. We compare different VLP frameworks on the same pre-training dataset (rows 1-2), and compare same VLP framework on different pre-training datasets (rows 2-4). Classification represents zero-shot image classification. We report R@M, AUC, CIDEr, FID, and Top-1 accuracy for five V+L downstream tasks respectively.

Method	Image-to-Text Retrieval			Text-to-Image Retrieval			R@M	Image-Text Matching	Image Caption	Generation	Classification
	R@1	R@5	R@10	R@1	R@5	R@10		AUC	CIDEr	FID	Top-1 Acc.
PRD2	53.92	75.67	82.01	43.97	71.19	80.28	67.14	73.89	239.91	19.21	32.27
R2D2	**64.51**	**81.02**	**85.92**	**56.63**	**78.22**	**84.49**	**74.45**	**80.82**	**243.29**	**17.58**	**39.96**
R2D2 w/o ET	64.14	78.48	84.96	55.32	77.38	83.01	73.81	80.31	243.01	17.82	39.72
R2D2 w/o MLM	63.72	80.19	85.14	55.73	77.29	83.70	73.57	80.01	242.90	17.91	39.54
R2D2 w/o TwD	63.08	79.51	84.69	54.69	76.74	83.53	73.03	79.98	242.64	18.16	38.76
R2D2 w/o TgD	63.87	80.43	85.39	55.97	77.02	83.23	73.52	80.39	243.01	17.90	39.29
R2D2 w/o FgD	63.39	79.86	85.01	54.92	76.83	83.45	73.11	80.28	242.83	18.01	38.92

Table 6. Effect of different components of R2D2. Note that we conduct ablation studies and report the average results on all downstream datasets. Generation and classification represent text-to-image generation and zero-shot image classification, respectively. R@* denotes the result for the image-text retrieval task. We report AUC, CIDEr, FID, and Top-1 accuracy for image-text matching, image caption, text-to-image generation, and zero-shot image classification tasks respectively.

plicity, we define R2D2$_{\text{ViT-L}}$ as R2D2 in the ablation study. We first train a restricted version of R2D2 using only the global contrastive pre-ranking and the two-way distillation strategy. We denote it as PRD2. This restricted setting is conceptually similar to CLIP [32]. R2D2 outperforms PRD2 on the downstream tasks, indicating the effectiveness of the proposed pre-ranking + ranking framework.

Effect of Enhanced Training (ET). From the third row of Table 6, R2D2 (with ET) performs slightly better than R2D2 w/o ET. Furthermore, R2D2 uses less computational resources than R2D2 w/o ET. R2D2 requires 154.0 GFLOPs and can run at 1.4 iterations per second (Iter/s), while without ET we get 168.8 GFLOPS and 1.1 Iter/s. This indicates that ET is able to both reduce the computational cost and improve the capability of the learning process.

Effect of Masked Language Modeling (MLM). Compared to R2D2 w/o MLM, R2D2 obtains better performance on all downstream tasks. MLM allows R2D2 to learn robust representations by masking data. These results indicate that MLM is indeed effective for downstream tasks.

Effect of Two-way Distillation (TwD). The proposed two-way distillation is composed of target-guided distillation (TgD) and feature-guided distillation (FgD). By analyzing the two components of TwD, we see that performing feature alignment is important, since the model w/o FgD shows a more noticeable drop in performance. Although milder, removing TgD also causes a reduction in performance. These results indicate that both components are relevant and TwD is an effective way to improve the generalization performance of the pre-trained model.

4.4. Further Experiments

Zero-shot Tasks. In this section, we conduct zero-shot transfer experiments. From Table 7, our R2D2$_{\text{ViT-L}}$ achieves the best performance on Flickr30k-CN, COCO-CN, MUGE, AIC-ICC, ICR, and IQR. For example, R2D2$_{\text{ViT-L}}$ achieves 80.5% R@M on COCO-CN, an absolute 5.3% gain over the previous best performance. These results demonstrate sound

	Method	Image-to-Text Retrieval			Text-to-Image Retrieval			R@M
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr30k-CN	CLIP$_{\text{ViT-L}}$ [32]	75.0	94.5	97.7	51.8	78.6	85.9	80.6
	FILIP$_{\text{ViT-L}}$ [46]	**78.9**	96.2	98.1	55.7	81.2	87.9	83.0
	Wukong$_{\text{ViT-L}}$ [14]	76.1	94.8	97.5	51.7	78.9	86.3	80.9
	R2D2$_{\text{ViT-L}}$	77.6	**96.7**	**98.9**	**60.9**	**86.8**	**92.7**	**85.6**
COCO-CN	CLIP$_{\text{ViT-L}}$ [32]	51.0	80.0	89.7	48.7	76.8	86.4	72.1
	FILIP$_{\text{ViT-L}}$ [46]	56.9	82.4	90.9	52.7	79.9	88.6	75.2
	Wukong$_{\text{ViT-L}}$ [14]	55.2	81.0	90.6	53.4	80.2	90.1	75.1
	R2D2$_{\text{ViT-L}}$	**63.3**	**89.3**	**95.7**	**56.4**	**85.0**	**93.1**	**80.5**
MUGE	CLIP$_{\text{ViT-L}}$ [32]	-	-	-	43.3	69.2	78.4	63.6
	FILIP$_{\text{ViT-L}}$ [46]	-	-	-	37.6	63.4	73.6	58.2
	Wukong$_{\text{ViT-L}}$ [14]	-	-	-	42.7	69.0	78.0	63.2
	R2D2$_{\text{ViT-L}}$	-	-	-	**49.5**	**75.7**	**83.2**	**69.5**
AIC-ICC	CLIP$_{\text{ViT-L}}$ [32]	16.8	32.0	39.8	9.7	21.1	27.5	24.5
	FILIP$_{\text{ViT-L}}$ [46]	20.6	37.0	45.4	11.3	24.3	31.4	28.3
	Wukong$_{\text{ViT-L}}$ [14]	18.2	34.5	42.4	8.8	20.3	27.3	25.3
	R2D2$_{\text{ViT-L}}$	**30.7**	**47.2**	**52.9**	**14.9**	**28.1**	**33.4**	**34.5**
ICR	CLIP$_{\text{ViT-L}}$ [32]	30.3	52.9	61.6	29.0	51.9	60.9	47.8
	FILIP$_{\text{ViT-L}}$ [46]	27.3	49.6	58.3	25.4	48.5	57.7	44.5
	Wukong$_{\text{ViT-L}}$ [14]	35.1	58.2	66.3	33.7	58.0	66.5	53.0
	R2D2$_{\text{ViT-L}}$	**58.0**	**80.5**	**85.2**	**55.9**	**78.2**	**82.4**	**73.4**
IQR	CLIP$_{\text{ViT-L}}$ [32]	24.3	47.1	56.2	22.2	45.2	54.8	41.6
	FILIP$_{\text{ViT-L}}$ [46]	21.9	43.2	52.8	19.9	42.0	52.0	38.6
	Wukong$_{\text{ViT-L}}$ [14]	26.1	48.9	58.1	24.9	48.1	57.7	44.0
	R2D2$_{\text{ViT-L}}$	**38.4**	**64.8**	**72.8**	**37.4**	**62.6**	**69.0**	**57.5**

Table 7. Zero-shot results of different methods on image-text retrieval task.

generalization ability of R2D2. The results of R2D2$_{\text{ViT-L}}$ on Flickr30k-CNA are the same as that of Flickr30k-CN, since we use the same test set for a fair comparison. In this way, we do not report the results of R2D2$_{\text{ViT-L}}$ on Flickr30k-CNA. In addition, the AUC scores of R2D2$_{\text{ViT-L}}$ on ICM and IQM are 89.8% and 84.5%, respectively.

Entity-conditioned Image Visualization. In this experiment, we visualize the attention map of images on COCO-CN. Specifically, we first extract an entity from the Chinese text and calculate the attention score of an image-entity pair. Here, we select the third layer of the text-image cross encoder following [21]. Figure 3 illustrates the visual explanations of eight images over eight different entities. It shows that R2D2 learns well to align text with the correct content inside the image. More analysis is shown in Appendix.

5. Related Work

5.1. Vision-Language Datasets

Chinese vision-language benchmark requires images and high-quality Chinese texts, which are hard to obtain and still rare for the research community's reach. To this end, existing public datasets [18, 25] use machine translation to adapt their English versions [3, 47] to Chinese, but the data quality is sacrificed due to machine translation errors. Newly reported datasets with Chinese texts [13, 14, 26] are proposed for Chinese VLP. However, they are either not publicly available or lack sufficient downstream tasks. In this paper, we propose a Chinese vision-language benchmark that covers a large-scale pre-training dataset and five high-quality downstream datasets.

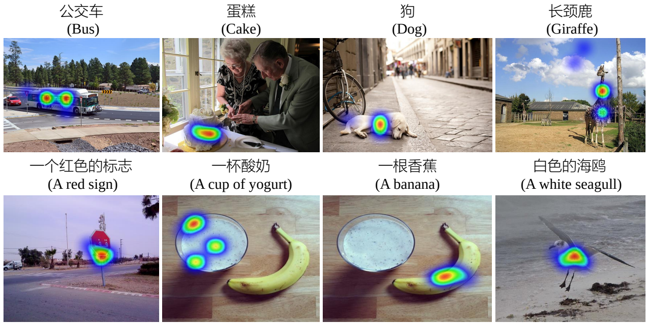


Figure 3. Entity-conditioned image visualization.

5.2. Vision-Language Pre-training Learning

Vision-Language Architecture. The vision-language pre-training architectures can be categorized as: single-stream and dual-stream. Most existing single-stream models [4, 19, 23, 29, 31] concatenate image and text as a single input to model the interactions between image and text within a transformer model [40]. On the other hand, popular dual-stream models [12, 16, 22, 28, 32, 43] aim to align image and text into a unified semantic space via contrastive learning. Besides, some works [20, 21] align the individual features of images and texts in a dual-stream architecture, and then fuse the features in a unified semantic space via a single-stream architecture. However, they ignore supervised signals from images. In addition, they use traditional masked language modeling (MLM) and local contrastive learning to conduct pre-training tasks, leading to potential inferior model performance. In this paper, we explore the effective signals via an image-text cross encoder and a text-image cross encoder while also maintaining the bottom dual-stream architecture. Moreover, we improve MLM with enhanced training and apply global contrastive learning to further improve performance.

Knowledge Distillation. The general purpose of knowledge distillation is to improve the student model's performance by simulating the output of the teacher network [1, 15, 45, 48]. Compared to previous works [21, 48], we propose target-guided distillation with a weighted momentum queue and feature-guided distillation to stabilize the model representations for vision-language pre-training.

6. Conclusion

In this paper, we introduce a large-scale Chinese cross-modal benchmark called ZERO and a vision-language framework named R2D2. ZERO includes a high-quality pre-training dataset, which is the largest Chinese cross-modal dataset, and five human-annotated downstream datasets, two of which are the largest Chinese vision-language downstream datasets and first proposed for Chinese image-text matching task. R2D2 employs a pre-training framework of pre-Ranking + Ranking with target-guided Distillation and feature-guided Distillation for cross-modal learning. After pre-training, R2D2 achieves state-of-the-art results on fine-tuning and zero-shot settings on twelve downstream datasets of five vision-language tasks. A limitation of our method is that it only treats Chinese text, and we will adopt R2D2 to English VLP learning in future work. We will release all datasets and models to promote the development of vision-language learning. We expect that the good cross-modal benchmark and framework will encourage a plethora of engineers to develop more effective methods in specific real-world scenarios.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 8

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[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*, 2015. 8

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2020. 2, 5, 8

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 7–10, 2016. 2

[6] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942, 2021. 1

[7] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, pages 657–668, 2020. 3, 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6, 13

[9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 4171–4186, 2019. 3, 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018. 8

[13] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):1–13, 2022. 1, 2, 6, 7, 8

[14] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*, 2022. 1, 2, 5, 6, 7, 8, 12, 13

[15] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *Advances in Neural Information Processing Systems Workshop*, 2014. 8

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021. 8

[17] 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*, 123(1):32–73, 2017. 2

[18] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557, 2017. 1, 2, 3, 8

[19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11336–11344, 2020. 8

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 8

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align

before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 2021. 2, 5, 8

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019. 8

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 8

[24] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2592–2607, 2021. 2

[25] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019. 1, 2, 8

[26] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3251–3261, 2021. 1, 2, 6, 8

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 2019. 8

[29] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 8

[30] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems*, 2011. 1, 2

[31] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 2, 8

[32] 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, 2021. 1, 2, 5, 6, 7, 8

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 12

[34] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Advances in Neural Information Processing Systems Workshop*, 2021. 1, 2

[35] 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, 2

[36] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe 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. 1

[37] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 2

[38] Shulong Tan, Meifang Li, Weijie Zhao, Yandan Zheng, Xin Pei, and Ping Li. Multi-task and multi-scene unified ranking model for online advertising. In *2021 IEEE International Conference on Big Data*, pages 2046–2051, 2021. 2

[39] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1, 8, 12

[41] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340, 2022. 1

[42] Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Cold: Towards the next generation of pre-ranking system. In *2nd Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD*, 2020. 2

[43] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 2021. 2, 8

[44] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. In *IEEE International Conference on Multimedia and Expo*, 2019. 2, 6, 12

[45] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classifica-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 8

[46] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. 6, 8

[47] 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*, 2:67–78, 2014. 3, 8

[48] Kaiyu Yue, Jiangfan Deng, and Feng Zhou. Matching guided distillation. In *European Conference on Computer Vision*, pages 312–328, 2020. 8

A. Details of ZERO-Corpus

We illustrate several representative examples of ZERO-Corpus in Figure A. Each sample contains one image and its corresponding attributes. For ease of understanding, we add an English translation version after each Chinese text. There are 3 types of text fields associated with each image: "Title", "Content" and "ImageQuery". "Title" and "Content" come from the source webpage containing the image, and the latter is also termed as surrounding text in other works. "ImageQuery" is the associated query string for the corresponding image. The average length of "ImageQuery", "Title", and "Content" is 5, 18, and 29, respectively. During pre-training, we randomly select one from the 3 text fields to construct an image-text pair, ensuring data diversity. In this way, the pre-trained model can flexibly fit different text lengths on various downstream tasks. For instance, the text length of AIC-ICC is about 18 words while the text length of MUGE is less than 10 words. We apply a series of filtering strategies to construct the ZERO-Corpus. For images, we filter out images with both dimensions smaller than 100 pixels or aspect ratio out of the range [1/4, 4]. In addition, we filter images that contain sensitive information, such as sexual, violent scenes, etc. For texts, we remove texts shorter than 2 words or longer than 128 words. Also, we remove texts that contain sensitive as in image filtering. We hope this dataset will bring help to the research community.

B. Examples of the Proposed Downstream Datasets

We illustrate examples of ICM, IQM, ICR, and IQR in Figure B. Moreover, Figure C highlights some cases of the difference between Flickr30k-CN and our proposed Flickr30k-CNA.

C. More Implementation Details

Selection Strategy of Negative Pairs in Fined-grained Ranking. We obtain hard negative samples by sampling in a mini-batch. Given an image in the mini-batch, we select the corresponding negative text by ranking the contrastive scores of the current batch. We choose the higher score except for the original positive text of the image. In this way, we construct one image-text negative pair for fined-grained ranking loss. The negative images of each text are similar to the description above.

Fine-tuning Strategy of Image-Text retrieval. We jointly optimize the GCPR loss (Equation 5) and the FGR loss (Equation 3). We extract the individual features of images and texts via our dual-stream encoder and compute the similarity of all image-text pairs. Then we take the top-K candidates and use two cross encoders to further calculate the corresponding similarity scores for ranking during inference. We use a mean operation for the outputs of the two cross encoders. Here, we adjust the K on different downstream datasets. We fine-tune the pre-trained model with 20 epochs on 7 downstream datasets, including Flickr30k-CN, COCO-CN, AIC-ICC, MUGE, ICR, IQR, and Flickr30k-CNA. K is set as 128, 256, 32, 64, 64, 64, 128, respectively. The batchsize is 32 and the learning rate is $1e^{-5}$.

For both image-to-text retrieval (TR) and text-to-image retrieval (IR) tasks, we report Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), and Mean Recall (R@M). For AIC-ICC and MUGE, we report their results on the validation sets, since their test sets are not released. For ICR and IQR, we also report the results on the validation sets in this paper, since we use the corresponding test sets to build a leaderboard. The test set of Flickr30k-CNA is also added to the leaderboard. For Flickr30k-CNA, we show the performance on the test set of Flickr30k-CN for a fair comparison in the main paper. For the remaining downstream datasets, we report the results on the test sets. Following [14], we select the first 10,000 images with the corresponding 50,000 texts when testing on AIC-ICC. In particular, we only provide IR scores on MUGE since it only has IR settings.

Fine-tuning Strategy of Image-Text Matching. This task predicts whether an image-text pair is matched or not. During fine-tuning, we only apply the FGR loss (Equation 3). We fine-tune the models with 5 epochs using a batchsize of 64. The initial learning rate is $1e^{-5}$. Additionally, we report the results on the validation sets of ICM and IQM.

Fine-tuning Strategy of Image Caption. Given an image, the goal of the image-caption task is to generate a caption to describe the image. Similar to Transformer [40], the image-caption model consists of an encoder and a decoder, where the encoder aims to extract the embedding of the given image and the decoder generates tokens of the caption. In specific, we use the image encoder and the text-image cross encoder of R2D2 to initialize the image-caption encoder and decoder, respectively. We fine-tune the image-caption model on the training split of AIC-ICC [44] with 20 epochs. The batchsize is 128 and the learning rate is $1e^{-4}$.

Fine-tuning Strategy of Text-to-Image Generation. Text-to-image generation requires the model to generate an image corresponding to the input text. Following DALL-E 2 [33], we build a generation model, including a CLIP-based module, a prior module and a decoder module. Specifically, the dual-stream weights of R2D2 are used to initialize the CLIP-based module. We fine-tune the CLIP-based module and fix it in the next step. Then, we train the prior module to generate image embeddings for given texts. Finally, we fix two former modules and train a diffusion decoder to invert the image embeddings to generate images. All three components of the generation model are fine-tuned on the ECommerce-T2I dataset with 20 epochs, respectively. The batchsize is 16 and the learning rate is $1e^{-4}$.

Fine-tuning Strategy of Zero-shot Image Classifica-

Title: 五大地缝奇观欣赏 (View of the five fissure wonders)

Content: 奉节地缝亦称天井峡地缝，全长有37公里，最大深度有229米，而最窄处仅2米、而峡谷高度达900米，形成气势宏伟的“一线天”，被岩溶专家称作“世界喀斯特峡谷奇中之稀”。峡谷上段较为开阔，但愈往下愈狭窄，上部宽10至30米，谷底宽仅1至30米，悬崖最深处达300米

(Fengjie fissure, also known as Tianjingxia fissure, has a total length of 37 kilometers and a maximum depth of 229 meters. The narrowest point is only 2 meters and the height of the canyon is 900 meters, forming a magnificent "one-line sky". The Fengjie fissure is called "the rarest karst canyon in the world" by karst experts. The upper part of the fissure is relatively open, but it becomes narrower as it goes down. The upper part is 10 to 30 meters wide, the bottom of the valley is only 1 to 30 meters wide, and the deepest cliff is 300 meters.)

ImageQuery: 天井峡地缝 (TianJingXia fissure)

Title: 英宠物狗戴墨镜穿潮装, 百变时装造型受热捧

(British pet dogs wear sunglasses and trendy clothes. The ever-changing fashion styles are popular.)

Content: 一只名叫托斯特(Toast)的查尔斯王小猎犬不用拥有专属于自己的漂亮手提包

(A King Charles Spaniel named Toast doesn't have its own fancy handbag.)

ImageQuery: 戴墨镜的狗, 戴墨镜的人, 狗戴墨镜, 墨镜狗狗, 戴墨镜的狗狗图片, 宠物戴墨镜, 漂亮的宠物狗造型, 宠物戴墨镜和围巾, 橙色的宠物狗, 小猎犬戴墨镜, 舔脚, 时装造型, 狗狗舔脚, 小狗戴墨镜, 狗狗戴墨镜 (Dog with sunglasses)

Title: 美呆了!25万盆鲜花齐聚小榄菊花展

(Stunningly beautiful! 250,000 pots of flowers gathered at the Xiaolan chrysanthemum exhibition.)

Content: 大立菊、盆景菊、悬崖菊

(Dali chrysanthemum, bonsai chrysanthemum, cliff chrysanthemum)

ImageQuery: 大立菊 (Dali chrysanthemum)

Title: 零基础学绘画-彩铅《紫红色百合花》 (Zero Basic Learning Painting - Color Lead "Fuchsia Lily")

Content: 最终的效果如图，能出这样的效果，真的是一层层涂出来的 (The final view is shown in the figure. To achieve such a view, it is painted layer by layer.)

ImageQuery: 彩铅百合,彩铅百合绘画大全 (Color lead lily, color lead lily painting Daquan)

Title: 茶百戏，一种能使茶汤纹脉形成物象的民间艺术

(Tea Baixi, a folk art that can make the veins of tea soup form objects.)

Content: 乌龙茶汤显现的茶百戏图

(Tea Baixi shown in Oolong tea soup)

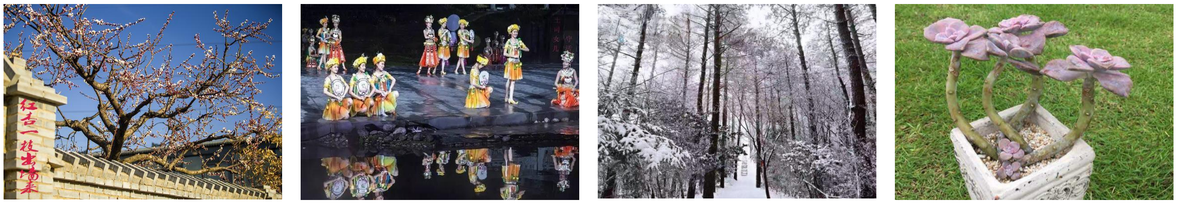
ImageQuery: 茶百戏 (Tea Baixi)

Figure A. Examples of ZERO-Corpus.

tion. Given an image, the zero-shot image classification task aims to predict the corresponding class label. Following [14], we use R2D2 to conduct zero-shot image classification task on ImageNet [8]. All the class labels in ImageNet are translated into Chinese.

D. More Cases about Image Visualization

In this experiment, we provide more cases of image visualization given an entity. From Figure D, R2D2 has the ability to capture the salient areas when given an image with complex backgrounds, such as the images of “A train” and “A bull”. Our R2D2 also precisely locates different objects within the same image, as shown in the images of “A cup of yogurt” and “Banana” in the main paper. Moreover, we analyze some bad cases in Figure E. We find that the attention score is disturbed when two adjacent entities are present in an image. This phenomenon is particularly evident for objects with similar colors or categories.



这么晴好的天，当然开得快！大家一定要抓住机会，去欣赏洛阳市这一年一度的杏花满山。
On such a sunny day, of course it drives fast! Everyone must seize the opportunity to appreciate the annual apricot blossoms in Luoyang City.

恩施民族服饰
Enshi National Costume

这场雨雪天气将持续到今天早上，预计平原地区的积雪将达到1-4cm。
The rain and snow will continue until this morning, with 1-4cm of snow expected in the plains.

紫乐用什么花盆
What flower pot does Zi Le use

Figure B. Image-text examples of ICM, IQM, ICR and IQR from left to right.

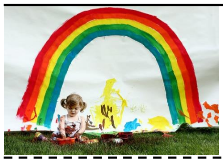
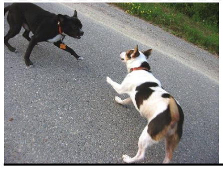
	**Flickr30k:** A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl. **Flickr30k-CN:** 一个小女孩在油漆前坐在一个彩虹的前面双手在碗里。 **Flickr30k-CNA:** 一个涂满染料的小女孩坐在画好的彩虹前，把她的手放在一个装颜料的碗里。
	Flickr30k: A man with reflective safety clothes and ear protection drives a John Deere tractor on a road. **Flickr30k-CN:**一个男人用反光安全服装和耳朵保护驱动的道路上约翰迪尔拖拉机。 **Flickr30k-CNA:**一个穿着反光安全服，带着耳护的男子在路上开着一辆约翰迪尔拖拉机。
	**Flickr30k:** A black dog and a white dog with brown spots are staring at each other in the street. **Flickr30k-CN:** 一只黑色的狗和一只棕色的白色狗在街上盯着对方。 **Flickr30k-CNA:** 一只黑狗和一只带有棕色斑点的白狗站在街上，互相盯着对方。

Figure C. Comparisons of Flickr30k, Flickr30k-CN and our proposed Flickr30k-CNA.

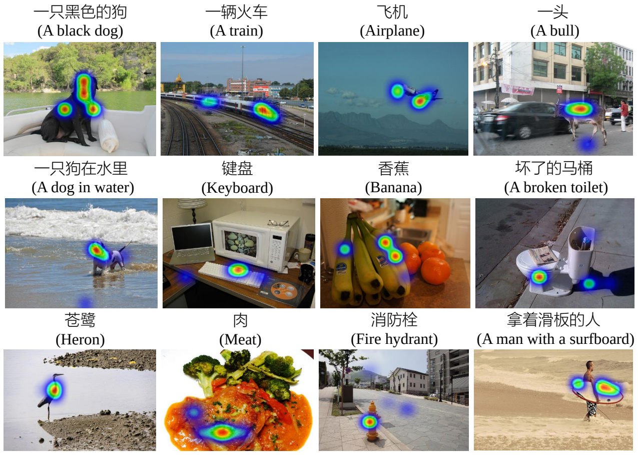


Figure D. More Examples of entity-conditioned image visualization.

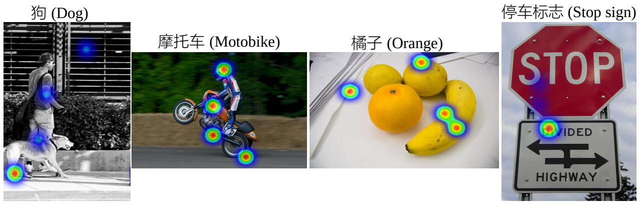


Figure E. Bad cases of entity-conditioned image visualization.